Faculty of Engineering and Information Technology
University of Technology Sydney

Application of Information Theory to RNA-sequencing Data Sets for Better Understanding of Human Cancers

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Chaowang Lan

February 2020

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Chaowang Lan declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: Jan. 30. 2020

Acknowledgments

Firstly and formost, I profoundly grateful to my chief supervisor Prof. Jinyan Li and my co-supervisor Prof. Gyorgy Hutvagner. They have guided and encouraged me to carry on through these four years and allowed me to grow as a good researcher. My research would have been impossible without their assistance and support. I have been extremely lucky to have two supervisors who cared about my research and responded my questions patiently.

I would like to express my sincere gratitude to Dr. Eilleen M.McGowan in the School of Life Sciences of UTS. Thank you for offering time to revise my paper and providing me lots of information on breast cancer. Without her help, I would not have been published paper on the high quality journal. Thank you so much for your knowledge in cancer research, writing skill, and support.

I really grateful to two my team members Dr. Hui peng and Yi Zheng, not only for providing valuable feedback and developing my ideas in my research, but also their help in my life in Australia. I also grateful to Demi Truong Aung for the effect on doing the wet-lab experiment to confirmed my prediction. When you told me that the wet-lab experiment is confirmed my prediction, I realised my research was very important for biological research.

Of course, I sincere thanks to Dr. Renhua Song, Dr. Jing Ren, and Dr. Shameek Ghosh for their help to discover the research topic at the beginning of my PhD study. I would like to thank the other team members: Yuangsheng Liu, Zhixun Zhao, Xiaocai, Zhang, Tang Tao, and Xuan Zhang, not only for giving me large amount of inspiring advise to me research in group meeting,

but also for giving me four years of interesting activities. I could hardly forget the wonderful time when we participated in these interesting activities.

I thank to my PhD scholarship funding source which is provided by China Scholarship Council. I also think to all staffs in Advanced Analytics Institute and School of Computer Science for offering conveniences in my research.

Finally, I deepest think to my parents and brother. It is because of their trust, encouragement, and unconditional support to me, I could overcome the difficulties encountered during my four years study. Thank your very much!

Chaowang Lan
Augest 2019 @ UTS

Contents

List of Figures

List of Tables

List of Publications

Below is the list of journal and conference papers associated with my PhD research:

Journal Papers Published

- **Lan, C.**, Peng, H, McGowan, E.M., Hutvagner, G., and Li, J., 2018.An isomiR expression panel based novel breast cancer classification approach using improved mutual information. BMC medical genomics, 11(6), pp.118

- Zhao, Z, Peng, H, **Lan, C.**, Zheng, Y, Fang, L,and Li, J., 2018. Imbalance learning for the prediction of N 6-Methylation sites in mRNAs. BMC genomics, 19(1), pp.574

- Liu, Y, **Lan, C.**, Blumenstein, M, and Li, J., 2017. Bi-level Error Correction for PacBio Long Reads. IEEE/ACM Transactions on Computational Biology and Bioinformatics, `https://doi.org/10.1109/TCBB.2017.2780832`

- Peng, H, **Lan, C.**, Zheng, Y, Hutvagner, G., Tao, D, and Li, J., 2016. Cross disease analysis of co-functional microRNA pairs on a reconstructed network of disease-gene-microRNA tripartite. BMC Bioinformatics, 18(1), 193:1-193:7

- **Lan, C.**, Chen, Q, and Li, J., 2016.Grouping miRNAs of similar functions via weighted information content of gene ontology. BMC

Bioinformatics, 17(S-19), pp.159-170

- Zheng, Y, **Lan, C.**, Peng, H, and Li, J., 2016. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp.2460-2463

- **Lan, C.**, Peng, H, Hutvagner, G., and Li, J.,Construction of Competing Endogenous RNA Networks from Paired RNA-seq Data Sets by Pointwise Mutual Information. (accepted by BMC genomics)

Abstract

This research utilizes information theory to study the regulatory roles of non-coding RNAs in human cancers. microRNAs (miRNA) are small non-coding RNAs binding to mRNAs to suppress protein expression. Long non-coding RNAs (lncRNA) can act as competing endogenous RNAs (ceRNAs) to compete with mRNAs to bind to miRNAs. LncRNAs, miRNAs, and mRNAs form the ceRNA networks, which play a vital role in regulating molecular pathways of human cancers. Furthermore, miRNA isoforms, which are called isomiRs, are also enable to regulate the gene expression and could be used to distinguish cancer subtypes. Therefore, constructing ceRNA regulatory networks and identifying isomiRs as cancer subtype biomarkers are very important for understanding the regulatory role of non-coding RNAs in cancers.

Current methods for constructing ceRNA networks and discovering biomarkers that faithfully classify different cancer subtypes have some limitations. Information theory is a powerful tool for better understanding the regulatory role of non-coding RNAs in human cancer. This thesis utilizes information theory for constructing ceRNA network and discovering human cancer subtype biomarkers in cancers. The novel contributions to the research field by this thesis are enlisted below:

- A competition rule-based pointwise mutual information is proposed to construct ceRNA networks.
- An improved mutual information and an information gain are developed to identify isomiRs as biomarkers for classifying different cancer

subtypes.

- A distribution-based method is proposed to flitter out the noisy data in RNA-seq data.

Three case studies have been performed to study the regulatory roles of non-coding RNAs in human cancers. (1) The first case study is to construct the competition relationships between lncRNA, miRNA, and mRNA in breast cancer by using pointwise mutual information. (2) The second case study is to utilize the improved mutual information to discover isomiR biomarkers for classifying different breast cancer subtypes. (3) The third case study applies the improved information gain to detect isomiR based biomarkers to classify different glioma subtypes.

Chapter 1

Introduction

1.1 Background

In this thesis, we manly focus on utilizing information theory to construct the RNA regulatory network and discover biomarkers for better understanding of the biological process of human cancers. This section briefly introduces three critical measures of information theory in biological research, relatived studies of RNA regulatory mechanism in human cancers, and the background of RNA-sequencing data.

1.1.1 Information theory and its application in biological research

Information theory was developed for studying the quantification, storage and transmission of information (Vinga 2013). There are three key measures in information theory. The first measure is entropy. Entropy measures the probability of the outcome of a random process. The smaller the entropy, the more likely the outcome of the random process. Therefore, the entropy could be applied to measure the important of the outcome in a random process. The second is mutual information. The mutual information measures the mutual dependence relationship between two variables. In reality, a variable

always has many events. Some researches studied the association relationship between two events in two variables. Thus, the third key measure of information theory, which is called pointwise mutual information, could be applied to find out the relationship between two events. The mutual information is constructed by the pointwise mutual information. The mutual information refers to the expected value of all events, while pointwise mutual information refers to single events.

Information theory had been widely used in many biological researches. For example, predicting the correlation between DNA mutations and disease (Milenkovic, Alterovitz, Battail, Coleman, Hagenauer, Meyn, Price, Ramoni, Shmulevich & Szpankowski 2010), analyse biology evolution (Danchin, Charmantier, Champagne, Mesoudi, Pujol & Blanchet 2011), and discovering the co-regulatory networks in biological process (Mousavian, Kavousi & Masoudi-Nejad 2016). Constructing the regulatory network and identifying biomarkers help us to understand the regulatory mechanism of molecules in cancer. Information theory could be applied to construct the regulatory network and identify biomarkers.

Constructing regulatory network is regarded as discovering the relationship between molecules. The pointwise mutual information and mutual information could be applied to measure the relationship between different kinds of molecules. Thus, the regulatory networks enable to be constructed by using pointwise mutual information or mutual information. The biomarker is the molecule which state or change state indicates biological processes. Identifying the biomarker can be viewed as selecting the most critical molecules that are involved in the regulation of gene expressions in cancers or discovering the relationship between molecules and different cancer subtypes. The entropy measures the probability of molecule as biomarkers to regulate the pathway of cancer and the mutual information could discover the relationship between molecules and cancer subtypes. So, the biomarker can be discovered by using the entropy and mutual information. The three key measures in information theory can be applied to construct the regulatory

network and discover the biomarker. Therefore, information theory is a powerful tool for understanding the role of molecules in human cancers.

1.1.2 The regulatory mechanisms of long noncoding RNA, microRNA (isomiR), and mRNA in cancers

The RNA can be divided into coding RNA, such as messenger RNA (mRNA) and non-coding RNA, e.g., long non-coding RNA (lncRNA). However, the coding RNA can translate into protein and non-coding RNA cannot. The protein translated by mRNA is the basis of living tissues and some of them participate in the development of cancer (Lodish, Berk, Zipursky, Matsudaira, Baltimore & Darnell 2000). The mRNA contains five regions: $5'$ cap, $5'$ untranslated region, coding sequence, $3'$ untranslated region ($3'$-UTR), and the poly(A) tail. The over-expressed mRNA could produce large amount of protein and then enhancing or inhabiting the development of cancers. For example, the over-expressed mRNA *PTBP3* promotes the growth and metastasis of breast cancer cell (Hou, Li, Chen, Chen, Liu, Li, Bai & Zheng 2018); the highly expressed mRNA *FOXA2* inhibits the proliferation, invasion, and tumorigenesis in glioma cell (Ding, Liang, Gao, Li, Xu, Fan & Chang 2017). Therefore, it critical to know the expression of mRNA for understanding the regulatory mechanism of mRNA in cancer.

Although mRNA is very important for the regulation mechanism of cancer, the mRNA expression is regulated by non-coding RNA. It implies that the non-coding RNA could regulate the gene expression in cancer. Therefore, the non-coding RNA plays an important role in cancer. The non-coding RNA is divided into two main groups: small non-coding RNA and long non-coding RNA.

There are many different types of small non-coding RNAs have been discovered, such as MicroRNA (miRNA), Piwi-associated RNAs (piRNAs), small nucleolar RNA (snoRNA), tRNA-derived small RNA (tsRNA), and

small interfering RNA (siRNA). In this thesis, we focus on the miRNA and its isoforms called isomiR. TMiRNAs are short (21-22nt long) regulatory RNAs. The miRNA could bind to the 3′-UTR of mRNA and therefore, suppress the protein expression. MiRNAs recognise their targets by binding to complementary sites between the seed region of the miRNA, which spans from the 2nd to 7th nucleotides from the miRNA 5′-end and the target mRNA (Kehl, Backes, Kern, Fehlmann, Ludwig, Meese, Lenhof & Keller 2017). The seed region of a miRNA is a key determinant of its targeting specificity, one nucleotide changed in the seed region could alter its target mRNA.

Mature miRNAs are generated from longer transcripts via several sequential processing steps (Li, Liao, Ho, Tsai, Lai & Lin 2012). First the primary miRNA transcripts (pri-miRNA) are cleaved by the Microprocessor complex that contains Drosha, an RNase III enzyme in the nucleus (Maher, Timmermans, Stein & Ware 2004). After transporting the cleaved precursor miRNAs (pre-miRNA) to the cytoplasm miRNAs are further processed by another RNase III enzyme, Dicer, to produce small miRNA duplexes (Hutvágner, McLachlan, Pasquinelli, Bálint, Tuschl & Zamore 2001). Alterations in miRNA maturation, such as the alternative and imprecise cleavage of Drosha and Dicer, or the turnover of miRNAs could result in miRNAs that are heterogeneous in length and/or sequence (Swierniak, Wojcicka, Czetwertynska, Stachlewska, Maciag, Wiechno, Gornicka, Bogdanska, Koperski, de la Chapelle et al. 2013, Neilsen, Goodall & Bracken 2012). These variants are called isomiRs (isoforms of miRNA) and can be divided into three main categories: 3′ isomiR (trimmed or addition of one or more nucleotides at the 3′ position), 5′ isomiR (trimmed or addition of one or more nucleotides at the 5′ position), and polymorphic isomiR (some nucleotides within the sequence are different from the wild type mature miRNA sequence) (Neilsen et al. 2012).

Different variant types of isomiRs could have different function from their wild type counterparts. The 3′ isomiR could contribute to changes

in complementarity between miRNAs and their targets and therefore, weakening or strengthening their regulatory power. The 5′ isomiRs could recognise novel target genes since the seed region of the 5′ isomiR is different from the wild type miRNA. The potential changes in the function of polymorphic isomiR is defined by the position of the substitute nucleotide. If the substitute nucleotide is occurred at the seed region, the isomiR could regulate a novel set of transcripts.

Both the miRNA or isomiR could regulate the gene expression, the miRNA (isomiR) and mRNA construct a miRNA(isomiR)-mRNA interaction network. This interaction network reflects the regulatory mechanism of miRNA (isomiR) and mRNA. It is critical to construct this interaction network to understand the regulatory mechanism of cancer. For instance, Li and colleagues analysis the miRNA-mRNA interaction network in breast cancer and they found the key RNAs in regulating the pathway of breast cancer with brain metastasis (Li, Peng, Gu, Zheng, Feng, Qin & He 2017). This finding provides a novel strategy for the treatment of breast cancer with brain metastasis. However, this interaction network cannot fully explain the regulatory mechanism of RNAs in cancer. This is because the long non-coding RNA also plays a vital role in regulating their interaction.

The lncRNA is defined as the non-coding RNA longer than 200 nucleotides. It was first identified in 2002 (Consortium, Team et al. 2002). Since lacking of functional annotation, lncRNA had been overlooked for a long time (Wei, Luo, Zou & Wu 2018). In 2005, researcher found that lncRNA could demarcate chromosomal domains of gene silencing and influence the gene expression in development and disease states (Rinn, Kertesz, Wang, Squazzo, Xu, Brugmann, Goodnough, Helms, Farnham, Segal et al. 2007). Later, researchers also found that lncRNA regulates a range of biological processes, such as cancer development, gene imprinting, and modulate the enzymatic activity (Gibb, Brown & Lam 2011, Quinn & Chang 2016, Marchese, Raimondi & Huarte 2017). These findings indicate that lncRNAs play important in regulating biological processes.

Recently studies show that lncRNAs could be competing endogenous RNAs to compete with mRNAs for binding to the same miRNAs (Salmena, Poliseno, Tay, Kats & Pandolfi 2011). For instance, *BRAFP1* can compete with gene *BRAF* for binding to the same miRNA hsa-miR-543 in lymphoma (Song, Liu, Liu & Li 2015). *PTENP1* can compete with gene *PTEN* for binding to the same miRNA hsa-miR-17-5p in hepatocellular carcinoma (Karreth, Reschke, Ruocco, Ng, Chapuy, Léopold, Sjoberg, Keane, Verma, Ala et al. 2015). An lncRNA can bind to many miRNAs and a miRNA is able to regulate multiple mRNAs. Therefore, these lncRNAs, miRNAs, and mRNAs construct a large and complex regulatory network that is called competition endogenous RNA networks. These ceRNA networks not only provide a reasonable justification for the presence of lncRNA, it also provides a new and global function map of lncRNA (Yang, Wu, Gao, Liu, Jin, Wang, Wang & Li 2016). Understanding this complex regulatory network is useful for detecting patterns for early cancer diagnosis (Sanchez-Mejias & Tay 2015) and developing new concepts for cancer treatment (Ebert, Neilson & Sharp 2007).

There are three common characteristics in the ceRNA network (Quinn & Chang 2016). First, the relative concentration of the ceRNAs. Second, the ceRNA is the primary target of the miRNA. Third, the relationships between the lncRNA, miRNA, and mRNA is the competition relationship. The competition relationship states that when the expression level of the ceRNA is very high, the ceRNA can compete for binding to the miRNA and decrease the expression level of the miRNA. Since miRNA has a low expression level, less number of miRNAs bind to its target mRNA. Therefore, the expression level of the mRNA becomes high. In contrast, when the expression level of the ceRNA is very low, the expression level of the miRNA will be high; a high expression level of miRNA leads to a low expression level of mRNA.

Although miRNA-mRNA interaction network and ceRNA network reflect the regulatory mechanism of the biological process, they have two different

aspects. The first is that the miRNA-mRNA interaction network is the regulatory relationship between two different types of RNAs. While the ceRNA network is the regulatory relationship between three different types of RNAs. The second is that the expression relationship between RNAs are different. The expression relationship in miRNA-mRNA interaction network is always negative correlation. However, in the ceRNA network, the expression relationship between lncRNA and mRNA is positive correlation and the expression relationship between miRNA and mRNA (or lncRNA) is negative correlation.

The coding RNA and non-coding RNA not only regulate the molecular pathway of cancer, they can be biomarkers for indicating different cancer subtypes as well. Cancer is a heterogeneous disease and could be divided into different subtypes (Kuijjer, Paulson, Salzman, Ding & Quackenbush 2018). The cancer subtype provides useful insight into disease pathogenesis and cancer treatment (Guo, Shang & Li 2019). The mRNA could be cancer subtype biomarker that indicates different cancer subtypes. This is because mRNA can direct regulate the biological pathway that related to cancer subtype. For example, Parker and colleagues defined the 50 genes that classified different breast cancer subtypes (Parker, Mullins, Cheang, Leung, Voduc, Vickery, Davies, Fauron, He, Hu et al. 2009). However, miRNAs and isomiRs provide a potentially better alternative biomarker for classifying cancer subtypes compared to mRNA since they are regulatory "hubs" of gene expression. Therefore, the changes in their expression could influence multiple downstream mRNAs and therefore diverse biological pathways. Many groups have found that miRNA or isomiR is a suitable biomarker for classifying different cancer subtypes (Chen & Wong 2017, Volinia & Croce 2013). For example, Telonis and colleagues demonstrated that isomiRs were able to classify two breast cancer subtypes (Telonis, Loher, Jing, Londin & Rigoutsos 2015). Telonis and colleagues stated that isomiRs could be biomarkers for classifying 32 different cancers (Telonis, Magee, Loher, Chervoneva, Londin & Rigoutsos 2017).

1.1.3 The introduction of RNA-sequencing data

Constructing the RNA regulatory network requires to understand the relationship between RNAs. The relationship between RNA is detected from the expression level of RNAs in biological tissues. The discovery of biomarkers also requires to quanntitative expression level of RNAs in biological tissues. Therefore, quantifying the expression level of RNAs in biological tissues is the foundation for understanding the regulatory mechanism of RNAs in cancers. RNA-sequencing (RNA-seq) uses the next-generation sequencing technology to provide RNA abundance and diversity of a biological tissue (Griffith, Walker, Spies, Ainscough & Griffith 2015). In general, the RNA-seq method consists of two main steps. The first step is that RNA isolation from the biological tissue. The second step is that RNA library preparation. In this process the cellular RNAs (mRNA, lncRNA, small RNAs) are enriched and transformed into cDNA a stable form of nucleic acid that allows the amplification of each RNA molecule using PCR (polymerase chain reaction).

Most of the sequencing technologies used to sequence RNAs generate vast amount of sequence fragment, which is called read, and the sequencing quality score of the read. The read and its quality score are written in a 'fastq' file. This 'fastq' file is called RNA-seq data. Reads stored this file is used to help to construct the genome of a species (Chang, Li, Liu, Zhang, Ashby, Liu, Cramer & Huang 2015), identify the biomarker of the disease (Sahraeian, Mohiyuddin, Sebra, Tilgner, Afshar, Au, Asadi, Gerstein, Wong, Snyder et al. 2017), discover variant RNAs (Richter, Hoffman, Manheimer, Patel, Sharp, McKean, Morton, DePalma, Gorham, Kitaygorodksy et al. 2019), detect gene fusion (Maher, Kumar-Sinha, Cao, Kalyana-Sundaram, Han, Jing, Sam, Barrette, Palanisamy & Chinnaiyan 2009), and calculate the expression level of RNAs. In this thesis, we focus on using RNA-seq data to discover the biomarker and construct the regulatory network in cancers. Therefore, we use this file to calculate the expression level of RNAs in biological tissues.

Calculating the expression level of RNA has four steps: (1) quality control. The RNA-seq file has many low quality reads, such as there are many nucleotides cannot be detected in a read. This low quality read is caused by the sequencing technology. Although the sequencing machine has very low error sequence ratio (the error sequencing error ratio of the second general sequencing machine is about 1%), the number of RNA detected from biological tissue is extremely huge and therefore, the RNA-seq files contains many low quality reads. Further, every RNA must induce adapters before sequencing. Therefore, some reads may have adapters in the RNA-seq data. The low quality read and adapter have negative influence on the downstream analysis. Thus, the low quality read and adapters must be removed. (2) Construct a genome reference database. The RNA-seq file contains large amount of read. However, the data is messy and we cannot identify the read is originated from the genome. The genome reference database provides a template for annotating the read. (3) Mapping the read to the reference genome database. (4) Selecting a suitable metric to calculate the expression level of the RNAs.

There are three main different metrics measuring the expression level of RNA: FPKM (Fragments per kilo base per million mapped reads), RPKM (Reads per Kilo base Million mapped reads), and RPM (Reads per million mapped reads). All these metrics use the number of read mapped to reference RNA and the total read of the RNA-seq file to calculate the expression level of RNA. However, FPKM and RPKM take the length of the RNA into consideration while RPM does not. The FPKM and RPKM are suitable for measuring expression level of long RNA, such as lncRNA and mRNA. The expression level of short RNA is measured by using RPM. The long RNA had been fragmented before sequencing. The longer the RNA, the more read this RNA had been fragmented. The length of the RNA influences the number of read mapped to the reference RNA and therefore, affecting its expression level. However, the short RNA, such as miRNA, is not fragmented before sequencing. Thus, the length of RNA cannot influence the number of

read mapped to the reference. Thus, the length of short RNA cannot affect its expression level.

The RNA-seq data could be download from many websites, for instance, TCGA (The Cancer Genome Atlas) website (`https://portal.gdc.cancer.gov/`), NCBI(The National Center for Biotechnology Information) website (`https://www.ncbi.nlm.nih.gov/`), and EMBL-EBI (The European Bioinformatics Institute) website (`https://www.ebi.ac.uk/`). The TCGA website provides large amount of RNA-seq data on human cancer tissues and a few RNA-seq data on human normal tissues. This RNA-seq data is always used to analyse the regulatory mechanism of RNAs in cancers. NCBI website and EMBL-EBI website offer the RNA-seq data for various tissues of multiple species.

1.2 Research aims and objectives

This thesis discusses two research topics: (1) understanding the regulatory relationship between RNAs in cancers and (2) identifying biomarkers for classifying different cancer subtypes in variety of cancers. We believe that application of the information technologies improves the understanding the regulatory role of non-coding RNAs in cancers and provide novel biomarkers for cancer subtypes.

Aim 1: The ceRNA network reflects the regulation relationship between lncRNA, miRNA, and mRNA. In information theory, the pointwise mutual information is used to analyse the relationship between variables, therefore we apply and improve this method to construct ceRNA networks in breast cancer.

Objectives:

- Calculating the expression level of lncRNA, miRNA, and mRNA from RNA-seq data of breast cancer tissue and normal tissue.
- Identifying lncRNA, miRNA, and mRNA that differentially expressed between breast cancer tissue and normal tissue. Using these cancer

related lncRNA, miRNA, and mRNA to construct the candidate ceRNA network.

- Utilizing improved the pointwise information to measure the competition relationship of the candidate ceRNA network. The candidate ceRNA network which has the high pointwise mutual information is the final ceRNA network.
- Applying the KEGG pathway to analyse the function of the ceRNA network in breast cancer.

Aim 2: The presence of isomiR in RNA sequencing data increases the information content of small RNA sequencing while providing a hidden and largely unresearched layer of regulation of gene expression. We reveal the biomarker potential of isomiRs using mutual information and information gain to distinct and characterise breast cancer and glioma subtypes, respectively. These two approaches also provide mechanistic insight about the mechanism of both cancers.

Objectives:

- Quantifying the expression level of isomiR from RNA-seq data.
- Using a null hypotheses method to remove the lowly expressed isomiR.
- Utilizing the mutual information and information gain to discover the biomarker for classifying breast cancer subtypes and glioma subtypes, respectively.
- Applying the KEGG pathway and wet-lab experiment to analysis the regulatory mechanism of biomarker in breast cancer subtypes and glioma subtypes.

1.3 Research contribution

We develop novel methods to construct ceRNA network in breast cancer and identify biomarkers for classifying breast cancer and glioma subtypes. Our

contributions are showed below:

- **Developed new method for discovering the competition relationship between lncRNAs, miRNAs, and mRNAs.**

  Chapter 3 presents a novel method for constructing the competition relationship between RNAs. The contributions of this research have four aspects: (1) using the competition regulatory mechanism to construct candidate ceRNA network. The competition regulatory mechanism is that if the lncRNA highly expressed (or down-regulated) in breast cancer, the miRNA down-regulated (or up-regulated) breast cancer, and mRNA up-regulated (or down-regulated) breast cancer. This competition rule obeys the regulatory mechanism of ceRNA network. (2) The competition relationship is a complex relationship: the relationship between lncRNA and mRNA is positive correlation and the relationship between mRNA and miRNA is negative correlation. Therefore, calculating the competition score of candidate ceRNA networks should consider this competition relationship. We combine the competition rule and the pointwise mutual information to calculate the competition score. The competition rule is that when the expression level of lncRNA is high (low), the expression level of miRNA is low (high) and the expression level of mRNA is high (low). Therefore, our method provides a suitable metric for measuring the competition relationship between RNAs. (3) The null hypothesis is applied to select the ceRNA network that has high competition score. (4) We construct the ceRNA network reveals the regulatory mechanism of RNA in the growth, development, and metastasis of breast cancer. Further, some ceRNA networks have the same lncRNA as ceRNA to compete with mRNA for binding to miRNA. This lncRNA could be an effective drug target for breast cancer treatment.

- **Proposed an improved mutual information to identify biomarker for classifying different breast cancer subtypes.**

Chapter 4 presents an improved mutual information method to identify biomarker of different cancer subtypes. The contributions in this research are that (1) this method extends the ability of mutual information for discovering important features in the dataset that the feature is continue data and label data is discrete data. Further, this method could classify multiple classes for discovering the feature. (2) The mRNA is the tradition biomarker for classifying different breast cancer subtypes. Our method finds out fewer isomiRs are required to classify different breast cancer subtypes compare with to the number of mRNA. The isomiR is more effective than mRNA for classifying different breast cancer subtypes.

- **A novel method is developed for discovering glioma subtype biomarker.**

  Chapter 5 proposes a new method for identifying isomiR for classifying different glioma subtypes. The information gain could be used to identify the isomiR that could be biomarker for classifying different cancer subtype. However, calculating the information gain requires suitable cut points for binning the data. In this research, we propose a distribution-based method to find out cut points for binning data.

- **Using a null hypothesis method to filter out the noisy data in RNA-seq data.**

  There is a common problem in Chapter 4 and 5: how to filter out the lowly expressed isomiR. The lowly expressed isomiR has limited influence the biological process of cancer and negative influence on the result. The lowly expressed isomiR is defined as the relative low expressed isomiR compare with other isomiR. So the lowly expressed isomiR is affected by the expression level of the entire isomiR. We apply a null hypothesis method to calculate a threshold for identifying the lowly expressed isomiR. This method is based on the total expression levels of the entire isomiR. Therefore, it provides a soft threshold to

remove the lowly expressed RNAs in different depth of sequencing.

1.4 Thesis Structure

This thesis is structured as fellow: Chapter 1 introduces the application of information theory in bioinformatics and the function of RNAs in cancer. It also describes the aims and contributions of our research. Chapter 2 presents the current methods for constructing RNA network and discovering biomarkers. Further, the bioinformatics tools and computational methods are illustrated in this Chapter. Chapter 3 proposes a novel method, which is based on competition rule and pointwise information method, for constructing the ceRNA network in breast cancer by using paired RNA-seq data. Chapter 4 displays an improved mutual information method to identify the isomiR biomarker that could classify different breast cancer subtype. Further, a 'soft' method, which is based on null hypothesis method, is developed for removing the noisy data. Chapter 5 designs a distribution-based method for finding out the cut point for binning data. Then applying the information gain to find out the isomiR that regulates glioma subtypes. Chapter 6 summaries the research and suggests the future.

Chapter 2

Literature Review and Background

This Chapter introduces the methods for constructing the RNA regulatory network in cancers and identifying the biomarkers for classifying different cancer subtypes. Section 2.1 presents the technologies of predicting miRNA targets. Predicting miRNA target is the foundation of constructing RNA regulatory network. Further, these technologies could be used to discovering the function of isomiR in regulating biological processes. Section 2.2 shows the main algorithms for constructing the RNA regulatory network. Section 2.3 describes the popular methods for discovering biomarker and their weaknesses and advantages. Section 2.4 displays the tools of discovering the function of the RNA regulatory network and biomarkers. The tool is used to annotate or validate the function RNA regulatory network and biomarkers in cancer.

2.1 Technologies for miRNA (isomiR) target Prediction

In the ceRNA network, the lncRNA to competes with mRNA for binding to miRNA. Thus, the miRNA could bind to both lncRNA and mRNA.

Further, the 5′ isomiR could regulate novel gene compare with its wild type miRNA to influence the biological processes. Therefore, identifying the target of miRNA is very important to construct the ceRNA network and understand the function of isomiR. There are large amount of technologies for predicting miRNA target have been developed and they can be classified as two categories: traditional method and high-throughput based method.

2.1.1 Traditional methods for predicting miRNA target mRNA

The traditional method predicts the miRNA target through the feature of sequence. There are two main features used in traditional methods: the free energy of target site and the perfect pairing of miRNA seed region to mRNA target (Liu & Wang 2019).

The free energy of the target site measures the stability of miRNA binds to mRNA. The lower free energy of the target site, the more stable the miRNA binds to mRNA. The AU content is defined as the ratio of the adenosine or uridine base in the target site and could influence the free energy of target site. The AU-rich region is more likely to be single stranded and relative to structural accessibility (Liu, Mallick, Long, Rennie, Wolenc, Carmack & Ding 2013). Therefore, the higher the AU content at the target site, the more stable for miRNA binds to mRNA. The perfect pairing of miRNA seed region to mRNA target have significance influence on the miRNA binds to mRNA. It had been found that the miRNA was likely to bind to mRNA which the seed region of miRNA was complementary to the mRNA sequence even although the other position of miRNA could not complementary to the mRNA sequence (Ellwanger, Büttner, Mewes & Stümpflen 2011). Therefore, the seed region is very important for predict the miRNA target. Most of the methods apply these two features for identifying the miRNA target. However, different methods use different strategies.

The miRanda algorithm applies the sequence alignment method to predict miRNA targets. This sequence alignment utilizes the dynamic programming

to search for maximal local target site of miRNA sequence and mRNA sequence (John, Enright, Aravin, Tuschl, Sander, Marks et al. 2004). Then calculating the complementarity score of the target site. There are some rules in calculating the complementarity score. For example, no mismatch at positions 2 to 4 and the complementarity score at first eleven positions have high weight. All these rules ensure that the seed region of miRNA could perfect complementarity to mRNA. Further, this method calculates the free energy of the target site. The candidate target site should have high complementarity score and low free energy.

Tragetscan is one of the popular method for predicting miRNA target (Lewis, Burge & Bartel 2005). This algorithm applies the sequence alignment to predict miRNA target. This algorithm focus on finding the binding site that complementarity to the miRNA seed region and has low folding free energy.

The traditional method for predicting miRNA target has a limitation: the features for predicting miRNA target is not validated. It means that some features used in the traditional method may have limit influence on predicting miRNA target and therefore, the accuracy rate of traditional method is very low.

2.1.2 High-throughput based methods for predicting miRNA target mRNA

With the development of the technology, the high-throughput data, such as CILP (Cross-linking immunoprecipitation) data and PAR-CLIP (photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation) data, are used to select the critical features that have significance effect on predicting miRNA target. These critical features are very useful for enhancing the accuracy rate of miRNA target prediction. Therefore, many methods which based on the high-throughput data are developed to predict miRNA target mRNA.

The latest version of Tragetscan algorithm applies the CLIP data to select

the feature that have significance in predicting miRNA target (Agarwal, Bell, Nam & Bartel 2015). This method selects 26 candidate features that include the feature of miRNA, feature of the site, and features of the mRNA. Not all these features have significance influence on predicting miRNA target. After analysing the CLIP data, 14 features are identified as the most critical features in predicting miRNA target. Finally, the context++ model is applied to calculate the score of miRNA binds to mRNA. The higher score, the miRNA is more likely to bind to the mRNA.

miRDB is a novel method that applies the CLIP data to find out the most important features for predicting miRNA target (Liu & Wang 2019). These important features include the base-pairing of the miRNA seed region, GC content, AU content and so on. Then the support vector machine (SVM) is applied to train the data. The output of the model is MirTarget score that is used to measure the significance of the miRNA binds to mRNA. The higher the MirTarget score the miRNA is more likely to bind to the mRNA. According to the results, the performance of this method is better than other methods.

2.1.3 Methods for predicting miRNA target lncRNA

Many methods have been developed to predict miRNA target mRNA. However, a few methods are developed for predicting miRNA target lncRNA. This is because people study the miRNA target mRNA prediction for a long time, while the study of predicting miRNA target lncRNA is still at the early stage. Further, many methods for predicting the miRNA target mRNA also could be applied for predicting the miRNA target lncRNA.

The DIANA-LncBase v1 database is the first extensive database that predicted the miRNA target lncRNA (Paraskevopoulou, Georgakilas, Kostoulas, Reczko, Maragkakis, Dalamagas & Hatzigeorgiou 2012). The database predicted the miRNA target lncRNA by using the DIANA-microT-CDS algorithm (Reczko, Maragkakis, Alexiou, Grosse & Hatzigeorgiou 2012). This algorithm is first used to predicted the miRNA target mRNA. However,

it also could be applied to predict miRNA target lncRNA. This algorithm combines protein coding sequence features, such as protein coding sequences conservation and Flanking AU content, and 3′-UTR features, for instance 3′-UTR conservation and accessibility of binding site, to predict the miRNA target. All these features are extraction by analysis the PAR-CLIP data.

The miRcode is a software that predicts the miRNA target lncRNA (Jeggari, Marks & Larsson 2012). This software predicts the miRNA target lncRNA based on seed complementarity and evolutionary conservation. Using the sequencing alignment method to find out the sequence region that the lncRNA sequence is complementary to seed region of miRNA. Then using the multiple alignment to discover the conservation of sequence region. The high conservation of the sequence region is the binding site.

These target prediction methods not only provide a strategy for annotation the function of 5′ isomiR, but also offer the foundation of constructing novel ceRNA network. However, all the methods for predicting miRNA target have high false positive rate, even though many methods contain large amount of features or some technologies to minimize the biased. Further, there is not a good strategy for discovering novel features. The novel feature used in all the methods is based on the researchers experience. Although the prediction model uses the novel features that have significance in predicting miRNA target, the selected novel features have limit to improve the performance of the method.

2.2 Algorithms for constructing RNA regulatory network

The ceRNA network reflects the regulatory mechanism of the biological process. Methods for constructing ceRNA could be divided into two categories: target-based method and expression-based method. In this

section, we will describe these two categories methods and discuss the advantages and weaknesses of these two methods.

2.2.1 Target-based method for constructing ceRNA network

In the ceRNA network, the lncRNA and mRNA must be the miRNA target. Therefore, many methods for constructing the ceRNA network through finding the miRNA target mRNAs and target lncRNAs. These methods are called target-based method.

The starBase is a ceRNA database (Li, Liu, Zhou, Qu & Yang 2013) that uses the target-based method to construct the ceRNA network. A total 108 CLIP data experiment data are applied to find the miRNA target sites. The miRNA, its target genes, and target lncRNA construct a candidate ceRNA pair. A hypergeometric test is applied to calculate the P-value of each ceRNA pair. The equation of calculating the P-value is below:

$$P - value = \sum_{i=a}^{\min(l,n)} \frac{\binom{l}{i}\binom{N-l}{n-i}}{\binom{N}{n}} \tag{2.1}$$

Where a is the number of miRNA share by mRNA and lncRNA, l is the number of miRNAs binds to the lncRNA, N is the number of miRNA in the dataset, n is the number of miRNA binds to mRNA. This p-value measure the probability of a ceRNA pair to cross-regulate each other. The higher the value, the more likely the ceRNA pair cross-regulate each other.

Liu et al. used a new way to construct the ceRNA network (Liu, Yan, Li & Sun 2013). They collected the miRNA target mRNA from TargetScan, miRanda, and PITA database. The miRNA target lncRNA were downloaded from miRanda. A hypergeometric test was also used to compute the p-value of the ceRNA pairs. The ceRNA pairs, which the p-value was higher than 0.95, were applied to construct the ceRNA network.

Das et al. proposed a target-based method for constructing ceRNA network (Das, Ghosal, Sen & Chakrabarti 2014). They downloaded the

miRNA target mRNA from the starBase database. The miRNA candidate target lncRNA were predicted by using miRCode (Jeggari et al. 2012). Then collecting the PAR-CLIP dataset for improving the prediction of miRNA target lncRNA. Using the miRNA, its target miRNAs, and its target lncRNAs to construct a ceRNA pair. Finally, the hypergeometric test was applied to calculate the p-value of each ceRNA pair.

All these target-based methods for constructing the ceRNA network have two commons: (1) using the miRNA target method to find out the miRNA target lncRNA and target mRNA. (2) Applying the hypergeometric test to measure the probability of the ceRNA pair to regulate each other. However, the difference between different methods is that using different methods for predicting miRNA targets. In starBase database, the miRNA target mRNA is predicted by using the CLIP dataset, while Das uses the CLIP data to predict the miRNA target lncRNA.

The miRNA target is very important for constructing the ceRNA network. The target-based method could identify the miRNA target and construct large and complex ceRNA networks to analysis the mechanism of RNA. However, the target-based method has many weaknesses in predicting ceRNA network. The first is that the ceRNA network constructed by using this method has high false positive rate. This is because the method for predicting miRNA target provides high false positive miRNA targets, even though CLIP dataset is applied to enhance the performance of prediction. The second is that although large amount of ceRNA networks are identified, these predicted ceRNA networks cannot be used to analysis regulation mechanism in a certain cancer. This is because the ceRNA network always occurs in a certain condition. For example, a ceRNA occurred in the lung cancer, but this ceRNA may not occurred in breast cancer. Thus, many predicted ceRNA networks must be removed when analysing the regulation mechanism of RNA in a certain cancer. The third is that this method does not take the relative concentration of the lncRNA into consideration. The relative concentration of lncRNA is very important for lncRNA act as ceRNA and is identified by

using the expression level of lncRNA. The forth is that this method does not use the expression data to discover the competition relationship between RNAs. Since the target-based method does not use the expression data, the expression relationship between RNAs is still unknown. The target-based method finds out the miRNA target mRNA and target lncRNA. However, it is does not implied that the lncRNA could compete with the mRNA to bind to the same miRNA.

2.2.2 Expression-based method for constructing ceRNA network

Since the limitation of the target-based method, the other method, which based on the expression data, had been developed to construct ceRNA network. The expression data provides the expression level of RNA in samples. The relative concentration of lncRNA and the competition relationship between lncRNA, miRNA, and mRNA can be detected by using the expression data. The critical processes in constructing the ceRNA network by using expression-based method are that finding the change relative concentration of lncRNA and measuring the competition relationship between lncRNA, miRNA, and mRNA.

Xia et al. proposed a method constructing the ceRNA network in gastric cancer method (Xia, Liao, Jiang, Shao, Xiao, Xi & Guo 2014). This method used the expression data of lncRNA in tumor and normal tissue to identify the overexpressed lncRNAs. The overexpressed lncRNA implied that the change relative concentration lncRNA was large and may become ceRNA to compete with mRNA to bind to the same miRNA. Then using the miRNA predict targets to construct the ceRNA network. The miRNA target lncRNA were collected from miRcode database and the miRNA target mRNA was derived from DIANA-TarBase database. This method considered the relative concentration lncRNA but not the competitive relationship between lncRNA, miRNA, and mRNA. Detecting the competitive relationship between lncRNA, miRNA, and mRNA requires

the expression data of lncRNA, miRNA, and mRNA.

There are many different method for measuring the competition relationship between lncRNA, miRNA, and mRNA. They mainly applied Pearson coefficient or information theory. Chiu et al. presented a Pearson coefficient-based method to construct the ceRNA network (Chiu, Hsiao, Chen & Chuang 2015). Chius method used the miRNA predicted targets, which were derived from TargetScan database, to construct a miRNA-target matrix. Then the Pearson coefficient was applied to calculate the pair-wised relationship of the miRNA-target matrix. Given the expression level of two genes $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ in n samples, the equation for calculating Pearson coefficient is below:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.2}$$

Where $\bar{x}$ and $\bar{y}$ are the average expression level of the gene x and y in n samples, respectively. However, the Pearson coefficient measures the relationship between two variables. The competition relationship between lncRNA, miRNA, and mRNA is triple relationship. Thus, the Pearson coefficient method should be improved to measure the competition relationship.

Wang et al. improved the Pearson coefficient to construct the ceRNA network (Wang, Ning, Zhang, Li, Ye, Zhao, Zhi, Wang, Guo & Li 2015). This method used the disease relatived lncRNA, miRNA, and lncRNA. Then applying the CLIP-data to find out the disease related miRNA targets. These miRNAs and miRNA targets were used to construct ceRNA pairs. Finally, utilizing the Pearson coefficient to measure the competition relationship between lncRNA, miRNA, and mRNA. If a ceRNA pair follow these three conditions: (1) the Pearson coefficient between lncRNA and miRNA was smaller than -0.5, (2) the Pearson coefficient between lncRNA and mRNA was larger than 0.5, and (3) the Pearson coefficient between mRNA and miRNA was smaller than -0.5. This ceRNA pair was applied to construct ceRNA network. Thus, the selected ceRNA pair by using Wangs method

must obey the competition relationship between lncRNA, miRNA, and mRNA.

The partial correlation is a Pearson coefficient-based method and can measure the triple relationship. Paci et al. constructed the ceRNA network by using the partial correlation (Paci, Colombo & Farina 2014). They downloaded the predicted miRNA target mRNA from TargetScan database. The lncRNAs which could perfect math to the seed region of miRNA were viewed as the miRNA target lncRNAs. A miRNA a miRNA target lncRNA, and a miRNA target mRNA construct a ceRNA pair. For given ceRNA pair, they used the partial correlation to calculate the competition relationship of the ceRNA pair. The equation of the partial correlation defined as:

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{(1-\rho_{XY})^2}\sqrt{(1-\rho_{ZY})^2}} \tag{2.3}$$

Where X is the mRNA, Y is lncRNA and Z is miRNA. ρ_{XY} is the Pearson correlation between X and Y. The partial correlation measures the correlation between X and Y after remove Z. Finally, using the partial correlation to calculate sensitivity correlation S:

$$S = \rho_{XY} - \rho_{XY|Z} \tag{2.4}$$

The ceRNA pairs, which the sensitivity correlation larger than 0.3, were used to construct ceRNA network.

Some methods apply information theory to measure the competition relationship between lncRNA, miRNA, and mRNA. Sumazin utilize the conditional mutual information to construct the ceRNA network (Sumazin, Yang, Chiu, Chung, Iyer, Llobet-Navas, Rajbhandari, Bansal, Guarnieri, Silva et al. 2011). The conditional mutual information could measure the relationship between three variables. It always applies to the discrete dataset. However, the expression data set of RNA is continuous dataset. In order to handle with this issue, the conditional mutual information estimators are used.

Zhang et al. applied the maximal information to construct the ceRNA network (Zhang, Fan, Jian, Chen & Lai 2015). They use two different

expression data to identify the overexpressed lncRNA. One expression data is from TCGA dataset and the other expression data is obtained from NCBI. The lncRNA, which is highly differentially expressed in one of the data, is used to construct the ceRNA network. Then, they collect the predicted miRNA target mRNA and lncRNA from miRTarBase and starBase. The miRNA targets construct the candidate ceRNA pairs. Finally, the maximal information-based nonparametric exploration statistics is applied to identify the relationship between two RNAs. The ceRNA pair, which has high maximal information coefficient, is applied to construct the final ceRNA network. This method not only considers the overexpressed lncRNA, but also the relationship between RNAs.

A few methods neither using Pearson coeffecient nor apllying information theory construct ceRNA network. Chuang et al. developed a novel method for identifying the ceRNAs in gliomblastoma by using the expression data of lncRNA, miRNA, and mRNA (Chiu, Chuang, Hsiao & Chen 2013). The miRNA targets were downloaded from TargetScan database. These miRNA targets were used to construct the candidate ceRNA network. Then using the miRNA program (miRP) enrichment was applied to measure the average expression level of miRNA in each candidate ceRNA network. It also an indicator for detecting competition relationship of the candidate ceRNA network. Finally, using the miRP enrichment to calculate the correlation and difference score of each candidate ceRNA network. The candidate ceRNA network, which has significance difference score and positive correlation, was viewed as the ceRNA network that involved in gliomblastoma.

Cupid et al. presented a simultaneous reconstruction method for constructing ceRNA network (Chiu, Llobet-Navas, Yang, Chung, Ambesi-Impiombato, Iyer, Kim, Seviour, Luo, Sehgal et al. 2015). This method focus on selecting miRNA targets. They collect three features of the miRNA binding site: (1) the predicted score of the miRNA binding site from TargetScam and miRanda, (2) the species-conservation scores, and (3) the relative distance from the $3'$ and $5'$ ends of the target $3'$-UTR. All these

features are applied to train a SVM classifier. The classifier is able to identify the consensus binding site. In the second step, the consensus binding site are used to assess the probability of the miRNA binds to its target. Finally, the expression data are used to measure the relationship between miRNA and its targets.

Compare with the target-based method, the expression-based method uses the expression level of RNA to measure the competition relationship between RNAs. However, these expression-based methods have their weaknesses. Xias method considers the overexpressed lncRNA but does not detect the competition relationship between RNAs. Many methods, such as Pacis method and Chuangs method, detect the competition relationship between RNAs while do not find out the overexpressed lncRNA. The change relative concentration of lncRNA is a very important for lncRNA competes with mRNA to bind to miRNA. Many methods, for example Chius method, use the Pearson coefficient to measure the competition relationship between lncRNA, miRNA, and mRNA. However, a miRNA can bind to multiple lncRNAs and mRNAs. The competition relationship between lncRNA, miRNA, and mRNA is non-linear. The Pearson coefficient is a suitable method for measuring the linear relationship rather than the non-linear relationship. Some methods, for instance Chius method, apply the paired-wise relationship between two RNAs to measure the competition relationship. The paired-wise relationship implies that these two RNAs are correlatived. The competition relationship is the relationship between three RNAs but not two RNAs. Thus, the paired-wise relationship between two RNAs is not suitable for measuring the competition relationship between three RNAs.

2.3 Methods for discovering biomarker in cancers

Biomarkers indicate the processes of the biological processes. The change expression level of the biomarker leads to the alternative of the

biological processes. Therefore, the expression level of biomarker should have significance difference between different biological processes. The biomarker could be applied to identify the tumor and normal samples or classify different cancer subtypes. There are many methods are developed for discovering the biomarker. The Fold change method and t-test method are two most popular methods for identifying tumor and normal samples.

The fold change method measures the expression change of RNA between tumor sample and normal sample (Grishin 2001). Give the expression level of RNA a in tumor samples and normal. The $\overline{Ex_a^T}$ and $\overline{Ex_a^N}$ are defined as the average expression level of a in tumor and normal sample. The fold change of the RNA a is calculated by this equation:

$$FC_a = \log_2 \frac{\overline{Ex_a^T}}{\overline{Ex_a^N}} \tag{2.5}$$

According to this equation, if the fold change of the RNA is larger than 0, it implies that the average expression level of the RNA in tumor sample is higher than the average expression level of the RNA in normal sample and we call this RNA is up-regulated in tumor sample. If the fold change of the RNA is lower than 0, it indicates that the expression level of the RNA in tumor sample is lower than the average expression level of the RNA in normal sample and we call this RNA is down-regulated in tumor sample. In general, if the absolute fold change of the RNA is larger than 1, this RNA is highly differentially expressed in tumor and normal sample. The highly differentially expressed RNA infer that this RNA may be a biomarker for identifying tumor and normal sample.

The fold change method has two weaknesses. The first is that this method is sensitive with the outliers. This method applies the average expression levels of RNA in tumor and normal samples. The sample, which the expression level of RNA is very large or very low, has significance influence the average expression level. However, this sample may be an outlier and should not be taken into consideration. Thus, the fold change of the RNA is influenced by these outliers. The second is that this method will miss

the RNA that is large differences but small ratios and induce some noisy RNA which is small differences but large ratios. For example, the average expression level of a RNA is 1000 in tumor sample and 600 in normal sample. This RNA has large difference between normal and tumor sample. However, the absolute fold change of this RNA is lower than 1. Besides, the average expression level of a RNA is 0.1 in tumor sample and 0.25 in normal sample. This RNA has very small difference between normal and tumor sample. However, the absolute fold change of this RNA is larger than 1. The small difference of the expression level of RNA may cause by the sequencing machine. Thus, this RNA may be a noisy RNA and cannot be a biomarker for classifying normal and tumor sample.

The t-test method determines whether there is a significance different between the average expression level of RNA in tumor and normal sample (Baldi & Long 2001). If the average expression levels of a RNA a in normal and tumor are $\bar{Ex}_a^N$ and $\bar{Ex}_a^T$, respectively. The number of the sample in normal sample and tumor are n_n and n_T. The variance of the expression level this RNA in normal and tumor samples are var_N and var_T. We always use this equation to calculate the T-test value:

$$T - value_a = \frac{\overline{Ex_a^N} - \overline{Ex_a^T}}{\sqrt{\frac{var_T^2}{n_T} + \frac{var_N^2}{n_N}}} \tag{2.6}$$

Then using the known distribution to calculate the P-value. In general, if the P-value is lower than 0.05, it implies that the RNA a is differentially expressed in normal and tumor samples.

In practice, researchers combine t-test and fold change to identify the biomarker of the RNA. The RNA which the fold change is larger than a threshold and the p-value is lower than 0.05, could be regarded as the tumor biomarker.

A RNA, which has high fold change and low p-value, is likely to be biomarker for identifying the tumor sample. This RNA has high probability play an important role for the development of cancer and could be used for diagnosing cancer. Therefore, researchers like to combine these two methods

to find out the biomarker for identifying tumor and normal sample. These two methods are used in data set that has two classes (normal sample and tumor sample). A cancer may have multiple subtypes, such as the breast cancer could be divided into four molecular subtypes. Therefore, t-test and fold change method cannot be used for discovering cancer subtype biomarker. Although fold change method and t-test method cannot be used for discovering multiple classes, many methods are developed for discovering the biomarker of multiple classes.

Fishers method is able to identify the biomarker for classifying different cancer subtypes (Gu, Li & Han 2011). Given a RNA X_i and m cancer subtypes. X_i^k is defined as a set of expression level of the RNA X_i in cancer subtype k. We state that $\bar{X}_i^k$ is the average expression level of RNA X_i in cancer subtype k and $\bar{X}_i$ is the average expression level of RNA X_i in all samples. The Fisher score of the RNA X_i is calculated by the equation:

$$Fisher(X_i) = \frac{\sum_{j=1}^{m} l_i(\bar{X}_i^j - \bar{X}_i)^2}{\sum_{k=1}^{m}\sum_{x\in X_i^k}(x - \bar{X}_i^k)^2} \tag{2.7}$$

The higher the Fisher score of the RNA, the RNA is more likely to be biomarker for classifying different cancer subtypes. The Fisher's method is based on the average expression level of the RNA in cancer subtype.

Hellinger distance is a method that measures the distributional divergence of two probability measures (Yin, Ge, Xiao, Wang & Quan 2013). The X and Y are defined as two probability measures and respect to a third probability measure θ. The Hellinger distance of these two probability measures is:

$$d_H(X,Y) = \sqrt{\int(\sqrt{X} - \sqrt{Y})^2 d\theta} \tag{2.8}$$

Specially, if these two probability measures follow the normal distributions $X \sim N(\mu_1,\sigma_1^2)$ and $Y \sim N(\mu_2,\ \sigma_2^2)$, the square Hellinger distance between these two probability measures is that

$$d_H^2(X,Y) = 1 - \sqrt{\frac{2\mu_1\mu_2}{\sigma_1^2+\sigma_2^2}} e^{-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}} \tag{2.9}$$

The Hellinger distance measures the distance between two variables. However, we can calculate the paired Hellinger distance between two difference cancer subtypes. Then using the total Hellinger distance to measure the distance between multiple variables. The RNA, which has the largest total Hellinger distance, is more likely to be a biomarker for cancer subtype classification.

Two information-based methods are used for finding the biomarker in classifying multiple classes. The first is information gain and the other is the mutual information. The information gain measures the information gained about a feature from observing another random variable (Quinlan 1986). In the decision tree algorithm, the information gain is applied to decide which feature is used to build the tree. The feature has higher information gain, this feature provides more information for classification. Thus, we use the information gain to measure the amount of information of a RNA in classifying different cancer subtypes. Calculating the information gain of a RNA (feature) in classifying different cancer subtypes requires the information entropy of RNA. Given a feature a and the dataset T. This dataset contains c classes. $Values(a)$ presents the attribute of feature a. T_i is a set of samples that belong to the classes i. $|S_a^T(j)|$ is a subset of dataset T which the attribute of feature a is equal to j. $|\cdot|$ is the total number of the sample. The equation of calculating the information gain $IG(T,\ a)$ of the feature a in dataset T is showed below:

$$IG(T, a) = H(T) - H(T|a) \tag{2.10}$$

$$H(T) = \sum_{i=1}^{c} -\frac{|T_i|}{|T|} \log_2 \frac{|T_i|}{|T|} \tag{2.11}$$

$$H(T|a) = \sum_{j \in Values(a)} \frac{|S_a^T(j)|}{|T|} H(S_a^T(j)) \tag{2.12}$$

The information gain measures how important the feature to classified the data. The higher the information gain, the more important the feature is.

The information gain has the maximum value. If a feature has the maximum information gain, this feature could perfect classify different classes. The information gain could measure the importance of the RNA for identifying cancer subtype. The higher the information gain of the RNA, this RNA is more likely to be a cancer subtype biomarker.

According to the information gain, the feature should have many different attributions. It implies that the information gain is used to the data set that the feature and label are discrete data. However, the expression data is the dataset that the feature is continuous data (expression level of RNA is continuous) and the label is discrete data. Therefore, the information gain cannot direct be applied. In order to calculate the information gain of the RNA in classifying different cancer subtypes, the expression level of RNA must be transformed to discrete data and we called this transformation is discretization. This transformation is that cluster the expression level of RNA into several groups. For example, clustering the expression level of RNA into two groups. The low expressed RNA are grouped into the first cluster and the high expressed RNA are grouped into the second cluster. Thus, this RNA has two attributes: highly expressed and lowly expressed. After the discretization, the information gain could be applied to measure the importance of the RNA in classifying different cancer subtypes.

There are three popular methods to discrete the continuous data: equal width, equal frequencies, and highest entropy. The equal width method separates the data into k equal size intervals. This method finds out the lowest and highest expression level of the RNA, which are defined as $minexp$ and $maxexp$, respectively. Then calculating the width of the interval is $w = (maxexp - minexp)/k$. Finally, the interval boundaries are: $minexp + w$, $minexp + 2w$, ..., $minexp + (k-1)w$. The expression levels of RNA are divided into k groups based on these interval boundaries.

The equal frequency method sorts the expression level of the RNA from low to high. Then dividing the RNA into k group that each group has approximately same number of the sample.

The lowest entropy method is that sorted the expression level of RNA from small to large and then finding out the entire cluster strategies. Then calculating the entropy of the entire cluster strategies. The cluster strategy that has the smallest entropy is viewed as the best cluster strategy. This best cluster strategy is applied to discrete the expression data.

Different discretization methods have their advantages and weaknesses. The equal width and equal frequencies method are easier than the lowest entropy method. However, the range of the expression level of biomarker in some cancer subtypes are very wide. While the range of the expression level in some cancer subtypes are very narrow. The range of the expression level of biomarker in different cancer subtypes do not have the equal width. Therefore, the equal width technology is not suitable for the expression data discretization. Further, for given a cancer subtype data, the number of the sample in different cancer subtypes is always different. Thus, the equal frequency method is also not suitable for discretising the expression data. The lowest entropy method could find out the best strategy to cluster the expression data. This is because the lowest entropy method discovers all the cluster strategies and then selecting the best strategy. However, the time consumption of this method increases dramatically with the growth of the number of sample and the number of cancer subtype. If the number of the sample is n and we want to divide the expression level of RNA into k groups. The total number of the cluster strategy is $\binom{n}{k}$. The number of the cluster strategy is very large and calculating the entropy of the entire cluster strategy requires large amount of time. In order to tackle with these weaknesses, we developed an improved method to discrete the continuous data. This method discretes the continuous data based on the distribution of the feature and the time complexity is slight increased with growth of the number of sample and the number of cancer subtype.

The other information-based method for identifying biomarker is the mutual information (Steuer, Kurths, Daub, Weise & Selbig 2002). The mutual information is a measure of how much information that one variable

has about another variable (Cover & Thomas 2012). This definition gives a way to quantify the relevance of a feature subset to the output (Vergara & Estévez 2014). Therefore, mutual information has been used as a criterion for feature selection in engineering especially in machine learning (Navot 2006). The mutual information can measure the relationship between the biomarker and the cancer subtype. The biomarker is the indicator of the cancer subtype and should have high correlative with the cancer subtype and therefore, have high mutual information. Given two datasets X and Y. The equation of the mutual information is that

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.13}$$

Where $P(x)$ is the probability of variable x, and $P(x,\ y)$ is the join distribution of the variable x and y. The mutual information is also used in the data set that the feature and label are discrete data. However, the estimate mutual information could apply to calculate the mutual information of data that the feature and label are continuous data. The estimate mutual information is that:

$$MI(X,Y) = \int_x \int_y f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy \tag{2.14}$$

Where $f(x)$ and $f(y)$ is the distribution function of dataset X and Y. $f(x,\ y)$ is their joint distribution if X and Y. The kernel density estimator is always used to calculate the distribution function and their join distribution function. The kernel density estimator of the distribution is that:

$$f(x) = \frac{1}{Nh\sqrt{2\pi}} \sum_{i=1}^{N} exp(-\frac{(x-x_i)^2}{2h^2}) \tag{2.15}$$

Where N is the number of the sample and h is the smoothing parameter. The join distribution function is that

$$f(x,y) = \frac{1}{Nh^2 2\pi} \sum_{i=1}^{N} exp(-\frac{d_i(x-x_i)^2}{2h^2}) \tag{2.16}$$

Where $d_i(x,\ y) = \sqrt{(x - x_i)^2 + (y - y_i)^2}$. Using the kernel density estimator, we could calculate the mutual information between two variables that are continuous data.

2.4 Applied mathematical methods and bioinformatics tool

2.4.1 Methods for selecting threshold

Selecting the threshold is a very important for constructing the ceRNA network and discovering the biomarker. For example, many methods are developed to measure the competition relationship between RNAs, the ceRNA pairs that have a high competition score are used to construct ceRNA network. Thus, a threshold is required to identify the ceRNA pairs that have a high competition score. In addition, the fold change method calculates the change expression of the RNA. The highly differentially expressed RNA is viewed as the biomarker. Identifying the biomarker needs a threshold to filter out the RNA that is small changed. Selecting a 'hard' threshold is very common in research.

The 'hard' threshold is that this threshold could be applied in every research or dataset. For example, in practices, if the P-value of the t-test between two samples is lower than 0.05, these two samples are significance different. This threshold 0.05 is a 'hard' threshold. This is because this threshold could be applied in every research or dataset. In addition, if the absolute value of Pearson coefficient between two variables is larger than 0.7, these two variables are highly correlative. Thus, the 0.7 is also a 'hard' threshold. Setting the 'hard' threshold required the experience and this 'hard' threshold could be applied in most of the threshold selection problem.

However, it may loss of information and sensitivity to the choice of the 'hard' threshold (Carter, Brechbühler, Griffin & Bond 2004). In order to handle with this issue, the 'hard' threshold may be changed based on the

data. For example, the fold change is very sensitivity to the expression level of RNA and it may neglect the high difference but low ratio RNA. Thus, we can select the threshold based on the expression level of RNA. The expression level of miRNA is very high in tissue. Therefore, the threshold for selecting the tumour relatived miRNA should be low. However, the lncRNA is lowly expressed in tissue. The threshold of selecting the tumour relatived lncRNA should be high.

The other strategy to select a threshold is that calculating the threshold through the dataset. Given a set of real values, these values are follow a certain distribution, such as normal distribution. If we want to set a threshold to find out the high real values in this dataset, this threshold could be calculated based on the distribution. We can calculate 95% confidence interval of these real values. The threshold of the 95% confidence interval is a good baseline to identify the very high or low value. The real value, which is higher than large threshold of the 95% confidence interval, is regarded as very high value.

2.4.2 Kyoto Encyclopedia of Genes and Genomes pathway

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database for analysing the gene function, especially for analysing the function of biomarker in cancer subtypes (Kanehisa & Goto 2000). A gene could regulate many KEGG pathway and a KEGG pathway is regulated by multiple genes. For example, $Lin28B$ gene regulates Hedgehog and Notch signalling pathway and the Wnt signalling pathway is regulated by 151 genes, such as $CCND1$ gene and $TP53$ gene.

Given a gene set g, these genes are enriched in some KEGG pathways. A hypergeometric test-based method is applied to measure the enrichment of

the gene set in a KEGG pathway. The pathway score is calculated by

$$S(g,p) = -\log_{10}(\sum_{m}^{n} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}) \tag{2.17}$$

Where N, M, n, and m are the total number of gene in human, the number of gene in the KEGG pathway p, the number of gene in the gene set g, and the number of gene in both gene set g and KEGG pathway. The lower the p-value, this gene set is more likely to regulate this KEGG pathway. In general, if the pathway score of is lower than 0.05, it implies that these genes are enriched in the KEGG pathway.

There are two websites contain the latest version of the KEGG pathway and could be used to analyse the gene enrichment in KEGG pathway. The first is Davide GO (`https://david.ncifcrf.gov/`) (Huang, Sherman & Lempicki 2008) and the second is enrichr website (`http://amp.pharm.mssm.edu/Enrichr/`) (Kuleshov, Jones, Rouillard, Fernandez, Duan, Wang, Koplev, Jenkins, Jagodnik, Lachmann et al. 2016). Both of these websites provide a user friendly interface to analyse the gene enrichment. Input the gene name into the website, we can easy to access the P-value of the gene enrich in the KEGG pathway.

2.4.3 Support vector machine

In section 2.3, we reviewed many methods measure the importance of the RNA in classifying different cancer subtypes. However, only a few RNAs could classify different cancer subtypes, we should find out these RNAs. A strategy for identifying these RNAs is that using a supervised learning model to calculate the accuracy of the RNA in classifying difference cancer subtypes. If this model has high performance, these RNAs are the critical biomarker for classifying cancer subtypes.

Support vector machine (SVM) is a famous supervised learning model (Cortes & Vapnik 1995). Given a data set of n smaples, it is defined as $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, x is the feature and y is the label.

The feature x is a p-dimensional vector and the label y is either 1 or -1. We want to find the maximum-margin hyperplane that divide these sample into two groups and each group contains the sample that have the same label. The hyperplane could be written as $w \cdot x - b = 0$. The w is the a vector to the hyperplane. Further, the distance between the hyperplane and the nearest sample from each group is maximized. This optimization problem can be written as follow:

$$min \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \parallel w \parallel^2$$

$$\text{suject to} \quad y_i(w \cdot x_i - b) \geq 1 - \xi_i \quad and \quad \xi_i \geq 0, \quad \text{for all } i$$

The hyperplane is completely determined by the samples that lie nearest to the hyperplane. Tshese samples are called support vectors. This is the SVM is used to classify two classes. In order to handle with multiple classes problem, the one-versus-all method, which is showed in Figure 2.1, is applied. This method extend the ability of SVM to classify multiple classes.

After training the model, most of data is redundant. This is because only the support vector determines hyperplane and the other data could not influence the hyperplane. Therefore, it performance better in small dataset. Further, SVM is unlikely overfitting compare with other machine learning methods.

2.5 Summary

This Chapter introduces the methods for predicting miRNA target mRNA and target lncRNA, constructing the ceRNA network, and identifying the biomarker. Further, some methods and tools in bioinformatics research are also described. In conclusion, the methods for predicting miRNA target mRNA and target lncRNA are the foundation of constructing ceRNA network. Constructing the ceRNA network should take the change of the lncRNA and the competition relationship between RNAs into consideration.

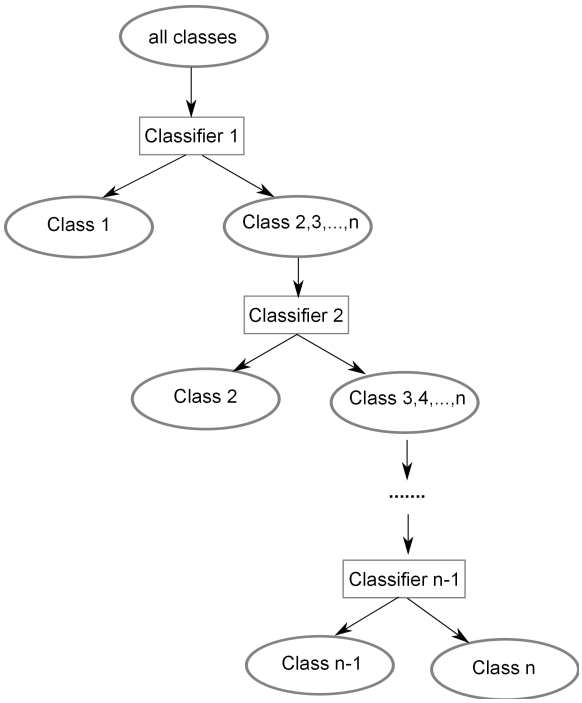

Figure 2.1: **The one-versus-all method for classified multiple classes in SVM.** If the dataset contains n classes. This method is that train a classifier which distinguish one class and the rest classes. Repeat this process until all the classes are distinguished.

Many methods could measure the importance of the RNA in cancer subtype classification. However, identify the most critical RNA for classifying different cancer subtypes requires machine learning to validate. Further, understanding the function of the biomarker should use the KEGG pathway.

Chapter 3

Construction of Competing Endogenous RNA Networks from Paired RNA-seq Data Sets by Pointwise Mutual Information

3.1 Introduction

In subsection 1.1.2, we describes the regulatory network in lncRNA, miRNA, and mRNA. When a lncRNA acts as a ceRNA to compete with an mRNA for binding to the same miRNA, this interplay between the lncRNA, miRNA, and mRNA is called a ceRNA crosstalk. An miRNA may have multiple target lncRNAs and it can also regulate several different mRNAs, therefore, there can exist many crosstalks mediated by this miRNA to form a ceRNA network. Such a network is useful for detecting cancer biomarkers (Li, Chen, Chen, Mo, Li, Shao, Xiao & Guo 2015), patterns for early diagnosis (Sanchez-Mejias & Tay 2015), and new concepts for cancer treatment (Ebert et al. 2007). Further, we also discussed three features in ceRNA network: (1) changes

in the ceRNA expression levels, (2)the lncRNA is the primary target of the miRNA, and (3)the relationships between the lncRNA, miRNA, and mRNA should obey a competition rule in the ceRNA network.

In section 2.2, we described two types of methods for constructing ceRNA network. The first is target-based method and the other is expression-based method. The target-based methods does not consider the relationship between RNAs. Some expression-based methods use the pairwised relationship between tow RNAs to construct ceRNA networks, the ceRNA network is not the relationship between lncRNAs, miRNAs, and mRNAs. Although other expression-based methods measure the relationship between lncRNAs, miRNAs, and mRNAs, they are not suitable for measuring the non-linear relationship. Therefore, a novel method is demanded to improve the predictions.

We propose a novel method for constructing ceRNA networks from paired RNA-seq data sets. This method identifies the over expressed lncRNAs from the lncRNA expression data of the normal and tumor samples. Thus, we can identify the ceRNA network related to breast cancer. Then, the competitive relationships between the lncRNAs, miRNAs, and mRNAs are established by using the expression levels of the lncRNAs, miRNAs, and mRNAs in the tumor samples. We combine the competition rule and pointwise mutual information to calculate a competition score for each of the ceRNA crosstalks. As an miRNA can have many ceRNAs and can bind to multiple mRNAs, the competitive relationship between lncRNA, miRNA, and mRNA is non-linear. Pointwise mutual information is suitable for measuring the complex point-to-point competitive relationship between RNAs.

3.2 Method

Our method for constructing ceRNA network has four steps. Firstly, it computes the expression levels of lncRNA, miRNA, and mRNA from breast cancer tumor tissues and normal tissues. Secondly, the predicted miRNA targets, differentially expressed RNAs, and the competition regulation

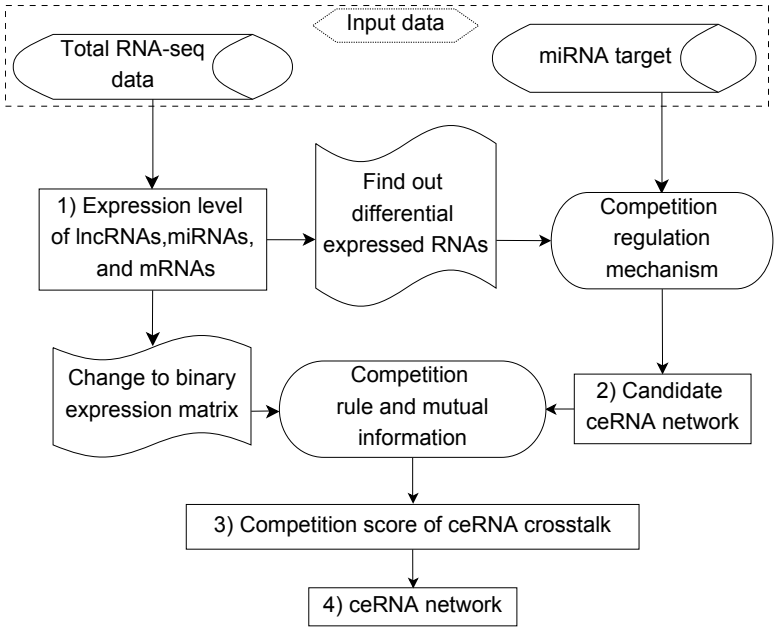

Figure 3.1: **The framework of our method**

mechanism are used to construct the candidate ceRNA networks. Thirdly, it combines the competition rule and the pointwise mutual information to compute the competition score of each ceRNA crosstalk. Finally, we select the ceRNA crosstalks which have significant competition scores to construct the ceRNA network. Figure. 3.1 shows the framework of our method.

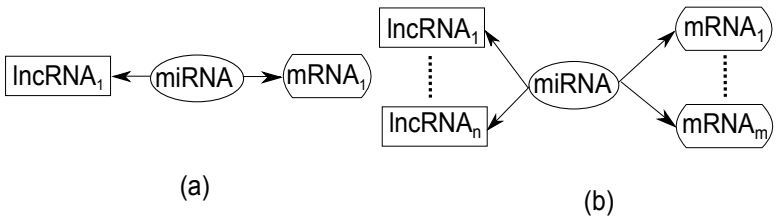

Figure 3.2: **The examples of ceRNA crosstalk and ceRNA network.** (a) A ceRNA crosstalk; (b) A ceRNA network

3.2.1 Definitions and Data Preprocessing

If a lncRNA *lnc* competes with an mRNA *mr* for binding to an miRNA *mir*, the triple of *lnc*, *mir*, and *mr* is called a ceRNA crosstalk denoted by $T = (lnc, mir, mr)$. We also say that ceRNA crosstalk $T = (lnc, mir, mr)$ is mediated by *mir*. For example, Figure. 3.2(a) is a ceRNA crosstalk $T = (lncRNA_1, miRNA, mRNA_1)$ mediated by $miRNA$.

All the ceRNA crosstalks mediated by the same miRNA as a whole is defined as a ceRNA network. In this thesis, one ceRNA network contains only one miRNA and the analysis is done with a set of individual networks. It is denoted by $N = (lnR, mir, mR)$, where lnR stands for the set of lncRNAs, *mir* is the miRNA, and the mR stands for the set of mRNAs. We also say ceRNA network $N = (lnR, mir, mR)$ is mediated by *mir*. For example, Figure. 3.2(b) is a ceRNA network, where $lnR = \{lncRNA_1, lncRNA_2, \ldots, lncRNA_n\}$ and $mR = \{mRNA_1, mRNA_2, \ldots, mRNA_m\}$.

The paired breast cancer RNA-seq data set was downloaded from the TCGA GDC data portal website (`https://portal.gdc.cancer.gov/cart`). This paired data set contains the expression levels of lncRNAs, mRNAs, and miRNAs of 102 tumor and normal tissue samples. These RNAs and their expression levels form an expression matrix. Table 3.1 is an example of expression matrix. Some RNAs expresses in only a few tissue samples. These low frequently expressed RNAs are not important for breast cancer study and may have noise affect to the result. Thus, these RNAs which are not expressed in half of the whole tissue samples were removed from the expression matrix. We transform the expression matrix to a binary expression matrix by using the equal frequency discretization method: for the same RNA expressed in all samples, if this RNA expression level of a sample is higher (lower) than the median RNA expression level of all the samples, this RNA is highly (lowly) expressed in this sample and is assigned with binary value 1 (0). This process was conducted using Weka3.8 (Frank 2014).

Let $I[R, S]$ denote the binary expression matrix, where R is the set of

Table 3.1: A matrix of expression levels of RNAs

	sa_1	sa_2	...	sa_s
lnc_1	40	50	...	70
...	...	...	...	...
lnc_n	10	15	...	33
mir_1	450	350	...	150
...	...	...	...	...
mir_k	500	700	...	600
mr_1	20	30	...	50
...	...	...	...	...
mr_m	65	85	...	25

RNAs from the original data set after the noise removal, and S is the set of samples. In the binary expression matrix, 1 represents that the expression level of the RNA is relatively high, 0 means that the expression level of the RNA is relatively low. Table 3.2 is the binary expression matrix transformed from Table 3.1.

For a given binary expression matrix $I[R, S]$, we define that r' is a RNA from R and sa' is a sample from S. $I[r', sa']$ is the value of the RNA r' of the sample sa' in the binary expression matrix $I[R, S]$. For example, in Table 3.2, $I[lnc_1, sa_1]$ is 0 and $I[mr_m, sa_2]$ is 1.

3.2.2 Constructing a candidate ceRNA network.

The target mRNAs and lncRNAs of the miRNAs were downloaded from the miRWalk2.0 database (`http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/holistic.html`). The miRWalk2.0 database contains the comparison results of binding sites from 12 existing miRNA-target prediction software tools (Dweep & Gretz 2015). It is a high quality database of miRNA targets. Also, this database contains the miRNA's target lncRNAs and target mRNAs. An miRNA (with p-value ≤ 0.05 and absolute fold change ≥ 2.0), its target lncRNAs (with p-value ≤ 0.05 and absolute fold

Table 3.2: The binary expression matrix of RNAs transformed from Table 3.1

	sa_1	sa_2	...	sa_s
lnc_1	0	0	...	1
...	...	...	...	...
lnc_n	0	0	...	1
mir_1	1	1	...	0
...	...	...	...	...
mir_k	0	1	...	1
mr_1	0	0	...	1
...	...	...	...	...
mr_m	1	1	...	0

change $\geq$ 3.0) and its target mRNAs (with p-value $\leq$ 0.05 and absolute fold change $\geq$ 2.0) are used to construct the initial ceRNA network. The differentially expressed lncRNA, miRNA, and mRNA are computed by using fold change (Grishin 2001) and the t-test method (Baldi & Long 2001).

Suppose a lncRNA *lnc*, an miRNA *mir*, and an mRNA *mr* form a ceRNA crosstalk. If *lnc* up-regulates in breast cancer samples, then the fold change of *lnc* should be larger than 0. According to the competition rule, the highly expressed lncRNA can lead to low expression of the miRNA, i.e., *mir* down-regulates and the fold change of *mir* should be smaller than 0. The low expression level of the miRNA increases the expression level of the mRNA. Therefore, *mr* up-regulates in breast cancer samples, and the fold change of *mr* should be larger than 0. Similarly, if *lnc* down-regulates and the fold change of *lnc* is smaller than 0, then *mir* up-regulates in breast cancer samples and the fold change of *mir* should be larger than 0. Then *mr* down-regulates in breast cancer tumor and the fold change of *mr* is smaller than 0. Based on this principle, we propose a competition regulation mechanism. This competition regulation mechanism is divided into a positive and a negative competition regulation facet:

- Positive competition regulation mechanism: the fold change of the

miRNA is larger than 0, and the fold changes of lncRNAs and mRNAs are smaller than 0.

- Negative competition regulation mechanism: the fold change of the miRNA is smaller than 0, the fold changes of lncRNAs and mRNAs are larger than 0.

Given the initial ceRNA network, we find the lncRNAs and mRNAs which follow the positive or negative competition regulation mechanism. Then the miRNA, lncRNAs, and mRNAs construct a candidate ceRNA network. We denote the candidate ceRNA network by $N' = (lncR,\ mir,\ mR)$, where $lncR$ and mR stand for the sets of lncRNAs or mRNAs which follow the competition regulation mechanism.

3.2.3 Computing the competition score

A candidate ceRNA network is formed by combining many ceRNA crosstalks. Some of these candidate ceRNA crosstalks may not satisfy the competitive relationship. Pointwise mutual information was proposed to measure the relationships between individual words in a corpus (Church & Hanks 1990). For given two words x and y, their pointwise mutual information is

$$PMI(x, y) = log\frac{P(x, y)}{P(x)P(y)}$$

Where $P(x, y)$ is the probability of observing two words together. $p(x)$ and $p(y)$ are the probability of observing x and y respectively. If the pointwise mutual information is high, these two words frequently co-occur and they likely to be a phrase. The equation of pointwise mutual information is similar with the equation of mutual information. However, the pointwise mutual ifnroamtion refers to single event while the mutual information reference to multiple events.

In this work, we apply it to measure the competitive relationships between RNAs in a ceRNA network, namely if a lncRNA can cross regulate an mRNA through an miRNA, the pointwise mutual information of this crosstalk should

be high. Traditional pointwise mutual information utilizes the probability coincidence or Gaussian kernel to measure the relationship between the variables; and only a positive or only a negative score between the variables is calculated. However, the competitions in a ceRNA crosstalk have both negative and positive relationships between the two RNAs. Therefore, the traditional pointwise mutual information needs to be refined for measuring the competition relationships between the RNAs in a ceRNA crosstalk. In this work, we calculate the pointwise mutual information based on our competition rule, as detailed below.

Given a candidate ceRNA network $N' = (lncR, mir, mR)$, where $lncR = \{lnc_1, lnc_2, \ldots, lnc_n\}$ and $mR = \{mr_1, mr_2, \ldots, mr_m\}$, any lncRNA $lnc_i \in lncR$, mir, and any mRNA $mr_j \in mR$ can form a ceRNA crosstalk $T = (lnc_i, mir, mr_j)$. We use a competition score to measure the reliability of each ceRNA crosstalk. The higher the competition score of the ceRNA crosstalk is, the more reliable the ceRNA crosstalk is.

Given a binary expression matrix $I[R, S]$, let lnc_i, mir, and mr_j be a lncRNA, an miRNA, and an mRNA of R, respectively, and let sa_l be one of the samples in S. If lnc_i, mir, and mr_j in sa_l are satisfied with one of these conditions:

- Condition 1: $I[lnc_i, sa_l] = 0$, $I[mir, sa_l] = 1$, and $I[mr_j, sa_l] = 0$.
- Condition 2: $I[lnc_i, sa_l] = 1$, $I[mir, sa_l] = 0$, and $I[mr_j, sa_l] = 1$.

we say that sa_l is the competition sample of $T = (lnc_i, mir, mr_j)$. For example, at Table 3.2, sa_1 is a competition sample of $T = (lnc_1, mir_1, mr_1)$, since $I[lnc_1, sa_1] = 0$, $I[mir_1, sa_1] = 1$, and $I[mr_1, sa_1] = 0$. In addition, we define that $supp^S(lnc_i, mir, mr_j)$ is the total number of the competition samples of $T = (lnc_i, mir, mr_j)$ in the sample set S.

The competition score of $T = (lnc_i, mir, mr_j)$ is computed by using pointwise mutual information (PMI):

$$PMI^S_{mir}(lnc_i, mr_j) = log \frac{P^S_{mir}(lnc_i, mr_j)}{P^S_{mir}(lnc_i) P^S_{mir}(mr_j)}$$

where $P^S_{mir}(lnc_i, mr_j)$, $P^S_{mir}(lnc_i)$, and $P^S_{mir}(mr_j)$ are computed by:

$$P^S_{mir}(lnc_i, mr_j) = \frac{supp^S(lnc_i, mir, mr_j)}{\sum_{i'=1}^{n}\sum_{j'=1}^{m} supp^S(lnc_{i'}, mir, mr_{j'})}$$
$$P^S_{mir}(lnc_i) = \frac{\sum_{j'=1}^{m} supp^S(lnc_i, mir, mr_{j'})}{\sum_{i'=1}^{n}\sum_{j'=1}^{m} supp^S(lnc_{i'}, mir, mr_{j'})}$$
$$P^S_{mir}(mr_j) = \frac{\sum_{i'=1}^{n} supp^S(lnc_{i'}, mir, mr_j)}{\sum_{i'=1}^{n}\sum_{j'=1}^{m} supp^S(lnc_{i'}, mir, mr_{j'})}$$

A positive PMI means the variables co-occur more frequently than what would be expected under an independence assumption, and a negative PMI means the variables co-occur less frequently than what would be expected.

3.2.4 Selecting a crosstalk which has a significant competition score

A competition score can be 0, negative, or positive. If the competition score of a ceRNA crosstalk is 0 or negative, it implies that there is no competitive relationship between the lncRNA, miRNA, and mRNA or the competitive relationship is less reliable than we would be expected. Such a ceRNA crosstalk should be discarded. A positive competition score indicates that the competitive relationship between these RNAs is more reliable than what we expected, and thus the ceRNA crosstalk is reliable to construct the ceRNA network. Further, the higher the competition score, the more reliable the ceRNA crosstalk is. Therefore, we should select those crosstalks which are reliable enough to construct the ceRNA network.

Suppose we are given t candidate ceRNA crosstalks and their competition scores are $\{PMI_1, PMI_2, \ldots, PMI_t\}$ which are all positive. A threshold θ is applied to distinguish low and high competition scores, and the problem is to reject the null hypothesis. The null hypothesis is that the competition score is small, that is, it implies there is no competing relationship in this crosstalk. If the competing score is very high, the null hypothesis can be rejected—it implies that this ceRNA crosstalk involves in regulating the biological

process. For a ceRNA crosstalk a, its significance level θ_a of the competition score is:

$$\theta_a = \frac{PMI_a - \overline{PMI}}{\sigma}$$

where $\overline{PMI}$ and σ are the average and standard deviation of the entire competition scores. The p-value of the ceRNA crosstalk a is $p_a = erfc(\theta_a/\sqrt{2})$ (Theiler, Eubank, Longtin, Galdrikian & Farmer 1992). If the p-value of a ceRNA crosstalk is lower than 0.05, this ceRNA crosstalk has significant competition score. We select those ceRNA crosstalks which have significant competition scores to construct the ceRNA network.

The novelty of our method is to apply competition regulation mechanism to construct candidate ceRNA networks and utilize the pointwise mutual information (PMI) to calculate the competition scores. The competition regulation mechanism, which is deducted from the competition rule, reflects the nature of the competition rule. Therefore, this regulation mechanism is a critical feature of the ceRNA network and can be applied to filter out many noisy eRNAs. Pointwise mutual information can measure both non-linear and linear relationship, and it is suitable for calculating the competition score of ceRNA crosstalks. Further, our method utilizes the pointwise mutual information to measure the point-to-point competitive relationships between lncRNA, miRNA, and mRNA, but not the pair-wise relationship between the two RNAs.

3.3 Results

We report two important ceRNA networks related to breast cancer and reveal their characteristics. We also report how these ceRNA networks play vital roles in KEGG pathways.

3.3.1 Two important ceRNA networks related to breast cancer

We applied the paired breast cancer RNA-seq data set in TCGA GDC data to construct the ceRNA network. Our method identified 352 mRNAs, 24 miRNAs, and 136 lncRNAs which are differentially expressed between the tumor and normal tissues. As there are 4 of these miRNAs which do not have any predicted target RNAs in the RNAwalker2.0 database, ceRNA networks mediated by the remaining 20 miRNAs which have target RNAs in the database are constructed. The 20 miRNAs are: hsa-miR-200a-5p, hsa-miR-203a-3p, hsa-miR-33a-5p, hsa-miR-21-3p, hsa-miR-183-5p, hsa-miR-144-5p, hsa-miR-145-5p, hsa-miR-184, hsa-miR-451a, hsa-miR-9-3-5p, hsa-miR-182-5p, hsa-miR-940, hsa-miR-375, hsa-miR-5683, hsa-miR-3677-3p, hsa-miR-429, hsa-miR-486-2-5p, hsa-miR-210-3p, hsa-miR-335-5p, hsa-miR-196a-2-5p, hsa-miR-21-5p, hsa-miR-378a-3p, hsa-miR-3065-5p, and hsa-miR-142-3p. The total number of candidate ceRNA crosstalks mediated by these 20 miRNAs is 75501.

To narrow down the study, we focus our analysis on two significant ceRNA networks: one is mediated by hsa-miR-451a, and the other is mediated by hsa-miR-375. These two miRNAs have a vital role in regulating the development of breast cancer as reported in literature (Camps, Saini, Mole, Choudhry, Reczko, Guerra-Assunção, Tian, Buffa, Harris, Hatzigeorgiou et al. 2014, Simonini, Breiling, Gupta, Malekpour, Youns, Omranipour, Malekpour, Volinia, Croce, Najmabadi et al. 2010), but their ceRNA networks have not been investigated previously. Our pointwise mutual information based method detected 132 candidate ceRNA crosstalks mediated by hsa-miR-451a and 1547 candidate ceRNA crosstalks mediated by hsa-miR-375. Of them, 25 candidate ceRNA crosstalks mediated by hsa-miR-451a have significant competition scores and only 273 candidate ceRNA crosstalks mediated by hsa-miR-375. We use these ceRNA crosstalks which have significant competition scores to construct the ceRNA networks. Figure. 3.3 is the ceRNA network mediated by hsa-miR-451a and Figure. 3.4

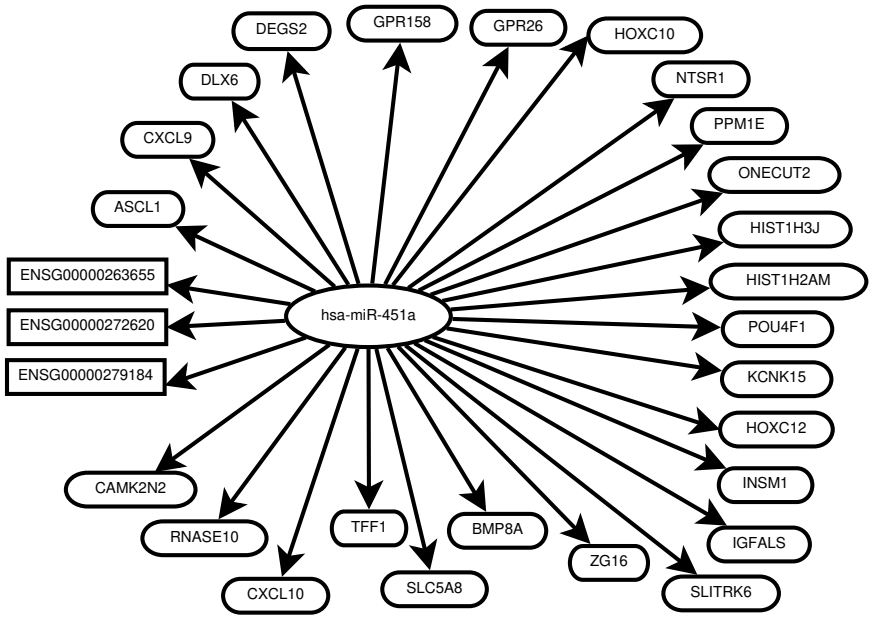

Figure 3.3: **A ceRNA network mediated by hsa-miR-451a**. The rectangle and oval boxes contain the names of lncRNAs and mRNAs, respectively

presents the ceRNA network mediated by hsa-miR-375. The Python source code of our algorithm to construct the network can be downloaded from website `https://github.com/ChaowangLan/ceRNA`.

3.3.2 Characteristics of the two ceRNA networks

The two ceRNA networks are satisfied with the three characteristics of ceRNA networks: (1) the expression level of every lncRNA between the normal and tumor samples is highly differential, (2) every lncRNA is a target of the miRNA, and (3) the expression levels of lncRNA, mRNA and miRNA follow the competition rule. The absolute fold change of these lncRNAs in ceRNA crosstalks mediated by hsa-miR-451a and hsa-miR-375 are larger than 3.0 and the p-values are smaller than 0.01. This means that these lncRNAs are over-expressed and satisfy the first point of characteristics of a ceRNA network. Table 3.3 presents the detailed expression fold change and the p-values of these lncRNAs.

When a lncRNA competes with an mRNA for binding to the same miRNA, the lncRNA and the mRNA both are the targets of the miRNA. We examined the seed regions of hsa-miR-451a to see whether its target mRNAs or lncRNAs are complementary to the seed region in sequence (Ellwanger et al. 2011). ENSG00000272620 is perfectly complementary to the seed region of hsa-miR-451a, and mRNA *DLX6* is complementary to the seed region of the hsa-miR-451a with one mismatch pair. This suggests that lncRNA ENSG00000272620 and mRNA *DLX6* should be very likely the targets of hsa-miR-451a. Figure. 3.5 shows the binding region of lncRNA ENSG00000272620 and hsa-miR-451a and the binding region of mRNA *DLX6* and hsa-miR-451a.

Table 3.4 shows the top 5 competition scores of the crosstalks mediated by hsa-miR-451a and hsa-miR-375, as calculated by our pointwise mutual information method. A different ceRNA network has a different competition score. Some of the ceRNA competition scores may be similar. For example, the largest competition score of the ceRNA crosstalk mediated by hsa-miR-

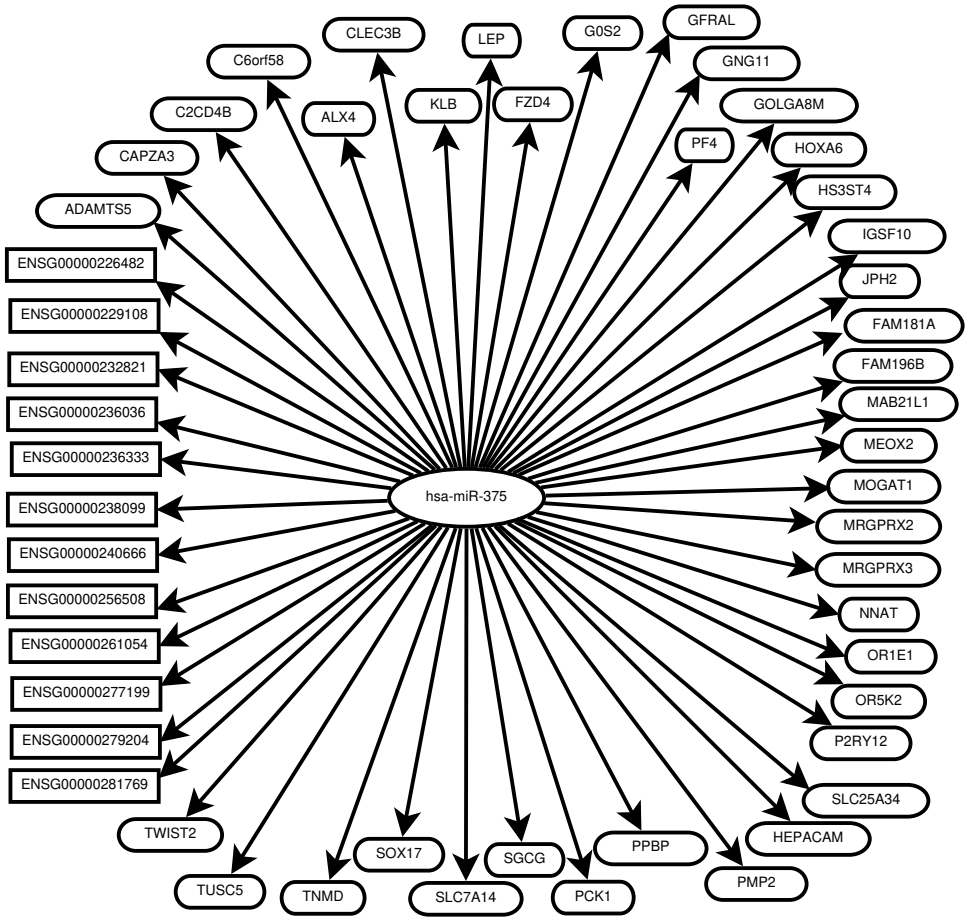

Figure 3.4: **The ceRNA network formed from the top** 50 **candidate ceRNA crosstalks mediated by hsa-miR-375.** Text words in the rectangle boxes are the names of the lncRNAs and text words in the oval boxes are the names of the mRNAs.

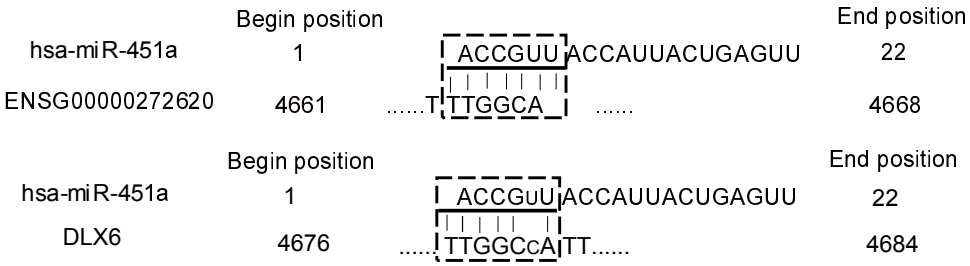

Figure 3.5: **The binding sites of lncRNA, miRNA, and mRNA.**

Table 3.3: Expression fold change ratios and p-values of the lncRNAs involved in the ceRNA networks mediated by hsa-miR-451a and hsa-miR-375

LncRNA	Fold change	p-value
ENSG00000226482	-4.19	$7.46 * 10^{-9}$
ENSG00000227260	-3.21	$3.18 * 10^{-26}$
ENSG00000229108	-3.76	$8.50 * 10^{-20}$
ENSG00000232821	-3.03	$2.33 * 10^{-12}$
ENSG00000236036	-3.25	$6.19 * 10^{-8}$
ENSG00000236333	-4.88	$2.11 * 10^{-22}$
ENSG00000238099	-3.20	$4.66 * 10^{-16}$
ENSG00000240666	-3.35	$1.41 * 10^{-18}$
ENSG00000256508	-3.55	$9.48 * 10^{-29}$
ENSG00000261054	-3.14	$2.17 * 10^{-18}$
ENSG00000277199	-3.48	$7.06 * 10^{-7}$
ENSG00000279204	-3.35	$4.92 * 10^{-18}$
ENSG00000281769	-4.10	$4.24 * 10^{-6}$
ENSG00000263655	4.62	$2.27 * 10^{-7}$
ENSG00000272620	3.86	$7.36 * 10^{-3}$
ENSG00000279184	3.31	$2.04 * 10^{-5}$

Table 3.4: Top-5 competition scores in the ceRNA crosstalks mediated by $hsa-miR-375$ and $hsa-miR-451a$

lncRNA	miRNA	mRNA	Score	P-value
ENSG00000277199	hsa-miR-375	*GFRAL*	0.35	$6.76 * 10^{-236}$
ENSG00000238099	hsa-miR-375	*C6orf58*	0.35	$8.48 * 10^{-228}$
ENSG00000279204	hsa-miR-375	*SOX17*	0.31	$1.51 * 10^{-184}$
ENSG00000229108	hsa-miR-375	*DUXA*	0.30	$2.56 * 10^{-171}$
ENSG00000277199	hsa-miR-375	*MEOX2*	0.30	$3.27 * 10^{-167}$
ENSG00000272620	hsa-miR-451a	*DLX6*	0.35	$8.88 * 10^{-45}$
ENSG00000279184	hsa-miR-451a	*ZG16*	0.32	$1.60 * 10^{-37}$
ENSG00000272620	hsa-miR-451a	*INSM1*	0.31	$3.89 * 10^{-35}$
ENSG00000272620	hsa-miR-451a	*NTSR1*	0.30	$4.92 * 10^{-33}$
ENSG00000272620	hsa-miR-451a	*GPR26*	0.30	$4.92 * 10^{-33}$

451a is equal with the competition score of the ceRNA crosstalk mediated by hsa-miR-375. But some competition score of the ceRNA crosstalk is not very similar. Such as the largest competition score of the ceRNA crosstalk mediated by hsa-miR-21-5p is 0.53 which is larger than the largest competition score of ceRNA crosstalk mediated by hsa-miR-451a. However, if two ceRNA crosstalks are mediated by the same miRNA, the higher competition score of the ceRNA crosstalk is, the more reliable the crosstalk is.

3.3.3 CeRNA networks and breast cancer treatment

The ceRNA crosstalks mediated by hsa-miR-375 or by hsa-miR-451a may regulate the development of breast cancer. These ceRNA crosstalks should be considered in the future for the treatment plan of breast cancer.

As suggested in the third row of Table 3.4, ENSG00000279204 competes with *SOX17* for binding to hsa-miR-375. *SOX17* is a member of the SRY-related HMG-box family that can regulate cell development (Kamachi &

Kondoh 2013). Fu. et al found that increasing the expression level of this gene can slow down the speed of breast cancer growth; but reducing the expression level of this gene can lead to poor survival outcomes in breast cancer patients (Fu, Tan, Wei, Zhu, Jiang, Zhu, Cai, Chong & Ren 2015). Thus *SOX17* can be a useful biomarker for breast cancer patients. It can be also understood that the expression of *SOX17* can be up-regulated with the increase of the expression of *ENSG00000279204*. A high expression level of *SOX17* would lead to decreased growth of breast cancer cell so as to improve the treatment of breast cancer patients.

The gene *MEOX2* is also called *GAX* or *MOX2*. This gene is down-regulated in breast cancer (Yu, Lee, Tan & Tan 2004). Recent research shows that *MEOX2* can up-regulate *p21* which is very important for breast tumor grading (Abbas & Dutta 2009). Highly expressed *p21* prevents the growth of breast cancer (Sheikh, Rochefort & Garcia 1995). As shown in the fifth line of Table 3.4, ENSG00000229108 competes with *MEOX2* for binding with hsa-miR-375. The high expression level of *MEOX2* can enhance the growth of breast cancer. Therefore, decreasing the expression level of ENSG00000229108 can reduce the expression level of *MEOX2*. Thus the high expression level of *MEOX2* would inhibit the growth of breast cancer.

In the last second line of Table 3.4, ENSG00000272620 competes with *NTSR1* for binding with hsa-miR-451a. *NTSR1* (Neurotensin Receptor 1) is a target of the Wnt/APC oncogenic pathways which is involved in cell proliferation and transformation (Souazé, Dupouy, Viardot-Foucault, Bruyneel, Attoub, Gespach, Gompel & Forgez 2006). Dupouy found that highly expressed *NTSR1* is associated with the size, the number of metastatic lymph nodes, and Scarff-Bloom-Richardson grading (Dupouy, Viardot-Foucault, Alifano, Souazé, Plu-Bureau, Chaouat, Lavaur, Hugol, Gespach, Gompel et al. 2009). These suggest that *NTSR1* is a promising target for breast cancer treatment. According to the predicted results, decreasing the expression level of ENSG00000272620 can decrease the expression level of *NTSR1*. Low expression level of *NTSR1* is beneficial for the treatment of

breast cancer.

Most breast cancer patients die because of the "incurable" nature of the metastasis breast cancer (Lu, Steeg, Price, Krishnamurthy, Mani, Reuben, Cristofanilli, Dontu, Bidaut, Valero et al. 2009). About 90% of breast cancer deaths are due to metastasis; indeed, only 20% of the metastatic breast cancer patients can survive more than 1 year (Neman, Choy, Kowolik, Anderson, Duenas, Waliany, Chen, Chen & Jandial 2013). Therefore, inhibiting breast cancer metastasis is very crucial for breast cancer treatment. Morini found that $DLX6$ involves in the metastasis potential of breast cancer (Morini, Astigiano, Gitton, Emionite, Mirisola, Levi & Barbieri 2010). Prest also pointed out that $TFF1$ can promote breast cancer cell migration (Prest, May & Westley 2002). These studies imply that $DLX6$ and $TFF1$ are highly related to breast cancer metastases. Therefore, decreasing the expression level of these two genes can inhibit breast cancer metastasis. According to our results, lncRNA ENSG00000272620 and ENSG00000279184 cross-regulate *DLX6* and *TFF1* via hsa-miR-451a, respectively. Decreasing the expression level of ENSG00000272620 and ENSG00000279184 can decline the expression levels of *DLX6* and *TFF1*. The low expression levels of these two genes would prevent the development of metastatic breast cancer.

3.3.4 Roles of ceRNA networks in KEGG pathways

Some lncRNAs can cross-regulate genes which are involved in KEGG pathways. Enrichr (`http://amp.pharm.mssm.edu/Enrichr/`), a gene enrichment analysis web server, is applied to find out these KEGG pathways (Kuleshov et al. 2016). 14 KEGG pathways are found with p-values lower than 0.05. Some of these KEGG pathways are the key pathway in breast cancer and may be a potential drug target for breast cancer treatment, such as the chemokine signaling pathway, the cytokine-cytokine receptor interaction, and the neuroactive ligand-receptor interaction (Park, Rogan, Tarnowski & Knoll 2012, Lazennec & Richmond 2010, Morales, Planet, Arnal-Estape, Pavlovic, Tarragona & Gomis 2011). All the KEGG pathways

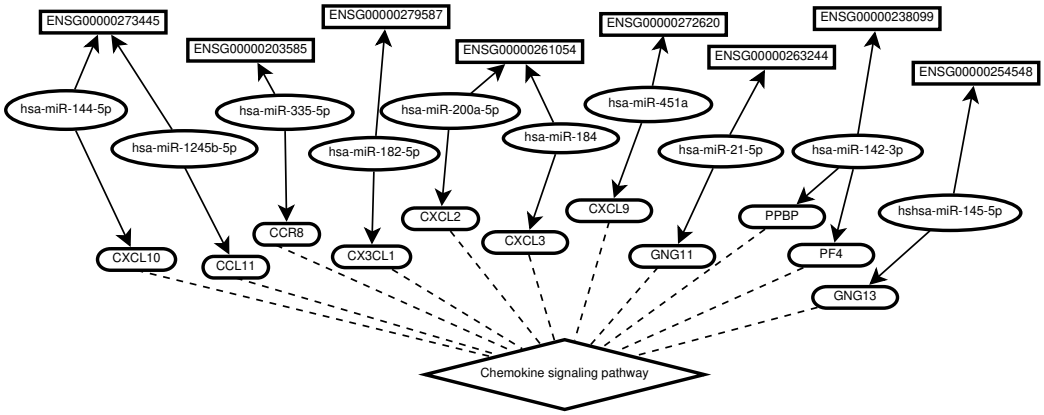

Figure 3.6: **The ceRNA networks involved in the chemokine signaling pathway.**

are presented in Table. 3.5. In this subsection, we focus on analyzing the chemokine signaling pathway.

The cross regulation between the lncRNAs and the genes involved in the chemokine signaling pathway is shown in Figure. 3.6, demonstrating 11 genes related to chemokine signaling pathway are involved in breast cancer. Of them, *CXCL10*, *CXCL9*, *CCL11*, *CCR8*, and *GNG13* up-regulate breast cancer, while the other genes download-regulate breast cancer. Chemokine signaling pathway expresses on the immune cells and regulates immune responder. However, new evidences show that the gene in the chemokine signaling pathway also plays a vital role in breast cancer progression (Lazennec & Richmond 2010). For example, *CXCL10* affects the tumor microenvironment and plays important role in breast cancer progression (Mulligan, Raitman, Feeley, Pinnaduwage, Nguyen, O'Malley, Ohashi & Andrulis 2013), *CXCL9* is identified as a biomarker in breast cancer (Ruiz-Garcia, Scott, Machavoine, Bidart, Lacroix, Delaloge & Andre 2010). Regulating these gene can inhibit the growth of breast cancer.

Table 3.5: KEGG pathways which can be regulated by ceRNA networks

KEGG name	P-value	Number of gene
Alcoholism	$3.62 * 10^{-19}$	28
Systemic lupus erythematosus	$4.48 * 10^{-19}$	25
Viral carcinogenesis	$5.04 * 10^{-5}$	13
Cytokine-cytokine receptor interaction	$1.84 * 10^{-4}$	14
Chemokine signaling pathway	$3.62 * 10^{-4}$	11
Transcriptional misregulation in cancer	$3.69 * 10^{-3}$	9
Salivary secretion	$4.06 * 10^{-3}$	6
Neuroactive ligand-receptor interaction	$7.92 * 10^{-3}$	11
Serotonergic synapse	$1.21 * 10^{-2}$	6
Oxytocin signaling pathway	$1.84 * 10^{-2}$	7
Morphine addiction	$1.93 * 10^{-2}$	5
Circadian entrainment	$2.28 * 10^{-2}$	5
Renin secretion	$2.32 * 10^{-2}$	4
Retrograde endocannabinoid signaling	$2.88 * 10^{-2}$	5

3.3.5 A ceRNA which may be an efficient drug target for breast cancer treatment

Two different miRNAs may have common target mRNAs and common target lncRNAs. A common target lncRNA can cross-regulate mRNAs through different miRNAs. Therefore, this common target lncRNA is an efficient drug target for cancer treatment. An example can be found in Figure. 3.7. The lncRNA ENSG00000261742 competes for binding to hsa-miR-21-5p, hsa-miR-33a-5p and hsa-miR-184 with HOXA5 and *EGR1*. *EGR1* is known to up-regulate *PTEN* which is a key tumor breast suppressor gene (Redmond, Crawford, Farmer, D'costa, O'brien, Buckley, Kennedy, Johnston, Harkin & Mullan 2010). It implies that increasing the expression level of EGR1 can suppress the development of breast cancer. The lowly expressed *HOXA5* lead to the functional activation of twist and promoting the development of breast cancer (Stasinopoulos, Mironchik, Raman, Wildes, Winnard & Raman 2005). Therefore, increasing the expression level of these two mRNAs are very important for breast cancer treatment.

Hsa-miR-21-5p, hsa-miR-33a-5p, and hsa-miR-184 can regulate the expression of these two mRNAs. However, only decreasing the expression level of one miRNA cannot enhance the expression levels of these two mRNAs, since the high expression of the other miRNA can decrease the expression of both mRNAs. In our results, increasing the expression of ENSG00000261742 can enhance the expression of these two mRNAs by decreasing the expression of these two miRNAs. Therefore, ENSG00000261742 is an efficient drug target for increasing the expression of both mRNAs. About all, this ceRNA is suggested to be an efficient drug target for breast cancer treatment.

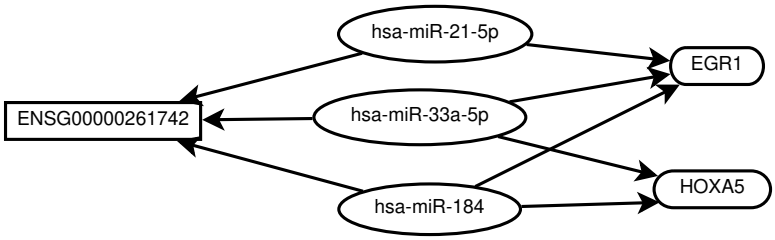

Figure 3.7: **A ceRNA network cross-regulates two mRNAs through three miRNAs.**

3.3.6 Comprehensive Comparison with Other Methods

We compared our prediction results with three existing methods. The first comparison is with Chen's method (Chen, Xu, Li, Zhang, Chen, Lu, Wang, Zhao, Xu, Li et al. 2017) which is based on the idea of Pearson correlation coefficient. The second comparison is with Paci's method (Paci et al. 2014) which is a partial correlation method. The third comparison is with Sumazin's method (Sumazin et al. 2011) which is based on the conditional mutual information. Our method, Chen's method, Paci's method, and Sumazin's method predicted total 30365, 106045, 15420, and 227755 ceRNA crosstalks, respectively. Sumanzin's method identified the most. Figure 3.8 is a Venn graph showing the common and unique ceRNA crosstalks predicted by these methods. Each boundary line encloses a number of ceRNA crosstalks predicted by one or more methods; and the intersection areas indicate the numbers of common ceRNA crosstalks.

Note that 26620 ceRNA crosstalks predicted by our method are also identified by one of the three existing methods. Our method have more common ceRNA crosstalks in comparison with Sumazin's method (21095 of our predicted ceRNA crosstalks) than the other methods (1168 and 16153 of our predicted ceRNA crosstalks are identified by the Paci's method and Chen's method, respectively). However, 3745 ceRNA crosstakls predicted by our method are not identified by the other methods.

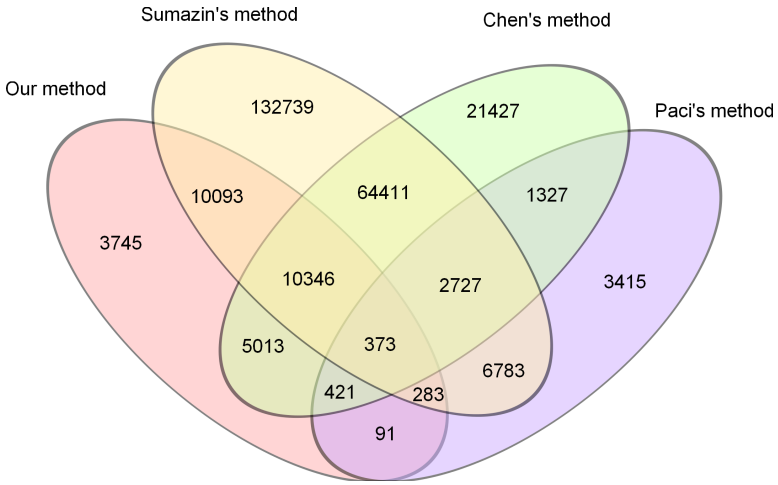

Figure 3.8: **The common and unique ceRNA crosstalks predicted by various methods.**

Some of these ceRNA crosstalks may regulate breast cancer processes. For example, the ceRNA crosstalk (ENSG00000272620, hsa-miR-451a, *GPR26*), which had been showed in Table 1, was able to be predicted by our method only. Gene *GPR26* is a member of G-protein-coupled receptors. The G-protein-coupled receptors can play key roles in tumorigenesis, angiogenesis, and metastasis (Singh, Nunes & Ateeq 2015). The ceRNA crosstalk (ENSG00000250266, hsa-miR-142-3p, *PF4*) is also predicted by our method but not identified by the other methods. The highly expressed lncRNA ENSG00000272620 may compete and cross regulate *GPR*26 for binding to hsa-miR-451a to influence breast cancer tumorigenesis, angiogenesis, and metastasis. The lowly expressed lncRNA ENSG00000250266 could not down-regulate the hsa-miR-142-3p and might lead to lowly expressed *PF4*. Lowly expressed *PF4* could not suppress breast cancer growth (Nafi, Idris & Jaafar 2017). LncRNA ENSG00000250266 may be a potential target for breast cancer treatment.

Many methods, including Paci's method, identify ceRNA networks only taking into account the expression data. These methods could find all the negative and positive expression relationships between the RNAs. It

seems that these methods are unbiased and preferable to identify ceRNA networks. However, the competition relationship between RNAs is a specific relationship (i.e., lncRNA and miRNA are negatively co-expressed; miRNA and mRNA are negatively co-expressed). Thus, ceRNA networks should all hold this specific relationship.

3.4 Conclusion and Discussion on Future Work

In this chapter, we proposed a novel method for constructing ceRNA networks from paired RNA-seq data sets. We first identify the differentially expressed lncRNAs, miRNAs, and mRNAs from the paired RNA-seq data sets. Then we derive the competition regulation mechanism from the competition rule and construct the candidate ceRNA crosstalks based on this rule. This competition regulation mechanism is another feature of the ceRNA network and is useful for constructing ceRNA networks. Finally, the pointwise mutual information is applied to measure the competitive relationship between these RNAs to select reliable ceRNA crosstalks to construct the ceRNA networks. The analysis results have shown that the function of ceRNA networks is related to the growth, proliferation, and metastatic of breast cancer. These ceRNA networks present the complex regulatory mechanism of the RNAs in breast cancer. In addition, the ceRNA networks suggest a new approach for breast cancer treatment. The ceRNA hypothesis is still in its infancy, many ceRNA networks have not been discovered yet. The mutations of miRNA may change existing or lead to new crosstalk. For example, the $5'$ variant of miRNA may bind to different target mRNA or lncRNA comparing to its wildtype miRNA since the shift of the seed region of the miRNA. Further, the ceRNA hypothesis illustrates the complexity of RNA regulatory network. By this hypothesis, some other complexity networks may exist. Our method for discovering ceRNA network from the RNA-seq data that contains the expression level of RNA (miRNA,

lncRNA, and mRNA) is limited to only the tumor and normal tissues, how to incorporate different tissues that have a matching RNA and miRNA sequencing data set to extend our analysis is a future direction of our research in this area.

A lncRNA that is not differentially expressed may contribute to the sponge mechanism as well (Conte, Fiscon, Chiara, Colombo, Farina & Paci 2017). In particular, the relative concentration of the ceRNAs and changes in the ceRNA expression levels are very important for discovering ceRNA networks (Salmena et al. 2011). Indeed, conditions like the relative concentration of ceRNAs and their microRNAs or other conditions not necessarily corresponding to differentially expressed RNAs can be applicable as starting points to discover ceRNAs. These will be some of our future work to enrich the ceRNA sponge hypothesis.

Chapter 4

An isomiR Expression Panel Based Novel Breast Cancer Classification Approach using Improved Mutual Information

4.1 Introduction

In subsection 1.1.2, we presented the miRNA and isomiR formation. The isomiR is the isoform of miRNA. It could be envisioned that the increased expression of miRNA variants, or individual isomiRs, lead to the loss or weakening of the function of the corresponding wild type mature miRNA or result in the regulation of a different transcriptome. Recent studies suggest that isomiRs probably play vital roles in a variety of cancers, tissues, and cell types (Chen & Wong 2017). For example, Juzenas and colleagues claimed that isomiRs are differentially expressed in different human blood cell types (Juzenas, Venkatesh, Hübenthal, Hoeppner, Du, Paulsen, Rosenstiel, Senger, Hofmann-Apitius, Keller et al. 2017). Telonis and colleagues showed that specific isomiRs could be superior cancer biomarkers compared to mature miRNAs when they used isomiRs to classify

32 different cancers (Telonis et al. 2017). Specifically, Telonis and colleagues demonstrated that miRNA-based analysis was unable to differentiate two specific subtypes of breast cancer while, in comparison, isomiRs were able to make clear distinctions between the two subtypes (Telonis et al. 2015). These findings suggest that isomiRs may play critical roles in differentiating subtypes of breast cancer and, furthermore, may provide novel insights into understanding the molecular mechanisms leading to the development of breast cancers.

Breast cancer is the most common cancer and the second leading cause of cancer-related deaths among women worldwide (Lynce, Blackburn, Cai, Wang, Rubinstein, Harris, Isaacs & Pohlmann 2018). Routine clinical evaluation and diagnosis of breast cancer is categorised into three major distinct molecular subtypes based on their hormone receptor status: estrogen receptor (ERα) and progesterone receptor (PR) positive, Herceptin 2 positive (HER2+), and triple negative (ER/PR/HER2 negative) (Patani, Martin & Dowsett 2013, Goldhirsch, Wood, Coates, Gelber, Thürlimann, Senn & members 2011, Ellsworth, Blackburn, Shriver, Soon-Shiong & Ellsworth 2017). However, the link between molecular mechanisms and disease prognosis defining the breast cancer subtypes is unclear (Taherian-Fard, Srihari & Ragan 2014). Understanding the mechanisms of breast cancer subtyping is clinically useful with respect to prognosis, prediction, and informed therapeutic choices (Santagata, Thakkar, Ergonul, Wang, Woo, Hu, Harrell, McNamara, Schwede, Culhane et al. 2014). Within the major breast cancer subtypes, gene expression profiling has been used to further classify these molecular subtypes with the potential to design more specific targeted therapies (Lehmann, Bauer, Chen, Sanders, Chakravarthy, Shyr & Pietenpol 2011). In addition, gene expression profiling has been found to be more predictive of treatment response. For example, in a study by Finn and colleagues they showed reclassification of breast cancer subtypes using an unbiased gene expression profiling technique predicted a better treatment outcome compared to the conventional breast cancer subtyping

(ER/HER2 status) (Finn, Dering, Ginther, Wilson, Glaspy, Tchekmedyian & Slamon 2007). In this study, a subset of three genes expressed in breast cancer were more likely to predict responsiveness to dasatinib, a small molecule specific kinase inhibitor. Dasatinib has been used in clinical trials for hard to treat metastatic breast cancer (Herold, Christina I and Chadaram, Vijaya and Peterson, Bercedis L and Marcom, P Kelly and Hopkins, Judith and Kimmick, Gretchen G and Favaro, Justin and Hamilton, Erika and Welch, Renee A and Bacus, Sarah and others 2011). However, most breast cancer clinical trial studies using dasatinib are inconclusive and potentially these studies would benefit from gene profiling to understand the lack of responsiveness.

Complex genetic diseases, such as breast cancer, inherently pose the problem to be characterised by a few biomarkers that faithfully characterise the subtypes of the disease. MiRNAs and isomiRs provide a potentially better alternative for classifying complex diseases compared to mRNA based biomarkering since they are regulatory "hubs" of gene expression. Therefore, the changes in their expression could influence multiple downstream mRNAs and therefore diverse biological pathways.

In this chapter, we present a novel method that applies isomiR expression profiles for improved classification of breast cancer types using small RNA sequencing data available in the TCGA database. Firstly, since the TCGA dataset has many lowly expressed isomiRs that have significant negative influence on the identification of biomarkers, these lowly expressed isomiRs should be removed. The traditional method for removing the lowly expressed isomiRs is by selecting a 'hard' threshold (Juzenas et al. 2017, Telonis et al. 2017). If the expression levels of an isomiR is lower than this 'hard' threshold, this isomiR is viewed as lowly expressed and should be removed. However, this 'hard' threshold may lead to a loss of information (Zhang & Horvath 2005). In order to tackle this disadvantage, a 'soft' method based on a null hypothesis method was applied, and this method was designed to remove these lowly expressed isomiRs. Secondly, we utilized an improved

mutual information method to calculate the weight of each isomiR, which measured the significance of the isomiR to classify different subtypes of breast cancer. The higher the weight of the isomiR, the more suitable the isomiR for classifying different subtypes of breast cancer. The traditional mutual information can only be used if both the feature and the label are continuous or discrete data. This improved mutual information can be applied to features if it is continuous data and the label is discrete data. Finally, a few isomiRs, which have high weights, were able to classify different breast cancer subtypes. In order to identify these key isomiRs, the SVM classification method was used.

Although there are many methods that have been designed for biomarker discovery, they can be divided into two major categories. The first category selects a set of biomarkers that can classify the data (Li, Cheng, Wang, Morstatter, Trevino, Tang & Liu 2017), such as support vector machine (SVM) (Zhang, Mo, Ghoshal, Wilkins, Chen & Zhou 2017), mutual information (Zheng & Wang 2018), and swarm optimizer (Gu, Cheng & Jin 2018). These methods do not calculate the weight of each biomarker and therefore, the importance of the biomarker in each breast cancer subtype classification is not known. The weight of the biomarker may reflect its regulatory importance in the molecular mechanism of the disease; therefore, it may be worth studying the potential role of gene regulation of highly weighed biomarkers. Another category of methods view the gene or isomiR as the feature and calculates the weight of each feature. The weight of the feature measures the importance of the feature in the classification. The top N features viewed as biomarkers. Information gain, t-test, and fold change methods are widely applied to identify biomarkers (Saeys, Inza & Larrañaga 2007). However, t-tests and fold change methods are not suitable for identifying biomarkers from the data that has more than two categories. Although the information gain can be applied to find biomarkers from multiple categories, this method is very time consuming. Other methods, such as Fisher (Gu et al. 2011) and correlation coefficient method (Weston,

Elisseeff, Schölkopf & Tipping 2003), can calculate the weight of each feature for data that comprises of more than two categories and is less time consuming than information gain. However, these methods also have their limitations. The Fisher method is based on the mean and standard deviation of the dataset and therefore, small data sets, confounded by outliers will negatively influence the results. If weights of the feature are calculated by the correlation coefficient method, it challenges the rank features based on their weights (Yin et al. 2013). Together, all these methods used for identifying biomarkers have their limitations. Therefore, a novel method is needed to identify unique, more discrete and effective biomarkers.

4.2 Method

Our method for identifying isomiR biomarkers in different subtypes of breast cancer is composed of three steps. Firstly, it computes the expression level of isomiRs in each breast cancer sample and removes the lowly expressed isomiRs. Secondly, it utilizes improved mutual information to calculate the weight of each isomiR. Finally, the third step selects the critical isomiRs for breast cancer subtype classification, for which the SVM classification method is applied. These key isomiRs are viewed as breast cancer subtype biomarkers. Figure 4.1 shows the framework of our methodology.

4.2.1 Data Source and Definitions

The expression profiles of isomiRs in breast cancer patients can be downloaded from TCGA GDC data portal website (`https://portal.gdc.cancer.gov/cart`). However, the website does not provide the name of each isomiR. The nomenclature used in this study for discrete isomiR was derived from its mature miRNA: the name of the isomiR comprises of the name of the corresponding wild type miRNA followed by a variant symbol, e.g hsa-miR-21-5p|3′t-2. The sign | separates the isomiR name into miRNA name and variant symbol. The variant symbol is divided into two parts by the sign (−).

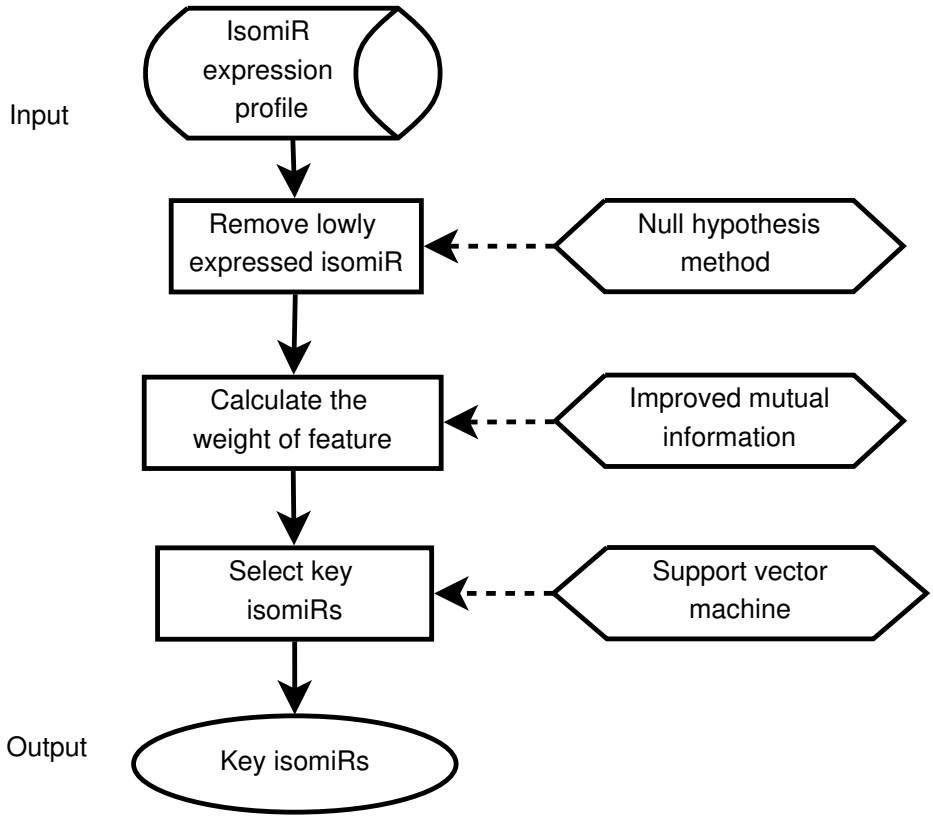

Figure 4.1: **IsomiR biomarker subtyping methodology.** The framework of the novel methodology designed for breast cancer biomarker subtyping is composed of three discrete steps from isomiR expression profiling to identification of key isomiRs used as novel biomarkers.

The first part indicates the variant type of the isomiR. 3′t (5′a) implies that this isomiR is 3′ trimming (5′ additional) isomiR. The second part denotes the number of the nucleotide that is trimmed or added. In addition, the number of reads are not suitable for analyze. Thus, we calculated the RPM (reads per million mapped reads) of each isomiR. The clinical information of the breast cancer patients was obtained from the website (`https://www.nature.com/articles/nature11412#supplementary-information`). Since the TCGA website does not provide the expression levels of polymorphic isomiRs, this kind of isomiR was not taken into consideration in this chapter. Although the

clinical information contained 824 breast cancer patients, only 698 patients had valid clinical information. In this chapter, we applied these 698 patients isomiR expression levels to identify biomarkers that classify breast cancer subtypes.

The traditional clinical classification method sorts breast cancer into three different subtypes based on the hormone receptor status. However, some breast cancer patients proved to be positive in both ERα/PR+ and HER2+ receptor status. These breast cancer patients were identified as ERα/PR+ or HER2+ breast cancer subtypes. However, it was not suitable to classify these breast cancer patients as ERα/PR+ or HER2+ breast cancer subtype patients. Therefore, these patients were reclassified as a fourth breast cancer subtype. Together, the breast cancer patients were classified into four subtypes and the number of patients in each subtype of breast cancer are shown in Table 4.1.

Table 4.1: Breast cancer subtype reclassification for isomiR identification.

Subtype name	ERα+HER2-	ERα-HER2+	ERα+HER2+	Triple negative
Number of patient	472	31	76	119

4.2.2 Removal of lowly expressed isomiR

A large amount of isomiRs were identified from the TCGA dataset. However, many isomiRs had to be removed since they were lowly expressed and had significant negative effects on the result. We defined in our dataset, that an isomiR was lowly expressed if the total expression level of the isomiR was relatively low in the dataset. The total expression level of isomiR was deemed the sum of the expression level of isomiR in all samples. In order to detect the distribution of total expression level of isomiRs, a histogram (Pearson 1895) of which the 'bin' of the bar graph equaled 1 was applied. Since the total expression level of isomiR was wide ranging, this histogram proved to be very large and therefore the complete histogram could not be displayed in

this research: the distribution of the total expression level less than 35 is shown in Figure 4.2.

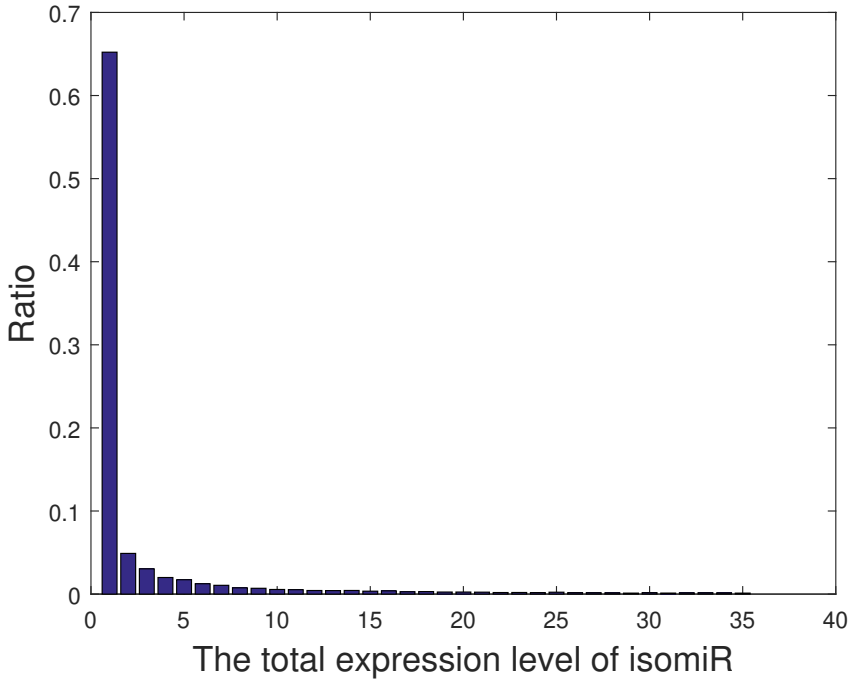

Figure 4.2: **The distribution of total expression levels of isomiRs.** The x-axis presents the total expression level. The ratio of the isomiRs was calculated using the number of the isomiRs in the bin divided the total number of isomiRs. For example, the ratio of the expression level isomiRs that lower than 1 is about 0.65. This implies that 65% of the isomiRs total expression level is lower than 1

According to Figure 4.2, about 65% of all isomiRs showed their total expression level was lower than 1. This implied that most of these isomiRs were lowly expressed. Further, it denoted that the distribution of the total expression level of isomiRs followed the exponential distribution. In order to remove these lowly expressed isomiRs, a null hypothesis method was applied. This null hypothesis states that: if the total expression level of an isomiR is very low, this isomiR is a noisy isomiR and should be removed. If the total

expression level of an isomiR is very high, the null hypothesis can be rejected and this isomiR is not a noisy isomiR. For given q isomiRs and the expression level of each isomiR in all breast cancer patients, we first calculated the total expression level of each isomiR. The total expression level of q isomiRs are denoted as $TE = \{te_1, te_2, \ldots, te_q\}$. The significance threshold θ of the competition score was calculated using the formula:

$$\theta = \frac{q * \overline{TE}}{\chi^2_{1-\alpha/2}(q)}$$

Where $\overline{TE}$ is the mean of all the total expression level of isomiRs, $\chi^2_{1-\alpha/2}(q)$ is the Chi-square with q degree of freedom, and α is the p-value. Here, the p-value was set at $P = 0.05$. Only the isomiRs whose total expression level was smaller than this significance threshold θ, being viewed as lowly expressed, were removed.

4.2.3 Calculating the weight of isomiR by improved mutual information

The mutual information is a powerful method in feature selection. Many mutual information-based feature selection methods have been developed and the performance has proven to be very good (Li, Cheng, Wang, Morstatter, Trevino, Tang & Liu 2017). However, these methods has some limitations. Although some methods select a set of features that are very important for classification, they do not provide the weight of the feature. Some methods are applied from the data of which both the feature and the label are discrete or continuous data. However, these methods were not deemed suitable for this type of research. Therefore, an improved mutual information was developed to calculate the weight of each isomiR. This improved mutual information calculated the weight of each isomiR and measured the relationship between features and labels.

For any given expression profile of isomiRs, this expression profile has m isomiR $X = \{x_1, x_2, \ldots, x_m\}$, n breast cancer patients $S = \{s_1, s_2, \ldots, s_n\}$,

and the subtype label of the patients $Y = \{y_1, y_2, \ldots, y_n\}$. x_τ^a is defined as the expression level of isomiR τ in the breast cancer sample a. The min-max normalization method is applied to scale the expression levels of each isomiR between 0 and 1. The mutual information between an isomiR x_τ and breast cancer subtype Y is:

$$I(x_\tau, Y) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_\tau^i, y_i)}{f(x_\tau^i) f(y_i)}$$

Where $f(x_\tau^i)$, and $f(y_i)$ are the density function of isomiR and label, respectively. $f(x_\tau^i, y_i)$ is the joint density function of isomiR and label. Since the expression level of isomiR is continuous data while the label is discrete data, the density function of isomiR and label should be calculated by different equations:

$$f(x_\tau^i) = \frac{1}{\sqrt{2\pi} n} \sum_{j=1}^{n} exp(-\frac{(x_\tau^i - x_\tau^j)^2}{2})$$

$$f(y_i) = \frac{1}{\sqrt{2\pi} n} \sum_{j=1}^{n} exp(-\frac{d(y_i, y_j)}{2})$$

Where $d(y_i, y_j)$ measures the distance between labels y_i and y_j. If these two labels are continuous data. The distance between two labels can be calculated by Euclidean distance. However, the label in this research is discrete data. The distance of two labels cannot be calculated by Euclidean distance. $d(y_i, y_j)$ is 0 if these two labels are the same, and it is 1 otherwise.

Since the improvement in calculating the distance between discrete labels, the mutual information is applicable for the dataset where the feature is continuous data and the label is discrete data. The joint density function $f(x_\tau^i, y_i)$ can be calculated by using two-dimensional Gaussian kernel estimate:

$$f(x_\tau^i, y_i) = \frac{1}{2\pi n} \sum_{k=1}^{n} exp(-\frac{D_k(x_\tau^i, y_i)}{2})$$

Where $D_k(x_\tau^i, y_i) = \sqrt{(x_\tau^i - x_\tau^k)^2 + d(y_k, y_i)}$.

This improved mutual information measured the relationship between features and labels. If the feature and the label have high co-relationship, the weight of the isomiR should be large. It implies that this isomiR is more important for the breast cancer subtype classification.

4.2.4 Identification of isomiR biomarkers that classify breast cancer subtypes

A few key isomiRs, which have the highest weights, can distinguish between the different subtypes of breast cancer. These key isomiRs can then be used as breast cancer biomarkers, and they can be identified through these processes: sorting isomiRs by using their weights from large to small, then using the different top N isomiRs to evaluate the performance in the classification of breast cancer subtypes. The performance of this type of breast cancer classification will be raised with the increasing number of selected isomiRs. If the performance of classification by using top N isomiRs is not significantly raised compared to the performance by using top $N+1$ isomiRs, it implies that these N isomiRs are key isomiRs and can be viewed as biomarkers.

In this chapter, the SVM (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg et al. 2011) classifier was applied to classify different subtypes of breast cancer. According to Table 4.1, different subtypes of breast cancer have variable numbers of patients. Around 68% of breast cancer patients are ERα+HER2-, while nearly 4.4% of breast cancer patients are ERα-HER2+. This dataset is an imbalanced dataset and the SMOTE method was used to balance the data (Chawla, Bowyer, Hall & Kegelmeyer 2002). The receiver operation characteristic (ROC) curve is very popular to judge the discrimination ability of various statistical methods (Hanley & McNeil 1982). The area under ROC curve (AUC) measures the performance of the classifier (Ferri, Hernández-Orallo & Flach 2011). Since this research is a multiclasses learning, macro-AUC of ROC was used to validate the performance of the classification (Zhang & Zhou 2014). Further, 5-fold cross-validation was

applied to evaluate the results.

4.3 Results and Discussion

4.3.1 Characterization of isomiRs identified in different subtypes of breast cancer

In this study, 20134 different isomiRs were identified in the small RNA sequencing results of 698 breast cancer patients. However, most of the isomiRs were lowly expressed. Thus, we removed the lowly expressed isomiRs by using the null hypothesis method that was described in the subsection 4.2.2. Finally, 435 isomiRs, whose total expression level was larger than the significance threshold, were viewed as highly expressed isomiRs. Among these highly expressed isomiR, 169 isomiRs were $5'$ variant isomiRs and 266 isomiRs were $3'$ variant isomiRs. These isomiRs are derived from 169 wild type miRNAs. The distribution expression of these isomiRs and their miRNAs across different breast cancer subtypes are shown in Figure 4.3 and Figure 4.4. In Figure 4.3, only the total expression level of the isomiRs, of which one nucleotide is added at $3'$ position, is larger than the expression level of wild type miRNA. While the expression level of the other $3'$ isomiRs is lowly expressed compare with wild type miRNAs. In Figure 4.4, the isomiR, which trimmed one nucleotide at the $5'$ position, has a similar expression level to the wild type miRNA. These two isomiRs (which added one nucleotide at $3'$ position and trimmed one nucleotides at the $5'$ position) may play vital roles in the breast cancer subtypes. Individual pre-miRNA may produce many different kinds of isomiRs and the expression level of isomiRs maybe higher than its wild type miRNA. Figure 4.5 displays the expression level of miRNA has-let-7d and its isomiRs across different breast cancer subtypes. We found that different kinds of isomiRs are produced during the miRNA maturation processes. Further, the expression level of isomiRs may be higher than the corresponding wild type miRNA.

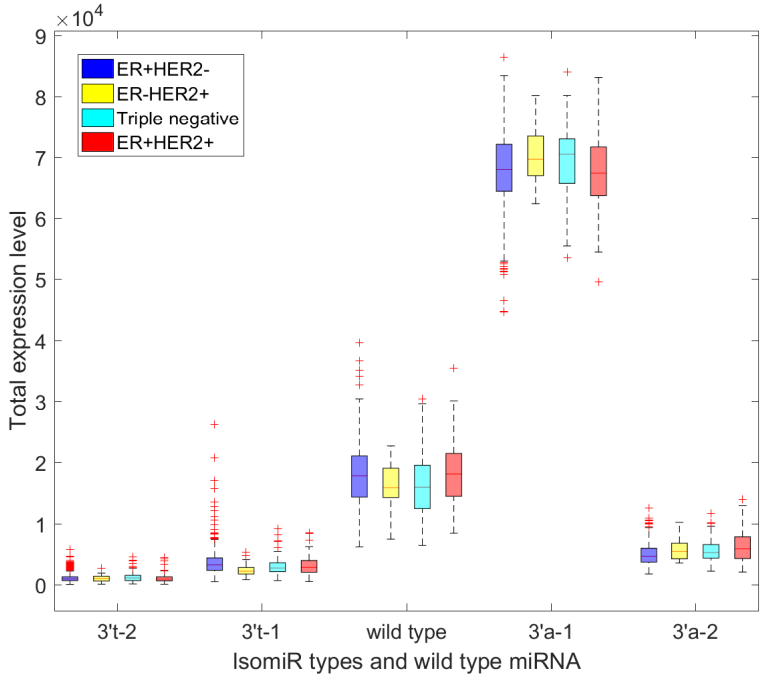

Figure 4.3: **The distributions of $3'$ isomiR and their wild type miRNAs across different breast cancer subtypes.** The x-axis is the variant symbol. The variant symbol is divided into two parts by the sign $(-)$. The left part of the sign $(-)$ is the variate type at $3'$ position. The right part of sign $(-)$ is the number of nucleotide added or trimmed at the $3'$ position.

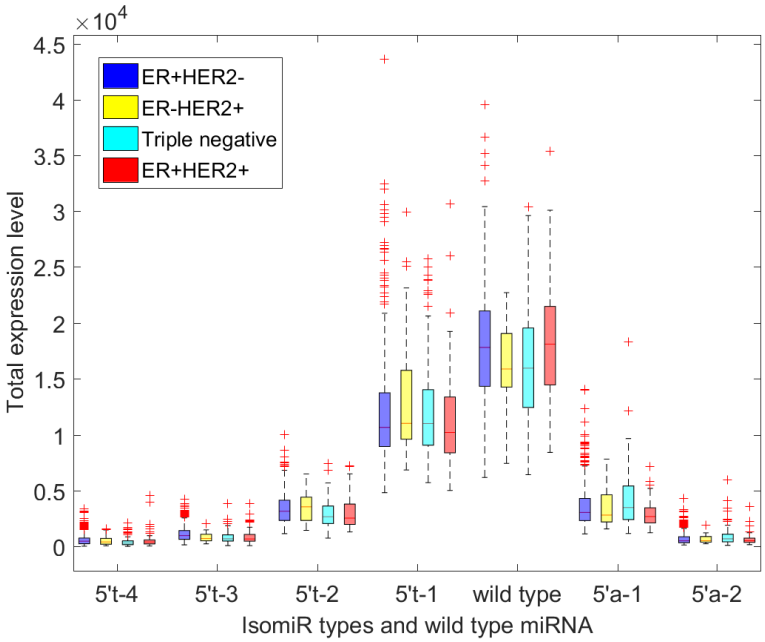

Figure 4.4: **The distributions of 5′ isomiR and their wild type miRNAs across different breast cancer subtypes.** The Y-axis is the total expression level of isomiR (or wild type miRNA). The x-axis is the variant symbol. The variant symbol is divided into two parts by the sign (−). The left part of the sign (−) is the variate type at 5′ position. The right part of sign (−) is the number of nucleotide added or trimmed at the 5′ position.

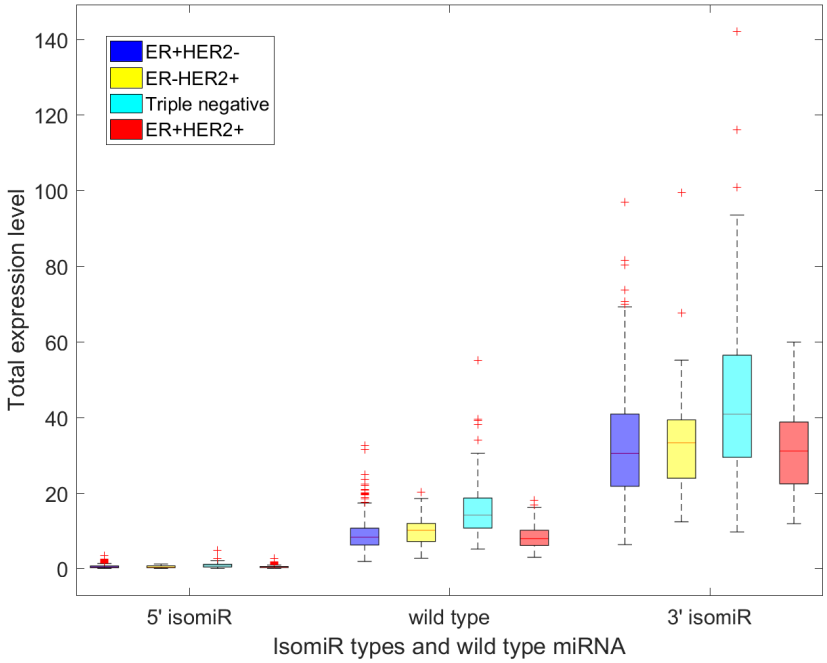

Figure 4.5: **The distributions of miRNA has-let-7d-5p and its isomiRs across different breast cancer subtypes.** The 3′ (5′) isomiR could have different lengths. The total expression level of 3′ (5′) isomiR is the sum of the expression level of different length of 3′ (5′) isomiR.

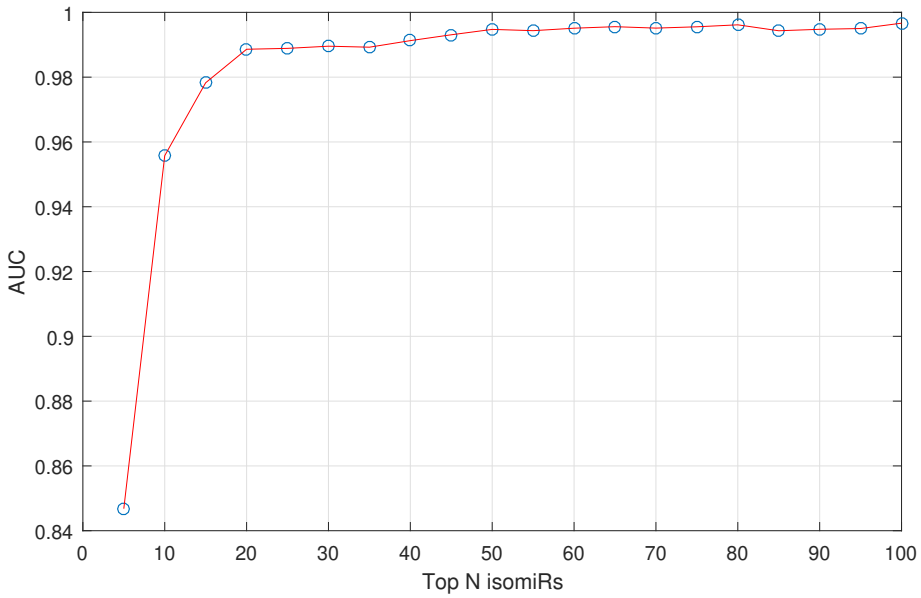

Figure 4.6: **The performance of classification by using different number of isomiR.** The x-axis is the number of isomiRs that are used to classify the breast cancer subtype. The Y-axis is the performance of the classification.

4.3.2 Identification of isomiRs that classify breast cancer subtypes

After the characterization of isomiRs in breast cancer, we calculated the weight of these isomiRs by using improved mutual information. Finally, we selected different numbers of isomiRs to compute their performance in the classification of breast cancer subtypes. The results and the Python source code of our algorithm can be downloaded from the website `https://github.com/ChaowangLan/isomiRbreastsubtype`.

According to Figure 4.6, with the increasing number of isomiRs selected, the performance of the classification was improved. However, when the number of isomiRs was more than 20, the performance of the classification plateaued. Therefore, the number of key isomiRs for breast cancer subtype classification was 20. These 20 isomiRs are viewed as breast cancer subtype biomarkers. These isomiRs and their weights are listed in the second and third column of Table 4.2.

Among the isomiRs that faithfully characterize breast cancer subtypes, 7 isomiRs were identified as 5′ variant isomiRs and the other isomiRs were identified as 3′ variant isomiR. Most of these isomiRs were highly expressed compared to their corresponding wild type miRNAs. We calculated the ratio of the expression levels of these isomiRs and their corresponding wild type miRNAs in different subtypes of breast cancer. These ratios are listed in Table 4.2. If the expression level of an isomiR was larger than the expression level of its corresponding wild type miRNA, the ratio was larger than 1. Among these 20 isomiRs, only hsa-mir-28-3p|3′a-2 and hsa-mir-22-3p|5′t-1 were lowly expressed compare to their corresponding wild type miRNAs, the other isomiRs were more abundant. These results denote that many of these isomiR biomarkers are more highly expressed compared to their corresponding wild type miRNAs.

Table 4.2: The 20 isomiR biomarkers, their weights, and their ratios

Rank	IsomiR name	Weight	Ratios			
			ERα+HER2-	ERα-HER2+	Triple negative	ERα+HER2+
1	hsa-mir-106b-5p|5′a-1	$1.86 * 10^{-3}$	19.53	25.56	37.89	16.27
2	hsa-mir-28-3p|3′a-2	$1.57 * 10^{-3}$	0.51	0.65	0.87	0.53
3	hsa-mir-93-5p|3′a-1	$1.46 * 10^{-3}$	128.75	161.48	260.60	126.86
4	hsa-mir-106b-3p|5′t-1	$1.45 * 10^{-3}$	2598.00	3570.00	6163.00	2890.00
5	hsa-mir-106b-3p|3′a-1	$1.37 * 10^{-3}$	2702.00	3643.00	6395.00	2987.00
6	hsa-mir-17-3p|3′a-1	$1.37 * 10^{-3}$	1233.00	1601.00	3776.00	1251.00
7	hsa-mir-197-3p|3′a-1	$1.19 * 10^{-3}$	6.18	10.81	12.87	6.15
8	hsa-mir-92a-1-3p|5′t-1	$1.14 * 10^{-3}$	1.80	2.26	3.47	1.75
9	hsa-mir-146b-5p|3′a-1	$1.13 * 10^{-3}$	5.38	7.46	9.93	6.50
10	hsa-mir-210-3p|5′a-1	$1.12 * 10^{-3}$	15.69	40.67	37.39	20.70
11	hsa-mir-146b-5p|3′a-2	$1.07 * 10^{-3}$	11.12	15.06	19.60	14.63
12	hsa-let-7i-5p|3′a-1	$1.03 * 10^{-3}$	1.03	1.46	1.72	1.10
13	hsa-mir-210-3p|3′a-1	$1.03 * 10^{-3}$	206.94	513.97	497.48	272.43
14	hsa-mir-106b-5p|3′a-1	$9.97 * 10^{-4}$	46.36	56.38	85.98	39.33
15	hsa-mir-532-5p|3′a-1	$9.60 * 10^{-4}$	11.98	21.97	17.88	13.88
16	hsa-mir-93-5p|3′t-2	$9.26 * 10^{-4}$	6.85	7.89	13.59	6.31
17	hsa-let-7d-5p|3′a-1	$8.80 * 10^{-4}$	2.96	4.32	5.49	3.15
18	hsa-mir-27a-3p|5′t-1	$8.51 * 10^{-4}$	62.11	104.63	106.78	59.38
19	hsa-mir-22-3p|5′t-1	$8.45 * 10^{-4}$	0.04	0.05	0.05	0.04
20	hsa-mir-93-5p|5′t-1	$8.45 * 10^{-4}$	1.80	2.02	3.18	1.58

4.3.3 Comparing the performance of improved mutual information to other feature selection methods

Many methods for feature selection have been developed. However, not all these methods are suitable for the dataset where feature is continuous data and label is discrete data. In this chapter, we focused on comparing the performance of our novel method with two popular feature selection methods. One is the Fisher score and the other is the Hellinger distance-based method (Gu et al. 2011, Cieslak & Chawla 2008). The AUCs, calculated using the three different methods and using different numbers of selected isomiRs, are presented in Figure 4.7. According to this figure, the AUCs show an increase with the raising of the number of selected isomiRs. However, if the number of selected isomiRs is larger than 30, the AUCs, which are calculated by these three methods, do not have significance changes. It indicates that the number of key isomiRs, by using these three methods, are lower or equal than 30. However, different methods identify different numbers of key isomiRs for breast cancer classification. The Fisher method identified 30 key isomiRs while the Hellinger method found 25 key isomiRs. Although fewer key isomiRs were discovered using Hellinger method, the AUC was found to be slightly lower than the Fisher method. Our method identified 20 key isomiRs that classify breast cancers, which is the lowest number of key isomiRs compared to the other methods mentioned, and the AUC was similar (nearly equal) to the Fisher method. It implied that our method can use fewer isomiRs as biomarkers to classify different subtypes of breast cancer

Since we applied all samples to detect the biomarker and SOMTE method to balance the data, the results are over-optimistic. However, all methods compared under the same condition. Tshe conclusion drawn from the comparison may not be affected.

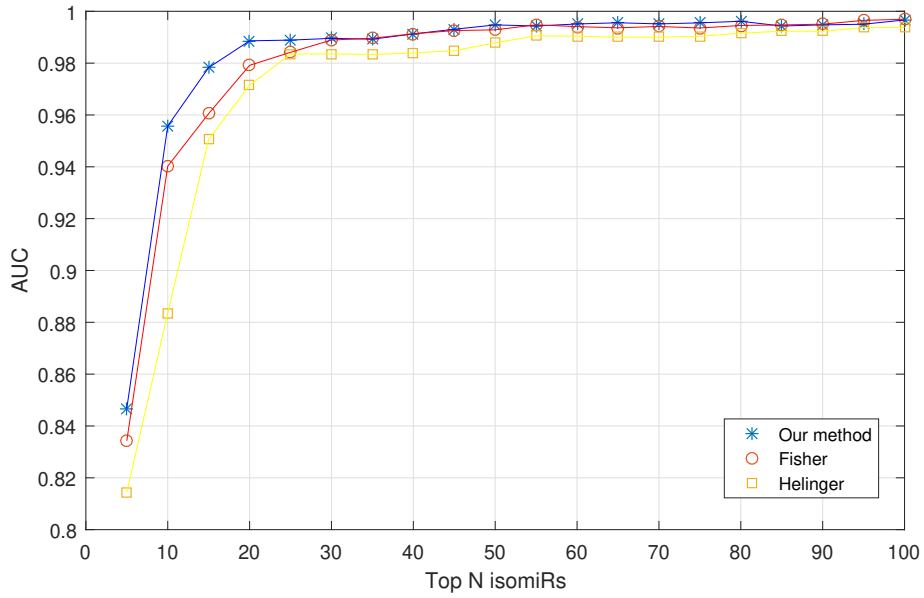

Figure 4.7: **Comparison of our isomiR panel based novel method classification with other feature selection methods.** The x-axis is the number of isomiRs that are used to classify the breast cancer subtype. Y-axis is the performance of the classification. The higher the AUC, the better the classification. Legend: the star represents the novel method described in this chapter. The circle, and cross sign are the Filter method, and the Hellinger method, respectively.

4.3.4 IsomiRs are superior biomarkers compared to protein coding gene expression-based approaches for the classification of different subtypes of breast cancer

Over the past decade, many studies have found that protein coding gene expression data can be used to classify breast cancer subtypes. For instance, Van and colleagues proposed that a 70-genes' expression profile can use for identifying different subtypes of breast cancer (Van De Vijver, He, Van't Veer,

Dai, Hart, Voskuil, Schreiber, Peterse, Roberts, Marton et al. 2002), Parker and colleagues defined the PAM50 genes, which are the most famous biomarkers for breast cancer subtype classification (Parker et al. 2009), and Neve and colleagues also applied genes expression data for the classification of different subtypes of breast cancer (Neve, Chin, Fridlyand, Yeh, Baehner, Fevr, Clark, Bayani, Coppe, Tong et al. 2006). Their research indicated that differentially expressed mRNAs can be used as breast cancer subtype biomarkers.

The TCGA database also provides the expression level of mRNAs in different subtype of breast cancer. Therefore, we can calculate if isomiR or gene expression-based profiling performs better for breast cancer subtype classification. Figure 4.8 presents the AUC by using isomiRs and mRNAs. According to the comparison in Figure 4.8, the performance of breast cancer subtype classification using the expression of five mRNAs is very high (the AUC is near to 0.89). Direct comparison of isomiRs and mRNA (gene expression) clearly show that fewer isomiRs are needed classify different subtypes of breast cancer compared to the number of mRNA (genes). With the increasing number of mRNA, the difference between the two classification methods is comparable, i.e. when the number of mRNA (gene classification) is more than 35, the AUC does not show any significant difference. Therefore, the number of key mRNA is 35, in comparison with isomiR, the key number is 20, showing fewer isomiRs can classify different subtypes of breast cancer. This experiment indicates that isomiRs also can be used as biomarkers for the classification of breast cancer subtypes and, importantly, fewer isomiRs can be used to classify different subtypes of breast cancer. These results strongly suggest that isomiRs are more suitable biomarkers compared to biomarkers based on protein coding gene expression profiles.

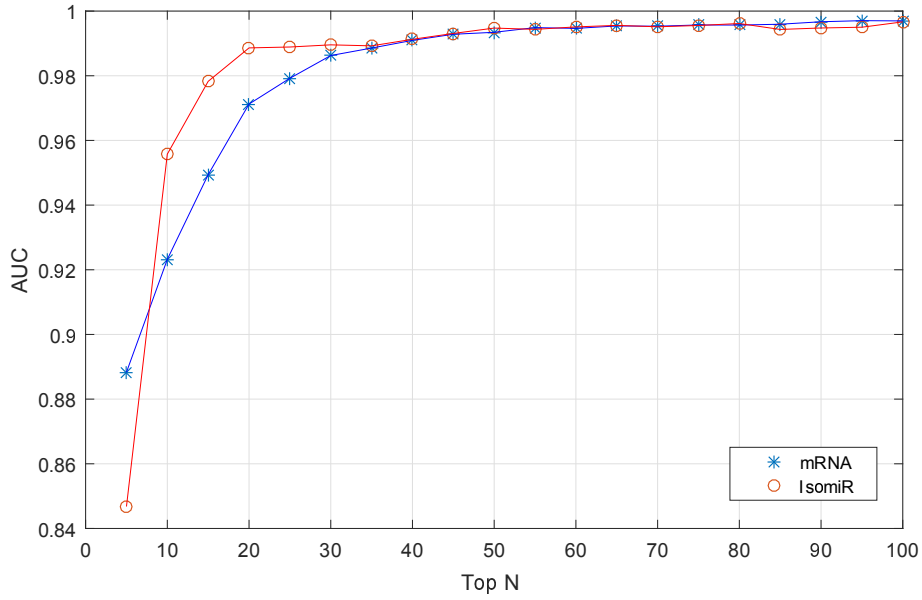

Figure 4.8: **Comparison of isomiR and gene classification for breast cancer subtyping.** The x-axis is the number of isomiR that are used to classify the breast cancer subtype. Y-axis is the performance of the classification. The higher the AUC, the better the classification. Legend: the star and circle present the classification using mRNA and isomiR, respectively.

4.3.5 IsomiRs may play important regulatory roles in different subtypes of breast cancer

Many studies have found that different categories of isomiRs have different functions in regulating biological processes. For example, 3′ isomiRs have low 3′ untranslated region stability and therefore, loose regulation of mRNAs (Burroughs, Ando, de Hoon, Tomaru, Nishibu, Ukekawa, Funakoshi, Kurokawa, Suzuki, Hayashizaki et al. 2010). 5′ isomiRs have slightly altered seed sequences compared to the corresponding wild type miRNAs; therefore, besides weakening the regulatory effect of the wild type miRNAs

they can target mRNAs that are significantly different from the wild type miRNA targeted transcriptome (Tan, Chan, Molnar, Sarkar, Alexieva, Isa, Robinson, Zhang, Ellis, Langford et al. 2014). Based on sequence similarities it is possible to predict potential mRNAs that are regulated by certain miRNAs (Agarwal et al. 2015, Betel, Koppal, Agius, Sander & Leslie 2010)and therefore, biological pathways that are influenced by miRNAs and their isomiRs. The elevated levels of isomiRs compared to their corresponding wild type miRNAs can also be used to predict changes in the regulation of gene expression in breast cancers that may well provide insight into the molecular mechanisms leading to breast cancer. We predicted that the presence of abundant 3′ isomiR develop weakened regulatory effects on transcripts that are regulated by the corresponding wild type miRNAs. Thus, mRNAs that are regulated by the wild type miRNAs should show elevated expression levels when the expression level of isomiRs were significantly elevated. These targets that may be affected by the accumulation of 3′ isomiRs can be obtained from the miRWalker2.0 website (`http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/holistic.html`). To predict potential targets for abundant 5′ isomiRs with modified seed sequence we used the miRDB website (`http://www.mirdb.org/`). In order to obtain the most likely targeted mRNAs, the score of the prediction target gene should be higher than 95 (the maximum score is 100).

After predicting the sets of potential mRNAs that are affected by the elevated miRNA isomiR levels, we wanted to characterize what molecular pathways may be changed in breast cancers. Enrichr (`http://amp.pharm.mssm.edu/Enrichr/`), which is a gene enrichment analysis web server, was applied to find out the Kyoto The KEGG pathway of these target genes (Kuleshov et al. 2016). 104 KEGG pathways were identified as significant pathways (the p-value of these pathways were lower than 0.05) from this website. In this chapter, we selected five pathways have been computed to be significantly affected by isomiRs to further discuss the function of the isomiRs in breast cancer. These 5 KEGG pathways are

presented in Table 4.3.

Table 4.3: Five KEGG pathways which are relative to breast cancer progresses and subtype classification

KEGG name	P-value	Number of gene
Pathways in cancer	$5.01 * 10^{-11}$	96
p53 signalling pathway	$1.29 * 10^{-6}$	24
MAPK signalling pathway	$1.20 * 10^{-5}$	56
Insulin signalling pathway	$3.16 * 10^{-3}$	29
Estrogen signalling pathway	$1.79 * 10^{-2}$	20

The first two KEGG pathways in Table 4.3 are very important for analysis of breast cancer outcome (Gasco, Shami & Crook 2002). This data suggests that isomiRs also play a vital role in breast cancer development. The clinical breast cancer classification is based on the hormone receptor status, some of these KEGG pathways are involved in regulating the hormone receptor status. For example, Neve's research highlights that up-regulation of genes involved in insulin/MAPK signaling predicts response to Herceptin (Neve et al. 2006). It implies that these two signaling pathways regulate the Herceptin status. According to the third and fourth line of Table 4.3, isomiRs were shown to influence 56 and 29 genes in MAPK and insulin signal pathways, respectively. Therefore, isomiRs could affect the Herceptin statue through these two pathways and lead to the development of different subtypes of breast cancer. We also identified the estrogen signalling pathway represented by 20 genes that is potentially affected by the isomiRs (Table 4.3). It implies that isomiRs could affect the expression of these genes to influence the estrogen receptor status. Above all, isomiRs may regulate the hormone receptor status via different KEGG pathways and therefore, affecting different breast cancer subtypes.

4.3.6 Assessing the role of individual isomiRs in the regulation of breast cancer specific pathways

Next, we focused on the further analysis of six isomiRs that have potential to characterise/classify breast cancer subtypes. Dressman and colleagues pointed out that there are 18 genes that may delineate the role of estrogen receptor in breast cancer (Dressman, Walz, Lavedan, Barnes, Buchholtz, Kwon, Ellis & Polymeropoulos 2001). Transforming growth factor-beta type III receptor (TGFBR3) and serpin family A member 3 (SERPINA3) are two of these genes. Accordingly, the miRwalker 2.0 database, TGFBR3 is one of the potential target genes of hsa-let-7i-5p and SERPINA3 is the target gene of hsa-mir-197-3p. However, if one nucleotide is added to the 3′ position of these two miRNAs, then there is a possibility that these isomiRs cannot efficiently bind to the gene TGFBR3 and SERPINA3, respectively. This is because the longer sequence alters the stability of the miRNA and cannot inhibit the expression level of its target gene. Alternatively, 3′ isomiRs could be a sign of actively turned over miRNA that may have weakened regulatory functions. In the ER negative breast cancer tumors, most hsa-let-7i-5p wild-type miRNAs are altered to isomiRs hsa-let-7i-5p|3′a-1 and hsa-mir-197-3p miRNAs are changed to isomiRs hsa-mir-197-3p|3′a-1. Therefore, they are predicted to have a weakened affect to inhibit the expression level of TGFBR3 and SERPINA3 and these two genes are highly expressed in the ER negative breast cancer subtype. Similarly, these two genes are lowly expressed in the ER positive breast cancer subtype. Table 4.4 displays the average expression level of these two isomiRs in different subtypes of breast cancer. According to the average expression levels of isomiRs hsa-let-7i-5p|3′a-1 and hsa-mir-197-3p|3′a-1 in different subtypes of breast cancer, these two isomiRs are highly expressed in the ER negative tumors (ERα-Her2+ and triple negative breast cancer subtype) and lowly expressed in ER positive tumors (ERα+Her2- and ERα+HER2+ breast cancer subtypes).

The 5′ variant isomiRs have distinct seed sequences compared to the corresponding wild type miRNAs; therefore, they may regulate a novel set

Table 4.4: The average expression level of isomiRs and miRNA in each breast cancer subtype.

IsomiR/miRNA name	Breast cancer subtype			
	ERα+HER2-	ERα-HER2+	Triple negative	ERα+HER2+
hsa-let-7i-5p|3′a-1	10.43	13.57	17.09	10.55
hsa-mir-197-3p|3′a-1	26.27	36.33	61.66	25.75

of transcripts relative to the wild type miRNAs. Table 4.5 presents the predicted target genes of some 5′ variant isomiRs by miRDB database. The dysregulation of estrogen signalling pathway leads to ER positive breast cancer and therefore, the genes involved in this pathway may be the most attractive target for ER positive breast cancer treatment In the first line of Table 4.5, hsa-miR-93-5p|5′t-1 may bind to gene SHC4. SHC4 is one of the gene involved in estrogen signalling pathway. The result implies that hsa-miR-93-5p|5′t-1 may regulate SHC4 and dysregulate the estrogen signalling pathway. Furthermore, three 5′ variant isomiRs, which exhibited in the last three lines of Table 4.5, potentially bind to MAPK14, MAPK8, and RAP1B, respectively. These three genes are the part of the MAPK signaling pathway, which affects the Herceptin status. These results revealed that 5′ variant isomiRs may bind to genes that regulate hormone receptor status and therefore, lead to different breast cancer subtypes.

Table 4.5: 5′ variant isomiRs' predicted target genes

isomiR	Predicted target mRNA	Score
hsa-miR-93-5p|5′t-1	SHC4	95
hsa-mir-27a-3p|5′t-1	MAPK14	97
hsa-miR-92a-1-3p|5′t-1	MAPK8	97
hsa-mir-106b-3p|5′t-1	RAP1B	95

4.4 Conclusion

In this chapter, we propose a novel method for identifying isomiR biomarkers for breast cancer subtyping from small RNA sequencing data. We first removed the lowly expressed isomiRs from the data sets. Then we calculated the weight of the isomiR by utilizing the improved mutual information. The improved mutual information measured the co-relationship between the expression level of isomiRs and breast cancer subtypes. The higher the co-relationship between isomiR's expression and breast cancer subtypes, the more important the isomiR for breast cancer subtype classification. Further, this improved mutual information can be applied to the data set that the feature is continuous data and the label is discrete data. While the traditional mutual information cannot. Finally, the SVM classifier was applied to find specific isomiR biomarkers for classification of the different breast cancer subtypes. This method, proved to be more effective and efficient in identifying fewer key isomiRs needed for breast cancer subtyping in comparison to the Fisher and Hellinger methods. Importantly, in this study, we describe the enhanced identification of isomiR biomarkers for classification of breast cancer subtypes and, in addition, isomiRs were found to be superior biomarkers compared to classification based on mRNA gene expression for this type of classification. Further, applying this improved methodology, we identified individual isomiRs that may be key in the regulation of specific breast cancer pathways. There is great potential in exploiting these novel isomiR regulatory mechanisms as drug-targets for more personalized subtype breast cancer specific therapies.

Discovery of unique biomarkers in different breast cancer subtype is a challenge in research, especially since the regulation mechanism of different breast cancer subtypes is not yet fully understood. Our research provides a new way to explore the mechanism of breast cancer subtypes.

Chapter 5

Identification of Glioma Subtypes Biomarkers through Information Gain

5.1 Introduction

Glioma is the most common primary central nervous system tumor in adults (Johnson, Dickerson, Connolly & Gephart 2018). The glioma subtype classification is based on histopathology and is the foundation to patient management and clinical investigations (Louis, Ohgaki, Wiestler, Cavenee, Burger, Jouvet, Scheithauer & Kleihues 2007, Coons, Johnson, Scheithauer, Yates & Pearl 1997). However, the histopathology-based classification of glioma subtypes has an issue: it is subjected to significant interobserver variability. Therefore, making the correct glioma subtype diagnosis is very challenging (Chen, Smith-Cohn, Cohen & Colman 2017). The glioma subtype in fact could be classified by some molecular biomarkers (Aldape, Zadeh, Mansouri, Reifenberger & von Deimling 2015). However, little is known about the molecular biomarker in classifying different glioma subtypes. Identifying the biomarker of glioma cancer subtype is very important for making accurate diagnosis and understanding the molecular

mechanism of different glioma cancer subtypes (Trabelsi, Chabchoub, Ksira, Karmeni, Mama, Kanoun, Burford, Jury, Mackay, Popov et al. 2017).

In this chapter, we study the role of isomiRs in classifying different glioma subtypes from RNA-seq data. 75, 417 isomiRs are identified from the RNA-seq data. However, most of the isomiRs are lowly expressed and have significantly negative influence on discovering the glioma subtype biomarker. Thus, the hypothesis method is applied to remove these lowly expressed isomiRs. Then, the information gain is used to identify the isomiR which could classify different glioma subtypes. Since the expression level of the isomiR is continuous data, the expression level of isomiR should transform to discrete data before calculating the information gain of the isomiR. Some of our prediction were further validated in cell lines using molecular biology approaches. Furthermore, using high throughput immunochemistry assays we showed that isomiR based prediction could reveal novel molecular markers to identify glioma subtypes.

5.2 Definition and Materials

The 16 glioma patient tissues were obtained from The Tumour Bank, The Children's hospital at Westmead. These glioma patient tissues are classified three categories: astrocytoma, ependymoma, and glioma cancer cell. The 16 small RNA libraries are prepared using NEBNext Multiplex Small RNA Library Pre Set 1/2 for Illumina and to be run on 1 HiSeq lane using 50bp single end reads. Since the availability of a spare lane of the sequencing machine, we got 2 lanes data for each patient tissue.

The isomiR can be divided into 6 variation types: 1.) $5'$ trimming (deleted nucleotides at the $5'$ end of the wild type miRNA); 2.) $5'$ addition (additional nucleotides at the $5'$ end of the wild type miRNA), 3.) polymorphic (nucleotide changes from wild type miRNA), 4.) $3'$ trimming (deleted nucleotides at the $3'$ end of the wild type miRNA), 5.) $3'$ templated addition (added nucleotides to $3'$ end of the wild type miRNA

Table 5.1: The type symbol and the variation form detail of isomiR

Variation Type	*typesyb*	*Detail*
3′ trimming	3′t	The length of trimmed sequence
3′ untemplated addition	3′a	The sequence added in the 3′ dicing site
3′ templated addition	3′g	The length of sequence mapped to the miRNA precursor
5′ trimming	5′t	The length of trimmed sequence
5′ addition	5′a	The sequence added in the 5′ dicing site
Polymorphic	ms	Position and original nucleotide/changed nucleotide

and the added nucleotides can be mapped to the miRNA precursor), and 6.) 3′ untemplated addition (added nucleotides to 3′ end of the wild type miRNA and the added nucleotides cannot be mapped to the miRNA precursor).

To discriminate different forms of isomiR, we use the symbol $miRNAname|typesyb-detail$. Where the $miRNAname$ is the wild type miRNA name that that produces the miRNA isoform. The *typesyb* is type symbol that presents one of the six types of isomiRs (showed in Table e 5.1) and the *Detail* describes the length of the addition or trimming, or the additional sequence at the 3 end, or the changed nucleotide. For instance, if an isomiR is produced from hsa-let-7c precursor with two nucleotides trimmed at the 3′ end, this isomiR is described by the symbol hsa-let-7c|3′t-2. Similarly, hsa-miR-4510|ms-6G/U means that this isomiR is generated from hsa-miR-4510 precursor with a nucleotide substitution (G change to U) at the sixth nucleotide.

5.3 Methods

5.3.1 Threshold selection

Given n total expression levels of isomiRs $T = \{t_1, t_2, \ldots, t_n\}$ which follow exponential distribution. The small total expression level of isomiR suggests that this isomiR has limit in regulating the biological process while the large total expression level of isomiR may have significance in regulating the biological process. The threshold θ is applied to distinguish the lowly and highly expressed isomiR. Here, the problem is to reject the null hypothesis. This null hypothesis is that the total expression level of isomiR is small, that is, it implies that this isomiR dysfunction the biological process. If the total expression level of isomiR is very high, the null hypothesis can be rejected it implies that this isomiR involved in regulating the biological process. The threshold is defined as:

$$\theta = \frac{n * \overline{T}}{\chi^2_{\alpha/2}(n)}$$

Where $\overline{T}$ is the mean of isomiRs' total expression level. $\chi^2_{\alpha/2}(2 * n)$ is the value of chi-square random variate with $2 * n$ degree of freedom that has probability level $\alpha/2$. In this research, the probability level α is 0.025. If the total expression level of an isomiR is larger than θ, this isomiR is viewed as highly expressed.

5.3.2 Information gain

The information gain is used to discover how many information we can obtain to classify the data. In this research, the information gain is utilized to measure the contribution of the isomiR for classifying different glioma subtypes. If the isomiR has a high information gain, this isomiR may play important role in classifying different glioma subtypes.

The information gain always used in the discrete data. However, the expression level of isomiR is continuous data. We should transform the continuous data into discrete data. The process of transformation is that

sorting the expression level of isomiR from small to large. Then finding $k-1$ cut points which divide the expression levels of the isomiR into k categories. The equal width method and equal frequent method are two classical methods for transforming continuous data into discrete data. These two methods are not suitable for our research since the distribution expression levels of isomiRs in different glioma subtypes are not equal and the number of the patient in different cancer subtypes are not equal.

In this paper, we propose a distribution-based supervise method to transform the continuous data to discrete data. For the expression level of isomiR a in n patients $D = \{d_1, d_2, \ldots, d_n\}$. Patients are classified into s cancer subtypes. We define that EX_a^{st} is a set of expression levels of isomiR a in the cancer subtype st. The median expression level of the isomiR in subtype group EX_a^{st} is M^{st}. Sorting all the subtype groups by using the median expression level of isomiR from small to large. The sorted subtype group list is presented by $\{EX_a^1, EX_a^2, \ldots, EX_a^s\}$. For any two adjacency glioma subtype group EX_a^i and EX_a^{i+1}. We find all the expression levels of isomiR a in EX_a^i which is smaller than M^{i+1} and V_{large}^i is denoted as the largest value in these expression levels. Similarly, we find all the expression level of isomiR in EX_a^{i+1} which is larger than M^i and V_{small}^i is denoted as the smallest value in these expression levels. The cut point between subtype group is EX_a^i and EX_a^{i+1} is $cp_i = \frac{(V_{large}^i + V_{small}^i)}{2}$. Given s subtype, we can obtain $s-1$ cut points. These cut points divide the expression levels of isomiR into s categories $C(a) = \{c_a^1, c_a^2, \ldots, c_a^s\}$. $D_a(j)$ is a set of patients that their expression levels of isomiR a classified in c_a^j. $|D_a(j)|$ is the number of patient in $D_a(j)$. $P(D^{st})$ and $P(D_a^{st}(j))$ are the ratios of the cancer subtype st in D and $D_a(j)$, respectively. The information gain of an isomiR a in classifying the n patients D is calculated by these equations:

$$IG(D, a) = H(D) - H(D|a) \tag{5.1}$$

$$H(D) = \sum_{i=1}^{s} P(D^i) \log(P(D^i)) \tag{5.2}$$

$$H(D|a) = \sum_{j \in C(a)} \frac{D_a(j)}{|S|} H((D_a(j)) \tag{5.3}$$

5.4 Result

5.4.1 IsomiRs are highly expressed in gliomas

Recent studies showed that in cancer the expression of isomiRs can supersede the expression of their corresponding wild type miRNAs (REF). In order to identify and quantify isomiRs in our glioma samples, we aligned sequencing reads to the miRNA reference sequences (miRbase) using Miraligner (Pantano, Estivill & Martí 2009) We use the output of this software to calculate the expression level of isomiRs. Our analysis identified 75,417 different forms of isomiRs in all glioma samples. The range of different isomiRs varies from 23,318 to 33,390 in individual gliomas (Figure 5.1). These isomiRs can be divided into 6 types. We calculate the total expression level of the 6 types of isomiR in each sample. Figure. 5.2 presents the expression level distribution of isomiR and wild type miRNA. The expression level of wild type miRNA, 3′ trimming miRNA and 3′ addition miRNA are the three most highly expressed miRNA (isomiR) in glioma patients. Specially, the total expression level of the 3′ trimming isomiR is larger than the expression level of wild type miRNA. It implies that the 3′ trimming isomiR may play important role in regulating the glioma.

The isomiR is widely existing in the glioma patients. However, the expression level of different types of isomiRs are different. Many 5′ variant isomiRs are likely to trimmed or added with one nucleotide (see in Figure 5.3). Similarly, 3′ variant isomiRs are more likely to trimmed or added with one nucleotide (shown in Figure 5.4 and 5.5). According to Figure 5.6, the nucleotide substitution could be found in all position of the wild type miRNA. However, the nucleotide substitution at the first position is most highly expressed in the polymorphic isomiR.

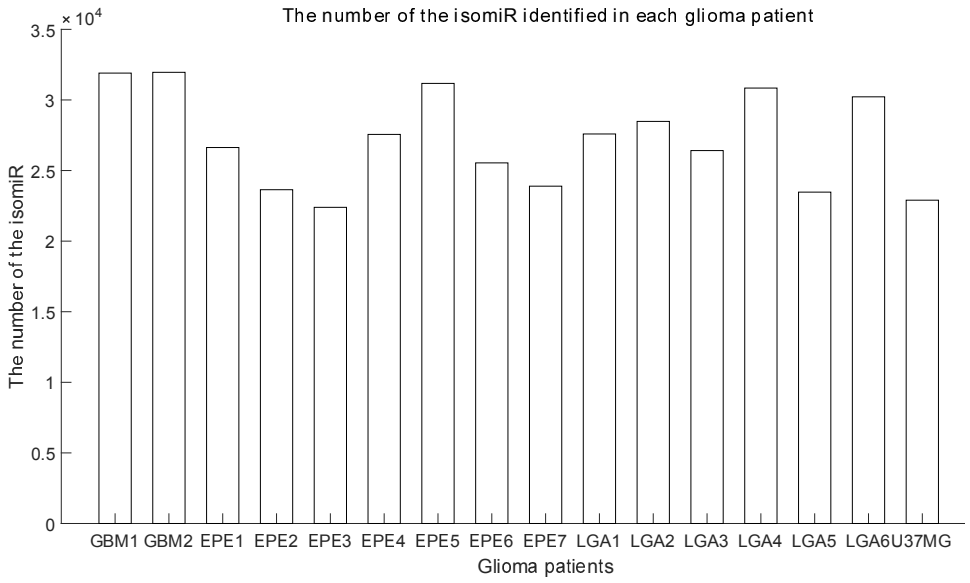

Figure 5.1: **The number of isomiR in each sample.** More than twenty thousand forms of isomiR are identified in each glioma patient. The x-axis presents the name of the label of glioma sample. GA: patients with astrocytoma subtype, EPE: patients with ependymoma subtype, and cell line is the patient with cell subtype.

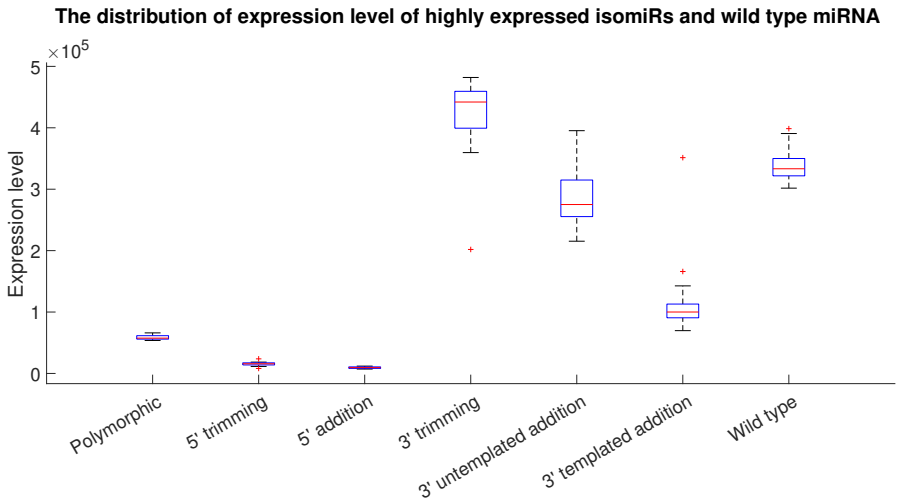

Figure 5.2: **The expression level of different types of isomiRs and miRNA.** We can found that the expression level of 3′ trimming isomiR is higher than wild type miRNA and the 3′ untemplated additional isomiR has comparable expression level to the wild type miRNA. The 3′ trimming and untemplated isomiRs may play important roles in regulating the gene pathway of glioma. The expression level of polymorphic, 5′ isomiR, and 3′ templated additional isomiR are related lowly expressed compare to other types of isomiRs and wild type miRNA.

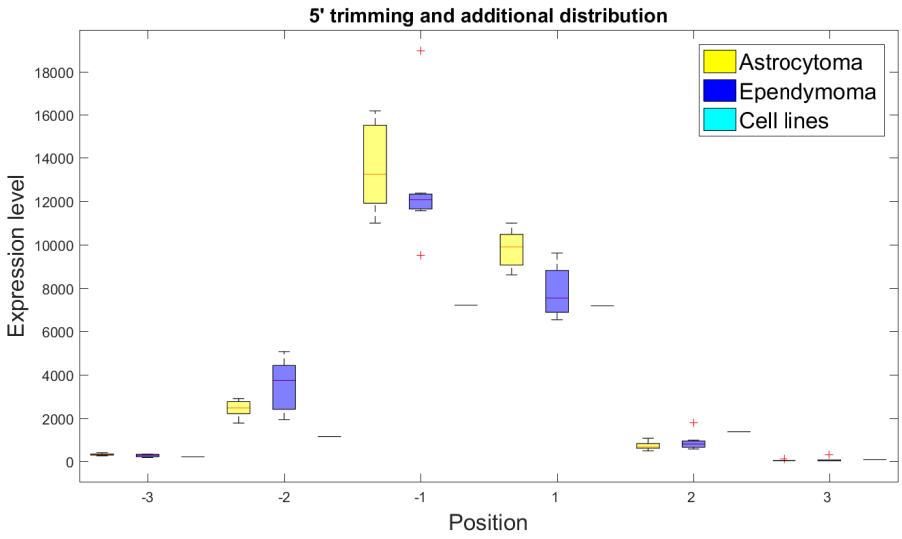

Figure 5.3: **The expression level distribution of 5′ isomiRs in different glioma subtypes.**The positive value in the x-axis means the isomiR is 5′ trimming isomiR and the negative value implies that the isomiR is 5′ added isomiR.

5.4.2 Selecting highly expressed isomiRs

Although there are large amount of isomiRs are discovered in glioma patient sample, many isomRs are lowly expressed, see in Figure 5.7. These lowly expressed isomiRs have negative influence on the result. Thus, these lowly expressed isomiRs should be removed.

An isomiR is highly expressed since its total expression level larger than a threshold θ in order for rejecting the null hypothesis. The method for calculating the threshold θ is showed in section 5.3. In this research, the threshold θ is 196.76. There are 2448 isomiRs are defined as the highly expressed isomiRs. The highly expressed isomiR account for 7.33% to 10.47% of the entire isomiR in a glioma sample. Although the number of the highly expressed isomiR is very small, the expression level of different types of isomiRs are very high.

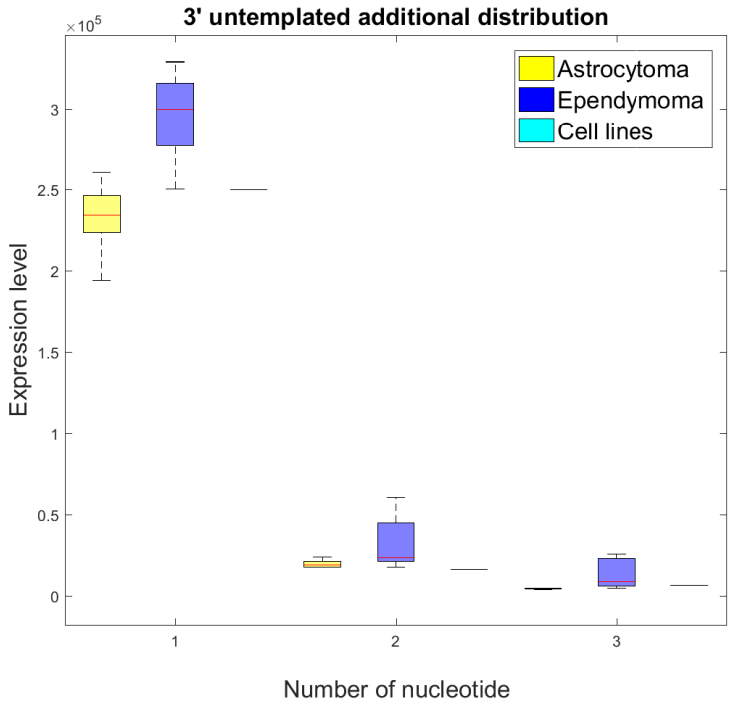

Figure 5.4: **The expression level distribution of 3′ trimming or templated additional isomiRs in different glioma subtypes.** The negative value in x-axis indicates that the isomiR is 3′ templated addition isomiR otherwise is 3′ trimming isomiR.

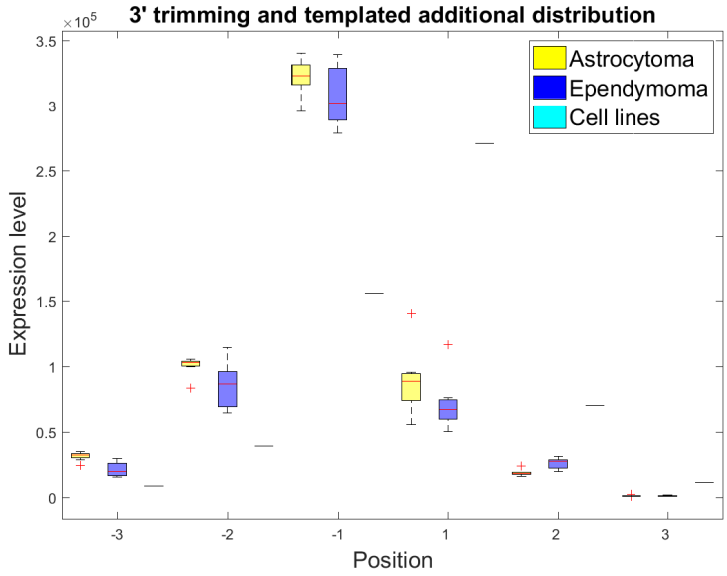

Figure 5.5: **The expression level distribution of 3′ untemplated addition isomiRs in different glioma subtypes.** The x-axis is the number of nucleotide added at the 3′ position.

5.4.3 IsomiRs could be biomarkers for classifying different glioma subtypes

The isomiR that differentially expressed in different glioma subtypes may reflect molecular differences between gliomas. In order to find out the differentially expressed isomiR, we apply the information gain. The method of calculating the information gain is showed in subsection 5.3.2. The information gain measures the importance of the isomiR in classifying different glioma subtypes. The higher the information gain, the isomiR is more likely to be a biomarker for classifying different glioma subtypes. The information gain has a maximum value, thus the isomiR that has maximum information gain is the most important for classifying different glioma subtypes. Finally, we found out 76 isomiRs that have maximum information gain. All these isomiRs are presented in Table A.1. These 76 isomiRs contain thirty-one polymorphic isomiRs, eight 3′ templated addition isomiRs,

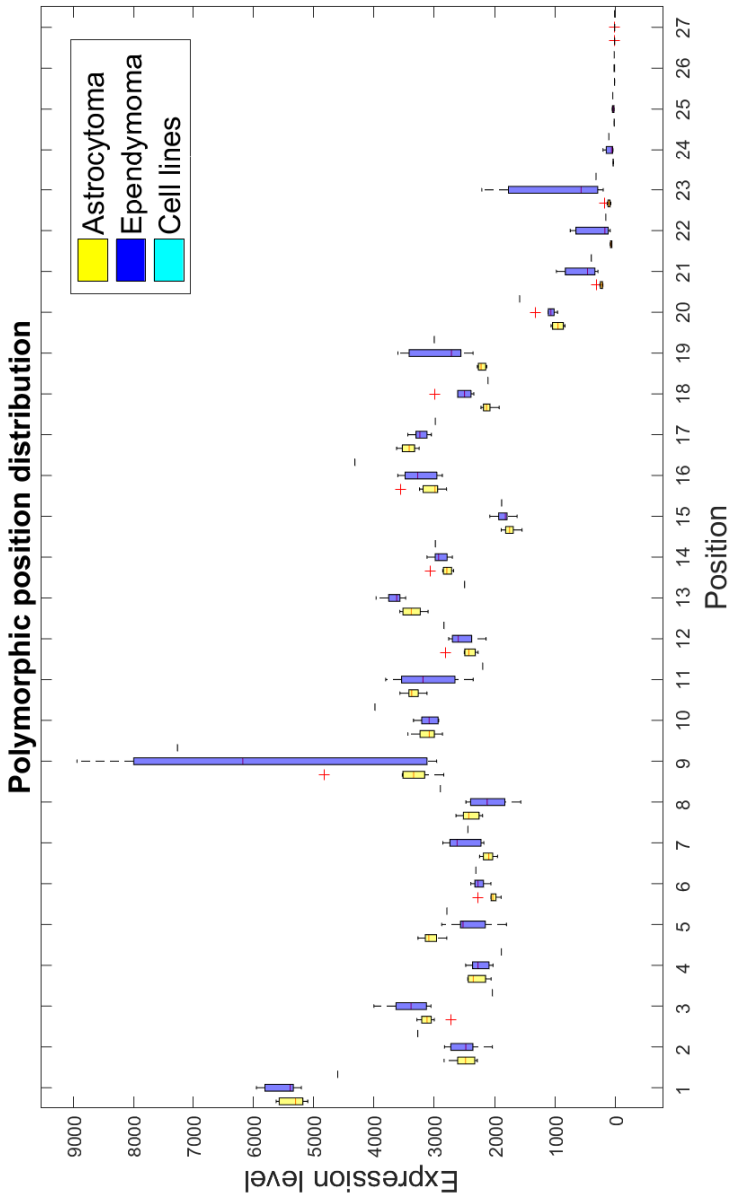

Figure 5.6: **The expression level distribution of polymorphic isomiRs in different glioma subtypes.** The y-axis is the total expression level of the isomiR. The x-axis is the nucleotide substitution position of the miRNA.

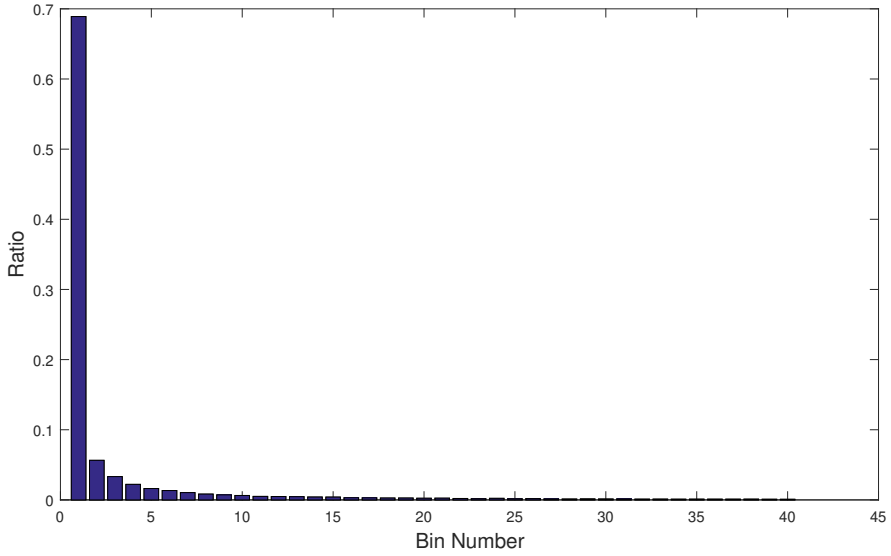

Figure 5.7: **The PDF of the total expression level of isomiRs.** The expression level of the isomiR is continuous data. A histogram of which the 'bin' of the bar graph equaled 1 was applied. X-asix is the bin number. Bin 1 is the number of the isomiR which its expression level is below 1 and bin 2 is the isomiR which its expression level is between 1 and 2. Since the total expression level of isomiR was wide ranging, this histogram proved to be very large and therefore, the complete histogram could not be displayed: the distribution of the total expression level less than 40.

eleven 3′ trimming isomiRs, nineteen 3′ untemplated addition isomiRs, five 5′ addition isomiRs, and two 5′ trimming isomiRs. Figure 5.8 presents the expression distribution of two isomiRs that have maximum information gain and we will discuss their functions in regulating the glioma subtypes in subsection 5.4.4. According to Figure 5.8, isomiR has-miR-138-5p|3′g-1 is highly expressed in cell line and lowly expressed in Ependymoma subtype. While isomiR has-miR-4510|ms-6G/U is highly expressed in Ependymoma subtype and lowly expressed in cell line subtype.

5.4.4 The role of isomiR in glioma cancer subtypes

The wet-lab experiment is applied to confirm the regulatory mechanism of the isomiR in glioma cancer subtypes. However, the wet-lab experiment is not done by myself. I give a brief description and conclusion of the wet-lab experiment in this chapter.

In the wet-lab experiment, we study the roles of two isomiRs has-miR-138-5p|3′g-1 and has-miR-4510|ms-6G/U in cancer. The has-miR-138-5p|3′g-1 is a 3′ isomiR and therefore, it weakens to regulate its two target mRNAs Cyclin D1 (*CCND1*) and Aurora kinase A (*AURKA*). The has-miR-138-5p|3′g-1 is differentially expressed in different glioma subtype. Thus, these two mRNAs also should differentially expressed in different glioma subtypes. The immunohistochemistry (IHC) approach is applied to detect whether these two mRNAs are deferentially expressed in different glioma subtypes. The results demonstrate that the expression levels of these two mRNAs are co-expression with has-miR-138-5p|3′g-1 and these two mRNAs are differentially expressed in different glioma subtype.

The isomiR has-miR-4510|ms-6G/U is a polymorphic isomiR and changed nucleotide is in the seed region. Thus, this isomiR binds to other mRNAs compare with its wild type miRNA. According to the target prediction of miRDB, this isomiR binds to *Lin28B* which regulates let-7 familly. The western blot results show that is the has-miR-4510|ms-6G/U could bind to Lin28B mRNA while the hsa-miR-4510 could not. It implies that the isomiR

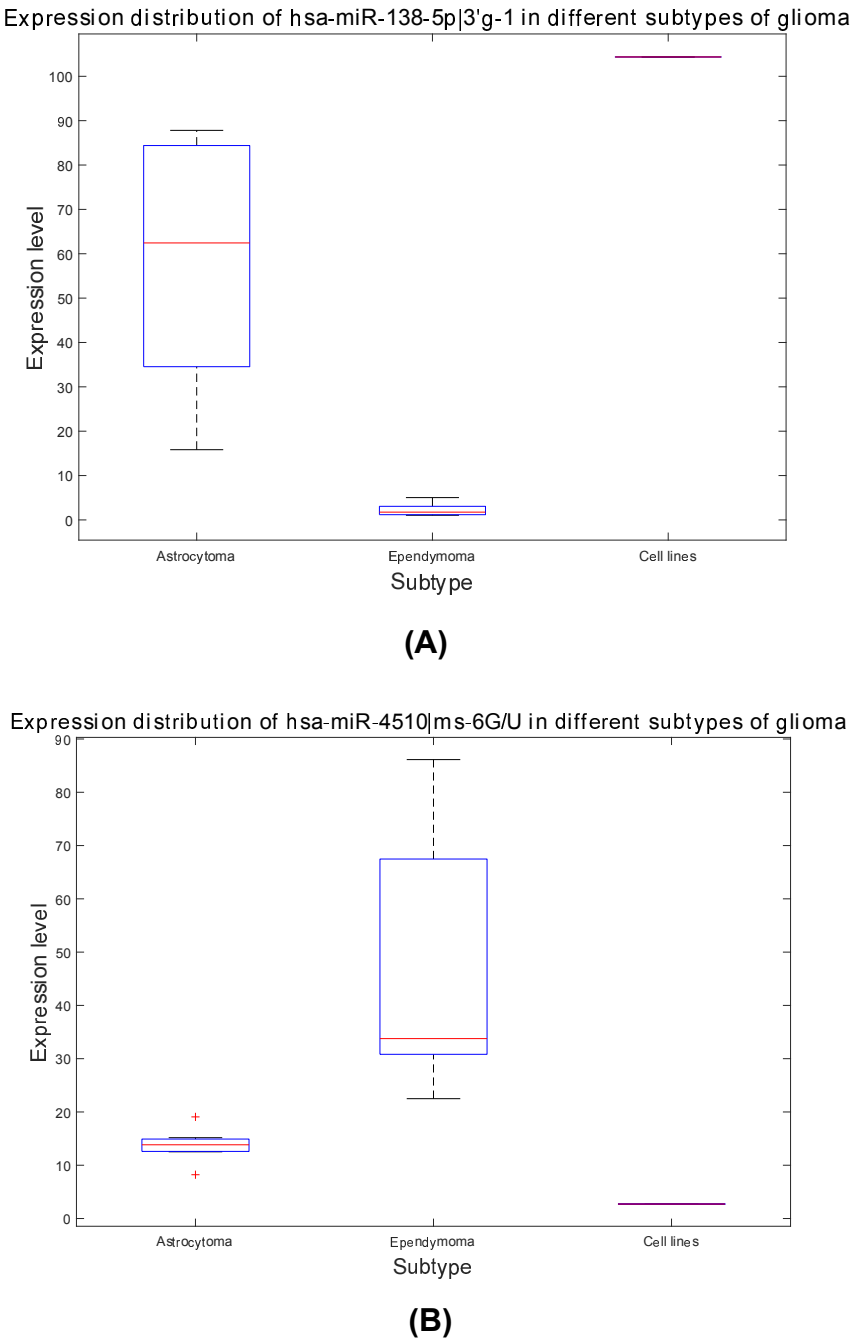

Figure 5.8: **The expression distribution of has-miR-138-5p|3′g-1 (A) and isomiR has-miR-4510|ms-6G/U (B) in different glioma subtypes.** The x-axis in this figure is the glioma subtype and the y-axis is the expression level of the isomiR.

has-miR-4510|ms-6G/U is become a novel let-7 tumour suppressor family and regulates let-7 maturation through targeting *Lin28B*.

5.4.5 Predicting molecular pathways of isomiRs in glioma subtypes

The isomiR provides a hidden unresarched layer of the gene regulation. 5′ isomiRs could regulate novel target genes and 3′ isomiRs weaken the function of regulating target genes compare to wild type miRNA. In order to detect the molecular pathways that influenced by isomiR, we find out genes that are regulated by isomiRs and then detecting the genes molecular pathway. The 76 isomiRs that have maximum information gain are key isomiRs in classifying different glioma subtypes. Thus, we focus on analyzing the molecular pathways that are regulated by these isomiRs. However, not all the isomiRs have significance regulate the molecular pathway. This is because the expression level of some isomiRs are relative low compared with their wild type miRNAs. The relative lowly expressed isomiRs have limited influence in regulating their target genes. In this chapter, the isomiR, which the total expression level is larger than its wild type miRNAs, significance regulates the gene and then affecting the molecular pathway. Finally, we found 12 isomiRs that their total expression level were higher than their wild type miRNAs. These 12 isomiRs include a polymorphic isomiR, three 3′ trimming isomiR, five 3′ templated additional isomiRs, a 5′ trimming isomiR, a 5′ additional isomiR, and a 3′ untemplated addition. These isomiRs are hsa-miR-4510|ms-6G/U, hsa-miR-338-3p|3′t-1, hsa-miR-99a-3p|3′g-1, hsa-miR-190b|3′g-2, hsa-miR-138-5p|3′g-1, hsa-miR-146b-3p|3′g-1, hsa-miR-146b-3p|5′t-1, hsa-miR-331-3p|3′t-1, hsa-miR-146b-3p|3′a-C, hsa-miR-29a-3p|5′a-C, hsa-miR-497-5p|3′g-1, hsa-miR-125b-2-3p|3′t-1.

Different kinds of isomiRs have different regulation mechanism to affect the gene expression. 3′ isomiRs loose regulation of mRNAs. And the 5′ isomiR and polymorphic isomiR, which the substitute nucleotide at the seed region, may target mRNAs that are different from the wild type miRNA

targeted transcriptome. Therefore, we apply the miRwalker2.0 (`http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/`) to discover the target gene of 3′ isomiR. While using the miRDB website (`http://mirdb.org/`) to predict the target gene of the 5′ isomiR and polymorphic isomiRs, which the substitute nucleotide at the seed region. The predicted target gene which the predicted score is higher than 95 is viewed as the target gene of the isomiR. The KEGG pathway of the gene is detected by using the Enrichr website (`http://amp.pharm.mssm.edu/Enrichr/enrich`).

5.4.6 Predicting general pathways that miss regulated due to elevated of 3′ isomiR expression

123 KEGG pathways that losed regulated by 3′ isomiRs are identified as significance pathway (the p-value of the pathway is lower than 0.05). In this chapter, we focus on analyzing 5 key glioma relatived KEGG pathways that are presented in Table 5.2. The first three KEGG pathways showed in this table are the KEGG pathway that relative to glioma or cancer. The PI3K-Akt signalling pathway, which is showed in the fourth line of Table 5.2, is very important in regulating cell growth, proliferation, and survival (Hemmings & Restuccia 2012). This pathway is elevated in many glioma cells (Haas-Kogan, Shalev, Wong, Mills, Yount & Stokoe 1998, Hu, Pandolfi, Li, Koutcher, Rosenblum & Holland 2005). We found that forty-three 3′ isomiRs weaken regulatory effects on genes relatived to PI3K-Akt signalling pathway. Thus, the gene relatived to PI3K-Akt signalling pathway will be activated. The activated of the PI3K-Akt signaling pathway allows glioma cell to apophasis (Cheng, Fan & Weiss 2009). Therefore, this pathway is a critical for the survival of the glioma cell. p53 signalling pathway regulates DNA replication and cell division (Harris & Levine 2005). Deregulated $p53$ pathway components enhance glioma cell invasion, proliferation and migration (Zhang, Dube, Gibert, Cruickshanks, Wang, Coughlan, Yang, Setiady, Deveau, Saoud et al. 2018). The 3′ isomiR weakens the regulatory effect on genes that regulate $p53$ signalling pathway and therefore, lead to

Table 5.2: 3′ isomiRs influence the KEGG pathways that relative to glioma

KEGG pathway	P-value	Number of gene
Pathways in cancer	$4.45 * 10^{-13}$	76
Glioma	$1.26 * 10^{-6}$	17
MicroRNAs in cancer	$1.57 * 10^{-5}$	37
PI3K-Akt signalling pathway	$5.34 * 10^{-6}$	43
p53 signalling pathway	$6.75 * 10^{-7}$	17

glioma cell invasion, proliferation and migration.

5.4.7 Predicting the subtype specific changes of individual targets of miRNAs based on isomiRs

The polymorphic isomiR that the substitute position at the seed region and the 5′ isomiR binds to novel genes and then affecting the KEGG pathways. There are 22 KEGG pathways that p-values are lower than 0.05 are affecting by the polymorphic isomiR that the substitute position at the seed region and the 5′ isomiR. We found that these isomiRs also regulate the gene relative to miRNA in cancer and PI3K-Akt signalling pathway, which is presented in the first two lines of Table 5.3. Further, four genes regulated by these isomiRs are involved in mToR signalling pathway, which is in the third line of Table 5.3. The mToR signalling pathway is critical for cell growth and survival (Vogt 2001). Regulating this pathway could influence glioma cell growth and survival. The polymorphic isomiR that the substitute position at the seed region and the 5′ isomiR regulate novel genes compare with wild type miRNA and influence KEGG pathway.

5.5 Conclusion

In this chapter, we applied the information gain to identify the isomiR for classifying different glioma cancer subtypes from RNA-seq data. We found that the 3′ trimming and untemplated isomiR were highly expressed compare

Table 5.3: 5′ isomiRs and polymorphic isomiRs effect the KEGG pathways that relative to glioma

KEGG name	P-value	Number of gene
PI3K-Akt signalling pathway	$1.08 * 10^{-4}$	20
MicroRNAs in cancer	0.0135	6
mTOR signalling pathway	0.017	4

with other types of isomiR. Most of the isomiRs were lowly expressed and should be removed from the dataset. Then, the information gain was applied to measure the significance of the isomiR in classifying different glioma subtypes. Since the information gain was used to the discrete data and the expression level of isomiR is continuous data. We transformed the continuous data into discrete data by using the distribution-based method. Finally, we obtained 76 isomiRs that had the maximum information and they may be very important for classifying different glioma subtypes. These isomiRs regulated the mRNA which relatived to the molecular pathway of glioma. The wet-lab experiment demonstrated that the isomiR hsa-miR-4510|ms-6G/U is a novel member of let-7 tumour suppressor family and regulates the let-7 maturation by targeting Lin28B mRNA. The hsa-miR-138-5p|3′g-1 isomiR regulates the *CCND1* and *AURKA* which are potential biomarkers for glioma subtype classification. Discovering the biomarker for classifying different glioma subtypes is challenge. Our research provides a new way to explore the glioma subtype biomarkers.

Chapter 6

Summary and Future Work

6.1 Summary

In this thesis, information theory is applied to study the regulation mechanism of non-coding RNA in human cancer. It is instigated to contribute more knowledge to the regulation mechanism of non-coding RNA in human cancers. This thesis mainly focuses on two critical regulatory mechanism of non-coding RNA in human cancers. The first is the ceRNA network and the other is the biomarkers for classifying different cancer subtypes. The pointwise mutual information is used to construct the ceRNA network in breast cancer, which reflects the competition relationship between lncRNAs, miRNAs, and mRNAs. The improved mutual information and the information gain are employed to discover the biomarker for classifying different breast cancer subtypes and glioma subtypes, respectively.

In Chapter 3, we propose a novel method to construct ceRNA network in breast cancer. The advantages of this method is that we apply the competition regulation mechanism is applied to remove the significance negative ceRNA crosstalk and combining the competition rule and pointwise mutual information to measure the competition relationship between lncRNA, miRNA, and mRNA. The results demonstrate that the ceRNA networks constructed by our method play critical roles in breast cancer

growth, development, and metastatic.

The Chapter 4 identifies the isomiR biomarker for classifying different breast cancer subtype by using improving mutual information. The traditional mutual information could be applied to the dataset which both the label and the feature are continuous data or both are discrete data. However, the expression level of the isomiR is continuous data and the breast cancer subtype is discrete data. Our method improved the mutual information method and it could be applied to the data set that feature is continuous data and label is discrete data. The results displays the improved mutual information is better than other feature selection methods for discovering isomiRs as biomarkers for classifying different breast cancer subtypes. Further, the isomiR is better a biomarker than mRNA for classifying differen breast cancer subtypes.

The Chapter 5 focuses on discovering the isomiR biomarker for classifying different glioma subtypes. In this research, we analyse the isomiR expression in different glioma subtypes and find that a few isomiR are highly expressed while large amount of isomiRs are lowly expressed. These lowly expressed isomiRs have significance negative influence on discovering biomarker. Therefore, a hypothesis method is applied to remove these lowly expressed isomiRs. The information gain is applied to measure the significance isomiR in classifying different glioma subtypes and 76 isomiRs has the maximum information gain. The wet-lab experiments reveal that the isomiR could regulate the mRNA and then influencing the molecular pathway of glioma subtypes.

6.2 Future Work

Following above our research, we can find that the information theories is a powerful tool in discovering the regulation mechanism of non-coding RNA in human cancers. However, the regulation mechanism of the non-coding RNA in cancers is not fully understood. Therefore, our future work will focus on

these research topics:

- **Constructing the ceRNA network which other RNAs as ceRNA**.
  The other RNAs, such as circular RNA (circRNA) and mRNA, also could be ceRNAs to compete with mRNA to bind to miRNA. Thus, the other RNAs, miRNA, mRNA are able to construct the ceRNA network. The ceRNA network, which the lncRNA acts as ceRNA, is a partition of the RNA regulatory network in breast cancers. Understanding the comprehensive regulatory mechanism of non-coding RNA in breast cancer should take all the RNA into consideration.

- **Discovering the ceRNA-ceRNA network interaction.**
  In this thesis, we define the ceRNA network is the interaction between a miRNA, its target mRNAs and target lncRNAs. One miRNA may co-regulate with the other miRNA, this co-regulate relationship connects two ceRNA networks and forms ceRNA-ceRNA interaction network. CeRNA networks may work synergistically during different developmental stages or tissues to control specific functions. Analysing ceRNA interactions in the context of tissue development will provide insights into the regulation of cell development, as well as the dysregulation of key mechanisms of pathogenesis (Xu, Feng, Han, Li, Wu, Shao, Ding, Li, Deng, Di et al. 2016). Cancers are always regulated by serval different ceRNA networks and two different ceRNA networks may regulate the same pathway. These two ceRNA networks are very important in regulating the mechanism of cancer.

- **The isomiR as the biomarker in classifying different stages of cancer.**
  The stages of cancer describes how far the cancer has grown. For example, in the stage I, cancer cells are very small in an area. However, the cancer is spread to other part of the body in stage IV. Understanding the stage of cancer help doctor to determine the

treatment and provides the possible outcome. IsomiR performs a good biomarker for classifying different cancer subtypes. It may be able to classify different stages of cancers.

- **The improvement of the method for calculating the number of key biomarker.**
  We can measure the weight of RNA for classifying difference cancer subtypes. The higher the weight of RNA, the more important for cancer subtype classification. However, it is challenge to determine the number of the key biomarker. The traditional method of selecting key biomarker is experts experience. In this thesis, we provide a new method (5-fold cross validation) to find out the key biomarker. However, my method has some limitations. For example, it requires to observe the trend of the classification result. We hope that the novel method could find out the number of key biomarker through calculation not through observation.

- **The improvement of the method for calculating the number of key biomarker.**
  Feature selection is a good method for identifying biomarkers for classifying different cancer subtypes. These biomarkers are useful to diagnosis cancer subtypes. However, not all biomarkers are suitable for cancer treatment. Identifying the biomarker that suitable for cancer treatment is a very interesting topic. In order to find out the treatment related biomarker, we should develop a novel method to find out the causal relationship between RNA and cancer subtypes.

Appendix A

Appendix: Long Table

Table A.1: 76 isomiRs that have maximum information gain and their average expression level in different glioma subtypes.

IsomiR name	Average expression level		
	Astrocytoma	Ependymoma	Cell line
hsa-let-7a-5p|ms-19A/G	14.66	72.14	4.59
hsa-let-7b-5p|ms-17G/C	10.85	33.02	2.34
hsa-let-7b-5p|ms-17G/T	6.73	21.50	1.03
hsa-let-7b-5p|ms-4G/A	13.30	36.43	8.06
hsa-let-7c-5p|3′a-A	825.95	3219.64	132.96
hsa-let-7c-5p|3′a-AC	5.96	28.74	1.31
hsa-let-7c-5p|3′a-AT	37.41	152.39	7.31
hsa-let-7c-5p|3′a-ATT	9.21	43.43	1.88
hsa-let-7c-5p|3′a-C	1653.48	6587.12	287.20
hsa-let-7c-5p|3′a-GC	6.49	44.35	1.50
hsa-let-7c-5p|3′a-GT	38.61	397.72	12.47
hsa-let-7c-5p|3′a-T	23.56	97.37	6.28
hsa-let-7c-5p|3′g-1	472.74	2163.26	98.17
hsa-let-7c-5p|3′t-3	120.45	337.97	20.25

IsomiR name	Average expression level		
	Astrocytoma	Ependymoma	Cell line
hsa-let-7c-5p|ms-10A/G	35.36	119.29	5.53
hsa-let-7c-5p|ms-10A/T	8.74	31.71	0.84
hsa-let-7c-5p|ms-11G/A	15.06	51.10	2.16
hsa-let-7c-5p|ms-13T/A	14.18	49.00	3.00
hsa-let-7c-5p|ms-13T/C	39.99	138.70	6.84
hsa-let-7c-5p|ms-13T/G	12.54	44.31	2.72
hsa-let-7c-5p|ms-14T/A	8.31	29.95	2.16
hsa-let-7c-5p|ms-16T/C	26.51	89.31	4.78
hsa-let-7c-5p|ms-17A/C	10.85	33.01	2.34
hsa-let-7c-5p|ms-17A/T	6.75	21.55	1.03
hsa-let-7c-5p|ms-18T/C	27.69	101.36	5.25
hsa-let-7c-5p|ms-1T/A	22.70	84.37	4.22
hsa-let-7c-5p|ms-1T/C	40.26	140.54	7.13
hsa-let-7c-5p|ms-1T/G	15.81	56.05	3.84
hsa-let-7c-5p|ms-2G/A	8.84	33.68	1.03
hsa-let-7c-5p|ms-3A/G	57.63	210.24	9.75
hsa-let-7c-5p|ms-3A/T	12.21	45.31	3.38
hsa-let-7c-5p|ms-4G/A	9.87	33.92	1.88
hsa-let-7c-5p|ms–1T/C	8.60	31.33	1.13
hsa-let-7e-5p|ms-7A/G	6.17	22.11	4.41
hsa-let-7i-5p|3′a-ATT	17.27	59.56	25.32
hsa-let-7i-5p|ms-4G/A	22.37	73.18	33.38
hsa-miR-100-5p|3′a-CGT	6.73	17.34	30.29
hsa-miR-100-5p|3′a-TA	9.09	21.34	49.70
hsa-miR-125b-2-3p|3′t-1	25.47	143.37	15.19
hsa-miR-125b-2-3p|5′a-A	18.79	74.01	9.19
hsa-miR-132-5p|3′t-2	54.66	13.48	5.34
hsa-miR-132-5p|3′t-3	34.45	6.89	2.16
hsa-miR-132-5p|5′a-A	36.61	13.94	2.72

IsomiR name	Average expression level		
	Astrocytoma	Ependymoma	Cell line
hsa-miR-138-5p|3′g-1	58.30	2.30	104.36
hsa-miR-146b-3p|3′a-C	3.02	37.92	0.94
hsa-miR-146b-3p|3′g-1	21.98	225.04	5.91
hsa-miR-146b-3p|5′t-1	20.89	204.31	5.81
hsa-miR-181d-5p|3′t-2	61.66	10.90	4.13
hsa-miR-190b|3′g-2	1.14	54.47	0.19
hsa-miR-24-3p|3′t-1	904.15	274.95	492.64
hsa-miR-26a-5p|3′a-TA	14.13	31.03	7.41
hsa-miR-26a-5p|3′a-TTT	10.06	28.12	4.59
hsa-miR-29a-3p|5′a-C	897.16	284.33	2352.29
hsa-miR-30a-5p|3′a-T	637.03	1219.70	195.88
hsa-miR-30a-5p|ms-20A/C	15.85	30.51	6.09
hsa-miR-331-3p|3′t-1	27.43	177.00	9.38
hsa-miR-338-3p|3′t-1	23.47	2.26	0.28
hsa-miR-340-3p|3′a-T	107.40	24.58	3.84
hsa-miR-340-3p|3′t-1	93.69	16.12	3.09
hsa-miR-340-5p|3′t-1	493.35	142.01	12.94
hsa-miR-340-5p|3′t-2	142.32	33.79	3.47
hsa-miR-4510|ms-6G/T	13.75	46.82	2.72
hsa-miR-483-3p|3′a-GCT	0.38	29.64	0.66
hsa-miR-497-5p|3′g-1	47.58	99.65	4.41
hsa-miR-497-5p|3′g-2	5.27	29.73	0.84
hsa-miR-497-5p|5′t-1	5.46	27.86	0.84
hsa-miR-92b-3p|ms-23G/A	7.88	172.89	18.38
hsa-miR-9-3p|ms-5A/G	31.42	8.40	0.28
hsa-miR-9-5p|3′a-AA	567.89	198.63	25.69
hsa-miR-9-5p|3′g-2	132.86	59.88	6.38
hsa-miR-9-5p|5′t-1	6794.71	1705.62	153.40
hsa-miR-9-5p|5′t-2	928.13	225.21	23.82

IsomiR name	Average expression level		
	Astrocytoma	Ependymoma	Cell line
hsa-miR-9-5p|ms-18G/T	20.57	10.01	0.84
hsa-miR-98-5p|ms-19T/G	12.31	43.64	2.16
hsa-miR-99a-3p|3′g-1	7.89	63.42	2.16
hsa-miR-99a-5p|3′a-CT	19.31	52.82	4.50

Appendix B

Appendix: List of Symbols

The following list is neither exhaustive nor exclusive, but may be helpful.

$3'UTR$	$3'$ untranslated region
$AUcontent$	Adenine and uracil content
AUC	Area under ROC curve
$AURKA$	Aurora kinase A
$BRAF$	B-Raf Proto-Oncogene
$BRAFP1$	V-Raf Murine Sarcoma Viral Oncogene Homolog B Pseudogene 1
$C6orf58$	Chromosome 6 Open Reading Frame 58
$CCND1$	Cyclin D1
$CLIP$	Cross-linking immunoprecipitation
$DLX6$	Distal-Less Homeobox 6
$DUXA$	Double Homeobox A
$EGR1$	Early Growth Response 1

$EMBL-EBI$	The European Bioinformatics Institute
$ER\alpha$	Estrogen receptor
$FOXA2$	Forkhead Box A2
$FPKM$	Fragments per kilo base per million mapped reads
$GCcontent$	Guanine and cytosine content
$GFRAL$	GDNF Family Receptor Alpha Like
$GPR26$	G Protein-Coupled Receptor 26
$HER2+$	Herceptin 2 positive
$HOXA5$	Homeobox A5
$isomiR$	MicroRNA isoform
IHC	Immunohistochemistry
$INSM1$	INSM Transcriptional Repressor 1
$KEGG$	Kyoto Encyclopedia of Genes and Genomes
$Lin28B$	Rotein Lin-28 Homolog B
$MAPK14$	Mitogen-Activated Protein Kinase 14
$MAPK8$	Mitogen-Activated Protein Kinase 8
$MEOX2$	Mesenchyme Homeobox 2
$NCBI$	The National Center for Biotechnology Information
$NTSR1$	Neurotensin Receptor 1
$PAR-CLIP$	Photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation

PCR	Polymerase chain reaction
$PF4$	Platelet Factor 4
PR	Progesterone receptor
$PTBP3$	Polypyrimidine Tract Binding Protein 3
$PTEN$	Phosphatase And Tensin Homolog
$PTENP1$	Phosphatase And Tensin Homolog Pseudogene 1
$RAP1B$	RAP1B
$RNA-seq$	RNA-sequencing
$RPKM$	Reads per Kilo base Million mapped reads
RPM	Reads per million mapped reads
$SERPINA3$	Serpin family A member 3
$SHC4$	SHC Adaptor Protein 4
$SOX17$	Sex Determining Region Y-Box 17
SVM	Support vector machine
$TCGA$	The Cancer Genome Atlas
$TFF1$	Trefoil Factor 1
$TGFBR3$	Transforming growth factor-beta type III receptor
$TP53$	Tumor Protein P53
$ZG16$	Zymogen Granule Protein 16
$ceRNA$	Competing endogenous RNA
$circRNA$	Circular RNA

$lncRNA$	Long non-coding RNA
$mRNA$	Messenger RNA
$miRNA$	MicroRNA
$miRP$	miRNA program
$piRNAs$	Piwi-associated RNAs
$pre-miRNA$	Precursor miRNAs
$pri-miRNA$	Primary miRNA transcripts
$siRNA$	Small interfering RNA
$snoRNA$	Small nucleolar RNA
$tsRNA$	tRNA-derived small RNA

Bibliography

Abbas, T. & Dutta, A. (2009), 'P21 in cancer: intricate networks and multiple activities', *Nature Reviews Cancer* **9**(6), 400–414.

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

Aldape, K., Zadeh, G., Mansouri, S., Reifenberger, G. & von Deimling, A. (2015), 'Glioblastoma: pathology, molecular mechanisms and markers', *Acta Neuropathologica* **129**(6), 829–848.

Baldi, P. & Long, A. D. (2001), 'A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes', *Bioinformatics* **17**(6), 509–519.

Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. (2010), 'Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites', *Genome Biology* **11**(8), R90.

Burroughs, A. M., Ando, Y., de Hoon, M. J., Tomaru, Y., Nishibu, T., Ukekawa, R., Funakoshi, T., Kurokawa, T., Suzuki, H., Hayashizaki, Y. et al. (2010), 'A comprehensive survey of 3 animal miRNA modification events and a possible role for 3 adenylation in modulating miRNA targeting effectiveness', *Genome Research* **20**(10), 1398–1410.

Camps, C., Saini, H. K., Mole, D. R., Choudhry, H., Reczko, M., Guerra-Assunção, J. A., Tian, Y.-M., Buffa, F. M., Harris, A. L., Hatzigeorgiou, A. G. et al. (2014), 'Integrated analysis of microRNA and mRNA

expression and association with HIF binding reveals the complexity of microRNA expression regulation under hypoxia', *Molecular Cancer* **13**(1), 28.

Carter, S. L., Brechbühler, C. M., Griffin, M. & Bond, A. T. (2004), 'Gene co-expression network topology provides a framework for molecular characterization of cellular state', *Bioinformatics* **20**(14), 2242–2250.

Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., Cramer, C. L. & Huang, X. (2015), 'Bridger: a new framework for de novo transcriptome assembly using RNA-seq data', *Genome Biology* **16**(1), 30.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), 'SMOTE: synthetic minority over-sampling technique', *Journal of artificial intelligence research* **16**, 321–357.

Chen, J., Xu, J., Li, Y., Zhang, J., Chen, H., Lu, J., Wang, Z., Zhao, X., Xu, K., Li, Y. et al. (2017), 'Competing endogenous RNA network analysis identifies critical genes among the different breast cancer subtypes', *Oncotarget* **8**(6), 10171.

Chen, L. & Wong, G. (2017), Novel tumor biomarker based on isomiR expression profiles, *in* 'Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on', IEEE, pp. 2328–2329.

Chen, R., Smith-Cohn, M., Cohen, A. L. & Colman, H. (2017), 'Glioma subclassifications and their clinical significance', *Neurotherapeutics* **14**(2), 284–297.

Cheng, C. K., Fan, Q.-W. & Weiss, W. A. (2009), 'Pi3k signaling in gliomaanimal models and therapeutic challenges', *Brain pathology* **19**(1), 112–120.

Chiu, H.-S., Llobet-Navas, D., Yang, X., Chung, W.-J., Ambesi-Impiombato, A., Iyer, A., Kim, H. R., Seviour, E. G., Luo, Z., Sehgal, V. et al.

(2015), 'Cupid: simultaneous reconstruction of microrna-target and cerna networks', *Genome Research* **25**(2), 257–267.

Chiu, Y.-C., Chuang, E. Y., Hsiao, T.-H. & Chen, Y. (2013), Modeling competing endogenous rna regulatory networks in glioblastoma multiforme, *in* '2013 IEEE International Conference on Bioinformatics and Biomedicine', IEEE, pp. 201–204.

Chiu, Y.-C., Hsiao, T.-H., Chen, Y. & Chuang, E. Y. (2015), 'Parameter optimization for constructing competing endogenous RNA regulatory network in glioblastoma multiforme and other cancers', *BMC Genomics* **16**(4), S1.

Church, K. W. & Hanks, P. (1990), 'Word association norms, mutual information, and lexicography', *Computational Linguistics* **16**(1), 22–29.

Cieslak, D. A. & Chawla, N. V. (2008), Learning decision trees for unbalanced data, *in* 'Joint European Conference on Machine Learning and Knowledge Discovery in Databases', Springer, pp. 241–256.

Consortium, F., Team, R. G. E. R. G. P. I. . I. et al. (2002), 'Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cdnas', *Nature* **420**(6915), 563.

Conte, F., Fiscon, G., Chiara, M., Colombo, T., Farina, L. & Paci, P. (2017), 'Role of the long non-coding RNA PVT1 in the dysregulation of the ceRNA-ceRNA network in human breast cancer', *PLoS One* **12**(2), e0171661.

Coons, S. W., Johnson, P. C., Scheithauer, B. W., Yates, A. J. & Pearl, D. K. (1997), 'Improving diagnostic accuracy and interobserver concordance in the classification and grading of primary gliomas', *Cancer: Interdisciplinary International Journal of the American Cancer Society* **79**(7), 1381–1393.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine Learning* **20**(3), 273–297.

Cover, T. M. & Thomas, J. A. (2012), *Elements of information theory*, John Wiley & Sons.

Danchin, É., Charmantier, A., Champagne, F. A., Mesoudi, A., Pujol, B. & Blanchet, S. (2011), 'Beyond DNA: integrating inclusive inheritance into an extended theory of evolution', *Nature Reviews Genetics* **12**(7), 475.

Das, S., Ghosal, S., Sen, R. & Chakrabarti, J. (2014), 'lnCeDB: database of human long noncoding RNA acting as competing endogenous RNA', *PLoS One* **9**(6), e98965.

Ding, B., Liang, H., Gao, M., Li, Z., Xu, C., Fan, S. & Chang, N. (2017), 'Forkhead Box A2 (FOXA2) inhibits invasion and tumorigenesis in glioma cells', *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics* **25**(5), 701–708.

Dressman, M., Walz, T., Lavedan, C., Barnes, L., Buchholtz, S., Kwon, I., Ellis, M. & Polymeropoulos, M. (2001), 'Genes that co-cluster with estrogen receptor alpha in microarray analysis of breast biopsies', *The Pharmacogenomics Journal* **1**(2), 135.

Dupouy, S., Viardot-Foucault, V., Alifano, M., Souazé, F., Plu-Bureau, G., Chaouat, M., Lavaur, A., Hugol, D., Gespach, C., Gompel, A. et al. (2009), 'The neurotensin receptor-1 pathway contributes to human ductal breast cancer progression', *PLoS One* **4**(1), e4223.

Dweep, H. & Gretz, N. (2015), 'MiRWalk2. 0: a comprehensive atlas of microRNA-target interactions', *Nature Methods* **12**(8), 697–697.

Ebert, M. S., Neilson, J. R. & Sharp, P. A. (2007), 'MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells', *Nature Methods* **4**(9), 721–726.

Ellsworth, R. E., Blackburn, H. L., Shriver, C. D., Soon-Shiong, P. & Ellsworth, D. L. (2017), 'Molecular heterogeneity in breast cancer: state of the science and implications for patient care', **64**, 65–72.

Ellwanger, D. C., Büttner, F. A., Mewes, H.-W. & Stümpflen, V. (2011), 'The sufficient minimal set of mirna seed types', *Bioinformatics* **27**(10), 1346–1350.

Ferri, C., Hernández-Orallo, J. & Flach, P. A. (2011), A coherent interpretation of AUC as a measure of aggregated classification performance, *in* 'Proceedings of the 28th International Conference on Machine Learning (ICML-11)', pp. 657–664.

Finn, R. S., Dering, J., Ginther, C., Wilson, C. A., Glaspy, P., Tchekmedyian, N. & Slamon, D. J. (2007), 'Dasatinib, an orally active small molecule inhibitor of both the src and abl kinases, selectively inhibits growth of basal-type triple-negative breast cancer cell lines growing in vitro', *Breast Cancer Research and Treatment* **105**(3), 319–326.

Frank, E. (2014), 'Fully supervised training of gaussian radial basis function networks in weka (computer science working papers, 04/2014)', *Department of Computer Science, The University of Waikato, Hamilton, NZ* .

Fu, D.-y., Tan, H.-s., Wei, J.-l., Zhu, C.-R., Jiang, J.-x., Zhu, Y.-x., Cai, F.-l., Chong, M.-h. & Ren, C.-l. (2015), 'Decreased expression of sox17 is associated with tumor progression and poor prognosis in breast cancer', *Tumor Biology* **36**(10), 8025–8034.

Gasco, M., Shami, S. & Crook, T. (2002), 'The p53 pathway in breast cancer', *Breast Cancer Research* **4**(2), 70.

Gibb, E. A., Brown, C. J. & Lam, W. L. (2011), 'The functional role of long non-coding rna in human carcinomas', *Molecular Cancer* **10**(1), 38.

Goldhirsch, A. ., Wood, W. C., Coates, A. S., Gelber, R. D., Thürlimann, B., Senn, H.-J. & members, P. (2011), 'Strategies for subtypes dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011', *Annals of Oncology* **22**(8), 1736–1747.

Griffith, M., Walker, J. R., Spies, N. C., Ainscough, B. J. & Griffith, O. L. (2015), 'Informatics for rna sequencing: a web resource for analysis on the cloud', *PLoS Computational Biology* **11**(8), e1004393.

Grishin, N. V. (2001), 'Fold change in evolution of protein structures', *Journal of Structural Biology* **134**(2), 167–185.

Gu, Q., Li, Z. & Han, J. (2011), Generalized fisher score for feature selection, *in* 'Twenty-Seventh Conference on Uncertainty in Artificial Intelligence', pp. 266–273.

Gu, S., Cheng, R. & Jin, Y. (2018), 'Feature selection for high-dimensional classification using a competitive swarm optimizer', *Soft Computing* **22**(3), 811–822.

Guo, Y., Shang, X. & Li, Z. (2019), 'Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer', *Neurocomputing* **324**, 20–30.

Haas-Kogan, D., Shalev, N., Wong, M., Mills, G., Yount, G. & Stokoe, D. (1998), 'Protein kinase b (pkb/akt) activity is elevated in glioblastoma cells due to mutation of the tumor suppressor pten/mmac', *Current Biology* **8**(21), 1195–S1.

Hanley, J. A. & McNeil, B. J. (1982), 'The meaning and use of the area under a receiver operating characteristic (ROC) curve.', *Radiology* **143**(1), 29–36.

Harris, S. L. & Levine, A. J. (2005), 'The p53 pathway: positive and negative feedback loops', *Oncogene* **24**(17), 2899.

Hemmings, B. A. & Restuccia, D. F. (2012), 'Pi3k-pkb/akt pathway', *Cold Spring Harbor Perspectives in Biology* **4**(9), a011189.

Herold, Christina I and Chadaram, Vijaya and Peterson, Bercedis L and Marcom, P Kelly and Hopkins, Judith and Kimmick, Gretchen G and Favaro, Justin and Hamilton, Erika and Welch, Renee A and Bacus, Sarah and others (2011), 'Phase II trial of dasatinib in patients with metastatic breast cancer using real-time pharmacodynamic tissue biomarkers of Src inhibition to escalate dosing', *Clinical Cancer Research* **17**(18), 6061–6070.

Hou, P., Li, L., Chen, F., Chen, Y., Liu, H., Li, J., Bai, J. & Zheng, J. (2018), 'PTBP3-mediated regulation of ZEB1 mRNA stability promotes epithelial–mesenchymal transition in breast xancer', *Cancer Research* **78**(2), 387–398.

Hu, X., Pandolfi, P. P., Li, Y., Koutcher, J. A., Rosenblum, M. & Holland, E. C. (2005), 'mTOR promotes survival and astrocytic characteristics induced by Pten/AKT signaling in glioblastoma', *Neoplasia* **7**(4), 356–368.

Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2008), 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Research* **37**(1), 1–13.

Hutvágner, G., McLachlan, J., Pasquinelli, A. E., Bálint, É., Tuschl, T. & Zamore, P. D. (2001), 'A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA', *Science* **293**(5531), 834–838.

Jeggari, A., Marks, D. S. & Larsson, E. (2012), 'miRcode: a map of putative microRNA target sites in the long non-coding transcriptome', *Bioinformatics* **28**(15), 2062–2063.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., Marks, D. S. et al. (2004), 'Human microRNA targets', *PLoS Biology* **2**(11), e363.

Johnson, E., Dickerson, K. L., Connolly, I. D. & Gephart, M. H. (2018), 'Single-cell RNA-sequencing in glioma', *Current Oncology Reports* **20**(5), 42.

Juzenas, S., Venkatesh, G., Hübenthal, M., Hoeppner, M. P., Du, Z. G., Paulsen, M., Rosenstiel, P., Senger, P., Hofmann-Apitius, M., Keller, A. et al. (2017), 'A comprehensive, cell specific microRNA catalogue of human peripheral blood', *Nucleic Acids Research* **45**(16), 9290–9301.

Kamachi, Y. & Kondoh, H. (2013), 'Sox proteins: regulators of cell fate specification and differentiation', *Development* **140**(20), 4129–4144.

Kanehisa, M. & Goto, S. (2000), 'KEGG: kyoto encyclopedia of genes and genomes', *Nucleic Acids Research* **28**(1), 27–30.

Karreth, F. A., Reschke, M., Ruocco, A., Ng, C., Chapuy, B., Léopold, V., Sjoberg, M., Keane, T. M., Verma, A., Ala, U. et al. (2015), 'The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo', *Cell* **161**(2), 319–332.

Kehl, T., Backes, C., Kern, F., Fehlmann, T., Ludwig, N., Meese, E., Lenhof, H.-P. & Keller, A. (2017), 'About miRNAs, miRNA seeds, target genes and target pathways', *Oncotarget* **8**(63), 107167.

Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W. & Quackenbush, J. (2018), 'Cancer subtype identification using somatic mutation data', *British Journal of Cancer* **118**(11), 1492.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A. et al. (2016), 'Enrichr: a comprehensive gene set enrichment analysis web server 2016 update', *Nucleic Acids Research* **44**(W1), W90–W97.

Lazennec, G. & Richmond, A. (2010), 'Chemokines and chemokine receptors: new insights into cancer-related inflammation', *Trends in Molecular Medicine* **16**(3), 133–144.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y. & Pietenpol, J. A. (2011), 'Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies', *The Journal of Clinical Investigation* **121**(7), 2750–2767.

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**(1), 15–20.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. & Liu, H. (2017), 'Feature selection: A data perspective', *ACM Computing Surveys (CSUR)* **50**(6), 94.

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H. & Yang, J.-H. (2013), 'StarBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data', *Nucleic Acids Research* **42**(D1), D92–D97.

Li, P., Chen, S., Chen, H., Mo, X., Li, T., Shao, Y., Xiao, B. & Guo, J. (2015), 'Using circular RNA as a novel type of biomarker in the screening of gastric cancer', *Clinica Chimica Acta* **444**, 132–136.

Li, S.-C., Liao, Y.-L., Ho, M.-R., Tsai, K.-W., Lai, C.-H. & Lin, W.-c. (2012), 'miRNA arm selection and isomiR distribution in gastric cancer', *BMC Genomics* **13**(1), S13.

Li, Z., Peng, Z., Gu, S., Zheng, J., Feng, D., Qin, Q. & He, J. (2017), 'Global Analysis of miRNA–mRNA Interaction Network in Breast Cancer with Brain Metastasis', *Anticancer Research* **37**(8), 4455–4468.

Liu, C., Mallick, B., Long, D., Rennie, W. A., Wolenc, A., Carmack, C. S. & Ding, Y. (2013), 'Clip-based prediction of mammalian microrna binding sites', *Nucleic Acids Research* **41**(14), e138–e138.

Liu, K., Yan, Z., Li, Y. & Sun, Z. (2013), 'Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis', *Bioinformatics* **29**(17), 2221–2222.

Liu, W. & Wang, X. (2019), 'Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data', *Genome Biology* **20**(1), 18.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D. & Darnell, J. (2000), The three roles of RNA in protein synthesis, *in* 'Molecular Cell Biology. 4th edition', WH Freeman.

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., Scheithauer, B. W. & Kleihues, P. (2007), 'The 2007 WHO classification of tumours of the central nervous system', *Acta Neuropathologica* **114**(2), 97–109.

Lu, J., Steeg, P. S., Price, J. E., Krishnamurthy, S., Mani, S. A., Reuben, J., Cristofanilli, M., Dontu, G., Bidaut, L., Valero, V. et al. (2009), 'Breast cancer metastasis: challenges and opportunities', *Cancer Research* **69**(12), 4951–4953.

Lynce, F., Blackburn, M. J., Cai, L., Wang, H., Rubinstein, L., Harris, P., Isaacs, C. & Pohlmann, P. R. (2018), 'Characteristics and outcomes of breast cancer patients enrolled in the National Cancer Institute Cancer Therapy Evaluation Program sponsored phase I clinical trials', *Breast Cancer Research and Treatment* **168**(1), 35–41.

Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N. & Chinnaiyan, A. M.

(2009), 'Transcriptome sequencing to detect gene fusions in cancer', *Nature* **458**(7234), 97.

Maher, C., Timmermans, M., Stein, L. & Ware, D. (2004), Identifyng microRNAs in plant genomes, *in* 'Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE', IEEE, pp. 718–723.

Marchese, F. P., Raimondi, I. & Huarte, M. (2017), 'The multidimensional mechanisms of long noncoding rna function', *Genome Biology* **18**(1), 206.

Milenkovic, O., Alterovitz, G., Battail, G., Coleman, T. P., Hagenauer, J., Meyn, S. P., Price, N., Ramoni, M. F., Shmulevich, I. & Szpankowski, W. (2010), 'Introduction to the special issue on information theory in molecular biology and neuroscience', *IEEE Transactions on Information Theory* **56**(2), 649–652.

Morales, M., Planet, E., Arnal-Estape, A., Pavlovic, M., Tarragona, M. & Gomis, R. R. (2011), 'Tumor-stroma interactions a trademark for metastasis', *The Breast* **20**, S50–S55.

Morini, M., Astigiano, S., Gitton, Y., Emionite, L., Mirisola, V., Levi, G. & Barbieri, O. (2010), 'Mutually exclusive expression of DLX2 and DLX5/6 is associated with the metastatic potential of the human breast cancer cell line MDA-MB-231', *BMC Cancer* **10**(1), 649.

Mousavian, Z., Kavousi, K. & Masoudi-Nejad, A. (2016), Information theory in systems biology. Part I: Gene regulatory and metabolic networks, *in* 'Seminars in Cell & Developmental Biology', Vol. 51, Elsevier, pp. 3–13.

Mulligan, A. M., Raitman, I., Feeley, L., Pinnaduwage, D., Nguyen, L. T., O'Malley, F. P., Ohashi, P. S. & Andrulis, I. L. (2013), 'Tumoral lymphocytic infiltration and expression of the chemokine CXCL10 in

breast cancers from the Ontario Familial Breast Cancer Registry', *Clinical Cancer Research* **19**(2), 336–346.

Nafi, S. N. M., Idris, F. & Jaafar, H. (2017), 'Cellular and Molecular Changes in MNU-Induced Breast Tumours Injected with PF4 or bFGF', *Asian Pacific Journal of Cancer Prevention: APJCP* **18**(12), 3231.

Navot, A. (2006), On the role of feature selection in machine learning, PhD thesis, Citeseer.

Neilsen, C. T., Goodall, G. J. & Bracken, C. P. (2012), 'IsomiRs–the overlooked repertoire in the dynamic microRNAome', *Trends in Genetics* **28**(11), 544–549.

Neman, J., Choy, C., Kowolik, C. M., Anderson, A., Duenas, V. J., Waliany, S., Chen, B. T., Chen, M. Y. & Jandial, R. (2013), 'Co-evolution of breast-to-brain metastasis and neural progenitor cells', *Clinical & Experimental Metastasis* **30**(6), 753–768.

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F. et al. (2006), 'A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes', *Cancer Cell* **10**(6), 515–527.

Paci, P., Colombo, T. & Farina, L. (2014), 'Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer', *BMC Systems Biology* **8**(1), 83.

Pantano, L., Estivill, X. & Martí, E. (2009), 'SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells', *Nucleic Acids Research* **38**(5), e34–e34.

Paraskevopoulou, M. D., Georgakilas, G., Kostoulas, N., Reczko, M., Maragkakis, M., Dalamagas, T. M. & Hatzigeorgiou, A. G.

(2012), 'DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs', *Nucleic Acids Research* **41**(D1), D239–D245.

Park, N. I., Rogan, P. K., Tarnowski, H. E. & Knoll, J. H. (2012), 'Structural and genic characterization of stable genomic regions in breast cancer: relevance to chemotherapy', *Molecular Oncology* **6**(3), 347–359.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z. et al. (2009), 'Supervised risk predictor of breast cancer based on intrinsic subtypes', *Journal of Clinical Oncology* **27**(8), 1160.

Patani, N., Martin, L.-A. & Dowsett, M. (2013), 'Biomarkers for the clinical management of breast cancer: international perspective', *International journal of cancer* **133**(1), 1–13.

Pearson, K. (1895), 'Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material', *Philosophical Transactions of the Royal Society of London* **186**(Part I), 343–424.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**(Oct), 2825–2830.

Prest, S. J., May, F. E. & Westley, B. R. (2002), 'The estrogen-regulated protein, TFF1, stimulates migration of human breast cancer cells', *The FASEB Journal* **16**(6), 592–594.

Quinlan, J. R. (1986), 'Induction of decision trees', *Machine Learning* **1**(1), 81–106.

Quinn, J. J. & Chang, H. Y. (2016), 'Unique features of long non-coding RNA biogenesis and function', *Nature Reviews Genetics* **17**(1), 47.

Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I. & Hatzigeorgiou, A. G. (2012), 'Functional microRNA targets in protein coding sequences', *Bioinformatics* **28**(6), 771–776.

Redmond, K., Crawford, N., Farmer, H., D'costa, Z., O'brien, G., Buckley, N., Kennedy, R., Johnston, P., Harkin, D. & Mullan, P. (2010), 'T-box 2 represses NDRG1 through an EGR1-dependent mechanism to drive the proliferation of breast cancer cells', *Oncogene* **29**(22), 3252–3262.

Richter, F., Hoffman, G., Manheimer, K., Patel, N., Sharp, A., McKean, D., Morton, S., DePalma, S., Gorham, J., Kitaygorodksy, A. et al. (2019), 'ORE identifies extreme expression effects enriched for rare variants', *Bioinformatics* .

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E. et al. (2007), 'Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs', *Cell* **129**(7), 1311–1323.

Ruiz-Garcia, E., Scott, V., Machavoine, C., Bidart, J., Lacroix, L., Delaloge, S. & Andre, F. (2010), 'Gene expression profiling identifies Fibronectin 1 and CXCL9 as candidate biomarkers for breast cancer screening', *British Journal of Cancer* **102**(3), 462.

Saeys, Y., Inza, I. & Larrañaga, P. (2007), 'A review of feature selection techniques in bioinformatics', *Bioinformatics* **23**(19), 2507–2517.

Sahraeian, S. M. E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P. T., Au, K. F., Asadi, N. B., Gerstein, M. B., Wong, W. H., Snyder, M. P. et al. (2017), 'Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis', *Nature Communications* **8**(1), 59.

Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. (2011), 'A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?', *Cell* **146**(3), 353–358.

Sanchez-Mejias, A. & Tay, Y. (2015), 'Competing endogenous RNA networks: tying the essential knots for cancer biology and therapeutics', *Journal of Hematology & Oncology* **8**(1), 30–39.

Santagata, S., Thakkar, A., Ergonul, A., Wang, B., Woo, T., Hu, R., Harrell, J. C., McNamara, G., Schwede, M., Culhane, A. C. et al. (2014), 'Taxonomy of breast cancer based on normal cell phenotype predicts outcome', *The Journal of Clinical Investigation* **124**(2), 859–870.

Sheikh, M. S., Rochefort, H. & Garcia, M. (1995), 'Overexpression of p21WAF1/CIP1 induces growth arrest, giant cell formation and apoptosis in human breast carcinoma cell lines.', *Oncogene* **11**(9), 1899–1905.

Simonini, P. d. S. R., Breiling, A., Gupta, N., Malekpour, M., Youns, M., Omranipour, R., Malekpour, F., Volinia, S., Croce, C. M., Najmabadi, H. et al. (2010), 'Epigenetically deregulated microRNA-375 is involved in a positive feedback loop with estrogen receptor α in breast cancer cells', *Cancer Research* **70**(22), 9175–9184.

Singh, A., Nunes, J. J. & Ateeq, B. (2015), 'Role and therapeutic potential of G-protein coupled receptors in breast cancer progression and metastases', *European Journal of Pharmacology* **763**, 178–183.

Song, R., Liu, Q., Liu, T. & Li, J. (2015), 'Connecting rules from paired miRNA and mRNA expression data sets of HCV patients to detect both inverse and positive regulatory relationships', *BMC Genomics* **16**(2), 11.

Souazé, F., Dupouy, S., Viardot-Foucault, V., Bruyneel, E., Attoub, S., Gespach, C., Gompel, A. & Forgez, P. (2006), 'Expression of neurotensin

and NT1 receptor in human breast cancer: a potential role in tumor progression', *Cancer Research* **66**(12), 6243–6249.

Stasinopoulos, I. A., Mironchik, Y., Raman, A., Wildes, F., Winnard, P. & Raman, V. (2005), 'HOXA5-twist interaction alters p53 homeostasis in breast cancer cells', *Journal of Biological Chemistry* **280**(3), 2294–2299.

Steuer, R., Kurths, J., Daub, C. O., Weise, J. & Selbig, J. (2002), 'The mutual information: detecting and evaluating dependencies between variables', *Bioinformatics* **18**(suppl_2), S231–S240.

Sumazin, P., Yang, X., Chiu, H.-S., Chung, W.-J., Iyer, A., Llobet-Navas, D., Rajbhandari, P., Bansal, M., Guarnieri, P., Silva, J. et al. (2011), 'An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma', *Cell* **147**(2), 370–381.

Swierniak, M., Wojcicka, A., Czetwertynska, M., Stachlewska, E., Maciag, M., Wiechno, W., Gornicka, B., Bogdanska, M., Koperski, L., de la Chapelle, A. et al. (2013), 'In-depth characterization of the microRNA transcriptome in normal thyroid and papillary thyroid carcinoma', *The Journal of Clinical Endocrinology & Metabolism* **98**(8), E1401–E1409.

Taherian-Fard, A., Srihari, S. & Ragan, M. A. (2014), 'Breast cancer classification: linking molecular mechanisms to disease prognosis', *Briefings in Bioinformatics* **16**(3), 461–474.

Tan, G. C., Chan, E., Molnar, A., Sarkar, R., Alexieva, D., Isa, I. M., Robinson, S., Zhang, S., Ellis, P., Langford, C. F. et al. (2014), '5 isomiR variation is of functional and evolutionary importance', *Nucleic Acids Research* **42**(14), 9424–9435.

Telonis, A. G., Loher, P., Jing, Y., Londin, E. & Rigoutsos, I. (2015), 'Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper

insights into breast cancer heterogeneity', *Nucleic Acids Research* **43**(19), 9158–9175.

Telonis, A. G., Magee, R., Loher, P., Chervoneva, I., Londin, E. & Rigoutsos, I. (2017), 'Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types', *Nucleic Acids Research* **45**(6), 2973–2985.

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B. & Farmer, J. D. (1992), 'Testing for nonlinearity in time series: the method of surrogate data', *Physica D: Nonlinear Phenomena* **58**(1-4), 77–94.

Trabelsi, S., Chabchoub, I., Ksira, I., Karmeni, N., Mama, N., Kanoun, S., Burford, A., Jury, A., Mackay, A., Popov, S. et al. (2017), 'Molecular diagnostic and prognostic subtyping of gliomas in tunisian population', *Molecular Neurobiology* **54**(4), 2381–2394.

Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J. et al. (2002), 'A gene-expression signature as a predictor of survival in breast cancer', *New England Journal of Medicine* **347**(25), 1999–2009.

Vergara, J. R. & Estévez, P. A. (2014), 'A review of feature selection methods based on mutual information', *Neural computing and applications* **24**(1), 175–186.

Vinga, S. (2013), 'Information theory applications for biological sequence analysis', *Briefings in Bioinformatics* **15**(3), 376–389.

Vogt, P. K. (2001), 'PI 3-kinase, mTOR, protein synthesis and cancer', *Trends in Molecular Medicine* **7**(11), 482–484.

Volinia, S. & Croce, C. M. (2013), 'Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer', *Proceedings of the National Academy of Sciences* **110**(18), 7413–7417.

Wang, P., Ning, S., Zhang, Y., Li, R., Ye, J., Zhao, Z., Zhi, H., Wang, T., Guo, Z. & Li, X. (2015), 'Identification of lncRNA-associated competing triplets reveals global patterns and prognostic markers for cancer', *Nucleic Acids Research* **43**(7), 3478–3489.

Wei, C., Luo, T., Zou, S. & Wu, A. (2018), 'The role of long noncoding RNAs in central nervous system and neurodegenerative diseases. front', *Behav. Neurosci* **12**, 175.

Weston, J., Elisseeff, A., Schölkopf, B. & Tipping, M. (2003), 'Use of the zero-norm with linear models and kernel methods', *Journal of Machine Learning Research* **3**(Mar), 1439–1461.

Xia, T., Liao, Q., Jiang, X., Shao, Y., Xiao, B., Xi, Y. & Guo, J. (2014), 'Long noncoding RNA associated-competing endogenous RNAs in gastric cancer', *Scientific Reports* **4**, 6088.

Xu, J., Feng, L., Han, Z., Li, Y., Wu, A., Shao, T., Ding, N., Li, L., Deng, W., Di, X. et al. (2016), 'Extensive ceRNA–ceRNA interaction networks mediated by miRNAs regulate development in multiple rhesus tissues', *Nucleic Acids Research* **44**(19), 9438–9451.

Yang, C., Wu, D., Gao, L., Liu, X., Jin, Y., Wang, D., Wang, T. & Li, X. (2016), 'Competing endogenous RNA networks in human cancer: hypothesis, validation, and perspectives.', *Oncotarget* **7**(12), 13479–13490.

Yin, L., Ge, Y., Xiao, K., Wang, X. & Quan, X. (2013), 'Feature selection for high-dimensional imbalanced data', *Neurocomputing* **105**, 3–11.

Yu, K., Lee, C. H., Tan, P. H. & Tan, P. (2004), 'Conservation of breast cancer molecular subtypes and transcriptional patterns of tumor progression across distinct ethnic populations', *Clinical Cancer Research* **10**(16), 5508–5517.

Zhang, B. & Horvath, S. (2005), 'A general framework for weighted gene co-expression network analysis', *Statistical Applications in Genetics and Molecular Biology* **4**(1).

Zhang, J., Fan, D., Jian, Z., Chen, G. G. & Lai, P. B. (2015), 'Cancer specific long noncoding rnas show differential expression patterns and competing endogenous rna potential in hepatocellular carcinoma', *PLoS One* **10**(10), e0141042.

Zhang, M.-L. & Zhou, Z.-H. (2014), 'A review on multi-label learning algorithms', *IEEE Transactions on Knowledge and Data Engineering* **26**(8), 1819–1837.

Zhang, S., Mo, Y.-y., Ghoshal, T., Wilkins, D., Chen, Y. & Zhou, Y. (2017), Novel gene selection method for breast cancer intrinsic subtypes from two large cohort study, *in* 'Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on', IEEE, pp. 2198–2203.

Zhang, Y., Dube, C., Gibert, M., Cruickshanks, N., Wang, B., Coughlan, M., Yang, Y., Setiady, I., Deveau, C., Saoud, K. et al. (2018), 'The p53 pathway in glioblastoma', *Cancers* **10**(9), 297.

Zheng, K. & Wang, X. (2018), 'Feature selection method with joint maximal information entropy between features and class', *Pattern Recognition* **77**, 20–29.