An Integrative Systems Biology Approach to Illustrate that Blood-Sourced MiRNAs are Not Diagnostic of Breast Cancer.

Irene Pylypenko
School of Computer Science
McGill University, Montreal

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science.

Acknowledgements

I dedicate this thesis to Dr. Chapados for his dedication and support throughout this work.

I would like to thank Drs. Mike Hallett and Vanessa Dumeaux for the intellectual contribution and previous work upon which this research was a continuation of. I thank Dr. Lund from Norway for providing us with the data thanks to funding from the European Research Council. I also want to thank the members of the Norwegian Breast Cancer Group (NBCG) for their participation in collecting blood samples used in this study. I would like to acknowledge contributions from Matthew Suderman, Ali Tofigh, Robert Lesurf, Sadiq Saleh, Sean Cory and Julie Livingstone during my time at McGill University.

This work could not have been completed without the financial support from the Natural Sciences and Engineering Research Council of Canada and the Canadian Institute of Health Research.

I would like to thank McGill University for accepting my application to study Systems Biology, and learn a little bit more about how biological systems naturally interaction with each other. My gratitude goes out to the administrative team at McGill for guiding me through the process of submitting this thesis.

Finally, I would like to express my gratitude for the endless support from my family, friends, teachers and mentors, especially Anna and Diana Pylypenko for always being there for me.

Abstract

Breast cancer (BC) is the most common cancer in women worldwide, and current detection technologies have limitations. MicroRNAs (miRNAs) are small 18-22 nucleotide single-stranded RNAs and there is some evidence suggesting that miRNAs in blood samples may be used as diagnostic biomarkers for BC. The Norwegian Women and Cancer (NOWAC) is a large prospective study that has collected blood and tumor biopsy samples from BC patients and healthy tissue from age-matched controls. Using the Illumina microarray system, miRNA and messenger RNA (mRNA) expression profiles were generated for 96 breast cancer cases with matched controls. We identified thirty-eight miRNAs that discriminate between breast cancer and matched healthy controls, some of which (miR-210, miR-335, miR-145, miR-15a/b) have been previously identified as potential diagnostic markers in blood samples of breast cancer patients. Then, we applied three different miRNA target prediction tools to look for potential gene targets based on a simple negative association miRNA-target model. In our clustering analysis, the predicted gene sets identified in the matched mRNA expression profile did not follow the classical negative association miRNA-target model. Thus, inconsistent with the hypothesized model, as also supported by previous studies. Functional analysis of identified miRNAs and their predicted target genes identified some gene pathways involved in breast cancer, such as B cell receptor signaling pathway, BRCA1 expression network, and p53 a common oncogenic signal pathway. Although miRNAs show promising results as diagnostic markers in blood samples of breast cancer patients, there is much work to be done in understanding the relationship between miRNAs and their target genes in order to identify viable miRNA biomarkers for breast cancer.

Resume

Le cancer du sein est le cancer le plus courant chez les femmes à travers le monde, et les technologies de détection actuelles ont leurs limites. Les microARN (miARN) sont de petits ARN monocaténaires de 18 à 22 nucléotides et certains éléments suggèrent que les miARN dans les échantillons de sang peuvent être utilisés comme biomarqueurs diagnostiques pour le cancer du sein. L'étude NOWAC sur les femmes norvégiennes et le cancer est une grande étude prospective qui a recueilli des échantillons de sang et de biopsies de tumeurs de patientes atteintes de cancer du sein et et de tissus sains provenant des témoins appariés selon l'âge. En utilisant le système de microréseaux Illumina, des profils d'expression de miRNA et d'ARN messager (ARNm) ont été générés pour 96 cas de cancer du sein avec des témoins appariés. Nous avons identifié trente-huit miARN qui discriminent entre le cancer du sein et les témoins sains appariés, dont certains (miR-210, miR-335, miR-145, miR-15a/b) ont déjà été identifiés comme marqueurs diagnostiques potentiels dans des échantillons de sang de patientes atteintes de cancer du sein. Ensuite, nous avons appliqué trois outils de prédiction de cible de miARN différents pour rechercher des cibles de gènes potentiels sur la base d'un simple modèle d'association miARN-cible négative. Dans notre analyse de regroupement, les ensembles de gènes prédits identifiés dans le profil d'expression d'ARNm apparié ne suivaient pas le modèle classique d'association miARN-cible négative, réfutant ainsi le modèle de miARN-cible supposé, tel que soutenu également par certaines études précédentes. L'analyse fonctionnelle des miARN identifiés et de leurs gènes cibles a identifié des voies de gènes impliqués dans le cancer du sein, telles que la voie de signalisation du récepteur des cellules B, le réseau d'expression BRCA1 et p53 une voie de signal oncogénique commune. Bien que les miARN montrent des résultats prometteurs en tant que marqueurs de diagnostic dans les échantillons de sang de patientes atteintes d'un cancer du sein, il reste encore beaucoup à faire pour comprendre la relation entre les miARN et leurs gènes cibles afin d'identifier des biomarqueurs viables du miARN pour le cancer du sein.

Table of Contents:

List of Figures:

List of Tables:

Chapter 1: Introduction

Introduction to Biomarkers

It is estimated that the human genome encodes approximately 19,000 protein-coding genes (Ezkurdia et al., 2014), about the same number as that for corn, but about twice as many as that for the common fruit fly. These 19,000 genes are encoded in about 1.5% of the genome. Some genes are expressed continuously, as they produce proteins involved in basic metabolic functions; some genes are expressed as part of the process of cell differentiation; and some genes are expressed as a result of cell differentiation. Gene regulatory mechanisms such as transcription factors or DNA methylation may control the rate of transcription by limiting the amount of mRNA that is produced from the nucleotide sequence of a particular gene. Once transcribed there are further opportunities for gene regulation, including regulation of mRNA decay and regulation of the translation of mRNA into protein. These forms of regulation are known as post-transcriptional regulation and play important roles in both normal physiology and organismal development. In the early 2000s, a novel class of ~21-nucleotide-long RNAs, known as microRNAs (miRNAs) emerged as key post-transcriptional regulators predicted to control the activity of ~50% of all protein-coding genes in mammals as shown in Figure 1.1 (Karp S., 2009).

Figure 1.1. The Central Dogma and miRNA (Karp S, 2009).

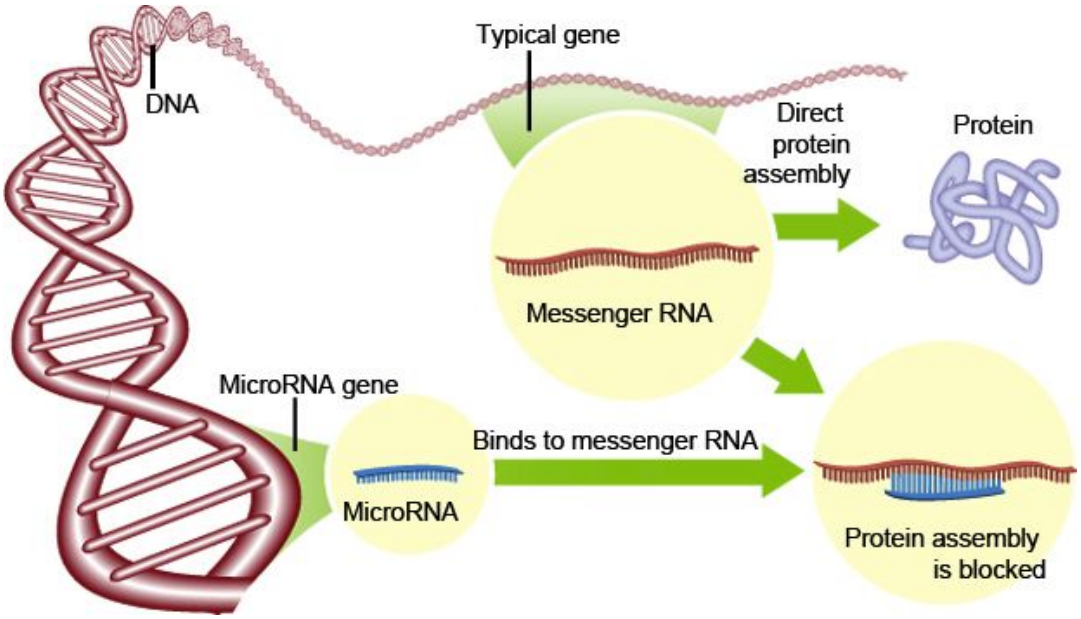

Figure 1.1. A microRNA gene is one-hundredth the length of a typical gene. The typical gene codes for messenger RNA, which in turn directs the assembly of a protein. MicroRNA genes can control this essential process, by coding for a microRNA strip that binds to the messenger RNA, effectively turning off production of the protein via translational repression or target degradation.

In 1993, Lee, Feinbaum and Ambros discovered that a nucleotide sequence in *C. elegans* did not code for a protein but instead produced a pair of short RNA transcripts. These RNA transcripts each regulated the timing of larval development by repressing the translation of lin-14, which encodes a nuclear protein (Lee et al., 1993). This regulation is due in part to sequence complementarity between lin-4 and unique repeats within a small region of the lin-14 mRNA (Figure 1.2), suggesting that lin-4 regulates lin-14 translation via an antisense RNA-RNA interaction. Loss-of-function of lin-4 results in the abnormal differentiation of specific cell lineages and affects later stages of development, thus providing the first evidence of miRNAs involved in cell differentiation and proliferation (Lee et al., 1993).

Figure 1.2. Complementarity between *lin4* and *lin-14* (Lee et al., 1993).

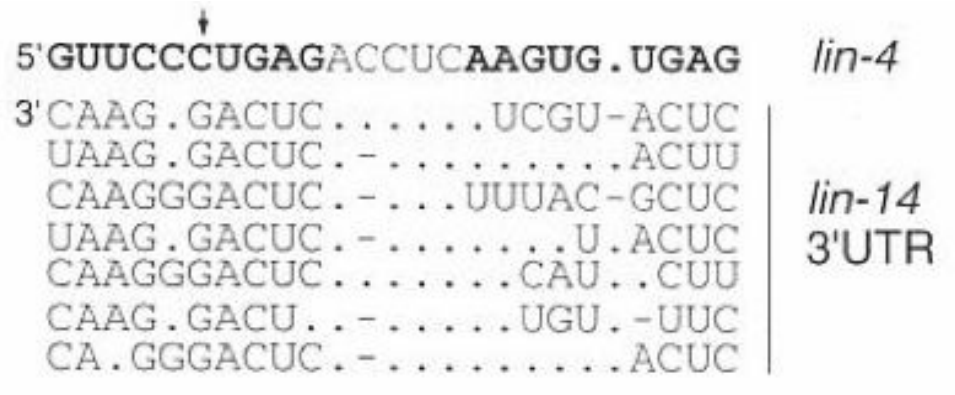

Figure 1.2. Complementarity between *lin-4* and seven copies of a repeated element in the 3'UTR of *lin-14* RNA. Dots indicate absence of a nucleotide; dashes indicate one or more non complementary nucleotides. Only *lin-4:lin-14* complementarity that is conserved between C. elegans and C. briggsae is represented.

The second miRNA to be identified was let-7, expressed later in worm development and complementary to a specific region of the chromosome that includes lin-14, lin-28, lin-41, lin-42, and daf-12 blocking their expression (Reinhart et al., 2000). Since the discovery of let-7, over 48,000 miRNAs have been identified in various organisms including viruses, worms, and primates, and humans (Kozomara et al., 2019). miRNA identification is done through two methods (i) random cloning and sequencing, like the let-7 or (ii) through computational prediction which identifies putative miRNAs (Krek et al., 2005). MiRNAs are commonly defined by the following criteria (Kim, 2005):

1. the final miRNA product is a single-stranded RNA of about 22-nucleotides;
2. the precursor forms a hairpin structure and the mature miRNA is present in one arm of the hairpin;
3. both the mature and the precursor miRNAs are usually evolutionarily conserved;
4. the precursor miRNAs should be experimentally observed when DICER (an essential miRNA processing enzyme) function is disturbed.

Most miRNAs originate from non-coding regions of the genome; however, miRNA-producing regions (called miRNA genes) have been found throughout the entire genome. Up to 50% of mammalian miRNA loci are found in close proximity to other miRNAs (Kim et al., 2009). This clustering suggests that miRNAs act together, as their close proximity allows miRNAs to be transcribed together. MiRNAs add a layer of complexity to gene regulation by base pairing with the target mRNAs, usually located in the 3'UTR region. This binding is often with a perfect complementary sequence, however many exceptions have been found. For example, miRNAs can bind to a specific mRNA with a base pair mismatch or a bulge (unpaired bases). The binding sequence of many miRNAs is composed of base pairs (bp) numbered 2 through 7 in the 5' region of the strand. This 6 bp segment is termed the seed region. Since the seed region is very short and genomes tend to be millions of base pairs long, there is a high chance for the complementary seed sequence to occur more than once. Thus, one miRNA likely targets multiple sites on the same mRNA, or multiple mRNAs (Friedman et al., 2009, Krek et al., 2005). It has been estimated that more than one-third of human genes are directly targeted by miRNAs (Friedman et al., 2009).

Further, once base paired with a target gene, miRNAs control target gene expression by either regulating mRNA degradation or mRNA translation (Huntzinger et al., 2011). However, studies (Levine et al., 2007, Mukherji et al., 2011) have uncovered new mechanisms that may be involved beyond the induction of mRNA degradation and the inhibition of translation for which miRNAs are best known. For example, it has been proposed that miRNAs counteract 'leaky' transcription by establishing thresholds in gene expression levels and induce correlations in the expression of their targets (Mukherji et al., 2011). Thus, although it is known that miRNAs bind to certain target genes, the mechanisms behind the post-transcriptional regulation is being explored.

The biogenesis of miRNAs in animals is a complex, multi-step process starting in the nucleus, passing through several post-transcriptional modifications, and ending in the cytoplasm (Figure 1.3). The canonical pathway initiates at transcription by RNA polymerase II to generate the primary transcripts (pri-miRNAs). The pri-miRNA is characterized by a hairpin RNA structure recognized by the nuclear RNAse III enzyme

Drosha, and its cofactor DGCR8 (Liu et al., 2009). Drosha and DGCR8 bind to create a complex, called the microprocessor complex, which cleaves the pri-miRNA to generate a shorter hairpin of ~65-75 nucleotides, called the pre-miRNA (Du et al., 2005). The pre-miRNA is then recognized by the nuclear export factor EXP5 responsible for exporting it from the nucleus to the cytoplasm. After exportation from the nucleus, the cytoplasmic RNase III DICER and other proteins TRBP and Argonaute catalyze the second processing step (dicing) to produce miRNA duplexes (Du et al., 2005). Finally, one strand of the duplex remains on the Argonaute protein as the mature miRNA, whereas the other strand is degraded (Figure 1.3). The miRNA biogenesis pathway is well studied in comparison to other small RNA pathways, although many questions remain unanswered. A more detailed understanding of the mechanism awaits the structures of the complexes, including Microprocessor, EXP5 and DICER –RISC in association with the substrate RNAs. Many protein factors are implicated in miRNA biogenesis, but their biochemical roles remain unknown (Kim et al., 2009).

Figure 1.3. MiRNA Biogenesis (Huang et al., 2011).

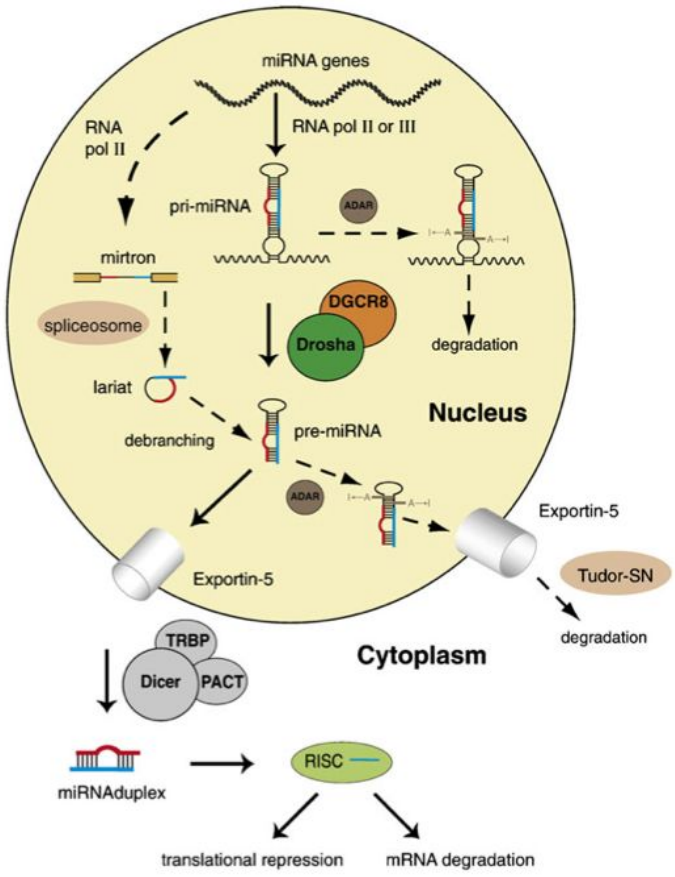

Figure 1.3. The biogenesis of miRNA is a multi-step process starting in the nucleus, passing through many post-transcriptional modifications, and ending in the cytoplasm. The pathway initiates at transcription by RNA polymerase II, generating a primary miRNA. The nuclear RNAse III enzyme Drosha, and its cofactor DGCR8 recognize the pri-miRNA, which work within a complex of several proteins known as the microprocessor. It then cleaves the pri-miRNA and exports it to the cytoplasm, where a second RNAse III enzyme, Dicer, makes the pair of cuts that defines the other end of the miRNA, generating the miR/miR* duplex. Finally, assembly of the mature, single stranded miRNA from the duplex into the RNA-induced silencing complex (RISC) completes the miRNA biogenesis.

Computational Identification of MiRNAs

As experimental approaches are often slow and costly, computational methods play important roles in the identification of new miRNAs. Traditionally, certain significant characteristics such as the hairpin-shaped stem loop structure, high evolutionary conservation, and high minimal folding energy (the energy released as the base pairs fold into its structure) were important features used by computational tools for the identification of miRNAs (Lindow and Gorodkin, 2007). For example, when a miRNA base pairs to a target mRNA, it forms an RNA duplex. This process of canonical base pair binding releases energy. Generally, the lower the free energy, the more bases are paired, and the more stable the RNA duplex is.

Lee and Ambros (2001) were the first to apply a computational approach to identify miRNAs. The took a comparative genomics approach by using bioinformatics tools with cDNA cloning to identify potential *C. elegans* miRNAs. They searched for sequences conserved (similar or identical base pairs) between the *C. elegans* and *C. briggsae* genomes that had characteristic pre-miRNA features and secondary structures similar to *lin-4* and *let-7*, the first two miRNAs identified. Since then, several tools have been developed to predict new miRNA genes based on either sequence and/or secondary structure similarity to known miRNAs (Lim et al., 2003; Wang et al., 2005). These methods described by Lim et al. (2003) and Wang et al. (2005) are based on previous findings that miRNAs tend to be evolutionary conserved (Krek et al., 2005) and filter out predicted hairpins that are not evolutionarily conserved in related species. However, excluding miRNAs that are unique to one organism may impair the identification of new miRNAs associated with that specific organism (Bentwich et al., 2005). Thus, the limiting factor with these initial methods, is the inability to discover new miRNAs.

Machine Learning Based Methods

In order to circumvent the main limitation of methods based solely on comparative genomics, machine learning based methods have been developed to predict *ab initio* miRNAs. In general, a machine learning algorithm is used to make a prediction on unseen data (test set), based on the features (attributes describing the data) it learns from an initial (training) dataset. In our case, these algorithms take in a set of features describing known miRNA sequences and structures, then classifies an unknown sequences as a candidate miRNA or a non-miRNA. This is an example of a binary classification machine learning task, a type of supervised machine learning. Common supervised machine learning algorithms include Support Vector Machines (SVM), Neural Networks, Hidden Markov Models (HMM), and Naive Bayes (Bishop, 2006), with SVM being the most popular choice for miRNA classification. To characterize their performance, two statistical parameters are commonly used: sensitivity and specificity. Sensitivity measures the percentage of correctly predicted targets out of total correct ones, and specificity measures the percentage of correctly predicted targets among overall predicted ones. Ideally, the performance of a method must be of high sensitivity and high specificity with a fair balance between them.

The seminal work of applying a machine learning based method to identify miRNAs was by Sewer et al. (2005), who compiled 40 distinct sequence and structural ‘‘markers’’ to describe a candidate pre-miRNA. The SVM classifier model was trained using 178 human pre-miRNAs as positive examples and 5395 random sequences from tRNA, rRNA, and mRNA genes as negative examples. They obtained a specificity of 91% and a sensitivity of 71% on the training set, then predicted 32 novel pre-miRNAs of pathogenic viruses, some of which were further confirmed experimentally by the same group. This work set the initial benchmark for pre-miRNA classification, after which many other machine learning based tools were developed. For an online compendium of miRNA prediction tools, see Lukasik et al. (2016).

One of the key challenges of predicting species-wide miRNAs, is the limited amount of annotated data per species. Thus, techniques that take into consideration the imbalance of positive and negative examples are applied to combined datasets from different species. One of the best performing methods is miRBoost (Tran et al., 2015), an ensemble method that applies a boosting technique with the SVM algorithm to address training data imbalance. It is trained on 187 novel and existing pre-miRNA features, with a positive data set of 2540 pre-miRNAs and a negative data set of 15688 pre-miRNAs. Not only is it much faster than most other methods, it achieves a good balance between sensitivity (88%) and specificity (91%) (Tran et al., 2015) in classifying pre-miRNAs.

In practice, machine learning based methods to identify *de novo* miRNAs face challenges that affect both their sensitivity and specificity. First, they require an adequate number of annotated miRNAs as training examples. However, the number of characterized and validated miRNAs is still relatively small, thus negatively affecting the true positive rate (sensitivity) of these methods. Second, these methods rely on genome annotations to reduce the number of falsely predicted putative miRNAs and thus increase the true negative rate (specificity). However, most sequenced genomes have not been well annotated, and many of them have few experimentally characterized miRNAs. Third, negative examples are crucial to train machine learning based tools since they affect both the specificity and sensitivity of the results. The challenge lies in selecting negative examples, which can effectively describe the complete negative space and define suitable features to distinguish non-miRNAs from miRNAs. A common method for selecting negative examples is randomly generating genome sequences, however this may not guarantee proper feature representation of real miRNAs.

Sequencing Technologies

As sequencing technologies have improved, methods based on high throughput experimental evidence have achieved great success in discovering novel miRNA genes.

Next Generation Sequencing (NGS) is based on massively parallel sequencing of millions of DNA or RNA molecule fragments, by fragmenting the genome into small pieces, randomly sampling for a fragment, and sequencing it using one of the NGS technologies, such as Illumina/Solexa, ABI/SOLiD, 454/Roche. For example, the Illumina/Solexa technology uses reversible fluorescent dye terminators as adapters that ligate to fragmented pieces of DNA/RNA to generate base calls as illustrated in Figure 1.4.

Figure 1.4. The Illumina sequencing workflow (Illumina Inc, 2008).

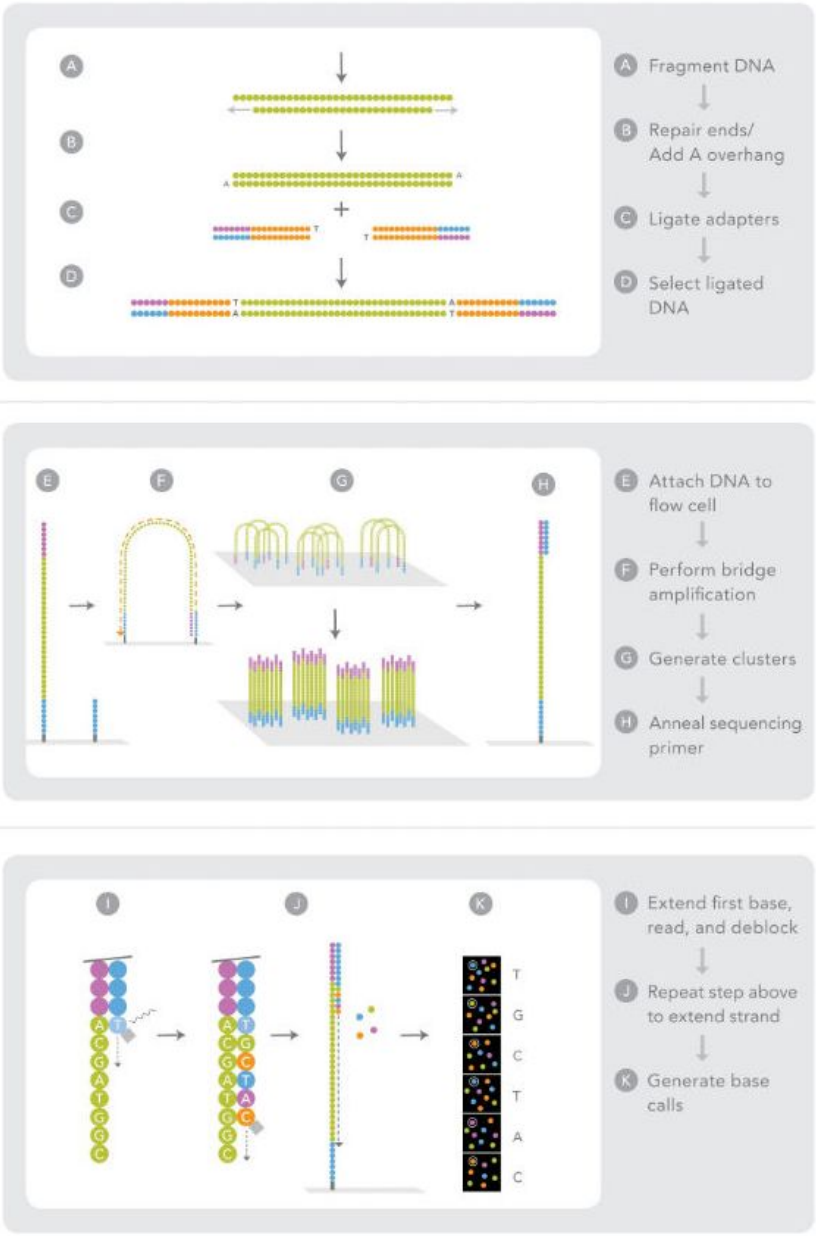

Figure 1.4.

The Illumina/Solexa sequencing method is based on reversible dye-terminators technology and engineered polymerases. First, the adapters are ligated to fragmented pieces of DNA. Then after the DNA is attached to flow cells and amplified, clusters are generated and sequencing primer is annealed. Finally, the process of extending the first base, reading and deblocking is repeated to extend the strand and generate base calls.

Further, these technologies have been adapted to sequence miRNAs and other small RNAs. The short length of these RNAs however, give the researchers fewer options for designing complementary sequences and the sequences often vary by as little as a single nucleotide, thus miRNAs are hard to distinguish from other small RNAs or degradation byproducts. To overcome this challenge, adapters are designed to capture small RNAs with a 5' phosphate group, for miRNA recognition. Computational tools are then used to analyze and understand biological implications of the sequence reads. The general steps to sequence data analysis involve read processing, annotating, and characterizing new features. In the read processing stage, the adaptor sequences are identified and removed. Then, the small RNA sequences are mapped back to reference genome sequences, and known miRNAs can be characterized by comparing them with known annotated miRNAs. In addition to measuring miRNA abundance levels from sequence reads, most computation tools have the ability to discover novel miRNAs, determine differentially expressed miRNAs and their associated mRNA gene targets.

miRDeep (Friedlander et al., 2008) and its variants miRDeep2 (Friedlander et al., 2012), miRDeep-P (Yang et al., 2011), miRDeep* (An et al., 2013), was one of the first to apply machine learning to NGS data for miRNA prediction. The core algorithm leverages Bayesian statistics to score the fit of sequenced RNAs to the biological model of miRNA biogenesis. The online pipeline tool predicts miRNAs from small RNA-seq data, provides a target prediction for both known and novel miRNA expression profiles and has a graphical interface to display RNA-seq reads and its predictions. miRDeep2 was tested on seven animal species and reported a high specificity of about 99% across all species, and varying sensitivity, from 71% on sea squirt data to 90% on anemone data (Friedlander et al., 2012). Further, the tool predicted numerous novel miRNAs, many of which are high-confidence candidates where the sequences were detected in at least two independent samples (Friedlander et al., 2012). Since, many tools have been developed and compiled online (Lukasik et al., 2016), including sRNAbench (Barturen et al., 2014) and deepSOM (Stegmayer et al., 2017). Chen et al., (2018) provide a comprehensive review of these tools.

MiRNA Databases

As identifying and characterizing *de novo* miRNA efforts mature, there is an inherent need to annotate and systemize these miRNAs. MiRBase (Kozomara et al., 2010) is a database with an online interface for access to miRNA sequence data, annotation and predicted gene targets. This registry provides a centralized system for assigning names for new miRNAs, thus providing a consistent naming system for miRNAs. Each entry in the database represents a predicted hairpin portion of a miRNA transcript, with information on the location and sequence of the mature miRNA sequence. In addition, it provides experimental evidence for each miRNA, and links the miRNAs to its target genes predicted by other tools. The latest miRBase release of 2018 has a total of 38,589 entries, representing 48,860 mature miRNA products in 271 species, 2654 of which are from the human genome (Kozomara et al., 2019).

Challenges persist with discovering and annotating *de novo* miRNAs. It is often difficult to distinguish between functional small RNAs and non- functional 'noise', and is reflected in the miRBase database. Even in commonly studied animal organisms, such as human and mouse, there's not enough information to support or refute the validity of 30% to 70% miRNA annotations (Kozomara et al., 2019). Evolutionary conservation confers compelling evidence for the functionality of predicted miRNAs, but these miRNAs will eventually need to be experimentally validated to prove their functionality.

The MiRNA Target Model Hypothesis

The biology of miRNAs and their functionality is a fairly young field, and thus few conclusions on how a miRNA targets a gene have been agreed upon. It has been documented that miRNAs mainly recognize complementary sequences in the 3'-untranslated regions (UTRs) of their target mRNAs (Bartel, 2004). However, later experiments have reported that they can also bind to their 5'UTR or the Open Reading Frame (ORF) (Lytle et al., 2007; Moretti et al., 2010). Typically, this binding down-regulates the expression of the gene by either blocking translation or attracting factors that degrade

the mRNA (Huntzinger et al., 2011). Thus, the miRNA-target model related to predicting an increase or decrease in expression has not been clearly defined. Most miRNA target gene-prediction methods follow the original model that miRNAs bind to the 3'UTR and down-regulate the target miRNA, although some methods attempt to incorporate more complexity such as allowing for multiple binding regions per target gene.

In early miRNA studies, investigators found that although target sites for miRNAs could be computationally identified in both 3'UTRs and 5'UTRs, the miRNA-mRNA duplex formation was far more pronounced in the 3'UTR region (Lewis et al., 2005). Thus, subsequent bioinformatic and experimental analysis has considered the 5'-end of the miRNA (the seed site) to be most important for the binding to the mRNA. Further, the target sites have been divided into three main classes, according to grade and localization of sequence complementarity (Brennecke et al., 2005). The first class is the dominant seed site targets, called the "5' seed-only" site. The second is the 5' dominant canonical seed site targets, called the "5' dominant" site, and the third is the 3' complementary seed site targets are called the "3' canonical" site. Considering that there are various rules regulating the interaction between a miRNA and its target mRNA, it is not surprising that each miRNA has the potential to target a large number of genes (Friedman et al., 2009).

Predicting Target Genes

Computational approaches play an important role in the identification of miRNA targets of specific genes. Several approaches have been used to successfully identify potential miRNA targets in mRNA sequences for experimental validation. The majority of first-generation methods are based on three major assumptions; 1) miRNAs are perfectly or near-perfectly complementary to their targets, 2) when the miRNA is bound to the target, the RNA-RNA duplex has a higher negative folding free energy, and 3) mature miRNAs are highly conserved from species to species (Yoon et al., 2006). First-generation methods include TargetScan (Lewis et al., 2005; Agarwal et al., 2015), DIANA-microT

(Kiriakidou et al., 2004; Paraskevopoulou et al., 2013) and miRanda (Enright et al., 2003, John et al., 2004). These are followed by machine learning based methods such as miRanda-miRSVR (Betel et al., 2010), PicTar (Krek et al., 2005) and HomoTarget (Ahmadi et al., 2013); and experimentally-driven tools such as PITA (Kertesz et al., 2007) and miRWIP (Hammell et al., 2008). Further methodologies, such as BCmicrO (Yue et al., 2012) and ComiR (Coronnello et al., 2012) combine existing algorithms.

TargetScan

TargetScan (Lewis et al., 2005) was the first method to explicitly use the concept of seed matches in predicting miRNA targets in vertebrates. The algorithm combines thermodynamics modeling of RNA-RNA duplex interactions with comparative sequence analysis to predict miRNA targets conserved across more than one genome. To accomplish this, the following algorithm is iterated, with inputs A) a miRNA conserved across multiple organisms and B) a set of orthologous 3'UTR sequences from these organisms :

For each 3'UTR region of each of the organisms whose comparative genomes are being used in the study:

1) Search the UTRs in the organism for segments of perfect Watson-Crick complementarity to bases 2-8 of the miRNA (the "miRNA seed"), the perfect complementarity to the seed is called a "seed match".

2) Extend each seed match with additional bases to the miRNA, allowing G:U pairs. The extension is in both the 3' and the 5' directions and stops when mismatches are found.

3) Optimize base pairing of the remaining 3' portion of the miRNA to the 35 bases of the UTR immediately 5' of each seed match using RNAfold, a secondary RNA structure prediction program (Hofacker et al., 1994).

4) Assign a folding free energy G to each miRNA-target site interaction using RNAeval, a free energy evaluator of RNA molecules with a fixed secondary structure (Hofacker et al., 1994).

5) Assign a Z score to each UTR, with the following equation:

$$Z = \sum_{k=1}^{n} e^{-G_k/T}$$

where,

n is the number of seed matches in the 5'UTR region

G_k is the calculated free energy (kcal/mol) of the interaction between the miRNA and its target for the k^{th} target evaluated in the previous step

T is a parameter that influences the relative weighting of UTRs as a function of the affinity and abundance of their target sites; *T* values are assigned by a trial-and-error method involving training and test sets of miRNAs.

6) Sort the UTRs in this organism by Z score, and assign a rank R to each, get the highest Z score

Until the Z score reaches a value higher than a predefined cut-off.

Following the publication of the TargetScan method, several improvements have been made (Agarwal et al., 2015). First, new organisms are constantly added to the working set, which improved the signal-to-noise ratios. Second, less conservative binding interactions with less than perfect pairings and bulges (insertion or deletion of a nucleotide), especially within a 5' region of the miRNA, are also predicted in the newest version of TargetScan. The conservation of seed regions among orthologous 3'UTRs within miRNA binding regions is important for the outcome score. The conservation level of the targets can be defined by the user as broadly conserved (across vertebrates) or highly conserved (across most mammals). The TargetScan research group (Friedman et al., 2009) have found a pattern of consecutive GC-rich base pairs in a set of known miRNA binding sites in *C. elegans,* and this pattern has been included in the scoring scheme of the algorithm. TargetScan ranks the prediction by two parameters: Context score and Probability of conserved targeting (Pct) (Lewis et al., 2005; Agarwal et al., 2015).

DIANA-microT

The DIANA-microT miRNA-target predicting algorithm (Kiriakidou et al., 2004) uses a 38 nucleotide-long frame that is moved along the 3'UTR and measures the minimum energy of potential miRNA binding sites (allowing for mismatches) after every shift. It compares this energy with the energy of a perfect complementary sequence bound to the 3'UTR region. The algorithm searches for sites with a canonical central bulge and it requires 7, 8, or 9 nucleotide-long complementarity within the 5' region of the miRNA. Both conserved and nonconserved sites are considered. Finally, a signal-to-noise ratio is computed for each miRNA; where the signal is the number of predicted targets of a single miRNA and the number of predicted targets of an artificial miRNA with randomized sequence in searched 3'UTR estimates the noise (Kiriakidou et al., 2004). The algorithm has been since published on a web server incorporating the latest miRBase version (Paraskevopoulou et al., 2013).

miRanda-miRSVR

The miRanda miRNA-target prediction algorithm was first developed using all known miRNAs of *D. melanogaster* (Enright et al., 2003), and then the three-step algorithm was extended to humans and other vertebrates (John et al., 2004). In the first step, the miRNAs are matched against the 3'-UTR regions of all possible targets allowing for wobbling (non-Watson-Crick base pairing), G:U base pairs and indels (insertions or deletions). The second step computes the thermodynamic stability of the miRNA:target duplex. The final step is a valuation of the evolutionary conservation of the miRNA:target duplex across two additional species. In miRanda, miRNAs with multiple binding sites within the 3'UTR region are promoted, which contributes to the increase in specificity, but it underestimates miRNAs with a single but perfect base pairing with their targets (John et al., 2004). Further, the method was expanded to include a Support Vector Regression (SVR) model, a variant of the SVM algorithm, to train on sequence and contextual features extracted from miRanda predicted target sites (Betel et al., 2010). These features include secondary structure accessibility of the site and conservation, without the need to define seed subclasses. It was trained on mRNA expression fold changes following miRNA

transfections. The miRNA target sites are ranked by the downregulation score, named mirSVR. This score is calibrated to correlate linearly with the extent of down regulation, and can be interpreted as an empirical probability of down regulation. This algorithm has identified a number of experimentally determined non-canonical and non-conservative sites (Betel et al., 2010). Unless otherwise specified, this algorithm will be referred to as miRanda for short.

PicTar

PicTar (Krek et al., 2005) is a machine learning algorithm that scans the alignments of 3' UTRs for near-perfect miRNA seed matches and filters the alignments according to their thermodynamic stability. Each predicted target is scored using a Hidden Markov Model (HMM – a simple example of a dynamic Bayesian network) maximum-likelihood fit approach. Thus, synergistic effects of multiple binding sites of several miRNAs acting together are accounted for in this model. PicTar utilizes miRNA sequence alignment to mRNAs of eight vertebrate species and it scores the candidate genes of each species separately to create a combined score for a gene (Krek et al., 2005).

HomoTarget

HomoTarget combines a Pattern Recognition Neural Network (PRNN) and Principal Component Analysis (PCA) in an architecture to model the relationship between miRNAs and their target mRNAs in humans (Ahmadi et al., 2013). This method incorporates twelve structural, thermodynamic and positional features of miRNA-mRNA binding sites to select target candidates.

PITA

Experimental studies suggest that target site accessibility is a critical factor for effective target gene repression (Long et al., 2007), where a strong secondary structure is formed by the 3'UTR of the target itself that prevents the binding of the miRNA. Kertesz et al. (2007) systematically examined and confirmed the site accessibility effect in an in-vivo

luciferase system and incorporated this effect into a thermodynamic model. They designed the genome-wide target prediction algorithm called Probability of Interaction by Target Accessibility (PITA), by combining this thermodynamic model with traditional seed-finding procedures with *Drosophila* datasets (Kertesz et al., 2007).

MirWIP

Experimental evidence of co-precipitation has also been included in predicting miRNA targets. For example, in a study with *C. elegans,* 3404 mRNA transcripts were recovered that specifically co-precipitated with miRNA-RISC complex proteins (Zhang L et al., 2007). Following, Hammell et al. (2008) developed a method based on this large data set of high-confidence miRNA-target interactions. This target prediction algorithm, MirWIP, scored miRNA-target sites by weighting site characteristics in proportion to their enrichment in the experimental data set. These important characteristics included structural accessibility of target sequences, total free energy of miRNA-target hybridization, and topology of base pairing to the 5' seed region of the miRNA (Hammell et al., 2008).

BCmicrO

BCmicrO (Yue et al., 2012) combines the prediction of six algorithms (TargetScan, miRanda, PicTar, mirTarget (Wang and Naga, 2008), PITA, and DIANA-microT) with a Bayesian Network. It is trained on positive and negative miRNA-target pairs of all algorithms and gives the probability of an mRNA being a target. BCmicrO was evaluated using mammalian miRNA-target pairs and protein expression data, showing higher sensitivity given the same specificity of each individual algorithm (Yue et al., 2012).

ComiR

ComiR (Coronnello et al., 2012) predicts whether a given mRNA is targeted by a set of miRNA. It applies miRNA expression to four targeting models (miRanda, PITA, TargetScan and mirSVR (Betel et al., 2010)) by identifying all binding sites of each miRNA

in a given mRNA 3'UTR. Then, it additively combines the individual target scores using a Support Vector Machine (SVM) trained on *Drosophila* Ago1 data. It gives a single probabilistic score, higher scores correspond to higher probability of an mRNA being a functional target of a particular set of miRNAs (Coronnello et al., 2012).

The seed hypothesis is an almost universally adopted early feature in miRNA-target prediction methods and is widely used to control for false positives. It was experimentally reinforced by a study that obtained the structure of an important component of the silencing complex bound to a RNA guide-strand, and lays down the biochemical basis for the role of seed sites (Wang et al., 2008). However, there have been certain experimentally confirmed targets that violate the seed rule by including mismatches or wobble G:U pairs (Lewis et al., 2005). Thus, the lack of consistent portrayal of miRNA targets is one of the greatest obstacles not only to the development of better prediction methods, but also to comparing and selecting a prediction tool.

It is clear that the growth of the quantity and quality of experimentally determined miRNA genes and their targets will be the driving force for the next generation of computational miRNA tools. New biological insights into the recognition between miRNA and its targets will inspire computational biologists to create new algorithms based on mechanistic understanding. Large-scale experiments will provide valuable data sets for both initial training and follow-up evaluation of computational methods.

Introduction to Breast Cancer and Potential Implications

MiRNAs play a critical role in multiple biological processes, including cell cycle control, cell growth, cell differentiation, apoptosis, and embryo development (Jiang et al., 2009). Just as miRNAs are involved in regulating the normal functioning of eukaryotic cells, deregulation of miRNA has also been associated with growth abnormality and disease. The basic components of the miRNA-complex have been implicated in human disease, such as Drosha enzyme, an essential miRNA biogenesis co-factor. This cofactor is encoded by the human gene DGCR8, which maps to chromosomal region 22q11.2 and is commonly deleted

in DiGeorge syndrome. This disorder affects one in 3,000 live births, and results in defects including heart defects, immunodeficiency, schizophrenia, and others (Landthaler et al., 2004).

The initial studies (Lee et al., 1993 ; Reinhart et al., 2000) showing evidence of miRNAs involved in regulating cellular differentiation and proliferation encouraged interest in studying miRNAs in cancers. The first study that associated miRNAs with cancer investigated blood samples from chronic lymphocytic leukemia (CLL) patients (Calin et al., 2002). CLL is a type of cancer in which the bone marrow makes too many lymphocytes. The authors investigated whether a tumour suppressor genes could be found in the region of chromosome 13q14, a genomic location that is lost in more than half of CLL patients. Instead, two miRNAs genes, miR-15a and miR-16-1 were found to be absent or down regulated in the majority (approximately 69%) of CLL patients when compared to normal tissue counterparts.

Furthermore, to question the extent of miRNA effects on the cancer genome, Calin et al. (2004) mapped all known miRNA genes on the human genome. They discovered that many of them are located in chromosomal loci prone to deletions or amplifications, as was found in many different human cancer types. In fact, further studies confirmed that chromosomal regions encompassing miRNAs involved in the negative regulation of a transcript encoding a known tumour suppressor gene are amplified in cancer development (Sevignani et al., 2007). This amplification results in the increased expression of miRNAs and consequently silences the tumour suppressor gene. Equally, miRNAs repressing oncogenes are often located in fragile loci, where deletions or mutations can occur and result in reduced miRNA levels and overexpression of the target oncogene. Consequently, alterations of miRNA expression are not rare occurrences, but rather very common in human cancers. Since these initial findings, many studies have provided evidence of miRNAs in various cancers, such as breast cancer (Iorio et al., 2005 ; Zhang et al., 2011 ; Kedmi et al., 2015 ; Yan et al., 2016 ; Thakur et al., 2016), ovarian cancer (Taylor et al., 2008 ; Hausler et al., 2010 ; Jeong et al., 2017), pancreatic cancer (Duell et al., 2017 ; Ho et al., 2010 ; Zhang J et al., 2014), and bone cancer (Huang et al., 2018).

Breast cancer biology

Breast cancer is the most common cancer for women across the world with an estimated 2,100,000 new cancer cases and 533,600 deaths each year (GBD, 2015) . In Canada, one in eight women are expected to be diagnosed with breast cancer in her lifetime. In 2017, about 26,000 Canadian women were diagnosed with breast cancer, with a mortality rate of about 5,000 the same year (Canadian Cancer Society, 2017). Like many cancers, breast cancer is graded with the TNM system (Hortobagyi et al., 2017), where stage 0 is the pre-cancerous or marker condition, and stage 4 is the metastatic cancer, with varying degrees in between. Identifying cancers at stage 0 allows more time for treatment and prevention of growth of the cancer, and thus may reduce the overall cancer mortality rate.

There are several ways to classify breast cancers, some of which indicate high risk of prognosis and treatment response. Breast cancer is usually classified by its histological appearance. The most common type of BC in women originates from the epithelium lining the ducts, and is known as ductal carcinoma (Eheman et al. 2009). There are two types of ductal carcinoma: ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC). DCIS is growth of pre-cancerous cells confined to the mammary ducts of the breast, and is considered stage 0 cancer. Whereas IDC is abnormal proliferation of cancerous cells invading the surrounding tissues, and is thus malignant.

At the cellular level, breast cancers are divided into three major categories, based on their expression of specific receptors as assessed by immunohistochemistry (IHC).

(i) ER-positive tumours display elevated expression of the estrogen receptor (ER) in approximately 80% of breast carcinomas, often in combination with overexpression of the progesterone receptor (PR) in 70-80% of cases (Lakhani et al., 2012);

(ii) HER2-positive tumours are characterized by amplification of the human epidermal growth factor receptor (HER2) in about 15-20% of breast carcinomas (Lakhani et al., 2012);

(iii) triple-negative tumours do not display increased expression of any of these three markers, and is the most heterogeneous group histologically, and genetically.

Furthermore, molecular classification of breast tumours have become possible based on analysis of gene expression profiles of breast cancer sample cohorts. Two luminal subtypes (A and B) exhibit ER positivity and have better survival than other subtypes. Luminal B tumors are characterized by increased expression of proliferation-associated genes and have a worse prognosis than luminal A tumors (Sorlie et al., 2003). The molecular HER2+ subtype highly overlaps with the analogous classical subtype and is characterized by proliferation genes. Finally, the basal-like subtypes is enriched for genes expressed in basal epithelial cells (Sorlie et al., 2001), and there is approximately 60% overlap between triple-negative and molecular basal and normal-like subtypes (Vuong et al., 2014). Meta-analysis of gene expression studies suggest that the prognostic impact of different signatures is related to the proliferation-associated genes (Wirapati et al., 2008). Further studies have yielded other molecular subgroups, including a molecular classification based on integrated genomic and transcriptomic profiling of 2,000 breast tumors yielding 10 novel subtypes of breast cancer with distinct clinical outcomes (Curtis et al., 2012; Ali et al., 2014).

Breast cancer detection

Currently, mammography is the standardized breast cancer screening technology used in clinical settings, with the aim to identify and treat breast tumours before they become symptomatic using low-energy X-ray imaging. Mammographic screening has been relatively successful, as it has increased early detection of breast cancer, and is believed to have contributed to an increased survival rate of 15% in Denmark (Jorgensen et al., 2010), 20% in the UK (Marmot et al., 2013), and more recently 9% in Ireland (Hanley et al., 2017). Unfortunately, systematic screening will result in some women receiving a cancer diagnosis (false positives), even if their cancer would not have metastasized leading to a poor prognosis. In Canada, a 25-year follow up study reported that up to 50% of mammographically-detected invasive breast cancers represent examples of overdiagnosis in women aged 40 to 59 (Miller et al., 2014). A review of seven trials that involved 600,000 women aged 39 to 74 reported biases in the studies and questioned the long-term effects of overdiagnosis and overtreatment due to systematic mammographic screening (Gøtzsche et

al., 2013). The sensitivity of mammography to detect BC is inversely correlated with the density of a woman's breasts, since the opacity of dense breast tissue is difficult for X-rays to traverse (Couzin, 2005). For this reason, mammography often fails to detect BC in young women and older women using menopausal hormone therapy, which affects the density of breast tissue. In some cases, mammography reports breast abnormalities that simply do not exist upon further investigation using follow-up mammographic tests, MRI, ultrasound, PET/CT scans, or needle/surgical biopsies (Orel et al., 1999). In addition, the risks associated with some of these scans outweigh the possible benefits as these procedures expose the patient to harmful radiation (Choosing Wisely, 2012), (Carlson et al., 2009), for example, Jacobsen et al. (2015) showed cumulative risks ranging between 9% to 45% after 8 screens in different populations. Finally, the emotional and psychological stress caused by such false positives is well-documented (Brodersen et al., 2013; Solbjør et al., 2018).

Thus, despite the relative success of mammographic screening in reducing breast cancer mortality, its limitations illustrate the need for more accurate detection tools to identify a potential cancer early, such as biomarkers.

MiRNAs as potential biomarkers of breast cancer

When searching for biomarkers as an early detection tool, there are many things to consider. First, the obvious being the accuracy (high specificity and sensitivity) and robustness of the test. Then, there's the practicalities of clinical and laboratory procedures, such as availability of samples, types of samples (frozen samples vs formalin-fixed paraffin-embedded (FFPE) tissues), and its clinical validity and utility (Harris et al., 2016). Finally, the ideal biomarker should be detectable by minimally invasive sampling procedures.

MiRNAs possess several features supporting their possible use as novel and robust diagnostic biomarkers. Due to their small size, miRNA levels are remarkably stable in tissue samples, serum and plasma. For example, Turchinovich et al. (2011) showed that extracellular miRNA remains stable in blood plasma for at least one month. These miRNAs are protected from RNase-dependent degradation by several mechanisms, including their inclusion in microvesicles, exosomes, and apoptotic bodies, as well as through the

formation of protein-miRNA complexes resistant to degradation (Chen et al., 2008). It has been demonstrated that miRNAs can be efficiently extracted and evaluated from formalin-fixed paraffin-embedded (FFPE) tissues. MiRNAs from FFPE showed improved stability and maintained the same expression profiles when compared with those from frozen samples (Xi et al., 2007).

In tumor tissue. Expressions of certain miRNAs are found to negatively correlate with breast cancer tumor development: miR-335 affects the upstream BRCA1-regulatory cascade (Heyn et al., 2011); miR-27a is associated with a reduced familial breast cancer risk (Yang et al., 2010); and miR-98 is decreased in ductal carcinoma breast cancers (Farazi et al., 2011). While other miRNAs are found to positively correlate with breast cancer tumor development: miR-210 correlates with hypoxic gene expression (a consequence of the growth of a malignant tumour) (Camps et al., 2008); miR-125b predicts poor survival by depression of its target gene ETS1 (Zhang Y. et al., 2011); miR-146a is linked to earlier onset of breast cancer by targeting BRCA1 and BRCA2 (Pastrello et al., 2010); and miR-21 is associated with advanced tumor stage (Yan et al., 2008).

In blood. Circulating miRNAs from blood samples are an especially attractive source of biomarkers, because of their non-invasive nature, and early signs of tumor development have been identified in blood; whereas tumour tissues require biopsies, and are not always available for molecular analysis (Guttery et al., 2013). A note on nomenclature, although most studies use circulating and blood-sourced miRNA interchangeably, technically circulating miRNAs can be found in other surrogate tissues, such as urine and any other body fluids. Blood-sourced miRNAs can also be extracted from whole blood cells, or plasma (whole blood medium without the white and red blood cells), or serum (remaining medium after clotting factors have been extracted). In one of the first blood-sourced RNA studies, Heneghan et al. (2010) surveyed a panel of 7 candidate miRNAs in whole blood RNAs from 148 IDC breast cancer patients and 44 age-matched and disease free controls. They found the expression of miR-195 to be significantly elevated in breast cancer patients, as compared to the control samples. In addition, they observed a significant reduction in

miR-195 in post-operative whole blood samples, compared to the pre-operative samples of the same patients. Schrauder at al. (2012) found miR-335, among other miRNAs, to be overexpressed in whole blood of early stage breast cancer patients compared to healthy controls, and to be involved in regulating target genes in several oncogenic signal-pathways, such as p53. Cuk et al. (2013) have found a panel of plasma microRNAs (miR-127-3p, miR-148b, miR-409-3p, miR-652 and miR-801) that can detect early stage BC. These studies suggest further inquiry into developing blood-sourced miRNA biomarkers for early breast cancer detection is required.

There are limitations to using blood-based miRNA-profiling of BC. The measurement of a miRNA profile represents a secondary response of blood cells during tumorigenesis, thus the main concern is the reduction of the testing accuracy compared to biopsy of breast tissue (Heneghan et al., 2010). However, Hausler et al. (2010) did indicate that the changes in the miRNA profile of blood cells from BC patients did reflect tumor-specific host-reactions, thus showing that tumor signals can be found and measurable in whole blood. In running these experiments, the high protein content of whole blood could be a problem for miRNA-extraction, thus many studies separate the sera and plasma. There are also discrepancies in previous studies, with miRNAs showing different expression directions and many associations are one-time studies without a thorough follow-up. As an example, Heneghan et al. (2010) showed a significantly higher expression of let-7a and miR-195 in whole blood cells of BC cases compared to controls, whereas Schrauder et al. (2012) could not reproduce the results. Possible reasons for the discrepancy are differences in sample handling, detection methods, and patient selection.

The need for an early detection biomarker is clear, but the limitations of blood-sourced miRNA profiling lead us to search for a more comprehensive technique in identifying early detection biomarkers for breast cancer. Thus, investigating biomarkers in both tumor tissue and blood is the next natural approach.

Previously, a comparison of blood mRNA profiles of BC patients vs. their controls across four NOWAC independent datasets identified a gene signature that reports the

presence of breast cancer (Dumeaux et al., 2015). The signature was specific to BC, classifying women with other non-breast carcinoma as negative. Pathway and gene set analysis revealed genes involved in immune processes, cell growth/proliferation, APP pathway, and MYC target genes - all important in cancer development and growth. This study shows that processes found deregulated in blood cells reflect a deficit in immune functions of BC patients. Thus, peripheral blood cell gene expression can be used to detect the presence of breast cancer.

Chapter 2: Methods

Microarray Technologies

Microarrays are a high-throughput technology that can measure the expression levels of large amount of DNA/RNA in parallel in a single experiment. The principle behind this technology is based on nucleic acid hybridization between target molecules and their corresponding complementary probes. Figure 2.1 illustrates a microarray experiment quantifying miRNA expression, based on relative dye intensities corresponding to miRNAs hybridized to the probes (Wei and Kangcheng, 2009). As fluorescently labeled miRNA strands are hybridized with stationary probes on the array, only strongly paired strands will remain after washing. Total strength of the fluorescent signal will depend on the amount of target sample binding to the probes present. Microarrays use relative quantization, in which the intensity of a feature is compared to the intensity of the same feature under a different condition and the miRNA target of the feature is known by its position.

Figure 2.1. A miRNA microarray experiment (Wei and Kangcheng, 2009).

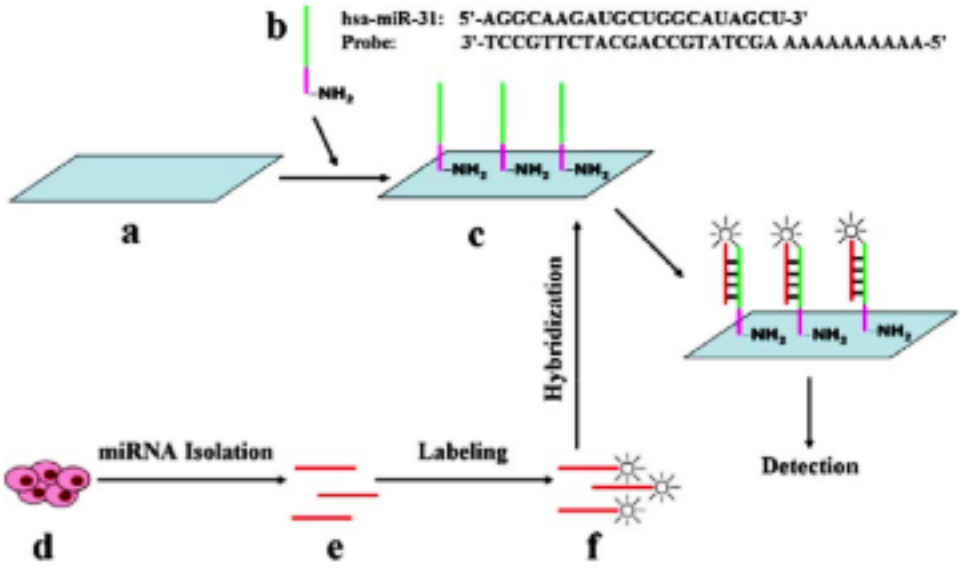

Figure 2.1. A miRNA microarray experiment starts with attaching miRNA probes of a linker and capture sequence to a glass plate. Then, the fluorescently labeled miRNA strands are hybridized with the stationary probes. Finally, fluorescent labels can be detected using the binding signals.

Experimental Data and Analysis

In this study, we analyze miRNA differential expression profiles from the NOWAC population study, and assess their predicted target genes in mRNA expression profiles from the same population. The Norwegian Women and Cancer (NOWAC) study is a national, population-based cohort study among about 170,000 women 30-70 years old, with questionnaire data on lifestyle and health collected at 4-6 year intervals (Dumeaux et al., 2008). At the time of the study, it was a unique combination of a biobank, a description of clinicopathological attributes, and outcomes for a large cohort of breast cancer patients and age-matched controls in a homogeneous population. The biobank comprises blood samples

collected both prior to, and at time of breast cancer diagnosis, with matched breast tissue samples. RNA expression profiles derived from these tissue samples using Illumina beadarray microarrays, were used in this study. We include data from an mRNA expression profile in our analysis described by Dumeaux et al. (2015).

MiRNA profiles from blood samples of breast cancer patients (n=95) were compared to profiles from healthy control samples (n=94). The Illumina humanRef-8 beadchip hybridized 12 samples per chip with probes targeting 1,145 human miRNAs (>97% coverage of the miRNA database at the time). Standard processing and analysis of miRNA expression profile was performed in R (v.2.12), an open-source programming language and software environment, and its associated Bioconductor packages (v.2.09) (Huber et al., 2015). For further information about the data processing, please refer to the Technical Supplements section. Further, we investigated if the miRNAs differentially expressed between breast cancer samples and normal samples. Class distinction was done using Linear Models for Microarray Analysis (Ritchie et al., 2015) package from BioConductor, following standard error adjusting methods (see Technical Supplements). For our final list of differentially expressed miRNAs, see Table 1.

Further, we included data from an mRNA expression profile in our analysis described by Dumeaux et al. (2015). In this study, we had both miRNA and mRNA expression profiles arrayed from matching tissues, allowing us to evaluate target mRNA expression levels. For each miRNA , we identified target genes in the predicted overlapping set of genes (N=612) on the mRNA microarray. Then, a class discovery approach (hierarchical clustering) was applied to these target genes to test if the genes group according to tissue type – cancer and healthy. In combining miRNA and mRNA expression profiles, we are assessing if the target mRNA expression follows the miRNA-target gene model. As previously discussed, the recognized model is that the complementary pairing of miRNAs to the mRNAs of protein-coding genes directs their post-transcriptional repression. Array data suggests that cells with higher miRNA expression should have lower target mRNA expression (Farh et al., 2005; Sood et al., 2006). Although discrepancies have

been reported in the literature, we will be using this simple and most robust model in the following analysis.

The miRNA-target model states that if a miRNA is highly expressed, then the expression of its target gene is expected to be low. This hypothesis describes a negative expression correlation relationship between a miRNA and its biological target gene. Using this hypothesis, we explore the expression correlation relationship between our miRNAs and their predicted targets. For each miRNA-predicted target pair, the correlation of their expression intensity across all samples is calculated using Pearson correlation. Pearson correlation is defined as the covariance of two variables divided by the product of their standard deviations. The result is a correlation coefficient for each miRNA-predicted target pair.

Measuring MiRNA Target Prediction Performance

There are studies that have compared the performance of a few methods with experimental validation, a summary of which is found in Table 3 (Witkos et al., 2011). Based on experimentally supported data sets, Sethupathy et al. (2006a) reported the performance of five individual programs, TargetScan, DIANA-microT, miRanda, and PicTar, and of various combinations of these programs. The specificity and sensitivity were calculated based on a set of experimentally validated mammalian targets from TarBase, a database of experimentally validated miRNA targets (Sethupathy et al., 2006b). They found that miRanda, TargetScan and PicTar have the highest sensitivity, and the intersection of all programs achieved the highest specificity but the lowest sensitivity. On the other hand, the union of all programs achieved the highest sensitivity but the lowest specificity. Thus, none of these three tools individually, nor a combination of, achieved a balance of high sensitivity and specificity. This study (Sethupathy et al., 2006a) shows that there is still much discrepancy between the different target prediction tools. Another comparative study by Baek et al. (2008) applied a quantitative-mass-spectrometry-based approach. They studied the average protein down-regulation of genes predicted by the algorithm to be miR-223 targets. The comparison between *in vivo* results and predictions *in silico* revealed that

TargetScan context scores correlated with protein down-regulation, thus revealing the prediction strength of TargetScan. Combining multiple methods has shown better sensitivity given the same specificity of each individual algorithm (Yue et al., 2012), and a machine learning method that automatically weights the multiple features has shown high specificity and sensitivity (Ahmadi et al., 2013).

Assessment of MiRNA Target Gene Predictions

Previous evidence has shown that there is no obvious one best choice of target prediction tools, and it is not clear if it is more effective to take the union or intersection of targets at the cost of not reaching an optimal balance between sensitivity and specificity (Sethupathy et al., 2006a). Many tools vary greatly in the selected features applied to predict targets, and thus a lack of overlapping predictions is not surprising. Therefore, in order to exhaustively cover the potentially significant features of target prediction, a set of methods should be selected with complementary features. For example, TargetScan and PicTar have been found to have a larger overlap of predicted targets because they both focus on strict seed matching and conservation, and thus may not be the best combination of methods. In addition, there are practical aspects of choosing a methodology to predict target genes. First, is relevancy judged by modernity and popularity – the search for miRNA targets field is young and fast moving, thus tools and methodologies must keep up with the sea of recent findings to stay relevant. Some of the earliest tools created in the early 2000s have been shortly abandoned, while some of the more recent ones are not yet widely accepted by the community. For example, TargetScan was first developed in 2005 and was version 7.2 was released in March 2018, in contrast, the last update for PicTar was in 2007. The second practical aspect is access to source code. Most target prediction tools have a web browser where the user can enter a miRNA and the tool outputs a list of predicted target genes. This user-interface is sufficient when studying a specific miRNA, but is not practical for high-throughput studies that may involve many miRNAs.

In summary, the following three criteria were applied in choosing miRNA target prediction tools for this study:

1) Complementary features
2) Up-to-date and maintained
3) Accessible source code

From many available target prediction tools described above three have been selected: TargetScan, miRanda and PITA. TargetScan emphasizes seed matching and sequence context; miRanda emphasizes looser complementarity and free energy binding; and PITA emphasizes target site accessibility energy. Cross-species conservation varies across the three methods. During the time of this experiment, TargetScan version 6.1 was used, released in March 2012, miRanda was last updated in August 2010, and PITA in August 2008. Lastly, these tools are easily accessible by downloading executable code – with TargetScan and PITA both written in Perl script.

TargetScan, miRanda, and PITA target prediction tools were applied to our list of 38 miRNAs differentially expressed between breast cancer and controls (Table 2) without any thresholds, cutoffs or any stringent criteria applied. Prediction agreement between tools was assessed using a hypergeometric test. The hypergeometric test is a test to see if a random variable follows the hypergeometric distribution, which is a discrete probability distribution describing the probability of success in a number of draws from a finite population containing the successes without replacement. First, the raw results of TargetScan, miRanda, and PITA target prediction tools will be assessed, followed by an investigation into tool-specific thresholds with the objective to reduce the number of false positives and thus identify viable miRNA target genes.

Benchmarking Against Curated Targets

There are two collections of experimentally validated miRNA targets. TarBase (Papadopoulos et al., 2009) is a database which houses a manually curated collection of experimentally supported miRNA targets in several animal species, plants and viruses. At the time of this experiment, we used TarBase version 5.0, which included about 1300 experimentally supported targets. MiRecords (Xiao et al., 2009) is a database of experimentally validated miRNA targets resulting from meticulously curated literature. MiRecords hosts 2286 records of interactions between 548 miRNAs and 1579 target genes in nine animal species. For further analysis, the scores of predicted targets of each tool were compared to the scores of curated targets. The two databases were merged, and the curated targets of the 38 miRNAs were extracted for comparison.

The results of each target prediction method are assessed individually to reduce the number of false positive target predictions. For each method, we plot the target score distributions of all predicted targets. This distribution is then compared to the target score distribution of targets from a curated miRNA-target gene database. If a predicted target is a true target, it is expected to have a score distribution similar to experimentally validated target genes found in curated databases.

The TargetScan method employs two measures for predicting target site efficacy: Context score and Pct. Context score, ranging between -0.6 and 0.2, is the sum of contributions of the following four features: site-type, 3' pairing, local AU and position contributions. The developers of TargetScan modeled the context score to be negatively correlated with target efficacy (Lewis et al., 2005). Thus, the lower the context score, the higher the effectiveness of the targeting. Pct is the probability of conserved targeting ranging between 0 and 1. It reflects the Bayesian estimate of the probability that a site is conserved due to selective maintenance of miRNA targeting rather than by chance. The developers of TargetScan have shown that Pct correlates with the effectiveness of targeting, measured by mRNA amount (Lewis et al., 2005).

The miRanda method incorporates the mirSVR score for predicting target efficacy (Betel et al., 2010). MiRanda uses a support vector regression model to train on mRNA

expression changes given various features, such as secondary structure accessibility of the site. MirSVR score, ranging between -1.5 and 0, measures the likelihood of target mRNA down-regulation. The score does not incorporate target conservation, and thus miRanda may identify non-conserved target sites. The developers of miRanda have shown that the lower the mirSVR score, the higher the effectiveness of the targeting (Betel et al., 2010).

The PITA method ranks the predicted miRNA targets by their free energy score (Kertesz et al., 2007). PITA is based on a thermodynamic model that incorporates measures of accessibility of target sites. The free energy score is the measured change of free energy between unbound 3'UTR of the mRNA and the hybridized state of the miRNA-mRNA duplex. If a given UTR has more than one target sites predicted, then the free energies are summed up together. The score ranges between -35 and 0, and since it is a measure of free energy, the lower the value, the stronger the binding of the miRNA to the given target site is expected to be (Kertesz et al., 2007).

Functional Analysis of Predicted Targets

Several tools are employed to assess the biological functionality of miRNAs and their target genes, such as GSEA and KEGG. Gene Set Expression Analysis (GSEA) is a computational method that determines whether an *a priori* defined set of genes shows statistically significant concordant differences between two biological states (Subramanian et al., 2005). Gene sets are defined based on prior biological knowledge, e.g. published information about biochemical pathways or co-expression in previous experiments. Currently, GSEA employs the MSigDB - a molecular signatures database with 3,272 curated gene sets. Gene sets are collected from various sources such as online pathway databases, publications in PubMed, and knowledge of domain experts. The objective of GSEA is to determine whether members of a gene set tend to occur toward the top (or bottom) of the given ranked gene list, in which case the gene set is correlated with the phenotypic class distinction. This over-representation at the extremes of the ranked gene list is reflected by an enrichment score, which corresponds to a weighted Kolmogorov-Smirnov-like statistic

(Hollander and Wolfe, 1999). GSEA has become a standard in the field to evaluate gene lists from a systems biology perspective.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of online databases of genomes, enzymatic pathways, and biological chemicals (Kanehisa et al., 2006), and is a major resource for pathway analysis. KEGG computerizes data and knowledge on protein interaction networks and chemical reactions that are responsible for various cellular processes. Then, it reconstructs protein interaction networks for all organisms whose genomes are completely sequenced. Lastly, it is utilized as reference knowledge for functional genomics and proteomics experiments. Both resources will be used to assess the biology of our predicted target gene lists.

Further, we explore the biological significance of negatively correlated target genes. For each miRNA-predicted target pair, the Pearson correlation of their expression intensity across all samples is calculated. Then, the targets are ranked according to their correlation coefficient, with the most negative target on the top of the list. This list is assessed by KEGG and GSEA to test for over-representation of gene sets.

Chapter 3: Results

MiRNA Expression Profiles

In our study, we have identified 38 experimentally supported and annotated miRNAs that are differentially expressed in blood samples of breast cancer patients compared to controls (Table 1). Evidence from other studies (Table 1) show that many of these miRNAs are differentially expressed between blood or tumor samples from cancer patients and healthy controls. While it's not a complete list, we focus on comparative studies (blood or tumor samples vs controls) with high-throughput data. Associations with these miRNAs may have been reported in non-cancer contexts, but they are not considered here.

Many of these miRNAs have been found in multiple cancers such as: lung (Chen et al., 2008), esophageal (Zhang C. et al., 2010), pancreatic (Duell et al., 2017; Zhang J. et al., 2014; Ho et al., 2010), osteosarcoma (Huang et al., 2018), ovarian (Jeong et al., 2017; Taylor et al., 2008), and prostate (Moltzahn et al., 2011). For example, miR-223 has been identified in lung cancer sera (Chen et al., 2008), esophageal sera (Zhang C. et al., 2010) and prostate sera (Moltzahn et al., 2011). Some miRNAs seem to be specific to breast cancer only, such as miR-145 (Kodahl et al., 2014; Mar-Aguilar et al., 2013; Thakur et al., 2016) and miR-335 (Schrauder et al., 2012; Heyn et al., 2011). Interestingly, only one miRNA, miR-210 is identified in breast cancers (Thakur et al., 2016; Ng et al., 2013) and other cancers, specifically pancreatic (Ho et al., 2010) and lymphoma (Lawrie et al., 2008). See Table 1 for a complete list.

Supported by previous BC studies, a total of eight miRNAs identified are differentially expressed between BC patients and healthy controls; four of which are upregulated in cancer samples (miR-15b, miR-335, miR-503, miR-637), and four are downregulated in cancer samples (miR-145, miR-210, miR-302c, miR-510) (Table 1). Previous studies support the upregulation of miR-15b in cancer samples, as its upregulated in invasive ductal carcinoma breast tissue (Sakurai et al., 2015); and the downregulation of miR-145, as its downregulated in primary breast carcinoma (Iorio et al., 2005). Otherwise

the results are inconsistent, or no evidence has been identified. Further, three of these miRNAs have been confirmed in previous blood-based studies. Of the upregulated miRNAs, miR-335 is previously identified in blood (Schrauder et al., 2012) and in serum (Wang et al., 2010). Of the downregulated miRNAs, miR-145 is previously identified in serum (Kodahl et al., 2014; Thakur et al., 2016), and plasma (Ng et al., 2013); and miR-210 is previously identified in serum (Thakur et al., 2016) and plasma (Ng et al., 2013).

Target Prediction Results

On average, Miranda predicted the most number of targets (an average of 8,140 targets per miRNA), followed by TargetScan (an average of 4,660), then PITA (an average of 1,694). The overlap between TargetScan and PITA is 632, $P(632 < overlap) > 0.999$. The overlap between TargetScan and miRanda is 2,516, $P(2,516 < overlap) > 0.999$. The overlap between miRanda and PITA is 821 $P(821 < overlap) = 0.99$. The overlap between all three methods is similarly not very large at 612 $P(612<x) = 99\%$ (Figure 3.1). Thus, we conclude that each method produces very different target prediction results for each miRNA.

Figure 3.1. Venn diagram of predictions by TargetScan, miRanda and PITA.

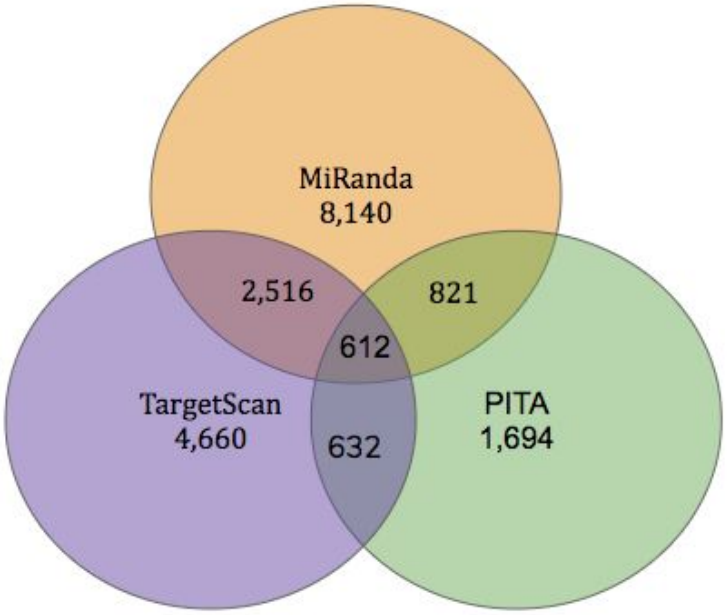

Figure 3.1. There is an overlap between miRNA target predictions by TargetScan, miRanda and PITA. The overlap between all three is 612 compared to each prediction of 4,660 (TargetScan), 8,140 (miRanda), and 1,694 (PITA). A hypergeometric test confirms that the overlap by the three tools is not significant P (612<x) = 99%.

MiRNA-Target Expression Correlations

To visualize the miRNA-target gene relationship, a heatmap of target genes is generated for each miRNA. We show two examples in Figure 3.2, miRNA-335 (Figure 3.2A) and miRNA-210 (Figure 3.2B). Previous studies have shown miR-335 to be over-expressed in cancer samples, and under-expressed in healthy samples (Schrauder at al., 2012). Thus, a significant proportion of its targets (N=81) are expected to be over-expressed in healthy samples and under-expressed in cancer samples. Figure 3.2A however does not illustrate this relationship. In contrast, mir-210 has been shown to be over-expressed in healthy

samples, and under-expressed in cancer samples (Camps et al., 2008, Thakur et al., 2016). Thus, it is expected that a significant proportion of its targets (N=123) to be over-expressed in cancer samples, and under-expressed in healthy samples. Figure 3.2B does not display this expression pattern either. Other miRNAs don't follow the negative association model, or any association for that matter as none of the target heatmaps we produced illustrate any grouping by disease-control status. Thus, we conclude that in our dataset, the miRNA target genes predicted by the three prediction methods are not found in the matching mRNA expression profiles.

Figure 3.2. Class distinction of target genes of miR-335 (A) and miR-210 (B).

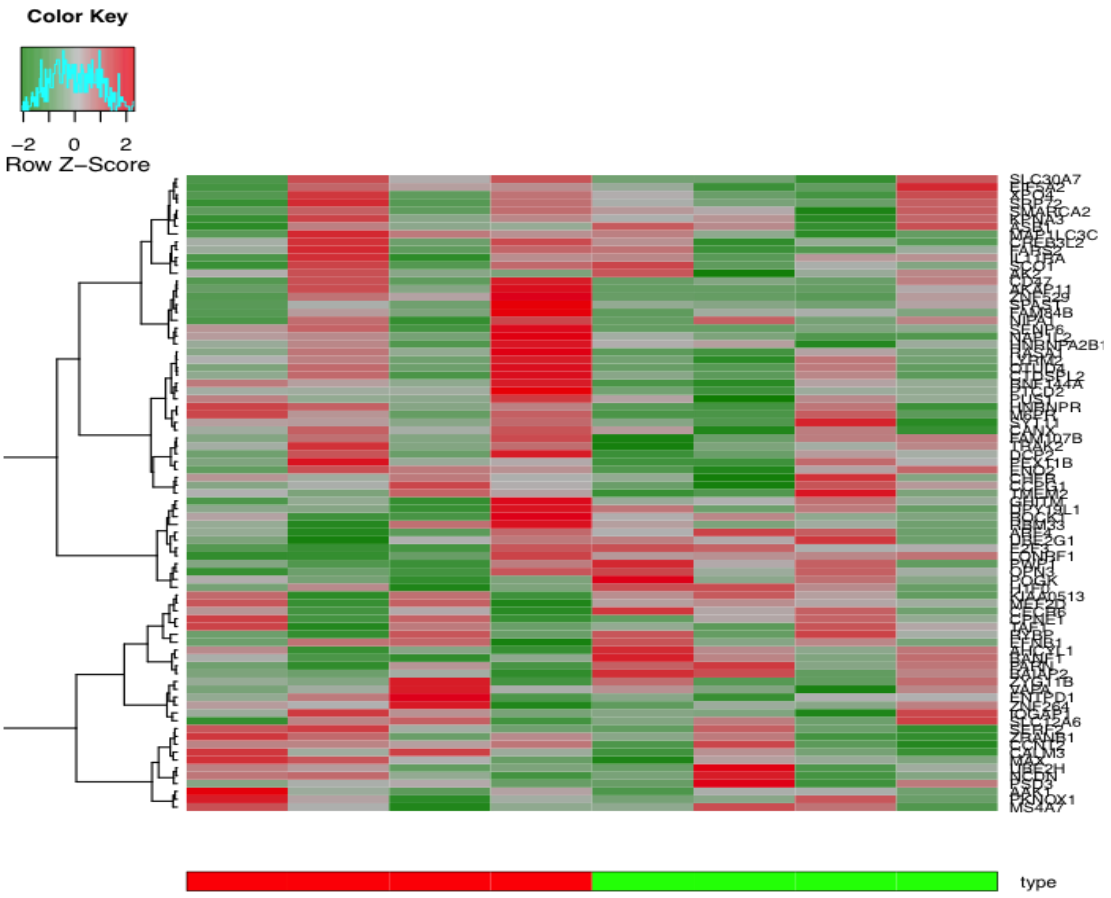

Figure 3.2.B.

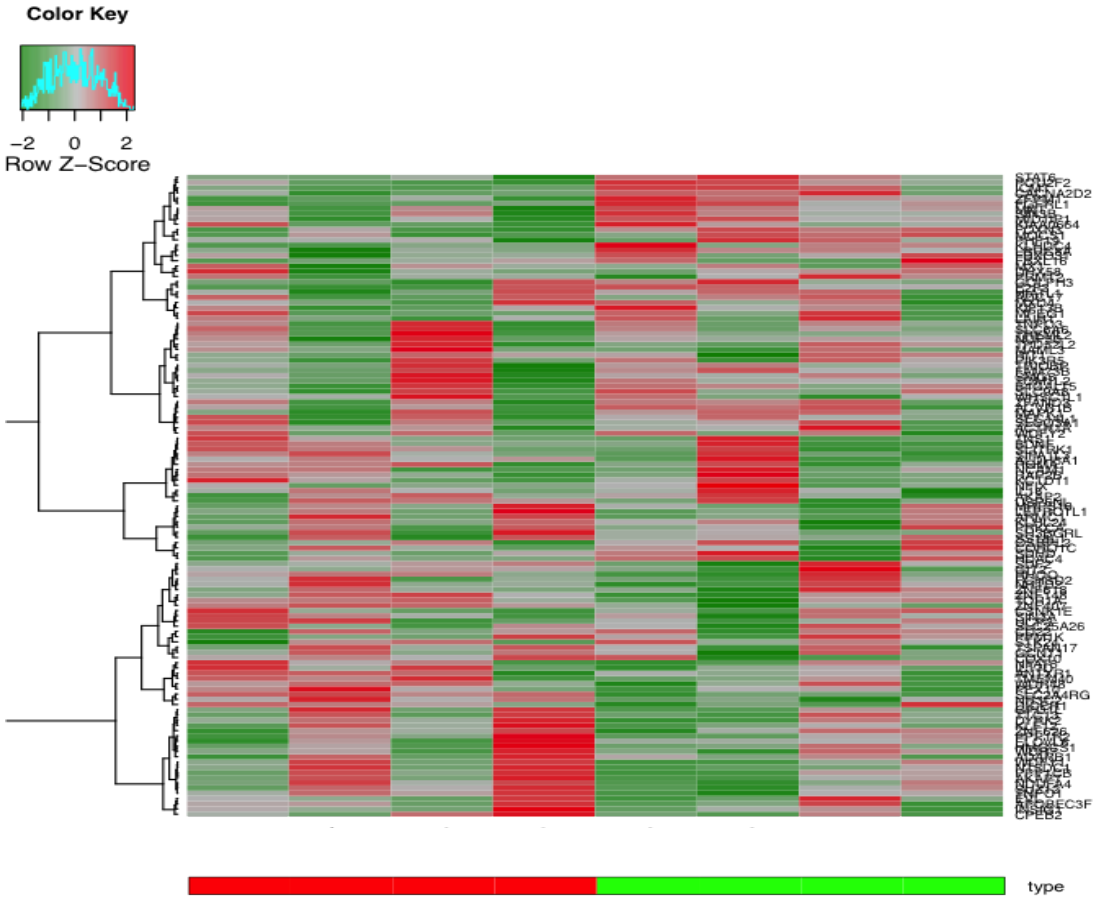

Figure 3.2. The target genes of miR-335 (A) and miR-210 (B) clustered based on the sample types: breast cancer and control samples. Heatmap colors represent mean centered fold change expression in log-space. Sample characteristics are represented in the boxes below each sample. The breast cancer samples are red, and control samples are in green. RNA concentration and expression mean is represented by a grey-red scale, where grey is low and red is high. All samples were hybridized on the same date and same slide.

Furthermore, we assess the density plots for a score threshold of each prediction method, as a quality control measure. If a good threshold is identified, we tested the results against the miRNA-target gene model, as previously described. We plotted the context

score density of all predicted targets by TargetScan with the context score density of curated targets from the databases described above (Figure 3.3). These density plots show that on average, curated targets have lower context scores than all predicted targets as expected. Figure 3.3 illustrates that no specific miRNA has significantly lower context scores than average and no particular miRNA follows the distribution of curated targets. Therefore, context score density plots do not illustrate any well-defined context score threshold to decrease the false positive target predictions.

Figure 3.3. Density graph for context scores of targets predicted by TargetScan.

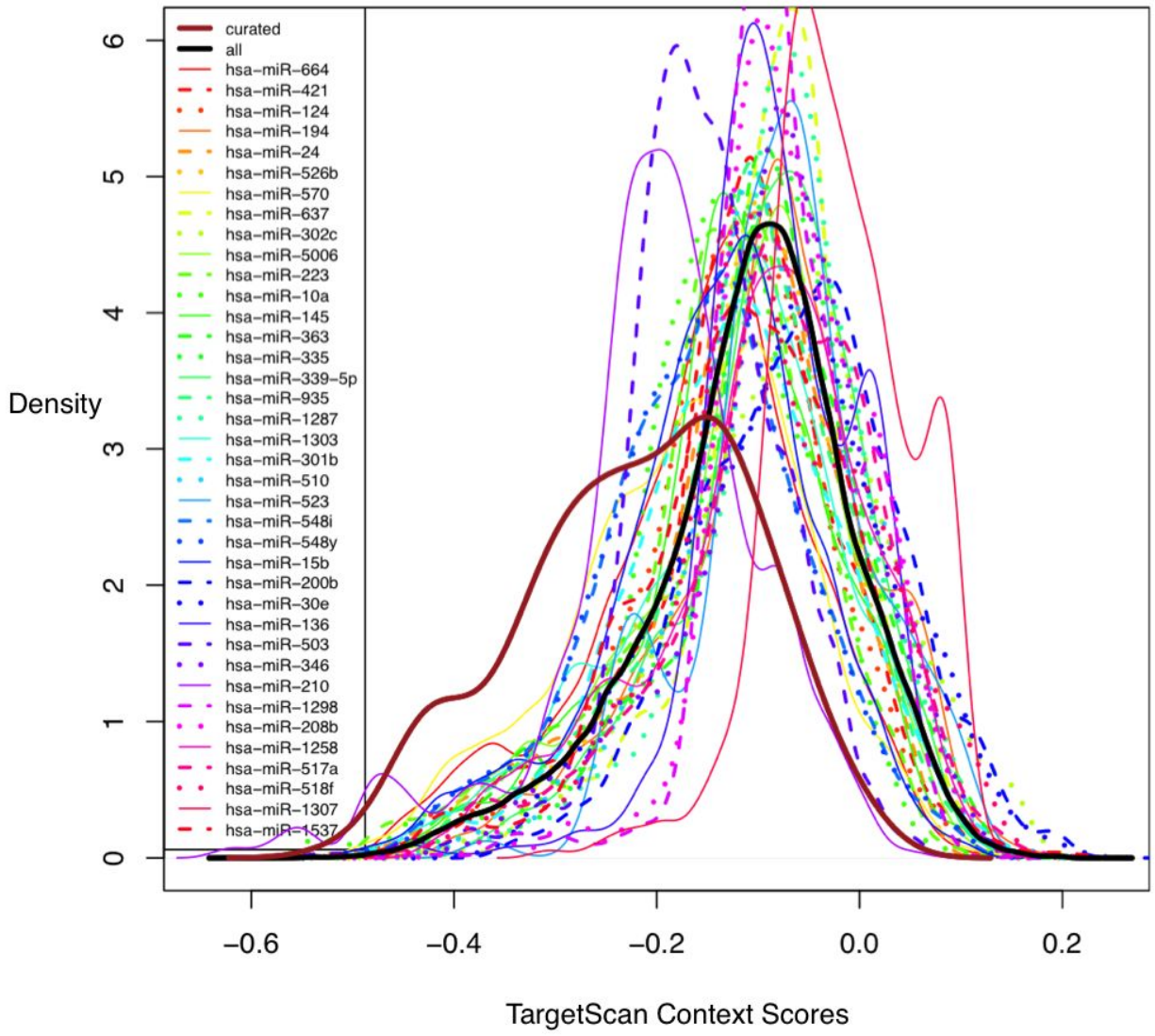

Figure 3.3. Density graph of TargetScan context scores, a metric for predicting target site efficacy. The context score densities of individual miRNAs are of various colors. The black thick line is the context score density of all miRNA-target pairs. The brown thick line is the context score density of curated miRNA-target pairs.

Further, we plotted the Pct score density of all predicted targets by TargetScan with the Pct density of curated targets (Figure 3.4). These density plots show that on average, curated targets have higher Pct scores than all targets, as expected. The density of curated targets increases significantly at a Pct of more than 0.6 – indicating it to be a candidate Pct threshold. Figure 3.4 illustrates a high percentage of targets to have a probability of conserved targeting of approximately 0, suggesting that most TargetScan predicted targets are not well conserved.

Figure 3.4. Density graph for Pct scores of targets predicted by TargetScan.

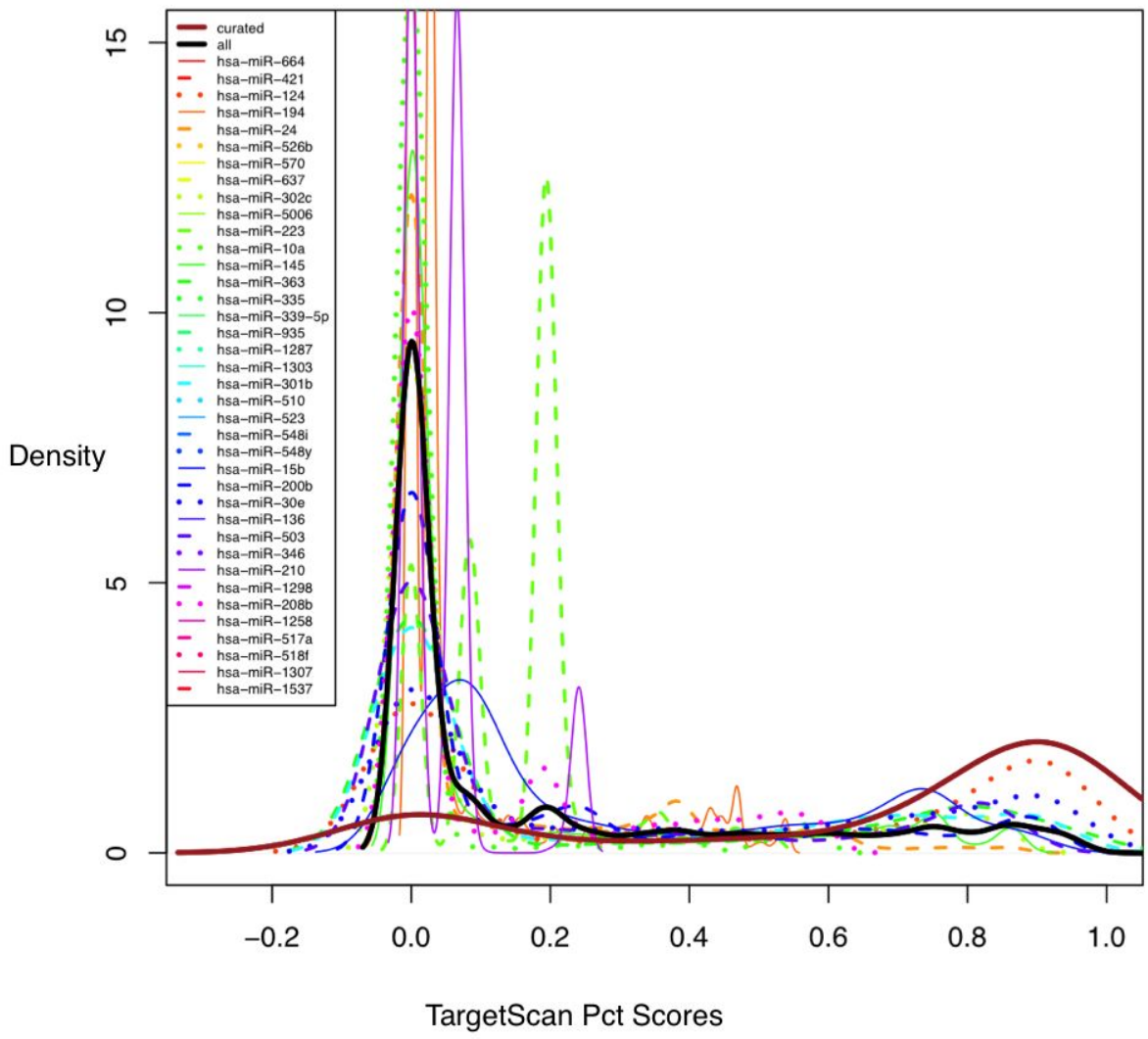

Figure 3.4. Density graph of TargetScan Pct scores, a metric for predicting target site efficacy. The Pct score densities of individual miRNAs are of various colors. The black thick line is the Pct score density of all miRNA-target pairs. The brown thick line is the Pct score density of curated miRNA-target pairs.

We noticed that the targets of two miRNAs have noticeably higher than average Pct scores: miR-30e and miR-208b. However, neither of these miRNAs has been previously implicated in cancer studies (Table 1). To further test the proposition that applying a Pct

threshold is significant in reducing the number of false positive predicted targets, we assess if the target expression profile follows the miRNA–target model described previously. Hierarchical clustering is applied to predicted targets with a Pct score of more than 0.6 to test if these genes group by tissue type. A negative expression correlation is expected as per the model - in samples where a miRNA is over-expressed, its targets should be under-expressed and vice-versa. To visualize this relationship, a heatmap of target genes is generated for each miRNA. For example, miR-30e is over-expressed in cancer samples, and under-expressed in healthy samples (Figure 3.5).

Thus, it is expected that a significant proportion of its predicted targets with the Pct threshold of >0.6 (N=469) to be over-expressed in healthy samples and under-expressed in cancer samples. Figure 3.5 illustrates this not to be the case. In contrast, miR-210 is over-expressed in healthy samples, and under-expressed in cancer samples (Figure 3.2). Thus, it is expected a significant proportion of its targets with the Pct threshold of >0.6 to be over-expressed in cancer samples, and under-expressed in healthy samples. However, none of its targets have a Pct threshold of >0.6. The other miRNAs follow this lack of agreement with the miRNA-target gene model.

Figure 3.5. Class distinction of miR-30e target genes predicted by TargetScan.

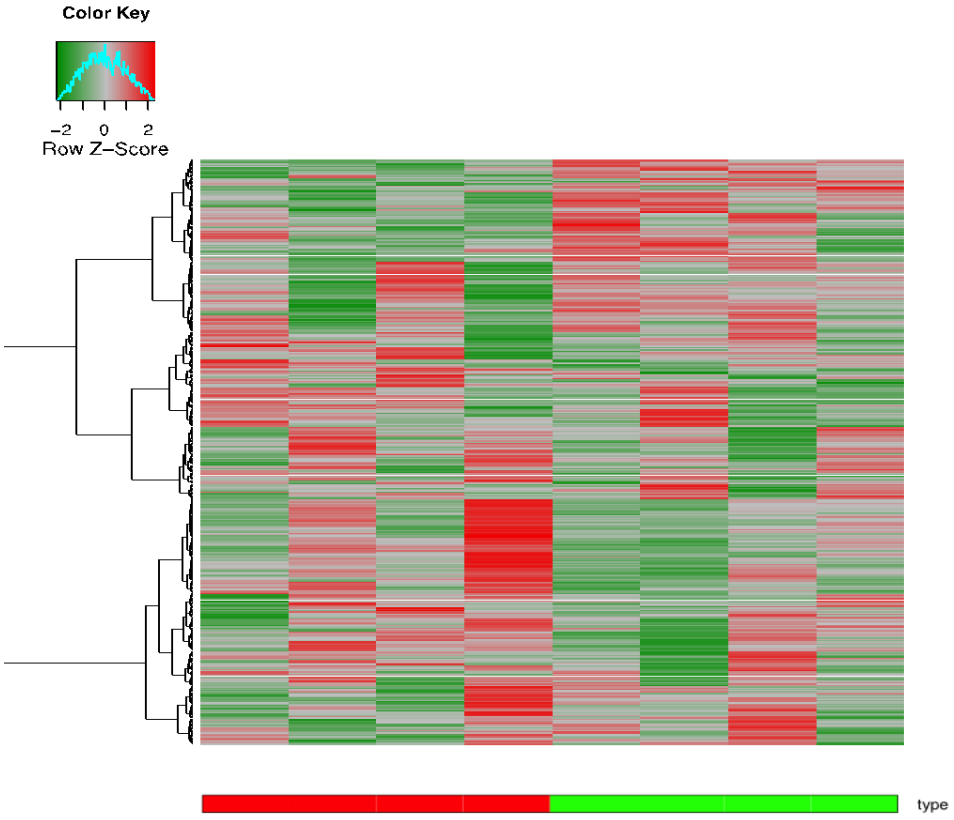

Figure 3.5. The miRNA target genes predicted by TargetScan with Pct cutoff >0.6 clustered based on the sample types: breast cancer and control samples. Heatmap colors represent mean centered fold change expression in log-space. The breast cancer samples are red, and control samples are in green.

We plotted the mirSVR score density of all predicted targets by miRanda with the mirSVR density of curated targets (Figure 3.6). These density plots show that the curated targets do not have a bias towards a lower or higher score, and are evenly distributed. In contrast, the average score hovers around 0, and thus is not significant. The mirSVR density plots do not illustrate any well-defined mirSVR score threshold to decrease the false positive target predictions.

Figure 3.6. Density graph for MirSVR scores of targets predicted by miRanda.

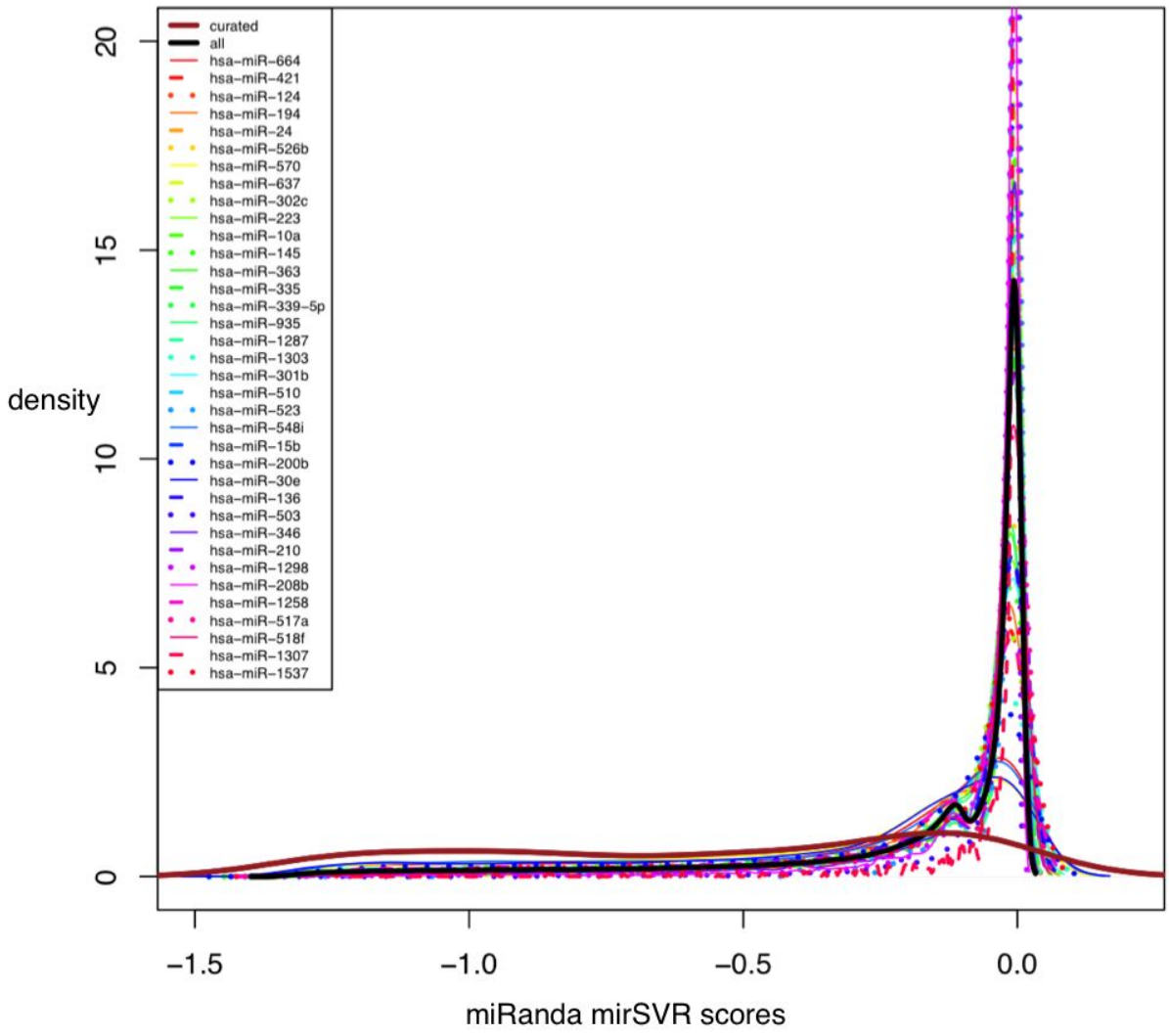

Figure 3.6. Density graph for miRanda MirSVR scores, a metric for predicting target site efficacy. The MirSVR score densities of individual miRNAs are of various colors. The black thick line is the MirSVR score density of all miRNA-target pairs. The brown thick line is the MirSVR score density of curated miRNA-target pairs.

We plotted the free energy score density of all predicted targets by PITA with the free energy score density of curated targets (Figure 3.7). These density plots show bimodal distributions, with peaks near the free energy scores of -5 and -10. One interpretation of these results is that the first peak represents actual targets, whereas the second peak with the higher free energy scores centered at -5 represents a large number of false positives. In support of this hypothesis, the developers of PITA suggest applying a threshold of -10 when filtering for functional targets. If, this were the case, then we would expect to see a significant portion of curated targets with scores of -10 or less. The density plot illustrates this to be the case, as the -10 peak is more than twice as high as the -5 peak. Therefore, -10 may be a viable threshold to reduce the number of predicted false positives. To test if applying a PITA score threshold is significant to reduce the number of false positive predicted targets, we assess if the target expression profile follows the miRNA–target model. Hierarchical clustering is applied to predicted targets with a PITA score of less than -10 to test if these genes group by tissue type. A negative correlation is expected as per the model; in samples where a miRNA is over-expressed, its targets should be under-expressed and vice-versa. To visualize this relationship, a heatmap of target genes is generated for each miRNA.

Figure 3.7. Density graph for PITA scores of targets predicted by PITA.

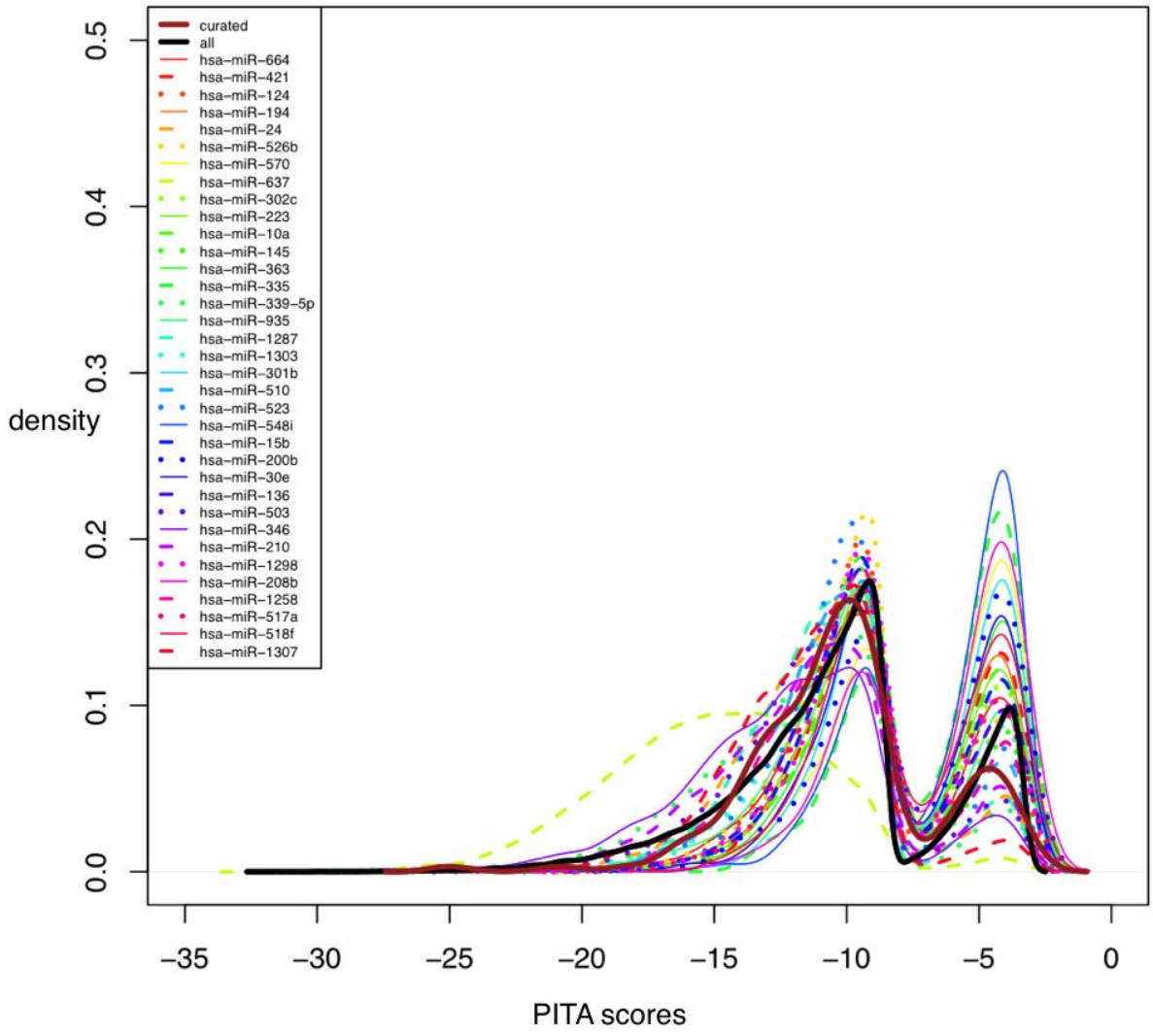

Figure 3.7. Density graph of PITA scores, a metric for predicting target site efficacy with a measure of the free energy. The PITA score densities of individual miRNAs are of various colors. The black thick line is the PITA score density of all miRNA-target pairs. The brown thick line is the PITA score density of curated miRNA-target pairs.

For example, miR-30e is over-expressed in cancer samples, and under-expressed in healthy samples (Figure 32). Thus, it is expected a significant proportion of its predicted targets with the PITA score threshold of < -10 (N=243) to be over-expressed in healthy samples and under-expressed in cancer samples. Figure 3.8A illustrates this not to be the case. In contrast, miR-210 is over-expressed in healthy samples, and under-expressed in cancer samples (Figure 3.2). Thus, it is expected a significant proportion of its targets with the PITA score threshold of < -10 (N=513) to be over-expressed in cancer samples, and under-expressed in healthy samples. Figure 3.8B does not display an expression pattern either. In addition, none of the predicted targets for the rest of the miRNAs within the threshold illustrate positive results.

Figure 3.8. Class distinction of miR-30e (A) and miR-210 (B) target genes predicted by PITA.

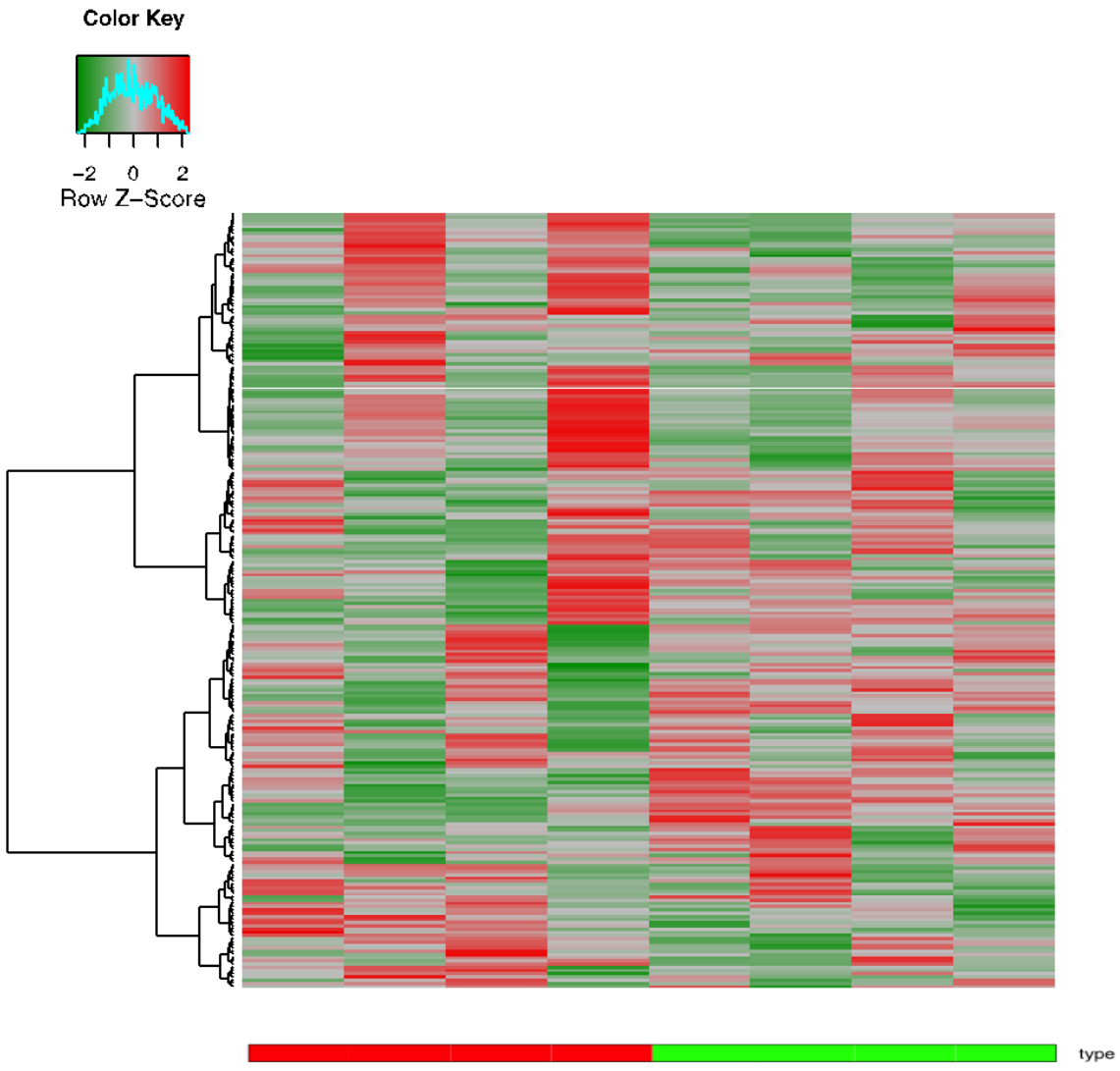

Figure 3.8.B.

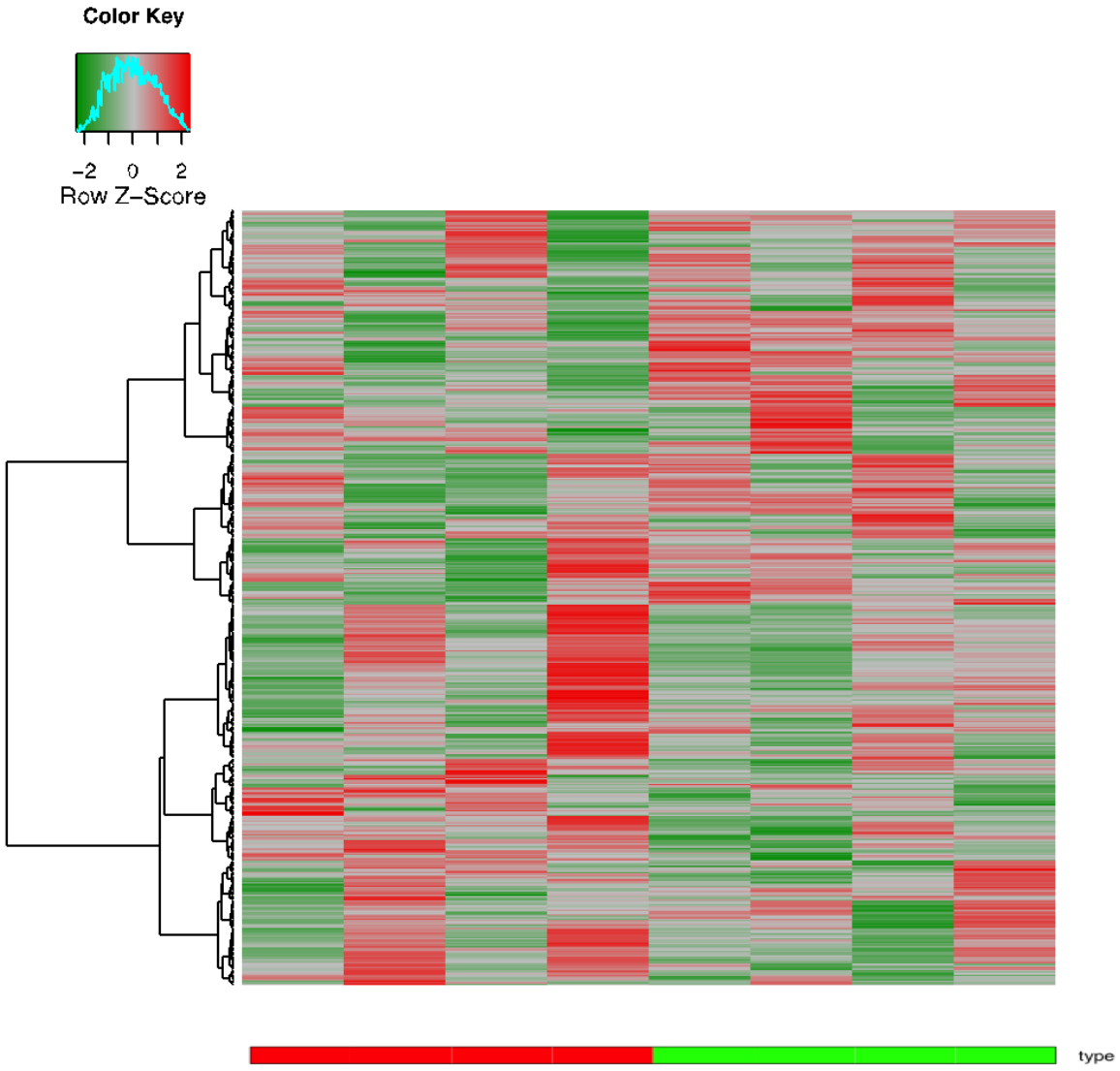

Figure 3.8. The PITA predicted target genes of miR-30e (A) and miR-210 (B), with a score cutoff < -10, clustered based on the sample types: breast cancer and control samples. Heatmap colors represent mean centered fold change expression in log-space. The breast cancer samples are red, and control samples are in green.

In conclusion to an effort to reduce false positives in the large target gene set predicted by TargetScan, PITA and miRanda, identifying a target prediction tool score cutoff, we have not identified a viable option. TargetScan's Pct score and PITA's score plots illustrated potential cutoffs, however the genes within those cutoffs did not fit the miRNA-target model described. To further explore the relationship between a miRNA and its target genes, we decided to employ the predicted targets of TargetScan for further analysis.

The correlation coefficients of 38 miRNAs were plotted against their density (Figure 3.9). The resulting correlation distributions fall into two categories: an approximately normal distribution centered on a correlation coefficient of 0, and a bi-modal distribution centered near -0.5 and 0.5. A normal distribution describes the coefficients being equally positive and negative, thus not supporting the correlation hypothesis. A bi-modal distribution also describes equal weight on both sides. Finally, none of the miRNAs illustrate a strong negative correlation with its predicted targets.

Figure 3.9. Density graph for correlation coefficients of the 38 miRNAs.

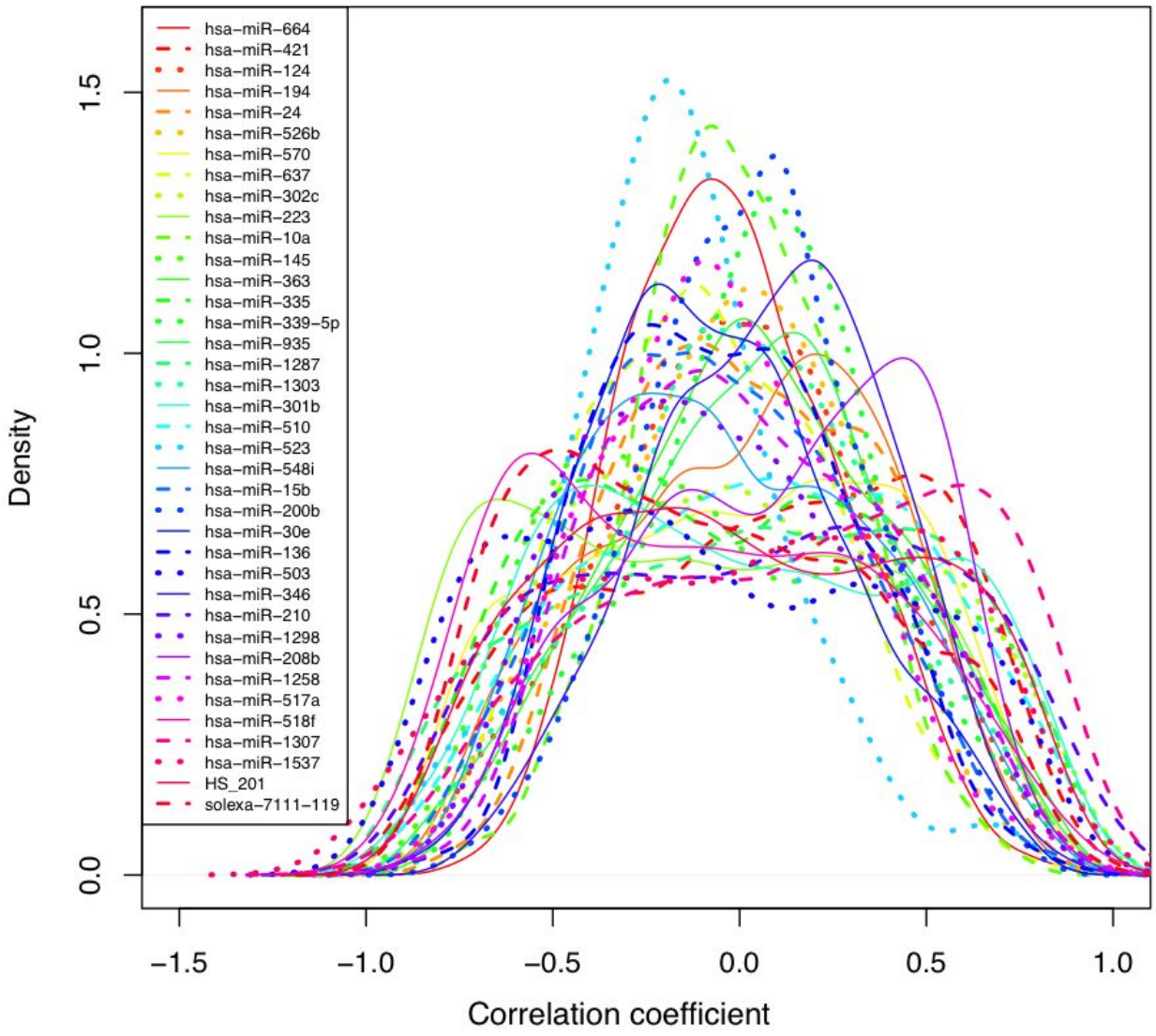

Figure 3.9. Density graph for MiRNA-Target expression correlation coefficients. The correlation coefficients densities of individual miRNAs are of various colors. The black thick line is the correlation coefficients density of all miRNA-target pairs. The brown thick line is the correlation coefficients density of curated miRNA-target pairs.

To further assess the negative correlation relationship between the miRNA and its predicted targets, a selected cutoff of < -0.5 is applied to the correlation coefficient. Hierarchical clustering is applied to predicted targets with correlation coefficient < -0.5 to test if these genes group by tissue type. To visualize the negative correlation relationship, a heatmap of target genes is generated for each miRNA. For example, miR-30e is over-expressed in cancer samples, and under-expressed in healthy samples (Figure 3.2). Thus, it is expected a significant proportion of its predicted targets with the correlation coefficient of < -0.5 (N=213) to be over-expressed in healthy samples and under-expressed in cancer samples. Figure 3.10A illustrates this not to be the case. In contrast, miR-210 is over-expressed in healthy samples, and under-expressed in cancer samples (Figure 3.2). Thus, it is expected a significant proportion of its targets with the correlation coefficient of > -0.5 (N=42) to be over-expressed in cancer samples, and under-expressed in healthy samples. Figure 3.10B does not support this hypothesis either. The same method was applied to other miRNAs, and the same observations were made.

Figure 3.10. Class distinction of predicted target genes of miR-30e (A) and miR-210 (B).

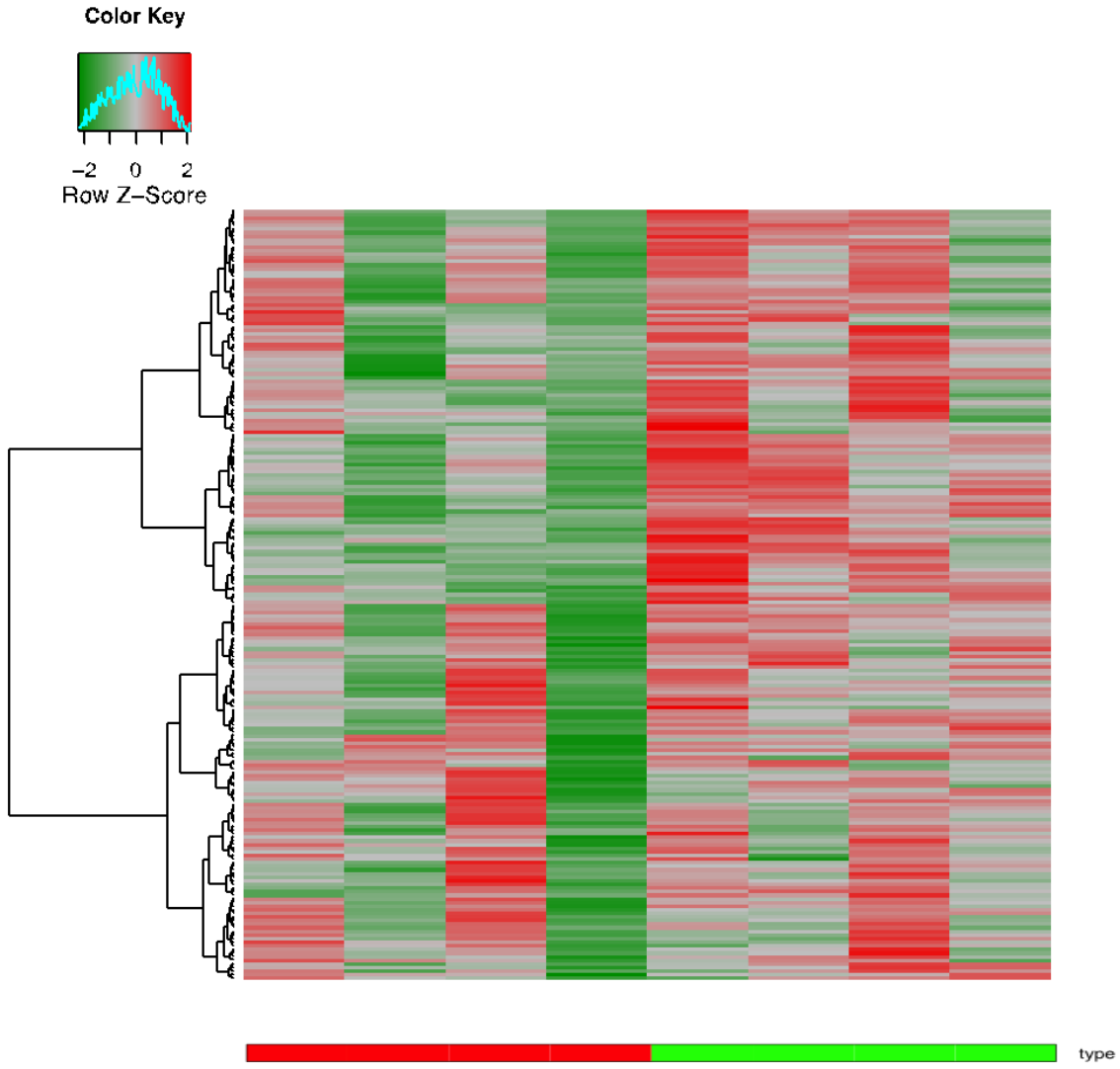

Figure 3.10.B.

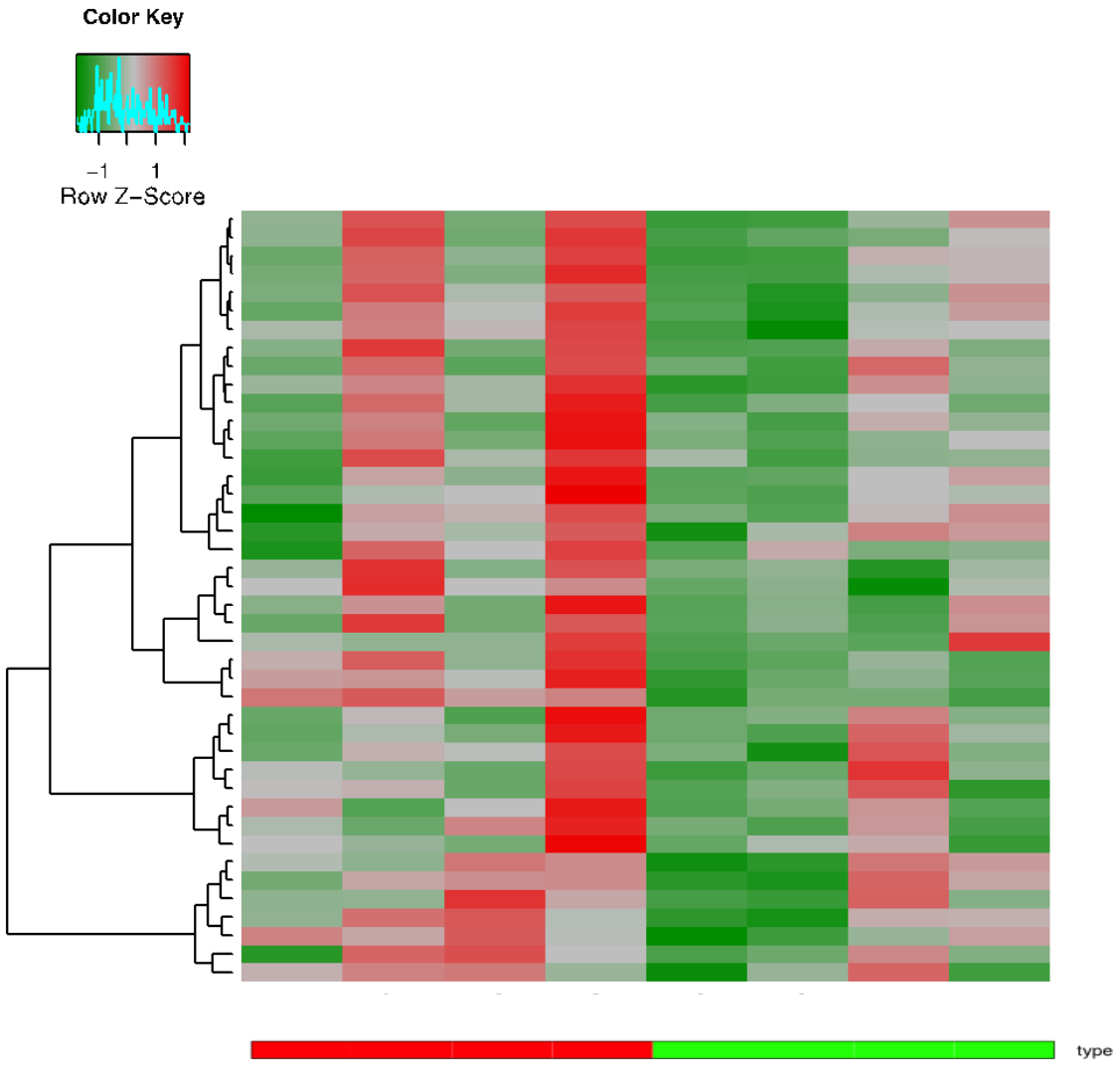

Figure 3.10. The predicted target genes of miR-30e (A) and miR-210 (B), with a correlation cutoff < -0.5, clustered based on the sample types: breast cancer and control samples. Heatmap colors represent mean centered fold change expression in log-space. The breast cancer samples are red, and control samples are in green.

After assessing the expression correlation between a miRNA and its predicted targets from a macro perspective, the results do not illustrate a strong interdependent

relationship. Considering all miRNAs, there is not a strong negative correlation with their predicted targets, and thus this data does not support the miRNA-target model hypothesized.

Biological Relevance of Predicted Targets

Further, we explored the biological relevance of miR-210 and its predicted target genes. miR-210 is over-expressed in normal samples, and its targets are expected to be under-expressed in normal samples. TargetScan predicted 233 targets for which we have gene expression for, and these targets are ranked according to their intensity.

Table 3 illustrates the KEGG terms over-represented in the negative correlation gene list targeted by miR-210 (p-value <0.05). One of the KEGG terms identified is important in breast cancer pathways. B cell receptor signaling pathway is a vital component of adaptive immunity; it controls the proliferation and differentiation of early B cells, which may lead to tumorigenesis (Jumma et al., 2005). Previously, miR-210 has been reported to be induced by Oct-2, a key transcriptional mediator of B cell activation, thus it has an inhibitory mechanism for the control of B cells and autoantibody production (Mok et al., 2013).

Table 4 illustrates GSEA gene sets over-represented at the top of the negative correlation gene list targeted by miR-210 (FDR <0.1). One gene set is identified as important in breast cancer pathways: genes constituting the PUJANA_BRCA1_PCC_NETWORK of transcripts are positively correlated with the expression of BRCA1 across a compendium of normal tissues (Pujana et al., 2007). Supporting this association, Volinia et al.(2012) found miR-210 up-regulated in invasive ductal carcinoma transition and identified BRCA1 as a protein coding gene inversely related to miR-210. This is consistent with our data, as in our study, miR-210 is up regulated in normal tissues and is expected to knock down targets in normal tissue.

Table 5 illustrates GSEA gene sets over-represented at the bottom of the positive correlation gene list targeted by miR-210 (FDR<0.1), and so is positively correlated with miR-210 expression. Two of the three gene sets identified are important in breast cancer

pathways. First, genes in the GOZGIT_ESR1_TARGETS_DN set are down regulated in ER+ breast-cancer cells (Gozgit et al., 2007). This gene set is consistent with our results, since these genes are expected to be up-regulated in normal cells when comparing to breast cancer cells which is what our data shows. Second, the p53 (PEREZ_TP53_TARGETS) pathway is commonly found across various cancers. Schrauder at al. (2012) identified miRNAs overexpressed in whole blood of breast cancer patients to be involved in regulating the p53 oncogenic signal-pathway. Thus, supporting our findings.

Finally, GSEA leading edge analysis identifies core genes that are over-represented between gene sets – genes that are most common within a set of gene sets. Thirty-three genes are identified to overlap between 3 or more gene sets. Figure 3.11 illustrates where these core genes are ranked among the rest of the miR-210 target genes. The plot is gene intensity vs. target gene rank, and a significant number of core genes are at the bottom of the gene rank with high-intensity. This result is surprising, because if these genes were true targets of miR-210, then their gene intensity is expected to be lower, thus showing inconsistencies in findings.

Figure 3.11. Intensity plot of core genes of miR-210 target gene sets.

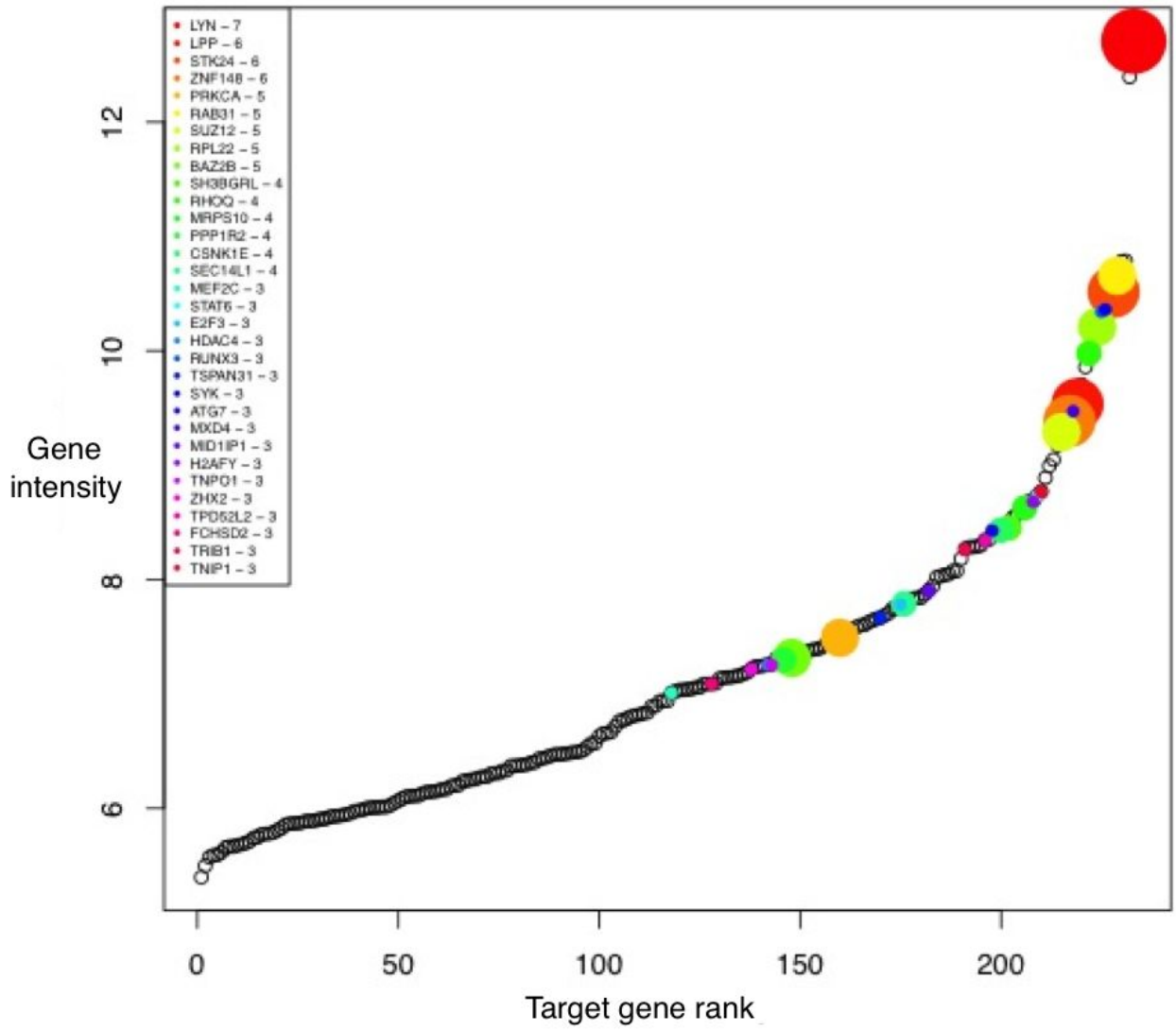

Figure 3.11. Gene intensity plotted against miR-210 target gene rank. The gene intensities are of various colors.

Chapter 4: Discussion

In this study, we identified differentially expressed miRNAs in blood samples of breast cancer patients compared to controls. There is experimental evidence that some of these miRNAs are associated with tumorigenesis in various cancers (Table 1). Although our results show that these miRNAs and their predicted target genes may be involved in cancer pathways, we conclude that there is not sufficient statistical power in this study to draw any conclusions.

First, we followed standard miRNA expression analysis procedures to identify 38 miRNAs (Table 1) that were differentially expressed between cancer cases and control cases. Then, we reviewed and compared different miRNA target methods in order to select the best tools to predict the target genes of our differentially expressed miRNAs. Three target prediction methods were selected, TargetScan (Agarwal et al., 2015), miRanda (Betel et al., 2010), and PITA (Kertesz et al., 2007). Further, we applied these tools to our list of miRNAs to predict a list of target genes. Each tool however, produced different results with little overlapping target genes. From a purely statistical perspective we are uncertain about the significance of these predictions.

Second, in comparing the overlapping target gene list with our mRNA dataset, we found that the target genes predicted by these methods are not found in the matching expression profiles. Although we experimented with attempts to reduce the false positive rate of the target predictions, benchmarking the predicted list against a curated database did not identify significant target prediction score thresholds.

Thirdly, we selected the predicted target genelist from TargetScan only for further functional analysis of the genelist. We found that the miRNA-mRNA expression profiles of this genelist did not identify a negative correlation pattern between the miRNA expression levels and their predicted target genes. Thus, we conclude that our datasets do not support the standard model we used in this study - the miRNAs down-regulating target genes hypothesis.

Lastly, for biological analysis, we selected the miR-210 target genes and found 233 targets for which we have the gene expression for. KEGG and GSEA analysis have shown that these genes may be involved in breast cancer related pathways, such as the B cell receptor signalling pathway (related to the immune system), BRCA1 (a breast cancer gene), and p53 (a common oncogene).

Differentially Expressed MiRNAs are Not Breast Cancer Specific

Although the list of 38 miRNAs (Table 1) differentially expressed between breast cancers and controls in our dataset, there is not sufficient supporting evidence from previous studies to consider the complete list as a BC biomarker. Some miRNAs have been previously found to be differentially expressed in multiple cancers: miR-136 in breast carcinoma tissue (Iorio et al., 2005) and ovarian cancer (Jeong et al., 2017), miR-194 in pancreatic carcinoma (Zhang J. et al., 2014), and miR-223 in lung cancer sera (Chen et al., 2008); whereas miR-335 (Wang et al., 2010; Shrauder et al., 2012) and miR-145 (Kodahl et al., 2014; Thakur et al., 2016) are breast cancer specific; and others (miR-1303, miR-339, miR-517b for examples) have no previous evidence. The literature search suggests that all 38 miRNAs together are not breast cancer specific.

A Panel of Diagnostic MiRNAs Show Little Consistency in Directionality

Further, we consider a small panel of miRNAs to be a candidate biomarker to test together specifically for early-stage breast cancer diagnosis. A panel of four miRNAs (miR-145, miR-210, miR-335 and miR-15a/b) were identified to have sufficient previous evidence to be considered as diagnostic of breast cancer in serum samples. Although eight miRNAs have been previously associated with breast cancers, only four of them have previous evidence from blood-sourced studies. This evidence suggests that there may be a breast cancer signal from blood samples, and thus can be applied for early diagnosis.

Regarding the directionality however, little consistency in how these miRNAs affect their target genes has been found in previous studies. Of the three miRNAs confirmed with previous blood BC studies, only miR-145 is consistent in its expression direction: it's

downregulated in our cancer expression profiles, and downregulated in BC serum (Kodahl et al., 2014; Thakur et al., 2016) and breast tumors (Iorio et al., 2005; Hu et al., 2015; Sempere et al., 2007; Sun et al., 2014). Thus, consistently shown to act as a tumor suppressor in breast cancers. As another miRNA to show consistency in expression direction, miR-15b is upregulated in our expression profiles of cancers, upregulated in invasive ductal carcinoma of breast tissues (Sakurai et al., 2015), and targets a tumor suppressor gene, MTSS1 (Kedmi et al., 2015). Thus, consistently shown to act as an oncogene in breast cancers. Although there is a lack of evidence of miR-15b in blood tissues specifically, its family member miR-15a has been previously found to be upregulated in BC serum (Kodahl et al., 2014). Since the family miRNA sequences are related, miR-15b is a potential oncogenetic biomarker in BC serum as well. Whereas miR-145 shows consistency in direction with our expression profiles, miR-210 and miR-335 don't. In our cancer expression profiles, miR-210 is downregulated, however the literature illustrates that miR-210 is consistently upregulated in BC serum and tumors; in serum (Thakur et al., 2016), in plasma (Madhavan et al., 2012; Jung et al., 2012; Ng et al., 2013) and in breast tumors (Foekens et al., 2008). These multiple studies suggests that miR-210 is oncogenic, as supported by it being identified as a hypoxic marker in BC (Camps et al., 2008), thus inconsistent with our data. In contrast, there is no strong evidence for a consistent expression direction of miR-335. In our cancer expression profiles, miR-335 is upregulated. Previously, it has been shown to be upregulated in whole blood (Schrauder et al., 2012) and downregulated in serum (Wang et al., 2010) from breast cancer patients. In further research, we recommend experimentally validating miR-145, miR-15a/b, miR-335, and miR-210, with PCR for example, to better understand their function and biological effects in blood samples from breast cancer patients.

The inconsistencies in the regulation direction in our data as compared to other studies, suggest that more data and better methods are needed for this type of work. We have previously showed concern over the quality of our data, and the need for standardized methods for collecting and analyzing miRNA data sets.

MiRNAs May be Predictive of Cancer Progression

The miRNAs' involvement in cancer progression exert a causal role at different steps of the tumorigenic process, some associated with several hallmarks of cancer (Goh et al., 2016). Our panel of diagnostic miRNAs (miR-210, miR-145, miR-335, and miR-15b) are involved in cancer progression pathways, and thus are suggested to be predictive of cancer progression.

In our analysis, TargetScan predicts miRNA-210 targets 233 genes that are also highly expressed in the mRNA expression profiles, some of which belong to the GOZGIT_ESR1_TARGETS_DN gene set shown to be downregulated in ER+ breast-cancer cells (Gozgit et al., 2007). In breast cancer cell lines, miR-210 inversely regulates FBXO31, a gene involved in DNA damage response and tumorigenesis (Tan et al., 2018); and in MCF-7 and T47D cell lines, it supports cancer migration (Liu et al., 2016), thus promoting cancer progression. Rothe et al. (2011) showed that the expression of miR-210 is related to tumor proliferation and poor prognosis. As a hypoxic marker (Camps et al., 2008), over-expression of miR-210 results in an increased hypoxic conditions which are associated with metastasis, leading to poor patient prognosis. These results are confirmed by a systematic review by Tang et al. (2015). Further, miR-335 has been shown to be involved in the regulatory networks of the breast cancer susceptibility gene BRCA1 (Heyn et al., 2011), by regulating the BRCA1 activators ERa, IGF1R, SP1 and the repressor ID4. This dual function of promoting and repressing a BC gene may explain the lack of consistency in the expression direction seen earlier in our miRNA expression profile. Lastly, miR-15b upregulates a BC gene, MTSS1, in breast tumors (Kedmi et al., 2015) directly impacting the tumor microenvironment. This may lead to regulation of cancer progression, as miR-15b is identified in invasive ductal carcinoma breast tissue (Sakurai et al., 2015).

In contrast to the above, miR-145 seems to have protective function - it significantly reduces BC cell migration by targeting FSCN-1 and inhibiting epithelial-mesenchymal

transition (Zhao et al., 2016), and regulating TGF-β 1 protein expression which contributes to tumor formation (Ding et al., 2017). Thus, miR-145 has tumor suppressor activity, inhibiting metastasis and thus cancer progression.

In conclusion, the evidence of tumor progression suggests that in addition to potential diagnostic biomarkers, miR-210, miR-335 and miR-15b are potential prognostic biomarkers of BC as well.

MiRNA Target Prediction is Challenging

The review of various target prediction methods and their results, suggest that miRNA target prediction is a challenging problem. The various miRNA target prediction methods have been trained on different miRNA features, assumptions, and evidence of miRNA-target interactions. Some methods are mainly miRNA seed-focused (Agarwal et al., 2015; Krek et al., 2005), and others are target site access focused (Kertesz et al., 2007). The MiRanda (Betel et al., 2010) method however takes into consideration both the miRNA seed features and the secondary structure thus accessibility of the target site, while measuring the thermodynamic stability of the duplex complex (the energetic likelihood of the miRNA-target interactions).

Further, some groups consider sequence conservation to be an important aspect of defining miRNAs and their targets (Agarwal et al., 2015; Betel et al., 2010), while others focus on experimentally derived evidence, such as mRNA/miRNA protein complex co-precipitation (Hammell et al., 2008). Evidence of co-precipitation suggests that there may be both functional and non-functional miRNA-target pairs, where the functional pairs manifest in co-precipitation experiments, but the non-functional pairs are dormant until a biological signal triggers functionality. The miRSVR scoring model correctly identified functional but poorly conserved target sites (Betel et al., 2010). The Betel group showed that imposing a miRNA-target interaction filter results in a reduced detection rate of true targets. This suggests that a true target does not need to interact with the miRNA and thus

does not need to be functional at the time of the experiment. This separates the concepts of a miRNA target definition and miRNA functionality.

These findings are significant because it impacts how we study miRNAs and their targets. It's important to note that some of the earlier target prediction methods were based on little experimental support and thus a lot of theoretical models and hypotheses. With time, as the public miRNA database grew, some groups kept improving their tools, while others quickly became outdated. Initially, many methods were miRNA-seed focused with a heavy weight on sequence conservation, because target prediction purely from conservation is fast and can be done in parallel. However, going through the methodical experiments to show co-precipitation takes a lot of time and more resources. Thus, currently there's an imbalance of information available regarding theoretically potential miRNA targets that can be functional or non-functional, versus the experimentally derived functional miRNA targets. These differences may explain the lack of overlapping genes predicted from our differentially expressed miRNAs.

More recently developed prediction tools, such as HomoTarget (Ahmadi et al., 2013), have shown better results combining twelve different features. Thus, for tools to have higher specificity and sensitivity, future efforts should focus on combining the various experimentally validated features of both the miRNA and its target, such as the seed sequence, secondary structures, thermodynamic stability, conservation, and co-expression. For example, we can automatically derive the weighted combinations of these features, with Neural Network algorithms, as done by Ahmadi et al. (2013).

Conclusion

In summary, we analysed a small data set of matched miRNA-mRNA expression profiles from breast cancer patients and controls from a large population-based cohort study to explore a diagnostic biomarker. We identified 38 differentially expressed miRNAs with previous supporting evidence in breast cancer and other cancers. Only a small panel of miRNAs are breast cancer specific, however they showed little consistency in

directionality and their biological effects. Although some predicted target genes of these miRNAs are associated with tumorigenesis and cancer progression, we conclude that there is not enough statistical power to draw any strong conclusions.

One of the main limitations of this study is the small data set, thus does not have enough statistical power to provide strong results with such high-dimension data. In addition, more modern technologies may provide stronger results. For example, Next Generation Sequencing has been shown to be more accurate than microarray studies, especially for small RNAs. Further, miRNA microarray expression processing and analysis techniques used in this study are similar to standard techniques applied to mRNA expression data. However, as miRNAs are smaller and require higher sensitivity than mRNA, new methods are being explored to process and analyze miRNA expression levels.

Overall, we found that there is little consistency between studies, and a general lack of cohesiveness in the field of miRNA research. As the MiRBase database has grown, and more miRNAs are being annotated and curated, it is difficult to rely on early evidence and hypotheses from the time miRNAs were first identified. As previously noted, many discrepancies in the database have been found, and its difficult to differentiate between small RNAs and non-functional noise from one-time experiments. Fromm et al. (2015) published a review discussing these concerns and presented a uniform system for annotating miRNAs. They showed that less than a third of the 1900 human miRNAs in MiRBase are robustly supported as mature and curated miRNAs, and established a new open access database - MirGeneDB (Fromm et al., 2015). Further, the inconsistencies in miRNA annotation make it difficult to build a comprehensive target prediction tools. Our review found that because different miRNA features were used to train the target prediction tools, its difficult to compare them and the quality of the prediction results.

In this study we investigated miRNAs as potential biomarkers for early diagnosis of breast cancer. Using various computational tools and methods, we explored if blood-sourced miRNAs discriminate between breast cancer and matched healthy controls. Our results show that although some miRNAs are differentially expressed between cancer samples and controls, they are not breast cancer specific and show little consistency in

their biological effect. However, a panel of miRNAs have been previously identified as potential biomarkers of breast cancer, and may be predictive of cancer progression.

We conclude that statistically speaking, there is little evidence in this study that blood-sourced miRNAs are diagnostic of breast cancer. However, the study of miRNAs come with natural limitations as listed above. In the future, further insight into the biogenesis of miRNAs and their relationship with target genes and gene networks will provide a stronger foundation for the development of an early diagnostic breast cancer biomarker.

Tables

Table 1. Differentially expressed miRNAs and their biological significance.

List of 38 experimentally supported and annotated miRNAs, and previous evidence showcasing their biological significance. Red miRNAs are upregulated in case samples and thus downregulated in controls, whereas the black miRNAs are downregulated in case samples and thus upregulated in controls.

	MirbaseID	Previous studies
1	hsa-miR-10a	Diagnostic in esophageal serum (Zhang et al., 2010), diagnostic for Coronary Artery Disease (Luo et al., 2016), diagnostic in plasma pancreatic cancer (Duell et al., 2017)
2	hsa-miR-124	Epigenetic inactivation in mammary carcinoma (Lehmann et al., 2007), tumor-suppressive in osteosarcoma (Huang et al., 2018)
3	hsa-miR-1258	Suppresses breast cancer brain metastasis (Zhang et al., 2011)
4	hsa-miR-1287	Diagnostic in ovarian whole-blood (Hausler et al., 2010)
5	hsa-miR-1298	Inhibits tumor growth (Zhou et al., 2016)
6	hsa-miR-1303	
7	hsa-miR-1307	
8	hsa-miR-136	Suppresses tumor metastasis in triple-negative breast cancer (Yan et al., 2016), inhibits cancer stem cell activity in ovarian cancer (Jeong et al., 2017)
9	hsa-miR-145	Downregulated in primary breast carcinoma tissue (Iorio et al., 2005), downregulated in serum BC (Kodahl et al., 2014), (Thakur et al., 2016)
10	hsa-miR-15b	Targets MTSS1 gene in breast (Kedmi et al., 2015), and found in plasma BC (Kumar et al., 2013)
11	hsa-miR-194	Contributes to tumor growth in renal cell carcinoma (Khella et al., 2013), pancreatic carcinoma (Zhang J et al., 2014)
12	hsa-miR-200b	Implicated in tumour metastasis (Gregory et al., 2008), Diagnostic of ovarian cancer sera samples (Taylor et al., 2008)
13	hsa-miR-208b	
14	hsa-miR-210	Upregulated in serum of BC (Thakur et al., 2016), and breast tumours (Foekens et al., 2008), a hypoxia marker in breast (Camps et al., 2008), and in pancreatic cancer serum (Ho et al., 2010), diagnostic in B-cell Lymphoma sera samples (Lawrie et al., 2008)
15	hsa-miR-223	Diagnostic in sera lung cancer (Chen et al., 2008), diagnostic in esophageal serum (Zhang et al., 2010) and prostate serum (Moltzahn et al., 2011)
16	hsa-miR-24	Diagnostic in oral carcinoma plasma (Lin et al., 2010) diagnostic in prostate serum (Moltzahn et al., 2011)
17	hsa-miR-301b	Hypoxia-responsive oncomiR in prostate cancer (Wang W et al., 2016)
18	hsa-miR-302c	Receptor status predictor in BC tissue (Lowery et al., 2009)
19	hsa-miR-30e	
20	hsa-miR-335	Diagnostic in BC serum (Wang et al., 2010) diagnostic in BC whole blood (Schrauder et al., 2012), regulates BRCA1 gene (Heyn et al., 2011)
21	hsa-miR-339	

22	hsa-miR-346	
23	hsa-miR-363	
24	hsa-miR-421	
25	hsa-miR-5006	
26	hsa-miR-503	Tumor suppressor in BC pathogenesis (Gong et al., 2014)
27	hsa-miR-510	Associated with invasive BC cells (Findlay et al., 2008)
28	hsa-miR-517a	
29	hsa-miR-517b	
30	hsa-miR-518f	
31	hsa-miR-523	Increased in plasma of leukemia patients (Madhavan et al., 2013)
32	hsa-miR-526b	Oncogenic in breast cancer by EP4 activation (Majumder et al., 2015), suppresses lung cancer (Zhang et al., 2015)
33	hsa-miR-548i	
34	hsa-miR-548y	
35	hsa-miR-570	Found in peripheral blood of gallbladder cancer (Li and Pu, 2015)
36	hsa-miR-637	Inhibits HER2 signaling (Leivonen et al., 2014)
37	hsa-miR-664	
38	hsa-miR-935	Promotes liver cancer cell proliferation (Liu et al., 2017)

Table 2. Summary of target prediction methods (Witkos et al., 2011).

Witkos et al. (2011) reviewed and assessed the performance of miRNA target prediction tools and methods. Tools include miRanda, Targetscan and its derivative TargetScanS, PicTar, Diana-MicroT, PITA and RNA22 with their features list and performance against experimentally validated miRNAs.

Target prediction algorithm	Features		Experimental evaluation results				Assessment	
	Parameters contributing to the final score	Cross-species conservation	Sethupathy *et al.* 2006	Baek *et al.* 2008	Alexiou *et al.* 2009		Advantages	Disadvantages
			sensitivity[1]	$\log_2$-fold change[2]	precision[3]	sensitivity[4]		
miRanda	complementarity and free energy binding	conservation filter is used	49%	0.14	29%	20%	- beneficial for prediction sites with imperfect binding within seed region	- low precision, too many false positives
TargetScan	seed match, 3' complementarity local AU content and position contribution[5]	given scoring for each result	21%	0.32[6]	51%	12%	- many parameters included in target scoring - final score correlates with protein downregulation	- sites with poor seed pairing are omitted
Target-ScanS	seed match type	only conservative sites are considered	48%	–	49%	8%	- simple tool for search of conserved sites with stringent seed pairing	- underestimate miRNAs with multiple target sites
PicTar	binding energy, complementarity and conservation	required pairing at conserved positions	48%	0.26	49%	10%	- miRNAs with multiple alignments are favored	- does not predict non-conservative sites
DIANA-microT	free energy binding and complementarity	dataset of conserved UTRs among human and mouse is used	10%	–	48%	12%	- SNR ratio and probability given for each target site - possibility of using own miRNA sequence as an input	- some miRNAs with multiple target sites may be omitted
PITA	target site accessibility energy	user-defined cut-off level	–	0.04[6]	26%	6%	- the secondary structure of 3'UTR is considered for miRNA interaction	- low efficiency compared to other algorithms
Rna22	pattern recognition and folding energy	not included	–	0.09	24%	6%	- allows to identify sites targeted by yet-undiscovered miRNAs	- low efficiency compared to other algorithms

[1] percentage of experimentally supported miRNA-target gene interactions predicted (used TarBase records for which a direct miRNA effect was examined).
[2] average protein depression of genes predicted by the algorithm to be miR-223 targets.
[3] proportion of correctly predicted target miRNAs to total predicted miRNA-mRNA interactions (data obtained from proteomic analyses carried out by Selbach *et al.*).
[4] proportion of correctly predicted target miRNAs to total correct miRNA-mRNA interactions (data obtained from proteomic analyses carried out by Selbach *et al.*).
[5] *position contribution* parameter promotes sites close to the 3'UTR ends.
[6] the final scoring correlates with the level of protein downregulation.

Table 3. Gene to KEGG test for over-representation in gene list targeted by miR-210. KEGG terms of gene lists that are predicted to be targeted by miRNA-210, with a p-value < 0.05. The list represented over-representation of genes involved in specific metabolic pathways. KEGG terms identified include the B cell receptor signaling pathways, an important component of adaptive immunity, also involved in the breast cancer pathways.

Gene to KEGG test for over-representation

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
650	0.006	20.875	0	2	12	Butanoate metabolism
4960	0.011	14.887	0	2	16	Aldosterone-regulated sodium reabsorption
4662	0.014	6.704	1	3	51	B cell receptor signaling pathway
5414	0.039	7.144	0	2	31	Dilated cardiomyopathy
4971	0.046	6.466	0	2	34	Gastric acid secretion

Table 4. GSEA test applied to the negative correlated gene list targeted by miR-210. GSEA test results of gene lists that are predicted to be negatively targeted by miRNA-210, with an FDR < 0.1. The list represented over-representation of genes involved in specific metabolic pathways. GSEA terms identified include the BRCA1 gene network of transcripts which are important in breast cancer pathways.

GSEA UPREGULATED

GS DETAILS	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
YAGI_AML_WITH_T_8_21_TRANSLOCATION	8	0.55	1.83	0.009	0.605	0.391	37
GEORGES_TARGETS_OF_MIR192_AND_MIR215	9	0.51	1.79	0.011	0.375	0.456	42
MULLIGHAN_MLL_SIGNATURE_2_UP	5	0.64	1.79	0.016	0.26	0.467	38
PUJANA_ATM_PCC_NETWORK	14	0.43	1.77	0.017	0.218	0.499	62
THUM_SYSTOLIC_HEART_FAILURE_UP	6	0.61	1.77	0.015	0.176	0.501	27
SATO_SILENCED_BY_METHYLATION_IN_PANCREATIC_CANCER_1	5	0.62	1.7	0.018	0.203	0.628	40
TIEN_INTESTINE_PROBIOTICS_24HR_UP	5	0.62	1.69	0.028	0.189	0.659	21
LOPEZ_MBD_TARGETS	8	0.5	1.67	0.036	0.182	0.693	53
NUYTTEN_NIPP1_TARGETS_DN	13	0.4	1.65	0.034	0.184	0.732	42
FULCHER_INFLAMMATORY_RESPONSE_LECTIN_VS_LPS_UP	5	0.61	1.63	0.032	0.184	0.76	41
PUJANA_BRCA1_PCC_NETWORK	11	0.42	1.53	0.057	0.268	0.901	62

Table 5. GSEA test applied to the positive correlated gene list targeted by miR-210, with an FDR < 0.1.

GSEA test results of gene lists that are predicted to be positively by miRNA-210, with an FDR < 0.1. The list represented over-representation of genes involved in specific metabolic pathways. GSEA terms identified include the ESR1 targets affected in ER+ breast-cancer cells.

GSEA DOWNREGULATED							
GS DETAILS	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
GOZGIT_ESR1_TARGETS_DN	5	-0.59	-1.64	0.053	0.802	0.717	45
PEREZ_TP53_TARGETS	9	-0.44	-1.57	0.065	0.563	0.833	37
HORIUCHI_WTAP_TARGETS_UP	5	-0.53	-1.43	0.085	0.67	0.958	51

References

Agarwal, V., Bell, G. W., Nam, J. W., & Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, *4*, e05005.

Ahmadi, H., Ahmadi, A., Azimzadeh-Jamalkandi, S., Shoorehdeli, M.A., Salehzadeh-Yazdi, A., Bidkhori, G., Masoudi-Nejad, A. (2013). HomoTarget: a new algorithm for prediction of microRNA targets in Homo sapiens. *Genomics*, Feb;101(2):94-100.

Ali, H.R., Rueda, O.M., Chin, S.F., Curtis, C., Dunning, M.J., Aparicio, S.A. (2014). Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol,* 15:431.

An, J., Lai, J., Lehman, M.L. (2013). miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Research,* 41(2):727–37.

Baek, D., Villen, J., Shin, C. (2008). The impact of microRNAs on protein output. *Nature*, 455(7209): 64-71.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell,* 116:281-97.

Barturen, G., Rueda, A., Hamberg, M., Alganza, A., Lebron, R., Kotsyfakis, M., Shi, B.-J., Koppers-Lalic, D., Hackenberg, M. (2014). sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments. *Methods Next Gener. Seq.,* 1 21 31.

Bishop, C. (2006). Pattern recognition and machine learning. *New York: Springer .*

Brennecke, J., Stark, A., Russell, R.B., Cohen, S.M. (2005). Principles of microRNA- target recognition. *PLoS Biol.*, 3: e85.

Brodersen, J., Siersma, V. (2013). Long-term psychosocial consequences of false-positive screening mammography – a cohort study with 3-year follow-up. *Annals of Family Medicine,* 11(2):106-15.

Calin, G.A., Sevignani, C., Dumitru, C.D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M., Croce, C.M. (2004). Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc Natl Acad Sci U S A*, Mar 2;101(9):2999-3004.

Camps, C., Buffa, F.M., Colella, S. (2008). Hsa-miR-210 is induced by hypoxia and is an independent prognostic factor in breast cancer. *Clin. Cancer Res*, 14(5),1340–1348.

Canadian Cancer Society's Steering Committee on Cancer Statistics. (2017). *Canadian Cancer Society*.

Carlson, R.W., Allred, D.C., Anderson, B.O., Burstein, H.J., Carter, W.B., Edge, S.B., Erban, J.K., Farrar, W.B., Goldstein, L.J., Gradishar, W.J., Hayes, D.F., Hudis, C.A., Jahanzeb, M., Kiel, K., Ljung, B.M., Marcom, P.K., Mayer, I.A., McCormick, B., Nabell, L.M., Pierce, L.J., Reed, E.C., Smith, M.L., Somlo, G., Theriault, R.L., Topham, N.S., Ward, J.H., Winer, E.P., Wolff, A.C. (2009). Breast cancer. Clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network : JNCCN,* 7 (2): 122–192.

Chen, X., Ba, Y., Ma, L., Cai, X., Yin, Y., Wang, K., Guo, J., Zhang, Y., Chen, J., Guo, X., Li, Q., Li, X., Wang, W., Zhang, Y., Wang, J., Jiang, X., Xiang, Y., Xu, C., Zheng, P., Zhang, J., Li, R., Zhang, H., Shang, X., Gong, T., Ning, G., Wang, J., Zen, K., Zhang, J., Zhang, C.Y. (2008). Characterization

of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Res*, Oct;18(10): 997-1006.

Chen, L., Heikkinen, L., Wang, C., Yang, Y., Sun, H., & Wong, G. (2018). Trends in the development of miRNA bioinformatics tools. *Briefings in bioinformatics*.

Choosing Wisely: an initiative of the ABIM Foundation, American Society of Clinical Oncology, "Five Things Physicians and Patients Should Question", archived from the original (PDF) on 31 July 2012, retrieved 14 August 2012.

Coronnello, C., Hartmaier, R., Arora, A., Huleihel, L., Pandit, K. V., Bais, A. S., Butterworth, M., Kaminski, N., Stormo, G. D., Oesterreich, S., Benos, P. V. (2012). Novel modeling of combinatorial miRNA targeting identifies SNP with potential role in bone density. *PLoS computational biology*, *8*(12), e1002830.

Couzin, J. (2005). Dissecting a hidden breast cancer risk. *Science*. Sept 9; 309, 3.

Cuk, K., Zucknick, M., Madhavan, D., Schott, S., Golatta, M., Heil, J., Marme, F., Burwinkel, B. (2013). Plasma MicroRNA Panel for Minimally Invasive Detection of Breast Cancer. *PLoS ONE*, 8 (10), no. e76729.

Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486:346–52.

Ding, Y.Z.C., Zhang, J., Zhang, N., Li, T., Fang, J., Zhang, Y., Zuo, F., Tao, Z., Tang, S., Zhu, W., Chen, H., Sun, X. (2017). miR-145 inhibits proliferation and migration of breast cancer cells by directly or indirectly regulating TGF- β 1 expression. *Int J Oncol*, 50:1701–10.

Du, T., & Zamore, P.D. (2005). microPrimer: the biogenesis and function of microRNA. *Development*, 132:4645–4652.

Duell, E.J., Lujan-Barroso, L., Sala, N., Deitz-McElyea, S., Overvad. K., Tjonneland, A., Olsen. A., Weiderpass, E., Busund, L.T., Moi, L., Muller, D., Vineis, P., Aune, D., Matullo, G., Naccarati, A., Panico, S., Tagliabue, G., Tumino, R., Palli, D., Kaaks, R., Katzke, V.A., Boeing, H., Bueno-de-Mesquita, H.B.A., Peeters, P.H., Trichopoulou, A., Lagiou, P., Kotanidou, A., Travis, R.C., Wareham, N., Khaw, K.T., Ramon Quiros, J., Rodríguez-Barranco, M., Dorronsoro, M., Chirlaque, M.D., Ardanaz, E., Severi, G., Boutron-Ruault, M.C., Rebours, V., Brennan, P., Gunter, M., Scelo, G., Cote, G., Sherman, S., Korc, M. (2017). Plasma microRNAs as biomarkers of pancreatic cancer risk in a prospective cohort study. *Int J Cancer*, Sep 1;141(5):905-915.

Dumeaux, V., Børresen-Dale, A.L., Frantzen, J.O., Kumle, M., Kristensen, V.N., Lund, E. (2008). Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer post genome cohort study. *Breast Cancer Res,* 10(1):R13.

Dumeaux, V., Ursini-Siegel, J., Flatberg, A., Fjosne, H.E., Frantzen, J.O., Holmen, M.M., Rodegerdts, E., Schlichting, E., Lund, E. (2015). Peripheral blood cells inform on the presence of breast cancer: a population-based case-control study. *Int J Cancer,* Feb 1;136(3):656-67.

Eheman, C.R., Shaw, K.M., Ryerson, A.B., Miller, J.W., Ajani, U.A., White, M.C. (2009). The changing incidence of in situ and invasive ductal and lobular breast carcinomas: United States, 1999-2004. *Cancer Epidemiol. Biomarkers Prev*, 18 (6): 1763–9.

Enerly, E., Steinfeld, I., Kleivi, K., Leivonen, S.K., Aure, M.R., Russnes, H.G., Rønneberg, J.A., Johnsen, H., Navon, R., Rødland, E., Mäkelä, R., Naume, B., Perälä, M., Kallioniemi, O.,

Kristensen, V.N., Yakhini, Z., Børresen-Dale, A.L. (2011). miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors. *PLoS One*, Feb 22;6(2):e16915.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C. (2003). MicroRNA targets in Drosophila. *Genome Biol*, 5:R1.

Esserman, L.J., Thompson, I.M., Reid, B. (2013). Overdiagnosis and overtreatment in cancer: an opportunity for improvement. *JAMA,* 310, 797-798.

Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., Tress, M.L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human molecular genetics,* 23.22: 5866-5878.

Farazi, T.A., Horlings, H.M., ten Hoeve, J.J., Mihailovic, A., Halfwerk, H., Morozov, P., Brown, M., Hafner, M., Reyal, F., van Kouwenhove, M., Kreike, B., Sie, D., Hovestadt, V., Wessels, L., van de Vijver, M.J., Tuschl, T. (2011). MicroRNA Sequence and Expression Analysis in Breast Tumors by Deep Sequencing. *Cancer Res*, 71(13); 4443–53.

Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., Bartel, D.P. (2005). The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science,* 310, 1817–1821.

Findlay, V.J., Turner, D.P., Moussa, O., Watson, D.K. (2008). MicroRNA-mediated inhibition of prostate-derived Ets factor messenger RNA translation affects prostate-derived Ets factor regulatory networks in human breast cancer. *Cancer Res*, Oct 15;68(20):8499-506.

Foekens, J.A., Sieuwerts, A.M., Smid, M., Look, M.P., de Weerd, V., Boersma, A.W. (2008). Four miRNAs associated with aggressiveness of lymph node-negative, estrogen receptor-positive human breast cancer. *Proc Natl Acad Sci USA*, Sep 2;105(35):13021-6.

Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnology,* 26:407–415.

Friedlander, M.R., Mackowiak, S.D., Li, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, 40(1):37–52.

Friedman, R.C., Farh, K.K., Burge, C.B., Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19: 92-105.

Fromm, B., Billipp, T., Peck, L.E., Johansen, M., Tarver, J.E., King, B.L., (2015). A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annual Review of Genetics*. 49: 213–42.

GBD. (2015). Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study. *Lancet*, Oct 8;388(10053):1545-1602.

Gong, J., Luk, F., Jaiswal, R., Bebawy, M. (2014). Microparticles Mediate the Intercellular Regulation of microRNA-503 and Proline-Rich Tyrosine Kinase 2 to Alter the Migration and Invasion Capacity of Breast Cancer Cells. *Front Oncol*, 4: 220.

Gøtzsche, P.C., & Jørgensen, K. (2013). Screening for breast cancer with mammography. *Cochrane Database of Systematic Reviews*, Issue 6. Art. No.: CD001877.

Gozgit, J.M., Pentecost, B.T., Marconi, S.A., Ricketts-Loriaux, R.S.J., Otis, C.N., & Arcaro, K.F. (2007). PLD1 is overexpressed in an ER-negative MCF-7 cell line variant and a subset of phospho-Akt-negative breast carcinomas. *British Journal of Cancer*, 97, 809–817.

Gregory, P.A., Bert, A.G., Paterson, E.L., Barry, S.C., Tsykin, A., Farshid, G., Vadas, M.A., Khew-Goodall, Y., Goodall, G.J. (2008). The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat Cell Biol,* May 10 (5): 593-601.

Guttery, D.S., Blighe, K., Page, K., Marchese, S.D., Hills, A., Coombes, R.C. (2013). Hide and seek: tell-tale signs of breast cancer lurking in the blood. *Cancer Metastasis Rev,* 32(1–2):289–302.

Hammell, M., Long, D., Zhang, L., Lee, A., Carmack, C.S. (2008). mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nat Methods,* 5:813–819.

Hanley, J.A., Hannigan, A., O'Brien, K.M. (2017). Mortality reductions due to mammography screening: Contemporary population-based data. *PLoS ONE,* 12(12): e0188947.

Harris, L.N., Ismaila, N., Mcshane, L.M., Andre, F., Collyar, D.E., Gonzalez-Angulo, A.M. (2016). Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American Society of Clinical Oncology Clinical Practice Guideline. *J Clin Oncol,* 34:1134–50.

Häusler, S.F., Keller, A., Chandran, P.A., Ziegler, K., Zipp, K., Heuer, S., Krockenberger, M., Engel, J.B., Hönig, A., Scheffler, M., Dietl, J., Wischhusen, J. (2010). Whole blood-derived miRNA profiles as potential new tools for ovarian cancer screening. *Br J Cancer,* Aug 24;103(5):693-700.

Heneghan, H.M., Miller, N., Lowery, A.J., Sweeney, K.J., Newell, J., Kerin, M.J. (2010). Circulating microRNAs as novel minimally invasive biomarkers for breast cancer. *Ann Surg,* Mar;251(3):499-505.

Heyn, H., Engelmann, M., Schreek, S., Ahrens, P., Lehmann, U., Kreipe, H., Schlegelberger, B., Beger, C. (2011). MicroRNA miR-335 is crucial for the BRCA1 regulatory cascade in breast cancer development. *Int J Cancer*, Dec 15;129(12):2797-806.

Ho, A.S., Huang, X., Cao, H., Christman-Skieller, C., Bennewith, K., Le, Q.T., Koong, A.C. (2010). Circulating miR-210 as a Novel Hypoxia Marker in Pancreatic Cancer. *Transl Oncol*, Apr;3(2):109-13.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie,* 125: 167-188.

Hollander, M., & Wolfe, D. A. (1999). Nonparametric Statistical Methods. (Wiley, New York)

Hortobagyi, G.N., D'orsi, C.J., Edge, S.B., Mittendorf, E.A., Rugo, H.S., Solin, L.J. (2017). *AJCC Cancer Staging Manual – Breast*, 8th ed. Chicago: Springer.

Hu, B., Ying, X., Wang, J., Piriyapongsa, J., Jordan, I.K., Sheng, J., Yu, F. (2014). Identification of a Tumor-Suppressive Human-Specific MicroRNA within the FHIT Tumor-Suppressor Gene. *Cancer Research*, 74(8); 2283-94.

Huang, J., Liang, Y., Xu, M., Xiong, J., Wang, D., Ding, Q. (2018). MicroRNA-124 acts as a tumor-suppressive miRNA by inhibiting the expression of Snail2 in osteosarcoma. *Oncol Lett*, Apr;15(4):4979-4987.

Huang, J.C., Babak, T., Corson, T.W., Chua, G., Khan, S. (2007a). Using expression profiling data to identify human microRNA targets. *Nat Methods,* 4:1045–1049.

Huang, T.H., Fan, B., Rothschild, M.F., Hu, Z.L., Li, K. (2007b). MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinformatics,* 8:341.

Huang, Y., Shen, X.J., Zou, Q., Wang, S.P., Tang, S.M., Zhang, G.Z. (2011). Biological functions of microRNAs: a review. *J Physiol Biochem,* Mar;67(1):129-39.

Huber, W., Carey, V.J., Gentleman, R., Anders, S. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods,* 12:115-121.

Huntzinger, E., and Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature Rev. Genet.,* 12, 99–110.

International Human Genome Sequencing Consortium. (2004). *Nature,* 431 (7011): 931-45.

Iorio, M.V., Ferracin, M., Liu, C.G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., Menard, S., Palazzo, J.P., Rosenberg, A., Musiani, P., Volinia, S., Nenci, I., Calin, G.A., Querzoli, P., Negrini, M., Croce, C.M. (2005). MicroRNA gene expression deregulation in human breast cancer. *Cancer Res*, 65(16):7065-70.

Iorio, M.V., Croce, C.M. (2012). MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Mol Med.,* Mar;4(3):143-59.

Jacobsen, K., K., Abraham, L., Buist, D. S., Hubbard, R. A., O'Meara, E. S., Sprague, B. L., Kerlikowske, K., Vejborg, I., Von Euler-Chelpin, M., Njor, S. H. (2015). Comparison of

cumulative false-positive risk of screening mammography in the United States and Denmark. *Cancer epidemiology,* 39(4), 656-63.

Jeong, J.Y., Kang, H., Kim, T.H., Kim, G., Heo, J.H., Kwon, A.Y., Kim, S., Jung, S.G., An, H.J. (2017). MicroRNA-136 inhibits cancer stem cell activity and enhances the anti-tumor effect of paclitaxel against chemoresistant ovarian cancer cells by targeting Notch3. *Cancer Lett*, Feb 1;386:168-178.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Research,* 37. D98–104.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., Marks, D.S. (2004) Human MicroRNA Targets. *PLoS Biol,* 2(11): e363.

Jørgensen, K.J., Zahl, P.H., Gøtzsche, P.C. (2010). Breast cancer mortality in organised mammography screening in Denmark: comparative study. *BMJ*, Mar 23;340:c1241.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research,* 34, D354–D357.

Karp, S. (2009). The Sea Change That's Challenging Biology's Central Dogma, Discover Magazine.

Kedmi, M., Ben-Chetrit, N., Körner, C., Mancini, M., Ben-Moshe, N.B., Lauriola, M., Lavi, S., Biagioni, F., Carvalho, S., Cohen-Dvashi, H., Schmitt, F., Wiemann, S., Blandino, G., Yarden, Y. (2015). EGF induces microRNAs that target suppressors of cell migration: miR-15b targets MTSS1 in breast cancer. *Sci. Signal.,* 8.368, ra29-ra29.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D. (2002a). The human genome browser at UCSC. *Genome Res.,* Jun;12(6):996-1006.

Kent, W.J. (2002b). BLAT - the BLAST-like alignment tool. *Genome Res.,* Apr;12(4):656-64.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat Genet*, 39:1278–1284.

Khella, H.W., Heba, W.Z. (2013). miR-192, miR-194 and miR-215: a convergent microRNA network suppressing tumor progression in renal cell carcinoma. *Carcinogenesis,* 34.10, 2231-2239.

Kim, V.N. (2005). MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol,* 6:376–385.

Kim, V.N., Han, J., Siomi, M.C. (2009). Biogenesis of small RNAs in animals. *Nature Rev. Mol. Cell Biol.,* 10, 126–139.

Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C. (2004). A combined computational-experimental approach predicts human microRNA targets. *Genes Dev,* 18:1165–1178.

Kodahl, A.R., Lyng, M.B., Binder, H., Cold, S., Gravgaard, K., Knoop, A.S. (2014). Novel circulating microRNA signature as a potential non-invasive multi-marker test in ER-positive early-stage breast cancer: a case control study. *Mol Oncol.,* 8:874–83.

Kozomara, A., Griffiths-Jones, S. (2010). MiRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39 (Database issue): D152–7.

Kozomara, A., Birgaoanu, M., Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Research,* 47:D155-D162.

Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L. (2005). Combinatorial microRNA target predictions. *Nat Genet*, 37:495– 500.

Kumar, S., Keerthana, R., Pazhanimuthu, A., Perumal, P. (2013). Overexpression of circulating miRNA-21 and miRNA-146a in plasma samples of breast cancer patients. *Indian J Biochem Biophys*. Jun;50(3):210-4.

Lakhani, S.R., Ellis, I.O., Schnitt, S.J., Tan, P.H., Van De Vijver, M.J. (2012). *WHO Classification of Tumours of the Breast*. Lyon: France International Agency for Research on Cancer.

Landthaler, M., Yalcin, A., Tuschl, T. (2004).The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. *Curr. Biol.,* 14,2162 -2167.

Lee, R.C., Feinbaum, R.L., Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell,* 75:843–854.

Lee, R.C., Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science,* 294:862–864.

Lehmann, U., Hasemeier, B., Romermann, D., Muller, M., Langer, F., Kreipe, H. (2007). Epigenetic inactivation of microRNA genes in mammary carcinoma. *Verh Dtsch Ges Pathol.,* 91:214-20.

Leivonen, S.K., Sahlberg, K.K., Mäkelä, R., Due, E.U., Kallioniemi, O., Børresen-Dale, A.L., & Perälä, M. (2013). High-throughput screens identify microRNAs essential for HER2 positive breast cancer cell growth. *Molecular oncology*, *8*(1), 93–104.

Levine, E., Zhang, Z., Kuhlman, T., Hwa, T. (2007). Quantitative characteristics of gene regulation by small RNA. *PLoS Biol.,* 5, e229.

Lewis, B.P., Burge, C.B., Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell,* 120:15–20.

Li, G., and Yabin, P. (2015). MicroRNA signatures in total peripheral blood of gallbladder cancer patients. *Tumor Biology,* 36.9: 6985-6990.

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S. (2003). The microRNAs of Caenorhabditis elegans. *Genes Dev,* 17:991–1008.

Lin, S.C., Liu, C.J., Lin, J.A., Chiang, W.F., Hung, P.S., Chang, K.W. (2010). miR-24 up-regulation in oral carcinoma: positive association from clinical and in vitro analysis. *Oral Oncol.,* Mar;46(3):204-8.

Lindow, M., Gorodkin, J. (2007). Principles and limitations of computational microRNA gene and target finding. *DNA Cell Biol,* 26:339–351.

Liu, X., Luo, G., Bai, X., Wang, X.J. (2009). Bioinformatic analysis of microRNA biogenesis and function related proteins in eleven animal genomes. *J Genet Genomics,* 36:591–601.

Liu, X., Li, J., Yu, Z., Li, J., Sun, R., Kan, Q. (2017). MiR-935 promotes liver cancer cell proliferation and migration by targeting SOX7. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics,* 25.3 : 427-435.

Long, D., Lee, R., Williams, P., Chan, C.Y., Ambros, V. (2007). Potent effect of target structure on microRNA function. *Nat Struct Mol Biol,* 14:287–294.

Lowery, A.J., Miller, N., Devaney, A. (2009). MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer. *Breast Cancer Res,* 11: R27.

Lukasik, A., Wójcikowski, M., Zielenkiewicz, P. (2016). Tools4miRs – one place to gather all the tools for miRNA analysis. *Bioinformatics,* 32: 17.

Luo, L., Chen, B., Li, S., Wei, X., Liu, T., Huang, Y., Lin, X. (2016). Plasma miR-10a: A Potential Biomarker for Coronary Artery Disease. *Dis Markers,* 3841927.

Lytle, J.R., Yario, T.A., Steitz, J.A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 50 UTR as in the 30 UTR. *Proc Natl Acad Sci USA,* 104: 9667-9672.

Marmot, M.G., Altman, D.G., Cameron, D.A. (2013). The benefits and harms of breast cancer screening: an independent review. *Br J Cancer,* 108:2205–40.

Miller, A.B., Wall, C., Baines, C.J., Sun, P., To, T., Narod, S.A. (2014). Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ,* 348:g366.

Moltzahn, F., Olshen, A.B., Baehner, L., Peek, A., Fong, L., Stöppler, H., Simko, J., Hilton, J.F., Carroll, P., Blelloch, R. (2011). Microfluidic-based multiplex qRT-PCR identifies diagnostic and prognostic microRNA signatures in the sera of prostate cancer patients. *Cancer Res,* Jan 15;71(2):550-60.

Moretti, F., Thermann, R., Hentze, M.W. (2010). Mechanism of translational regulation by miR-2 from sites in the 50 untranslated region or the open reading frame. *RNA,* 16: 2493-2502.

Morozova, O., Marra, M.A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics,* 92: 255-264.

Mukherji, S., Ebert, M.S., Zheng, G., Tsang, J.S., Sharp, P.A., Oudenaarden, A. (2011). MicroRNAs can generate thresholds in target gene expression. *Nature Genet,* 43, 854–859.

Nam, S., Li, M., Choi, K., Balch, C., Kim, S., Nephew, K.P. (2009). MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Research*, Jul;37.

Olson, A.J., Brennecke, J., Aravin, A.A., Hannon, G.J., Sachidanandam, R. (2008). Analysis of large-scale sequencing of small RNAs. *Pac Symp Biocomputing,* 126-136.

Orel, S.G., Kay, N., Reynolds, C., Sullivan, D.C. (1999). BI-RADS categorization as a predictor of malignancy. *Radiology*, 211:845-50.

Papadopoulos, G.L., Reczko, M., Simossis, V.A., Sethupathy, P., Hatzigeorgiou, A.G. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Research*, Jan; 37 D155-8.

Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Vlachos, I.S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T., Hatzigeorgiou, A.G. (2013). DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Research,* Jul;41 (Web Server issue): W169-73.

Pastrello, C., Polesel, J., Della Puppa, L., Viel, A., Maestro, R. (2010). Association between hsa-mir-146a genotype and tumor age-of-onset in BRCA1/BRCA2-negative familial breast and ovarian cancer patients. *Carcinogenesis,* 31:2124–6.

Pujana, M.A., Han, J.D., Starita, L.M., Stevens, K.N., Tewari, M., Ahn, J.S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., Elshamy, W.M., Rual, J.F., Levine, D., Rozek, L.S., Gelman, R.S., Gunsalus, K.C., Greenberg, R.A., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Solé, X., Hernández, P., Lázaro, C., Nathanson, K.L., Weber, B.L., Cusick, M.E., Hill, D.E., Offit, K., Livingston, D.M., Gruber, S.B., Parvin, J.D., Vidal, M. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.,* 39(11).

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat Genet.,* 32:S496-S501.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature,* 403:901–906.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K. (2015). Limma powers expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43.

Rothé, F.I.M., Chaboteaux, C., Haibe-Kains, B., Kheddoumi, N., Majjaj, S., Badran, B., Fayyad-Kazan, H., Desmedt, C., Harris, A.L., Piccart, M., Sotiriou, C. (2011). Global microRNA expression profiling identifies miR-210 associated with tumor proliferation, invasion and poor clinical outcome in breast cancer. *PLoS One*, 6:e20980.

Sakurai, M., Masuda, M., Miki, Y., Hirakawa, H., Suzuki, T., & Sasano, H. (2015). Correlation of miRNA Expression Profiling in Surgical Pathology Materials, with Ki-67, HER2, ER and PR in Breast Cancer Patients. *The International Journal of Biological Markers*, *30*(2), 190–199.

Sethupathy, P., Megraw, M., Hatzigeorgiou, A.G. (2006a). A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat Methods,* 3:881–886.

Sethupathy, P., Corda, B., Hatzigeorgiou, A.G. (2006b). TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, 12(2): 192-7.

Sevignani, C., Calin, G.A., Nnadi, S.C., Shimizu, M., Davuluri, R.V., Hyslop, T., Demant, P., Croce, C.M., Siracusa, L.D. (2007). MicroRNA genes are frequently located near mouse cancer susceptibility loci. *Proc Natl Acad Sci U S A*, May 8;104(19):8017-22.

Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S. (2005). Identification of clustered microRNAs using an ab-initio prediction method. *BMC Bioinformatics,* 6:267.

Schrauder, M.G., Strick, R., Schulz-Wendtland, R., Strissel, P.L., Kahmann, L., Loehberg, C.R., Lux, M.P., Jud, S.M., Hartmann, A., Hein, A., Bayer, C.M., Bani, M.R., Richter, S., Adamietz, B.R., Wenkel, E., Rauh, C., Beckmann, M.W., Fasching, P.A. (2012). Circulating microRNAs as potential blood-based markers for early stage breast cancer detection. *PLoS One*, 7(1):e29770.

Smith, T.F., & Waterman, M.S. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology,* 147 (1): 195-197.

Smyth, G.K. (2004). Linear models and empirical Bayes models for assessing differential expression in microarray experiments. *Statistical Applications in Genetic and Molecular Biology*, 3.

Solbjør, M., Forsmo, S., Skolbekken, J. A., Siersma, V., & Brodersen, J. (2018). Psychosocial consequences among women with false-positive results after mammography screening in Norway. *Scandinavian journal of primary health care*, *36*(4), 380-389.

Sood, P., Krek, A., Zavolan, M., Macino, G., and Rajewsky, N. (2006). Cell type-specific signatures of microRNAs on target mRNA expression. *Proc. Natl. Acad. Sci. USA,* 103, 2746–2751.

Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A,* 98:10869–74.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A,* 100:8418–23.

Stegmayer, G., Yones, C., Kamenetzky, L. (2017). High class imbalance in pre-miRNA prediction: a novel approach based on deepSOM. *IEEE/ACM Trans Comput Biol Bioinform*, 14(6):1316–26.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P. (2005). Gene set

enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A,* Oct 25;102(43):15545-50.

Tan, Y., Liu, D., Gong, J., Liu, J., Huo, J. (2018). The role of F-box only protein 31 in cancer. *Oncology Letters,* Apr;15(4): 4047-4052.

Taylor, D.D., Gercel-Taylor, C. (2008). MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer. *Gynecol Oncol*, Jul;110(1):13-21.

Thakur, S., Grover, R.K., Gupta, S., Yadav, A.K., Das, B.C. (2016). Identification of specific miRNA signature in paired sera and tissue samples of indian women with triple negative breast cancer. *PLoS One*, 11:e0158946.

Tran, V.T., Tempel, S., Zerath, B., Zehraoui, F., Tahi, F. (2015). miRBoost: boosting support vector machines for microRNA precursor classification. *RNA,* 21(5):775–785.

Turchinovich, A., Weiz, L., Langheinz, A., Burwinkel, B. (2011). Characterization of extracellular circulating microRNA. *Nucleic Acids Res*, Sep 1; 39(16):7223-33.

Volinia, S., Galasso, M., Sana, M.E., Wise, T.F., Palatini, J., Huebner, K., Croce, C.M. (2012). Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. *Proc Natl Acad Sci U S A*, Feb 21;109(8):3024-9.

Vuong, D., Simpson, P.T., Green, B., Cummings, M.C., Lakhani, S.R. (2014). Molecular classification of breast cancer. *Virchows Arch*, Jul;465(1):1-14.

Wang, F., Zheng, Z., Guo, J., Ding, X. (2010). Correlation and quantitation of microRNA aberrant expression in tissues and sera from patients with breast tumor. *Gynecol Oncol*. Dec;119(3):586-93.

Wang, W., Liu, M., Guan, Y., Wu, Q. (2016). Hypoxia-responsive mir-301a and mir-301b promote radioresistance of prostate cancer cells via downregulating NDRG2. *Medical science monitor: international medical journal of experimental and clinical research.* 22 : 2126.

Wang, X., Zhang, J., Li, F., Gu, J., He, T. (2005). MicroRNA identification based on sequence and structure alignment. *Bioinformatics.* 21:3610–3614.

Wang, X., El Naqa, I. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*. 24 (3): 325-10.

Wang, Y., Sheng, G., Juranek, S., Tuschl, T., Patel, D.J. (2008). Structure of the guide-strand-containing argonaute silencing complex. *Nature*. 456, 209–213.

Wei, L., Kangcheng, R. (2009). MicroRNA detection by microarray. *Anal Bioanal Chem*. 394:1117-1124.

Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B. (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res*. 10:R65.

Witkos, T.M., Koscianska, E., Krzyzosiak, W.J. (2011). Practical Aspects of microRNA Target Prediction. *Curr Mol Med.* Mar;11(2):93-109.

Wu, X., Watson, M., (2009). CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics*. 25(6), 832–3.

Xi, Y., Nakajima, G., Gavin, E. (2007). Systematic analysis of microRNA expression of RNA extracted from fresh frozen and formalin-fixed paraffin-embedded samples. *RNA.* 13:1668-74.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.* 37: D105-D110.

Yan, L., Huang, X-F., Shao, Q., Huang, M-Y., Deng, L., Wu, Q-L., Zeng, Y-X., Shao, J-Y. (2008). MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. *RNA*. 14: 2348-2360.

Yan, M., Li, X., Tong, D., Han, C., Zhao, R., He, Y., Jin, X. (2016). miR-136 suppresses tumor invasion and metastasis by targeting RASAL2 in triple-negative breast cancer. *Oncol Rep*. Jul;36(1):65-71.

Yang, R., Schlehe, B., Hemminki, K., Sutter, C., Bugert, P., Wappenschmidt, B. (2010). A genetic variant in the pre-miR-27a oncogene is associated with a reduced familial breast cancer risk. *Breast Cancer Res Treat*, 121:693–702.

Yang, X., Li, L. (2011). miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics,* 27(18):2614–5.

Yoon, S., De Micheli, G., (2006). Computational identification of microRNAs and their targets. *Birth Defects Res C Embryo Today*, Jun;78(2):118-28.

Yue, D., Guo, M., Chen, Y., & Huang, Y. (2012). A Bayesian decision fusion approach for microRNA target prediction. *BMC genomics*, 13 Suppl 8(Suppl 8), S13.

Zhang, C., Wang, C., Chen, X., Yang, C., Li, K., Wang, J., Dai, J., Hu, Z., Zhou, X., Chen, L., Zhang, Y., Li, Y., Qiu, H., Xing, J., Liang, Z., Ren, B., Yang, C., Zen, K., Zhang, C.Y. (2010). Expression profile of microRNAs in serum: a fingerprint for esophageal squamous cell carcinoma. *Clin Chem,* Dec; 56(12):1871-9.

Zhang, J. (2014). Upregulation of miR-194 contributes to tumor growth and progression in pancreatic ductal adenocarcinoma. *Oncology reports,* 31.3 1157-1164.

Zhang, L., Ding, L., Cheung, T.H., Dong, M.Q., Chen, J. (2007). Systematic identification of *C. elegans* miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Mol Cell,* 28:598–613.

Zhang, L., Sullivan, P.S., Goodman, J.C., Gunaratne, P.H., Marchetti, D. (2011). MicroRNA-1258 suppresses breast cancer brain metastasis by targeting heparanase. *Cancer Res*, 71(3):645-54.

Zhang, Y., Yan, L.X., Wu, Q.N., Du, Z.M., Chen, J., Liao, D.Z. (2011). miR-125b is methylated and functions as a tumor suppressor by regulating the ETS1 proto-oncogene in human invasive breast cancer. *Cancer Res*, 71:3552–62.

Zhang, Z. (2015). By downregulating Ku80, hsa-miR-526b suppresses non-small cell lung cancer. *Oncotarget,* 6.3 1462.

Zhao, H.K.X., Xia, X., Wo, L., Gu, X., Hu, Y., Xie, X., Chang, H., Lou, L., Shen, X. (2016). MiR-145 suppresses breast cancer cell migration by targeting FSCN-1 and inhibiting epithelial-mesenchymal transition. *Am J Transl Res*, 8:3106–14.

Zhou, Y., Dang, J., Chang, K.Y., Yau, E., Aza-Blanc, P., Moscat, J., Rana, T.M. (2016). MiR-1298 Inhibits Mutant KRAS-Driven Tumor Growth by Repressing FAK and LAMB3. *Cancer Res*, Oct 1;76(19):5777-5787.

Technical Supplements

Data Processing

First, the miRNA microarray data was pre-processed with background correction (subtracting background noise signal) (Quackenbush, 2002). In the sample quality control step, samples with missing values, those with a negative mean correlation, and samples below an RNA quality threshold (RIN score < 7) were removed. Next, the quality of probes was examined via the Signal-to-Noise ratio (SNR), where probes with consistently low SNR (<2) across all samples were removed. Then, Principal Component Analysis (PCA) identified some samples found to be outliers with very low raw intensity correlation, and they were excluded. Finally, in order to compare the intensity levels between samples, they were quantile normalized to make the intensity distributions identical. After preprocessing, we identified a major confounding effect via histogram clustering - most patient samples were hybridized on one day and controls on another day. Thus, for expression analysis we continued with 4 patient samples and 5 control samples.

Technical Methodology

The standard errors were moderated using an empirical Bayes model adjusting high variability genes down and low variability genes up (Smyth, 2004). Using a p-value threshold of 0.1 adjusted for multiple tests, fifty-five miRNAs were differentially expressed between breast cancer and control samples. The heatmap in Figure 2.2 illustrates the grouping of samples according to disease status based on differentially expressed miRNAs.

Figure 2.2. Class distinction of differentially expressed miRNAs.

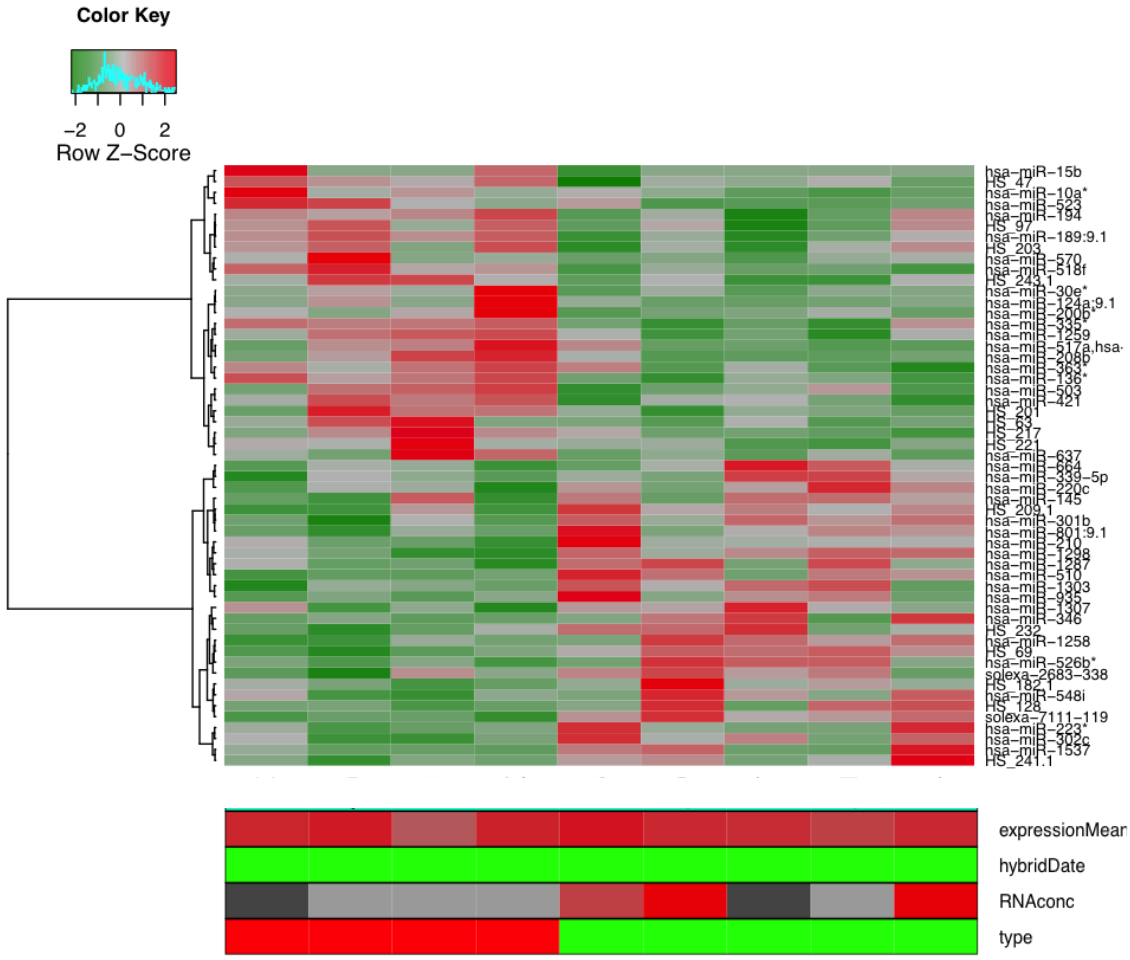

Figure 2.2. The differentially expressed miRNAs clustered based on the sample types: breast cancer and control samples. Heatmap colors represent mean centered fold change expression in log-space with Z-scores, meaning that red colored miRNAs are significantly differentially expressed higher in their sample types. Sample characteristics are represented in the boxes below each sample. The breast cancer samples are red, and control samples are in green. RNA concentration and expression mean is represented by a grey-red scale, where grey is low and red is high. All samples were hybridized on the same date and same slide.

MiRNA Annotation

Sixteen of the fifty-five miRNAs were not annotated in miRBase (Table 1), some of which were *de novo* miRNAs - newly discovered miRNAs that are not yet fully annotated nor validated by the scientific community. Thus, the "solexa-####-###" miRNA annotation in Table 1 is from Illumina's own sequencing efforts and the "HS_###" annotation is from sequencing efforts by Berezikov et al. (2006), which have not been validated. The sequences of the 16 unannotated miRNA were matched against the miRBase database, and only two of them were identified to be annotated miRNAs – hsa-miR-5006 and hsa-miR-548y.

To test the hypothesis that the rest of the 14 unannotated miRNAs are true miRNAs, we used the BLAT (Kent et al., 2002a) search tool on the USCS Genome Browser (Kent, 2002b) to locate the oligonucleotide sequence in the genome in order to inspect if their sequence characteristics follow the defining traits of a miRNA. BLAT is a pairwise sequence alignment tool for DNA/RNA. None of these unannotated miRNAs are conserved within mammals, nor did the secondary structure prediction method in BLAT identify the specific stem loop structure that defines precursor miRNAs. Figure 2.3 presents an example of differences observed between a true miRNA, miR-15b, and our hypothetical candidate miRNA, HS_128. Since there was no clear evidence that these 14 miRNAs are true miRNAs; they were excluded from further analysis, thus leaving us with 38 (of 49) (Table 2).

Figure 2.3. Sequence characteristics of HS_128 (A) and miR-15b (B).

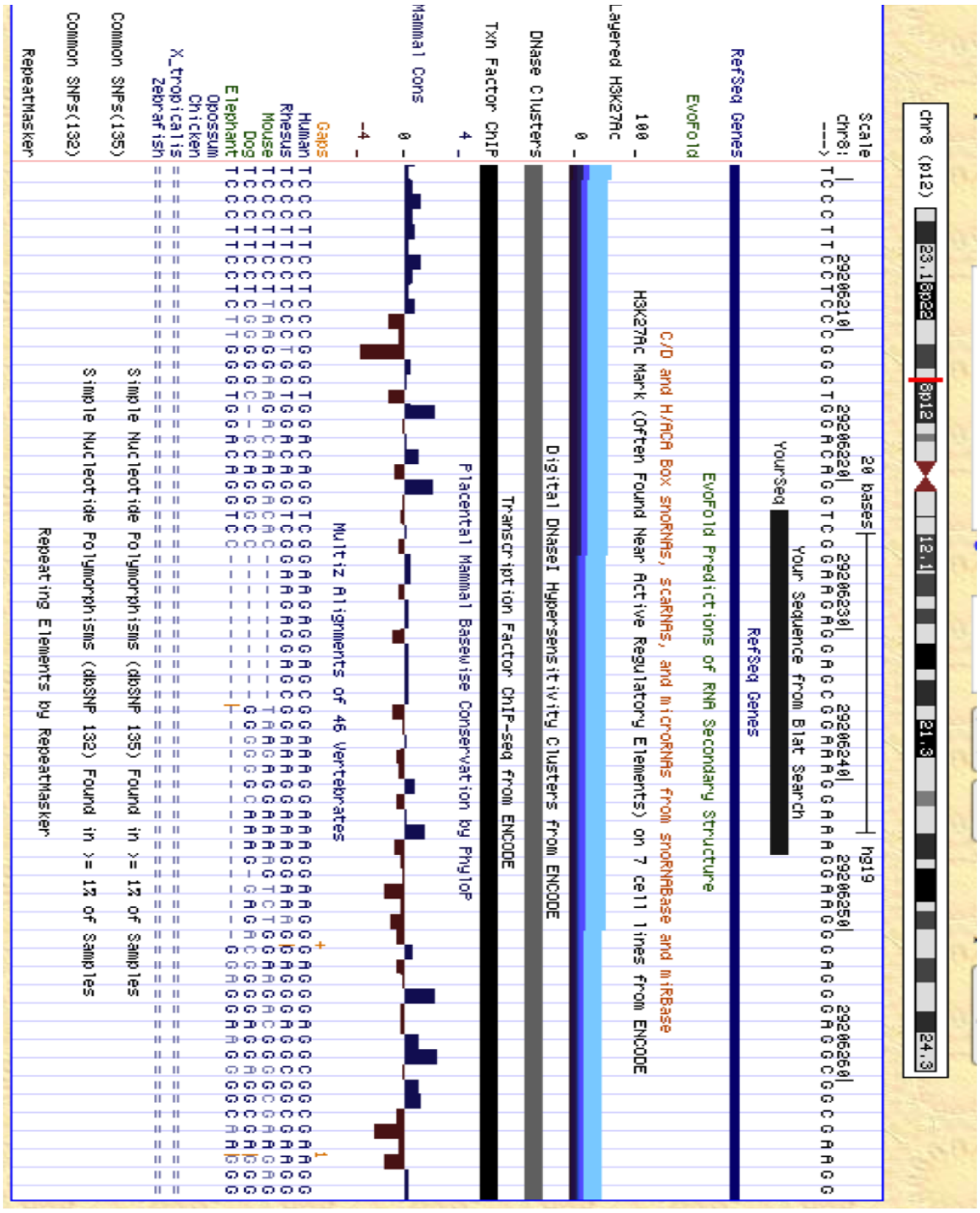

Figure 2.3B.

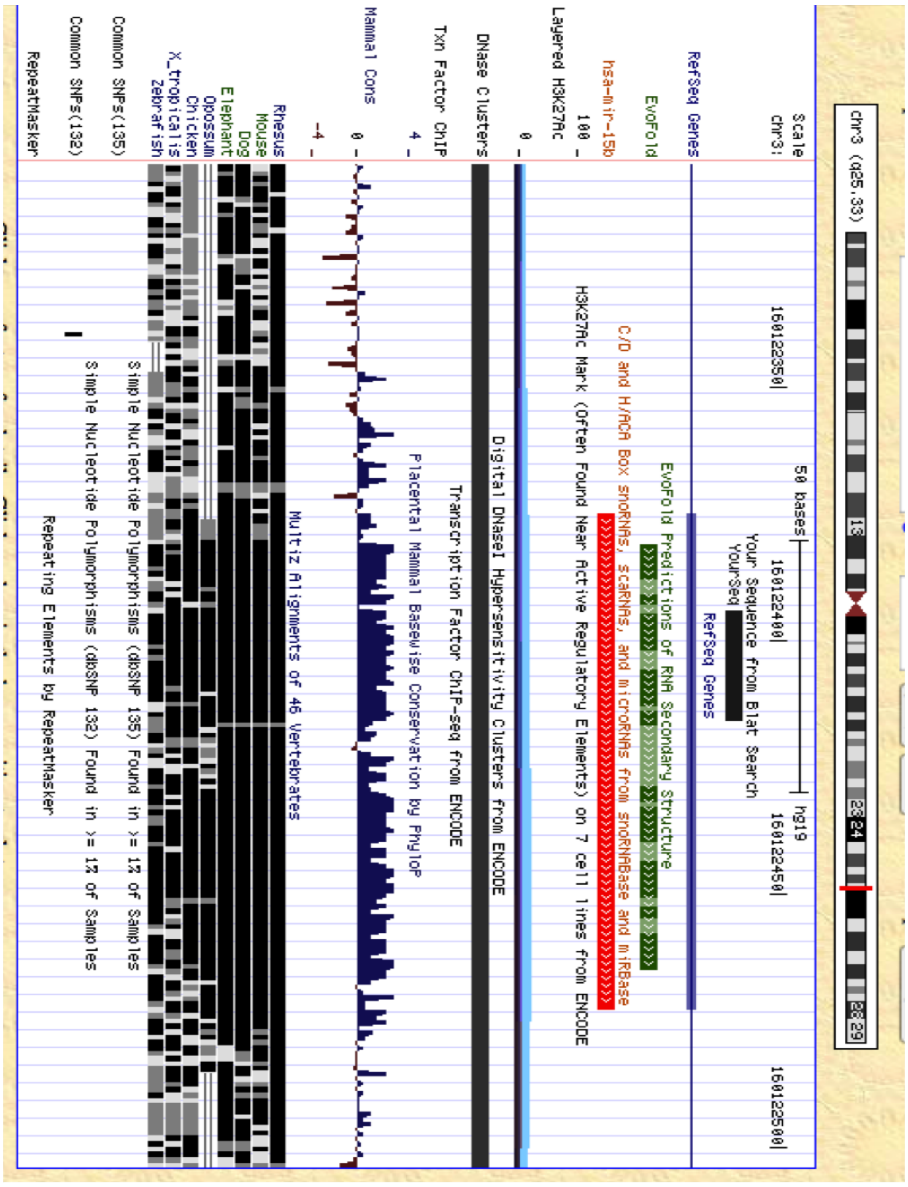

Figure 2.3. Sequence alignment of HS_128 (A) and miR-15b (B) against a database of animal genomes, miRBase, SNPs, etc. to inspect the sequence characteristics and test of they follow the defining traits of a miRNA. The miR-15b (B) sequence is found to be conserved, as per the Multiz alignments of 47 Vertebrates, and placental mammal basewise conservation by Phylop. The sequence has been annotated as has-miR-15b from the miRBase database, and was predicted to have a RNA secondary structure, which follows

the definition of a miRNA. Conversely, the sequence HS_128 (A) is not conserved at all, and has no annotation, and lacks an RNA secondary structure. Thus, it is concluded that this sequence is not a miRNA.

MiRNAs and the Importance of Conservation

As per the definition of what is a miRNA, the sequence must be conserved across multiple species. However, we found inconsistencies in our study, as there are many inconsistencies on other studies, making miRNAs challenging to study. In our case, we use HS_128 and miR-15b as examples to illustrate this point. For example, the sequence alignment of HS_128 (A) and miR-15b (B) against a curated database of animal genomes, miRBase, SNPs, etc. showed that HS_128 is not a real miRNA. The miR-15b (B) sequence is found to be conserved, as per the Multiz alignments of 47 Vertebrates, and placental mammal basewise conservation by Phylop. Also, the sequence has been annotated as has-miR-15b from the miRBase database, and the EvoFold Prediction was able to predict RNA secondary structure. All these characteristics follow the definition of a miRNA. Conversely, the sequence HS_128 (A) is not conserved at all, and has no annotation, and lacks an RNA secondary structure. Thus, it is concluded that this sequence is not a miRNA.

Differentially Expressed MiRNAs

List of differentially expressed miRNAs with their matching mirbaseIDs. Not all miRNAs were identified in the mirbase database of annotated and experimentally curated miRNAs.

	Name	**MirbaseID**
1	HS_128	
2	HS_201	hsa-miR-5006
3	HS_203	
4	HS_217	
5	HS_221	
6	HS_232	

7	HS_47	
8	HS_63	
9	HS_69	
10	HS_97	
11	solexa-2683-338	
12	solexa-7111-145	hsa-miR-548y
13	hsa-miR-10a	hsa-miR-10a
14	hsa-miR-124	hsa-miR-124
15	hsa-miR-1258	hsa-miR-1258
16	hsa-miR-1287	hsa-miR-1287
17	hsa-miR-1298	hsa-miR-1298
18	hsa-miR-1303	hsa-miR-1303
19	hsa-miR-1307	hsa-miR-1307
20	hsa-miR-136*	hsa-miR-136
21	hsa-miR-145	hsa-miR-145
22	hsa-miR-1537	hsa-miR-1537
23	hsa-miR-15b	hsa-miR-15b
24	hsa-miR-189	hsa-miR-24
25	hsa-miR-194	hsa-miR-194
26	hsa-miR-200b*	hsa-miR-200b
27	hsa-miR-200b	
28	hsa-miR-208b	hsa-miR-208b
29	hsa-miR-210	hsa-miR-210
30	hsa-miR-223	hsa-miR-223
31	hsa-miR-301b	hsa-miR-301b
32	hsa-miR-302c	hsa-miR-302c
33	hsa-miR-30e*	hsa-miR-30e
34	hsa-miR-335	hsa-miR-335

35	hsa-miR-339	hsa-miR-339
36	hsa-miR-346	hsa-miR-346
37	hsa-miR-363	hsa-miR-363
38	hsa-miR-421	hsa-miR-421
39	hsa-miR-503	hsa-miR-503
40	hsa-miR-510	hsa-miR-510
41	hsa-miR-517a	hsa-miR-517a
42	hsa-miR-518f	hsa-miR-518f
43	hsa-miR-523	hsa-miR-523
44	hsa-miR-526b*	hsa-miR-526b
45	hsa-miR-548i	hsa-miR-548i
46	hsa-miR-570	hsa-miR-570
47	hsa-miR-637	hsa-miR-637
48	hsa-miR-664	hsa-miR-664
49	hsa-miR-935	hsa-miR-935