Faculty of Engineering and Information Technology
University of Technology, Sydney

Rule Mining on MicroRNA Expression Profiles for Human Disease Understanding

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

by

Renhua Song

July 2016

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

Acknowledgments

First, and foremost, I would like to express my gratitude to my chief supervisor, Assoc. Prof. Jinyan Li, and to my co-supervisors Assoc. Prof. Paul Kennedy and Assoc. Prof. Daniel Catchpoole. I am extremely grateful for all the advice and guidance so unselfishly given to me over the last three and half years by these three distinguished academics. This research would not have been possible without their high order supervision, support, assistance and leadership.

The wonderful support and assistance provided by many people during this research is very much appreciated by me and my family. I am very grateful for the help that I received from all of these people but there are some special individuals that I must thank by name.

A very sincere thank you is definitely owed to my loving husband Jing Liu, not only for his tremendous support during the last three and half years, but also his unfailing and often sorely tested patience. My deepest thanks is also extended to every member of my family, especially my little son Boao Liu, for his love, understanding and obedience. Special thanks are also owed to my parents Congchun Song and Guiyun Zhang for all their tremendous assistance and unfailing support.

My sincere appreciation and gratitude goes to Dr. Qian Liu and Dr. Yun Zheng, for all their freely given invaluable advice and insightful discussion throughout my research. My heartfelt admiration and thanks is also extended to Dr. Gyorgy Hutvagner and Dr. Hung Nguyen in the Centre for Health Technologies of UTS, Dr. Kotagiri Ramamohanarao in the University of

Melbourne, Dr. Limsoon Wong in the National University of Singapore, Dr. Tao Liu in the Children's Cancer Institute Australia. Their respective assistance in terms of biological knowledge was invaluable.

I would like to thank of my colleagues at the Advanced Analytics Institute (AAI), for their selfless support over the course of my PhD candidature. and for all the fun that we shared in the last three and half years.

Thanks and praises are also due to Almighty God, my Lord and Saviour, who has bestowed great blessing upon me throughout my life. During the last three and half years, while conducting my research and preparing to write this dissertation, I have come to realise the great relevance of the words of the apostle Saint Paul who proclaimed, "I can do all things through Christ who strengthens me".

Finally, to anyone whom I have not mentioned, please forgive me. I can most definitely assure you that you have occupied a unique and special place in my thoughts. Thank you!

Renhua Song
July 2016 @ 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

- Zheng, Y., Ji, B., **Song, R.**, Wang, S., Li, T., Zhang, X., Chen, K., Li, T. and Li, J., 2016. Accurate detection for a wide range of mutation and editing sites of microRNAs from small RNA high-throughput sequencing profiles. Nucleic acids research, p.gkw471. `http://dx.doi.org/10.1093/nar/gkw471`

- Tee, A.E., Liu, B., **Song, R.**, Li, J., Pasquier, E., Cheung, B.B., Jiang, C., Marshall, G.M., Haber, M., Norris, M.D. and Fletcher, J.I., 2016. The long noncoding RNA MALAT1 promotes tumor-driven angiogenesis by up-regulating pro-angiogenic gene expression. Oncotarget, `http://dx.doi.org/10.18632/oncotarget.6675`

- Tee, A.E., Liu, P., Maag, J., **Song, R.**, Li, J., Cheung, B.B., Haber, M., Norris, M.D., Marshall, G.M., Dinger, M. and Liu, T., 2015. The long noncoding RNA MALAT1 promotes hypoxia-driven angiogenesis by upregulating pro-angiogenic gene expression in neuroblastoma cells. Cancer Research, 75(Supplement 15), pp.146-146

- Liu, Q., **Song, R.** and Li, J., 2015. Inference of gene interaction networks using conserved subsequential patterns from multiple time

course gene expression datasets. BMC Genomics, 16(Supplement 12): S4

- **Song, R.**, Catchpoole, D.R., Kennedy, P.J. and Li, J., 2015. Identification of lung cancer miRNAmiRNA co-regulation networks through a progressive data refining approach. Journal of Theoretical Biology, 380, pp.271-279

- **Song, R.**, Liu, Q., Liu, T. and 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(Supplement 2): S11

- **Song, R.**, Liu, Q., Hutvagner, G., Nguyen, H., Ramamohanarao, K., Wong, L. and Li, J., 2014. Rule discovery and distance separation to detect reliable miRNA biomarkers for the diagnosis of lung squamous cell carcinoma. BMC Genomics, 15(Supplement 9): S16

Conference Papers Presented

- Liu, Q., **Song, R.** and Li, J. "Inference of gene interaction networks using conserved subsequential patterns from multiple time course gene expression datasets." `www.jsbi.org/giw-incob2015/`, Tokyo, Japan. **Oral Presentation**

- **Song, R.**, Liu, Q., Liu, T. and Li, J. "Connecting rules from paired miRNA and mRNA expression data sets of HCV patients to detect both inverse and positive regulatory relationships." `http://apbc2015.mbc.nctu.edu.tw/`, HsinChu, Taiwan. **Oral Presentation.**

- **Song, R.**, Liu, Q., Hutvagner, G., Nguyen, H., Ramamohanarao, K., Wong, L. and Li, J. "Rule discovery and distance separation to detect reliable miRNA biomarkers for the diagnosis of lung squamous cell carcinoma."`http://incob.apbionet.org/incob14/`, Sydney, Australia. **Oral and Poster Presentation.**

Papers to be Submitted/Under Review

- **Song R.**, Zheng, Y., Catchpoole, D.R., Kennedy, P.J. and Li, J. "A Novel Framework for Inferring Self-regulation miRNAs." **Under Revision**

Abstract

This research employs rule mining methods to study the important roles of miRNAs in human diseases. From past experience and from reviewing the literature, rule mining is a widely used data mining technique for the discovery of interesting relationships in large data sets. MicroRNAs (miRNAs) are endogenous and highly conserved non-coding RNA molecules. They can inhibit and/or promote the post-transcriptional expression of target messenger RNAs (mRNAs). miRNAs thus play a pivotal role in a cell's differentiation, proliferation, growth, mobility, and apoptosis, as well as in viral replication and proliferation. This has inspired many research works aimed at detecting miRNAs' functions in human disease. However, with the current deluge of miRNA data, previous works have suffered from limitations in terms of handling the relationship between various molecules. Firstly, they usually identify single miRNAs as biomarkers, and always produce low sensitivity and specificity. Secondly, intensive research largely depends on the inverse expression relationships between miRNAs and mRNAs to discover miRNA-mRNA regulatory modules. Finally, the miRNA-miRNA co-regulations and miRNA self-regulations have not been well investigated. As a result, rule mining is a powerful new technology with great potential to help researchers focus on the most important miRNAs for understanding human diseases. This thesis reports our past and current research outcomes in this area. The contributions of the thesis are as follows:

- A novel rule mining method is proposed to detect the significant miRNA biomarkers.

- A "change to change" method is proposed to mine both positive and negative regulatory relationships from paired miRNA and mRNA expression data sets.
- A progressive data refining approach is proposed to identify the lung cancer miRNA-miRNA co-regulation network.
- A novel framework is proposed to detect the self-regulation miRNAs.

The research was conducted through four case studies. (1) The first case study was on lung squamous cell carcinoma for accurate diagnosis of this disease through the reliable miRNA biomarkers identified by a novel rule discovery method. (2) The second case study was on paired miRNA and mRNA expression data of HCV patients to detect both positive and negative regulatory modules. (3) The third case study was on lung cancer data sets for the computational methods to identify miRNA-miRNA co-regulation networks and miRNA-miRNA co-regulatory relationships. (4) The fourth case study was on multiple data types to infer self-regulation miRNAs in humans through an integrative rule mining framework and approach. All the results have been verified by the existing literature and databases.

Chapter 1

Introduction

1.1 Background

In this thesis, we mainly focus on designing rule mining methods to study microRNAs (miRNAs) and their functions for helping fight against human diseases such as lung cancer, hepatitis C virus (HCV) and leukemia. In biological systems, miRNAs always work with other molecules, mostly with messenger RNAs (mRNAs) and Transcription Factors (TFs). Accordingly, this section thus briefly introduces rule mining, the related studies of the three types of molecules and their relationships, and the studied human diseases.

1.1.1 Rule Mining

Rule mining is a well-known data mining technique which is widely used for discovery of interesting relations in large data sets. Mining rules from a date set is a challenging problem that has attracted considerable interest because a rule provides a concise statement of potentially useful information that is easily understood. The original motivation for seeking strong rules came from the need to analyse supermarket transaction data to examine customer behaviour in terms of the purchased products (Brin, Motwani, Ullman & Tsur 1997).

With the current deluge of biological data, one of the central problems in biological knowledge discovery is the development of good measures of interestingness of discovered patterns. With such measures, a biological expert needs to manually examine only the more interesting rules, instead of each of a large number of mined rules. Therefore, rule mining methods have become indispensable to biological investigations. Rule mining can be developed for the analysis of a wide range of biological data including miRNA expression data. It aims to discover frequent patterns in data sets using some measures of interestingness. Patterns in the data can be represented in many different forms, including units of knowledge called rules. Each rule has a form:

If *set of conditions* **then** *action*.

The left side and also the right side of the rule may involve a single attribute value or a conjunction of attributes values and their domains of different attributes.

1.1.2 microRNAs, mRNAs and TFs, and their Relationships

microRNAs and their Roles

microRNAs (miRNAs) were first discovered in 1993 by Lee et al. (Lee, Feinbaum & Ambros 1993) during a study of the gene lin-14 in C. elegans development. miRNAs are a class of small (19-25 nucleotides) and endogenous non-coding RNAs. Until the early 2000s, miRNAs were identified as a distinct class of biological regulators in gene regulation with conserved functions. A miRNA is complementary to a part of one or more mRNAs. miRNAs can regulate gene expression at a post-transcriptional stage, and can control fundamental cellular processes such as differentiation, cell growth, proliferation and apoptosis (He & Hannon 2004). miRNAs have the potential to regulate at least 20-30% of all human transcripts (Calin & Croce 2006). They have also been shown to control the expression of oncogenes and

tumour-suppressor genes (Zhang, Pan, Cobb & Anderson 2007*a*). The human genome may encode over 1000 miRNAs, targetting about 60% of mammalian genes and it is abundant in many human cell types.

Human miRNA biogenesis is a multiple-step process. miRNA origin affects the nuclear pathway to classify intergenic miRNA and coding-intronic miRNA. A intergenic miRNA gene is first transcribed to a primary miRNA (pri-miRNA) by Polymerase (Pol) II enzyme (Lee, Kim, Han, Yeom, Lee, Baek & Kim 2004).

miRNAs that are organised in clusters have the same transcriptional regulation (Altuvia, Landgraf, Lithwick, Elefant, Pfeffer, Aravin, Brownstein, Tuschl & Margalit 2005), because they form the same long precursor transcript, and this is then cleaved to a stem loop intermediate termed miRNA precursor (pre-miRNA) consisting of a single-stranded RNA molecule by Drosha RNase III endonuclease in animals (Lee, Jeon, Lee, Kim & Kim 2002).

In contrast, miRNA located within an intron of a protein coding gene is transcribed by pol II as part of the pre-mRNA. Finally, pre-miRNAs are further exported to the cytoplasm by Exportin-5 and the loop is cleaved by Dicer, another RNAse III enzyme that releases a double stranded RNA miRNA:miRNA* (the opposed sequence of mature miRNA at the stem arm of the secondary structure), and mature miRNAs are released for regulating targeted gene expression (Bartel 2004*a*).

The characteristics of miRNAs make them play an key role in many diseases including cancer, cardiovascular disease, and immune disorders. miRNA expression profiles can be used to distinguish normal cells from disease cells in patients. Biological markers (biomarkers) referring to a measured characteristic are widely used as indicators of some biological state or condition, because they are often measured and evaluated to examine normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (shown in Figure 1.1). The inherent stability of miRNAs makes them an ideal candidate for biomarkers, which

can classify human cancers (Lu, Getz, Miska, Alvarez-Saavedra, Lamb, Peck, Sweet-Cordero, Ebert, Mak & Ferrando 2005).

Exploring miRNA functions is important for diagnostics and therapeutics. Aberrant miRNA expressions have been linked to many diseases, and have recently been intensively investigated to discover miRNA biomarkers for the diagnosis of diseases including lung cancer (Raponi, Dossey, Jatkoe, Wu, Chen, Fan & Beer 2009, Shen, Todd, Zhang, Yu, Lingxiao, Mei, Guarnera, Liao, Chou & Lu 2010, Tan, Qin, Zhang, Hang, Li, Zhang, Wan, Zhou, Shao & Sun 2011*a*). The inherent stability of miRNAs in serum and the reliability and reproducibility of expression analysis (Alevizos, Alexander, Turner & Illei 2011, Gilad, Meiri, Yogev, Benjamin, Lebanony, Yerushalmi, Benjamin, Kushnir, Cholakh & Melamed 2008, Ludwig & Weinstein 2005, Mitchell, Parkin, Kroh, Fritz, Wyman, Pogosova-Agadjanyan, Peterson, Noteboom, O'Briant & Allen 2008, Mraz, Malinova, Mayer & Pospisilova 2009) make them ideal candidates for biomarkers (Alevizos et al. 2011, Bartels & Tsongalis 2009, Shen et al. 2010, Tan, Qin, Zhang, Hang, Li, Zhang, Wan, Zhou, Shao & Sun 2011*a*, Yang, Li, Yang, Wang, Zhou, Jiang, Ma & Wang 2010, Yu, Todd, Xing, Xie, Zhang, Liu, Fang, Zhang, Katz & Jiang 2010).

The function of miRNAs is shown in gene regulation. miRNAs regulate one or more mRNAs mainly via two main mechanisms: target mRNA cleavage and 'translational repression' (Carrington & Ambros 2003). In plants, miRNAs are usually complementary to coding regions of mRNAs and the perfect or near perfect base pairing with the target RNAs promotes cleavage of the RNAs. In animals, the 5' miRNA region is usually complementary to a site in the 3' UTR of the target site with imperfect base pairing. This 5' miRNA region (nucleotides 2-7) is called the 'seed region' which is short. So miRNAs are predicted to regulate large numbers of genes. Moreover, animal miRNAs may initially block protein translation of the target mRNA.

Cancer is a multistage process in which normal cells experience genetic

changes that progress them through a series of pre-malignant states into invasive cancer that can spread throughout the body. The dysregulation of genes involved in cell proliferation, differentiation and apoptosis has a close relationship with cancer initiation and progression.

Genes linked with cancer development are characterised as oncogenes and tumour suppressors. miRNAs as oncogenes and tumour suppressors play a vital role in the regulation of numerous metabolic and cellular pathways by controlling cell proliferation, differentiation and survival (Zhang, Pan, Cobb & Anderson 2007*b*).

Messenger RNAs and their Roles

Messenger RNAs (mRNAs) (Dreyfuss, Kim & Kataoka 2002) are a large family of RNA molecules. mRNA is first transcribed from DNA by RNA polymerase, and then translated into a polymer of amino acids and a protein, with each sequence of three nitrogen-containing bases in the mRNA specifying the incorporation of a particular amino acid within a protein. A mRNA encodes a protein (or more than one protein in bacteria). mRNAs promote the amino acid sequence of the protein products of gene expression. They can transport genetic information from DNA in the nucleus to the sites of protein synthesis in the ribosome. Finally, proteins are the leader actors within the cell, carrying out the duties specified by the information encoded in genes.

Transcription Factors and their Roles

In molecular biology and genetics, Transcription Factors (TFs) (Wang, Lu, Qiu & Cui 2010) are proteins involved in the process of converting, or transcribing, DNA into RNA. TFs include a wide number of proteins, excluding RNA polymerase, that initiate and regulate the transcription of genes. A distinct characteristic of TFs is that they can bind to specific DNA sequences called enhancer or promoter sequences of DNA adjacent to the genes that they regulate. In addition, TFs can carry out this function

alone or with other proteins in a complex. Therefore, some TFs bind to a DNA promoter sequence near the transcription start site and help form the transcription initiation complex. Other TFs bind to regulatory sequences, such as repressor sequences, and can block transcription of the related gene.

Relationships of miRNAs, mRNAs and TFs

miRNAs play important regulatory roles via the RNA-interference pathway by targetting mRNAs for cleavage or translational repression. Accumulating studies demonstrate that complex diseases may arise from cooperative effects of multiple dysfunctional miRNAs or systematical function of miRNAs. Thus, identifying abnormal functions which are cooperatively regulated by multiple miRNAs is very useful for understanding the pathogenesis of complex diseases. miRNAs affect the stability and translational efficiency of target mRNAs by binding to their 3' untranslated regions (UTRs) to inhibit expression (Hobert 2008*a*). These direct effects are amplified by modulation of gene transcription pathways. A miRNA can have many target mRNAs, and a mRNA can be regulated by multiple miRNAs, forming complicated many-to-many regulatory modules between miRNAs and mRNAs (Filipowicz, Bhattacharyya & Sonenberg 2008).

Consequently, indirect mRNA modulatory effects of miRNAs to increase or decrease mRNAs greatly outnumber direct target suppressions, because among the miRNAs predicted to target mRNAs, many are transcription factors. Therefore, the variation of a certain miRNA affects the expression of the transcription factors that in turn regulate the transcription of correlated miRNAs, forming miRNA-mediate-miRNA regulatory modules.

Moreover, miRNAs act in transcription to modulate protein expression at the post-transcriptional level and can be considered in terms of post-transcription factors. TFs act in transcription to modulate protein expression at transcriptional level. Therefore, cell phenotype is the result of two distinct but similar mechanisms that affect gene expression at two different levels. Chen *et al.* compared the evolution of transcriptional regulation and post-

transcriptional regulation that is mediated by microRNAs, in plants and animals, paying attention to the evolution of the individual regulators and their binding sites (Chen & Rajewsky 2007), forming TF-mRNA-miRNA networks by analysing multiple genome profiles simultaneously.

1.1.3 Human Disease Studied by this Work

Altered expression profiles of miRNAs are linked to many diseases including lung cancer. This study focus on two types of disease (lung cancer and HCV infection) known to be associated with miRNA deregulation.

Lung Cancer

Lung cancer occurs when abnormal cells in one or both lungs grow in an uncontrolled cell growth way. Lung cancer is the biggest cancer killer in Australia. Lung cancer is often diagnosed at a late stage with poor prognosis (Jemal, Siegel, Ward, Murray, Xu, Smigal & Thun 2006, Minna, Roth & Gazdar 2002); and it is also the leading cause of cancer-related deaths worldwide (Minna et al. 2002). According to the National Cancer Institute, by the end of 2012 there were 226,160 new lung cancer diagnoses and 160,340 lung-cancer related deaths in the USA. According to the World Health Organisation (WHO); cancer is the cause of 13% of all global deaths (Judice & Geetha 2013). Lung cancer can be broadly classified into two main types based on the cancer's appearance under a microscope: non-small cell lung cancer and small cell lung cancer. Non-small cell lung cancer (NSCLC) accounts for 80% of lung cancers, while small cell lung cancer accounts for the remaining 20%. The ability to diagnose early-stage lung cancer patients is vital for improving their survival rate of these patients. The Chest X-ray has been applied for its early detection, but it has low sensitivity (Fontana, Sanderson, Taylor, Woolner, Miller, Muhm & Uhlenhopp 1984, Frost, Ball Jr, Levin, Tockman, Baker, Carter, Eggleston, Erozan, Gupta & Khouri 1984, Liu, Li & Tsykin 2009*b*).

Hepatitis C Virus

Hepatitis C virus (HCV) is a positive sense single-stranded RNA Hepacivirus in the family of Flaviviridae (Jopling, Yi, Lancaster, Lemon & Sarnow 2005). HCV is capable of infecting the human liver to develop a contagious and potentially life-threatening liver disease, Hepatitis C. It is estimated that HCV has infected an approximately 170 million people worldwide (He, Tan, Tareen, Vijaysri, Langland, Jacobs & Katze 2001), thereby causing a serious public health problem. Treatment for Hepatitis C patients is on the cutting edge of medicine. However, the treatment effect is not good. In fact, the most commonly used antiviral combination of pegylated interferon (IFN) and ribavirin (Su, Pezacki, Wodicka, Brideau, Supekova, Thimme, Wieland, Bukh, Purcell & Schultz 2002) achieves a sustained virological response for only 55% of the patients (Murakami, Aly, Tajima, Inoue & Shimotohno 2009). Another two agents Boceprevir and Telaprevir inhibiting non-structural protein 3 (NS3) protease in HCV were newly approved in 2011 but with uncertain effect. No vaccine is available against HCV infection (Wilby, Partovi, Ford, Greanya & Yoshida 2012).

1.2 Research Questions

miRNAs have many important features and functions:

- Target 3' untranslated regions of mRNAs
- Regulate post-transcriptional genes for degradation
- Target 1-3% of all eukaryotic genes
- Regulate 30% of protein-coding genes
- Are involved in many physiological processes

Dysregulation of miRNA expression profiles has been demonstrated in most tumours, implying that miRNAs may be involved in the development

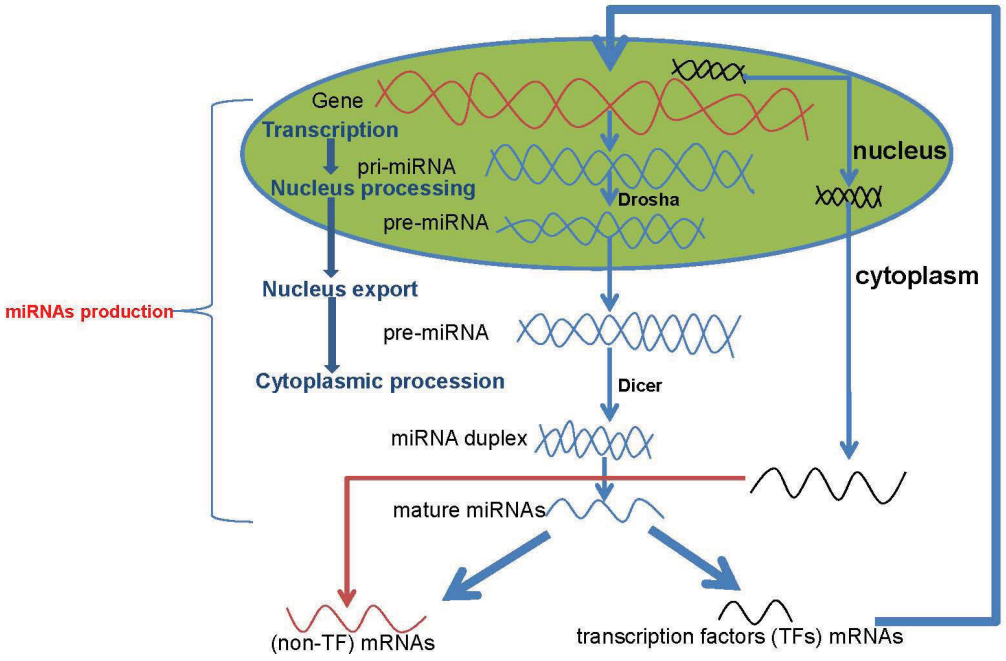


Figure 1.1: **Gene regulation of miRNAs and TFs.** miRNAs regulate biological processes in proliferation, metabolism, differentiation, development, apoptosis, cellular signaling and even cancer development and progression. TFs are fundamental players of gene expression regulation at the transcriptional level. miRNAs regulate target gene expression at the post-transcriptional level.

of cancer and other diseases. Classification is critical to successful treatment and sufficient biomarkers are quite important. Early stage detection and treatment can control disease progression. Therefore, accessible, reliable and non-invasive biomarkers can be medically valuable and can provide some relevant insights into disease biology. We have received evidence that recent works usually identified the individual miRNAs as biomarkers.

miRNAs can bind to partially complementary sites in the 3' untranslated regions of target genes, and regulate protein production of the target transcript. Different combinations of miRNAs are expressed in different cell types and may coordinately regulate cell-specific target genes. However, the miRNA-mRNA regulatory modules are often based on an inverse relationship between miRNAs and mRNAs.

miRNAs and mRNAs constitute an important part of gene regulatory networks, influencing diverse biological phenomena. miRNAs are widely believed to regulate complementary mRNA targets. Co-regulation analysis of multiple miRNAs is useful for understanding complex post-transcriptional regulations in humans. Complex diseases are associated with several miRNAs rather than a single miRNA. It is still a challenging work to discover co-regulation miRNAs and self-regulation in cancers at a systematic level, which are widely neglected.

Based on these facts, how do we identify the significant miRNA biomarkers associated with prognosis, diagnosis and progression in cancers? In addition, the systematical function of miRNA in human diseases, i.e. miRNA-mRNA regulatory modules, is also important. The post-transcriptional and transcriptional regulation in human diseases, co-regulation and self-regulation miRNAs are still under intensive investigation.

Rule mining is a data mining technique that is used to find associations between two or more random variables. It has been used extensively in relational databases. Therefore, we can discover reliable miRNA biomarkers, regulatory modules and networks.

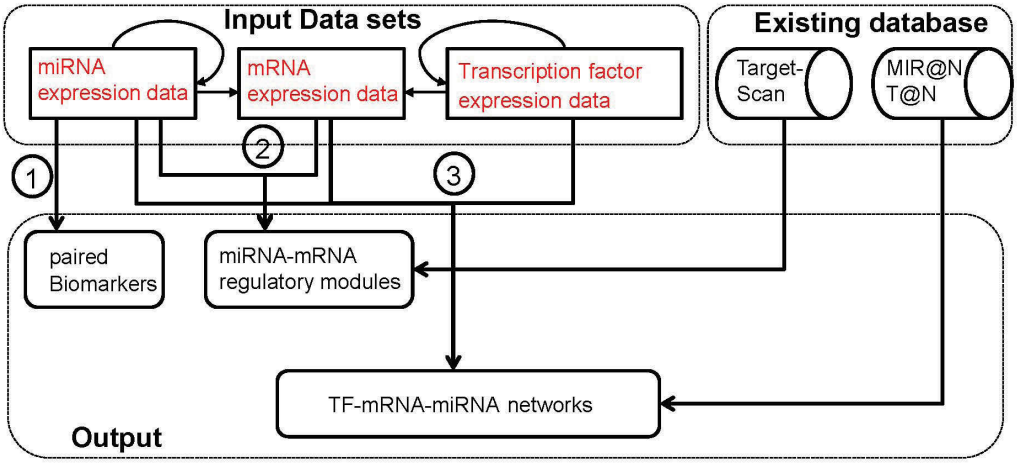


Figure 1.2: **A graphical description framework of research plan.** ①: to identify paired biomarkers instead of individual miRNA biomarkers. ②: to discover both positive and negative miRNA-mRNA regulatory modules, and miRNA-miRNA co-regulation modules. ③: to study the relationships among miRNAs, mRNAs and TFs.

By taking all of the above aspects into account, from different levels and heterogeneous data sources, a graphical description framework of data mining on mRNA expression profiles for human disease understanding can be specifically summarised in Figure 1.2. The thesis mainly focuses on two research questions by applying a rule mining approach in this domain:
1. How do we identify the significant miRNA biomarkers associated with prognosis, diagnosis and progression in cancers?
2. How do we identify the uncovered systematical functions of miRNAs in human disease?

1.3 Research Contributions

The purpose of this thesis is to study miRNA expression profiles in understanding human disease using rule mining methods. Considering the characteristics of miRNAs in human disease, my research topic targets the

following objectives:

- **Contribution 1: Rule Discovery and Distance Separation to Detect microRNA Biomarkers for SCC Diagnosis**
  Chapter 4 presents a rule mining method to detect 2- and 3-miRNA groups, together with specific expression ranges of these miRNAs, to form simple linear discriminant rules for biomarker identification and biological interpretation. Our method is based on a novel committee of decision trees to derive 2- and 3-miRNA 100%-frequency rules. This method is applied to a data set of lung miRNA expression profiles of 61 squamous cell carcinoma (SCC) samples and 10 normal tissue samples. A distance separation technique is used to select the most reliable rules which are then evaluated on a large independent data set. The results indicate that rule discovery followed by distance separation is a powerful computational method to identify reliable miRNA biomarkers. The visualization of the rules and the clear separation between the normal and cancer samples by our rules will help biology experts for their analysis and biological interpretation.

- **Contribution 2: Rule Discovery for Detecting Both Inverse and Positive miRNA-mRNA Regulations in HCV Patients**
  Chapter 5 presents a 'change-to-change' method to detect both inverse and positive regulatory relationships from a paired miRNA and mRNA expression data set of HCV patients. Our study uncovered many novel miRNA-mRNA regulatory modules. We followed the biological principle that inverse expression relationships and positively regulated miRNA-mRNA pairs can both exist in many-to-many regulatory modules. We detected 100%-frequency rules from the most differentially expressed miRNAs and then mined 100%-frequency rules from the relevant target mRNAs expression data for each miRNA rule. We integrated the miRNA rules and their mRNA rules to construct miRNA-mRNA regulatory modules. Many detected miRNAs and mRNAs can be supported by recent work in the literature. We

also detected novel positive and inverse regulatory relationships. The detected miRNA-mRNA regulatory modules will provide new insights into the regulation of host responses and the pathogenesis of HCV infection. We conclude that our rule discovery method is useful for integrating binding information and expression profile for identifying HCV miRNA-mRNA regulatory modules and can be applied to the study of the expression profiles of other complex human diseases.

- **Contribution 3: Identification of lung cancer miRNA-miRNA Co-regulation network through a refining approach**
  Chapter 6 presents a novel integrative approach to the discovery of miRNA-miRNA co-regulating networks which can progressively refine various data and computational analysis results. Applied to three lung cancer miRNA expression data sets of different subtypes, our method has identified a miRNA-miRNA co-regulating network and co-regulating functional modules common to lung cancer. We find that the co-regulating network is scale free and that lung cancer related miRNAs have more synergism in the network. We also confirm that known lung cancer related miRNAs have more synergism than lung cancer unrelated miRNAs. Kyoto encyclopedia of genes and genomes (KEGG) pathway enrichment analysis and transcription factor analysis have all demonstrated the biological relevance of the miRNA-miRNA co-regulation network to lung cancer. According to our literature survey and database validation, many of the results are biologically meaningful for understanding the mechanism of the complex post-transcriptional regulations in lung cancer.

- **Contribution 4: A Novel Framework for Inferring Self-regulation miRNAs for Understanding their Mechanisms**
  Chapter 7 presents a study of self-regulating miRNAs through combining the relationships among miRNAs, TFs and target genes. We design a novel framework (called SRmiR) to integrate multiple data types for

exploring self-regulated miRNAs for understanding their mechanisms. Particularly, SRmiR is aimed at discovering the self-regulation miRNAs specific to humans, by using heterogeneous data, including miRNAs, mRNAs and TFs. We define a self-regulated miRNA if the miRNA regulates a TF and together with this thee is a TF-target interaction with one of more target genes. Firstly, we collected human miRNAs from miRBase and obtained the promoter regions of all the miRNA primary transcripts. Secondly, we collected ChIP-seq datasets representing unique regulatory transcription factors from ENCODE at UCSC. Thirdly, we discovered the potential miRNA-TF relationships between TFs and miRNAs by comparing the miRNAs' promoter regions and transcription factor binding sites (TFBS). After that, we also obtained the miRNA-target relationships between miRNAs and genes, and TF-gene relationships bewteen TFs and genes based on the miRNA-TF relationships. We also discussed the FFL involving these genes as Transcription Factors and targets.

1.4 Thesis Structure

The thesis is structured (Figure 1.3) as follows:

Chapter 1 introduces the background of the whole thesis starting with a brief discussion on the background and research contributions regarding the rule mining on miRNA expression profiles for human disease understanding. Finally, It illustrates the structure of the thesis for ease of reading and understanding. **Chapter 2** provides the literature review of various miRNA biomarker studies, including microarray-based and statistics-based methods. In addition, we reviewed the methods for discovering the relationships between miRNAs and mRNAs. Lastly, the miRNAs' co-regulation and miRNA-mRNA-TF regulations are reviewed. It outlines the research questions and the limitations of the existing methods. **Chapter 3** describes our proposed rule mining methods that were specifically constructed to study

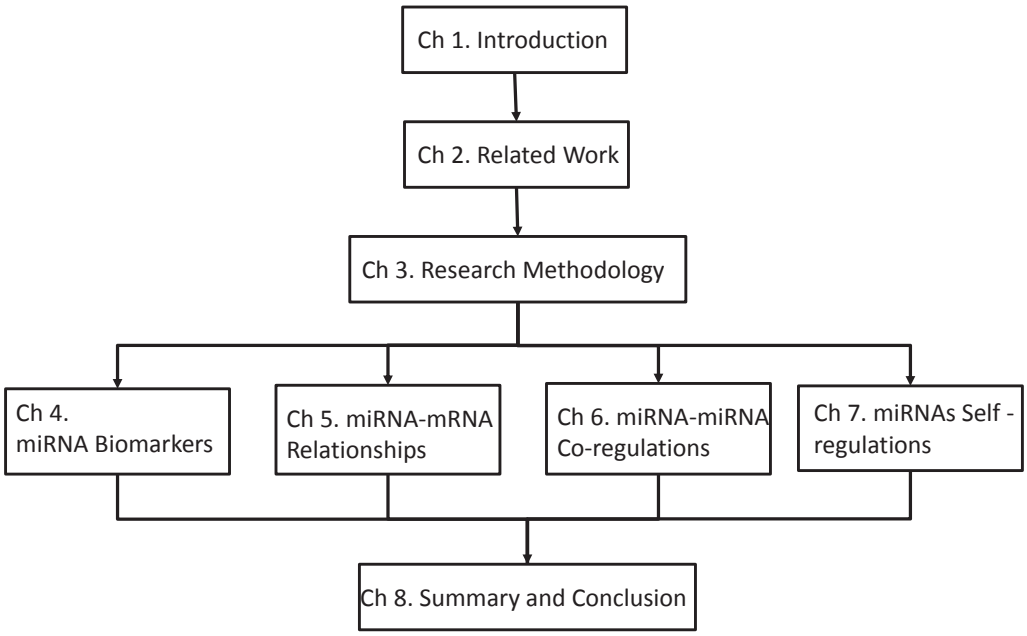


Figure 1.3: **Thesis Structure.** Ch. 1 introduces the research background. Ch. 2 provides the literature review. Ch. 3 describes research methodology. Ch. 4 presents a novel method to miRNA biomarkers for SCC diagnosis. Ch. 5 presents a 'change-to-change' method to detect both inverse and positive regulatory relationships from a paired miRNA and mRNA expression data set of HCV patients. Ch. 6 presents a novel integrative approach to the discovery of miRNA-miRNA co-regulating networks. Ch. 7 presents a study of self-regulating miRNAs through combining the relationships among miRNAs, TFs and target genes. Ch. 8 provides a final summary of this research and also suggests some future directions.

the miRNA expression profiles. It also explains the basic knowledge, the computational methods, bioinformatics methods and accuracy measurement. **Chapter 4** Proposed a novel rule to discover reliable miRNA biomarkers for cancers, and developed a novel approach to find the minimal number of miRNAs that can be used to distinguish between healthy and cancer tissue samples. **Chapter 5** proposes a "change to change" method to derive discriminatory rules for detecting both inverse and positive regulatory relationships. Specifically, rules from paired miRNA and mRNA expression data of human disease samples and controls are connected to identify the many-to-many miRNA-mRNA regulatory modules involved in cancers. **Chapter 6** designs an integrative computational method to identify a miRNA-miRNA co-regulation network common to three lung cancer miRNA expression data sets of different subtypes. **Chapter 7** proposes a robust methodology for mining big regulatory modules especially the self-regulation miRNAs in the pre-and post-transcriptional level from miRNAs, mRNAs and TFs sequence data. **Chapter 8** provides a final summary of this research and also suggests some future directions.

Chapter 2

Related Work

This chapter introduces the related work. The miRNA biomarkers studies are introduced in Section 1, and then current miRNA-mRNA regulation relationship studies are introduced in Section 2. Section 3 reviews miRNA-miRNA co-regulation study. Section 4 introduces the relationship between miRNAs, TFs and target genes. Section 5 describes the limitation of existing methods. A summary is shown in the last section.

2.1 miRNA Biomarkers

A biomarker can be a substance that is introduced into an organism as a way to examine organ function or other aspects of health (Dimri, Lee, Basile, Acosta, Scott, Roskelley, Medrano, Linskens, Rubelj & Pereira-Smith 1995). The perfect candidate marker has to overcome the insufficient sensitivity, specificity, robustness and low predictive power. Furthermore, its detection can indicate a particular disease state especially cancers, for example, the presence of an antibody may indicate an infection (Issaq, Waybright & Veenstra 2011).

More specifically, a biomarker indicates a change in expression or state of a protein that correlates with the risk or progression of a disease, or with the susceptibility of the disease to a given treatment. Genomic and proteomic

technologies have significantly increased the number of potential DNA, RNA and miRNA biomarkers under study (Hennessey, Sanford, Choudhary, Mydlarz, Brown, Adai, Ochs, Ahrendt, Mambo & Califano 2012, Tan, Qin, Zhang, Hang, Li, Zhang, Wan, Zhou, Shao & Sun 2011*a*). In this study, we made a comparison of DNA biomarkers, RNA biomarkers and miRNA biomarkers for cancer research. Finally, we chose to pay more attention to the miRNA biomarkers due to their characteristics: highly abundant, stable and quantifiable 1.1.2.

DNA Biomarkers: Circulating DNA and tumour cells were among the first markers evaluated for cancer staging. Increased serum DNA concentrations are associated with cancer and with other conditions such as sepsis and autoimmune disease (Calin & Croce 2006).

RNA Biomarkers: However, most DNA markers are evaluated individually. Many high-throughput technologies have been developed to assess mRNA expression comprehensively (Alevizos et al. 2011).

miRNA Biomarkers: As a single miRNA can regulate hundreds of genes and may act as a master regulator of processes, selected subsets of miRNAs can be used as biomarkers of physiologic and pathologic states. A recent study showed that the expression of as few as two miRNAs could accurately discriminate acute lymphoid from acute myeloid leukemia (Bartels & Tsongalis 2009).

Another feature that makes miRNAs excellent candidates for biomarker studies is their remarkable stability and resistance to degradation, especially compared with mRNA.

Biological experts have been able to isolate miRNA from archived clinical specimens, including urine, saliva and formalin-fixed paraffin embedded tissues (Tan, Qin, Zhang, Hang, Li, Zhang, Wan, Zhou, Shao & Sun 2011*a*). Since miRNAs are often highly conserved, they could be advantageous for practical applications in further research.

2.1.1 Identification of miRNA Biomarkers by qRT-PCR

As miRNAs are promising biomarker candidates, more and more studies have focused on identifying the significantly and differentially expressed miRNAs as biomarkers for cancers diagnosis and prognosis (Kosaka, Iguchi & Ochiya 2010). Accordingly, we will introduce some studies on miRNA biomarkers for cancer diagnosis and prognosis by qRT-PCR. For example, in 2008, miRNAs are introduced as a new class of biomarkers for cancer by Nakasa et al. (Nakasa, Miyaki, Okubo, Hashimoto, Nishida, Ochi & Asahara 2008). In this work, serum miRNAs were purified from patient serum and selected miRNAs were quantified by the Taqman-based real-time polymerase chain reaction (PCR).

Furthermore, Chen et al. (Chen, Ba, Ma, Cai, Yin, Wang, Guo, Zhang, Chen, Guo et al. 2008) identified specific expression patterns of serum miRNAs for lung cancer, providing evidence that serum miRNAs contain fingerprints for various diseases. They used Solexa (Bentley, Balasubramanian, Swerdlow, Smith, Milton, Brown, Hall, Evers, Barnes, Bignell et al. 2008) to validate two non-small cell lung cancer-specific serum miRNAs in an independent trial of 75 healthy donors and 152 cancer patients. miRNAs were demonstrated to be robust and therefore were viewed as improved biomarker for several diseases. In addition to robustness, miRNAs are also detectable in almost all body fluids and excretions, so they can serve to provide a new set of diagnostic tools for a variety of diseases.

In order to diagnose NSCLC, particularly at an early stage, Xie et al. (Xie, Todd, Liu, Zhan, Fang, Peng, Alattar, Deepak, Stass & Jiang 2010) firstly identified 12 miRNAs (miRNAs) the aberrant expressions of which in primary lung tumours are associated with early-stage NSCLC. After that, they extended the previous research by investigating whether the miRNAs could be used as potential plasma biomarkers for NSCLC. They used the real-time quantitative reverse transcription PCR, and then evaluated the diagnostic value of the plasma miRNAs in a cohort of 58 NSCLC patients and 29 healthy

individuals. The altered miRNA expressions were reproducibly confirmed in the tumour tissues and were stably present and reliably measurable in plasma (Taylor & Gercel-Taylor 2008). Finally, of the 12 miRNAs, five displayed significant concordance of the expression levels in plasma and the corresponding tumour tissues. The study shows that altered expressions of the miRNAs in plasma can serve to provide potential blood-based biomarkers for NSCLC (Shen et al. 2010).

miRNAs are reported to be present in the blood of humans and have been increasingly suggested as biomarkers for disease. Wang et al. (Wang, Zhu, Zhang, Li, Li, He, Qin & Jing 2010) discovered that cardiac-specific miR-208a in plasma might be a novel biomarker for early diagnosis of myocardial injury in humans.

Furthermore, many specific miRNAs are identified for various diseases. (Ai, Zhang, Li, Pu, Lu, Jiao, Li, Yu, Li, Wang et al. 2010) elevated miRNA-1 as a potential novel biomarker for acute myocardial infarction. MiR423-5p was has been viewed as a circulating biomarker for heart failure (Tijsen, Creemers, Moerland, de Windt, van der Wal, Kok & Pinto 2010). miRNA miR-155 has been viewed as a biomarker for early pancreatic neoplasia (Habbe, Koorstra, Mendell, Offerhaus, Ryu, Feldmann, Mullendore, Goggins, Hong & Maitra 2009). Table 2.1 shows the dysregulated miRNAs in some diseases.

Several improved methods for the identification of miRNA biomarkers are urgently needed to decrease the morbidity and mortality caused by other diseases. miRNAs are frequently dysregulated in cancer and have shown promise as markers for prostate cancer diagnosis and prognosis. Mitchell et al. (Mitchell, Parkin, Kroh, Fritz, Wyman, Pogosova-Agadjanyan, Peterson, Noteboom, O'Briant, Allen et al. 2008) used biological analysis method qRT-PCR for human plasma and serum samples from healthy donors or patients with cancer, and discovered that miR-141 (a miRNA expressed in prostate cancer) can distinguish patients with prostate cancer from healthy controls. Their results established the measurement of tumour-derived miRNAs in serum or plasma as a key approach for the blood-based detection of human

cancers.

Recent evidence has indicated that miRNAs circulate in a stable, cell-free form in the bloodstream and that an abundance of specific miRNAs in plasma or serum can serve as biomarkers of cancer and other diseases (Iorio, Ferracin, Liu, Veronese, Spizzo, Sabbioni, Magri, Pedriali, Fabbri, Campiglio et al. 2005, Iorio, Visone, Di Leva, Donati, Petrocca, Casalini, Taccioli, Volinia, Liu, Alder et al. 2007, Dahiya, Sherman-Baust, Wang, Davidson, Shih, Zhang, Wood III, Becker & Morin 2008, Calin, Dumitru, Shimizu, Bichi, Zupo, Noch, Aldler, Rattan, Keating, Rai et al. 2002, Lowery, Miller, McNeill & Kerin 2008). Measurement of circulating miRNAs as biomarkers is associated with some special challenges, including those related to pre-analytic variation and data normalisation.

Heneghan et al. (Heneghan, Miller, Lowery, Sweeney & Kerin 2009) presented a comprehensive and timely review of the role of miRNAs in cancer: addressing miRNA function, their putative role as oncogenes or tumour suppressors, with a particular emphasis on breast cancer. They described the potential role of miRNAs in breast cancer management, particularly in improving current prognostic tools and achieving the goal of individualized cancer treatment. Mattie et al. (Mattie, Benz, Bowers, Sensinger, Wong, Scott, Fedele, Ginzinger, Getts & Haqq 2006) demonstrated that optimised high-throughput miRNA expression profiling offers novel biomarker identification from typically small clinical samples such as breast and prostate cancer biopsies, after the comparison of microarray and qRT-PCR measured miRNA levels from two different prostate cancers to assess novel miRNA biomarkers.

2.1.2 Microarray Analysis of miRNA Biomarkers

miRNA profiling of circulating tumour exosomes can potentially be used as diagnostic markers for biopsy profiling (Taylor & Gercel-Taylor 2008). miRNA microarray data analysis has revealed the existence of a novel biomarker for successfully poorly differentiated tumours (Miska, Alvarez-

Saavedra, Townsend, Yoshii, Šestan, Rakic, Constantine-Paton & Horvitz 2004). For example, Raponi et al. (Raponi et al. 2009) identified miRNA expression profiles in lung cancer that would better predict prognosis from 61 SCC samples and 10 matched normal lung samples on MirVana miRNA Bioarrays (version 2, Ambion). Fifteen differentially expressed miRNAs were identified between the normal lung and the cancerous lung, after comparison with a previously identified 50-gene prognostic signature. Their results indicated that miRNAs might have greater clinical utility in predicting the prognosis of patients with squamous cell lung carcinomas than mRNA-based signatures.

Lu et al. (Lu et al. 2005) used a new, bead-based flow cytometric miRNA expression profiling method and presented a systematic expression analysis of 217 mammalian miRNAs from 334 samples, including multiple human cancers. They found that the miRNA profiles were surprisingly informative, reflecting the developmental lineage and differentiation state of the tumours. They also observed a general down-regulation of miRNAs in tumours compared with normal tissues.

Furthermore, they successfully classified poorly differentiated tumours using miRNA expression profiles, whereas messenger RNA profiles were highly inaccurate when applied to the same samples. These findings highlighted the potential of miRNA profiling in cancer diagnosis. Meanwhile, Yang et al. (Yang, Kaur, Volinia, Greshock, Lassus, Hasegawa, Liang, Leminen, Deng, Smith et al. 2008) performed miRNA microarray to detect the miRNAs associated with chemotherapy response in ovarian cancer and found that let-7i expression was significantly reduced in chemotherapy-resistant patients. In addition, they also validated this result by stem-loop real-time reverse transcription-PCR.

Table 2.1: miRNAs with altered expression in malignancy.

Tumour type	Increased expression	Decreased expression
Breast Cancer (Iorio et al. 2005, Mattie et al. 2006)	miR-21, miR-29b-2	miR-125b, miR-145, miR-10b, miR-155, miR-17-5p, miR-27b
Ovarian Cancer (Iorio et al. 2007, Dahiya et al. 2008)	miR-141, miR-200(a-c), miR-221	let-7f, miR-140, miR-145, miR199a, miR-424
CLL (Calin et al. 2002)		miR-15, miR-16
Hepatocellular (Iorio et al. 2005, Lowery et al. 2008)	miR-18, miR-224	miR-199a, miR-195, miR-200a, miR-125a
Pancreatic (Lu et al. 2005, Lowery et al. 2008)	miR-221, miR-376a, miR-24, miR-100, miR-103, miR-107, miR-301,miR-21, miR-125b	miR-375
Prostate (Lu et al. 2005)	let-7d, miR-195, miR-203	miR-128a
Gastric (Lowery et al. 2008, Lu et al. 2005)	miR-223, miR-21, miR-103	miR-218-2
Lung (Lu et al. 2005, Lowery et al. 2008)	miR-17-92 cluster, miR-17-5p	let-7 family

2.2 miRNA-mRNA Regulatory Modules

As introduced in Section 1.2, miRNAs are widely believed to regulate complementary mRNA targets and miRNA-mRNA regulatory modules are important to understand human deseases.

2.2.1 Existing Databases Based on Sequence Data

After the significant miRNA-mRNA regulatory relationship discovery, elucidating closely related miRNAs and mRNAs may be viewed as an essential first step towards the discovery of their combinatorial effects on understanding complex cellular systems. In the meantime, in order to predict miRNAs' target mRNAs by their binding sequence information, at present, some databases are publicly available and commonly used to predict miRNA targets, such as such as TargetScan (Lewis, Shih, Jones-Rhoades, Bartel & Burge 2003), TargetScanS (Lewis, Burge & Bartel 2005*a*), miRanda (John, Enright, Aravin, Tuschl, Sander & Marks 2004),RNAhybrid (Rehmsmeier, Steffen, Höchsmann & Giegerich 2004), PicTar (Krek, Grün, Poy, Wolf, Rosenberg, Epstein, MacMenamin, da Piedade, Gunsalus, Stoffel et al. 2005), and DIAN-AmicroT (Kiriakidou, Nelson, Kouranov, Fitziev, Bouyioukos, Mourelatos & Hatzigeorgiou 2004), etc.

All the predicted and experimentally confirmed databases are shown in appendix Table A.1 (Agarwal, Bell, Nam & Bartel 2015, Wong & Wang 2014, Krek et al. 2005, Lewis, Burge & Bartel 2005*b*, Rehmsmeier et al. 2004, Betel, Wilson, Gabow, Marks & Sander 2008, Griffiths-Jones, Grocock, Van Dongen, Bateman & Enright 2006, Maragkakis, Reczko, Simossis, Alexiou, Papadopoulos, Dalamagas, Giannopoulos, Goumas, Koukis, Kourtis et al. 2009, Kertesz, Iovino, Unnerstall, Gaul & Segal 2007, Huang, Babak, Corson, Chua, Khan, Gallie, Hughes, Blencowe, Frey & Morris 2007, Miranda, Huynh, Tay, Ang, Tam, Thomson, Lim & Rigoutsos 2006, Liu, Yue, Chen, Gao & Huang 2010, Sturm, Hackenberg, Langenberger & Frishman 2010, Yousef, Jung, Kossenkov, Showe & Showe 2007, Vejnar &

Zdobnov 2012, Wang, Ning, Wang, Li, Ye, Zhao, Li, Huang & Li 2013) . These methods can be divided into three main categories: (1) rule-based algorithms which use expression level data and data-driven algorithms. Rule-based algorithms are those that filter the result of their prediction according to a set of rules which are driven from biological evidence. (2) Algorithms which use expression level data of miRNAs and target mRNAs. (3) Data-driven algorithms are computational methods which apply knowledge discovery techniques in order to predict the potential mRNA target for a miRNA.

Rule Based Algorithms

These algorithms consist of a set of defined rules which are tested for each given target. Each rule is an evidence that can contribute to a target gene transcript being labelled as a potential target. The most common rules used in these algorithms are: the seed match condition, minimum free energy and conservation ratio. These algorithms are usually proceeded by testing the rules in a specific order. Since testing a rule is a filtering step, the order of testing the rules affects the performance of the algorithms. These algorithms also perform post filtering steps which might include:

- Number of total base pairs cut off: if the number of matched nucleotides on predicted target mRNA is less than a cut off, the prediction will be discarded.

- Gap permission: if the number of gaps on the predicted binding site is more than a cut off, the prediction will be discarded.

- Bulge permission: if the bulge size is longer than a specific length, the prediction will be discarded.

- Binding region filtering: if the predicted binding site is not located in the 3' region of the target transcript, the prediction will be discarded.

- Conservation level cut off: if the predicted binding site is not highly conserved, the prediction will be discarded.

The most famous rule based algorithms are reviewed for examples in the following subsections.

TargetScan The TargetScan (`http://www.targetscan.org/`) predicts biological targets of miRNAs by searching for the presence of conserved 8mer and 7mer sites that match the seed region of each miRNA (Lewis et al. 2005*b*). As an option, non-conserved sites are also predicted. Also identified are sites with mismatches in the seed region that are compensated by conserved 3' pairing (Friedman, Farh, Burge & Bartel 2009*a*). Conserved targetting has also been detected within open reading frames (ORFs).

In the TargetScan algorithm, there are different types of target sites for a given miRNA seed region including 6mer, 7mer-m8, 7mer-1A and 8mer with the following definitions:

- 6mer: An exact match to position 2-7 of the mature miRNA
- 7mer-m8: An exact match to positions 2-8 of the mature miRNA (the seed + position 8)
- 7mer-1A: An exact match to positions 2-7 of the mature miRNA (the seed) followed by an 'A'
- 8mer: An exact match to positions 1-8 of the mature miRNA

Each binding site has a different binding stability ranked as 8mer>7mer-1A>7mer-m8>6mer. Each seed region match is given a score according to the type of binding site which corresponds to a target site type. The next step in the algorithm is to find the conservation ratio of the candidate sites.

TargetScan has three different conservation definitions for binding site including poorly conserved, conserved and broadly conserved. If the candidate site on the target mRNA sequence is conserved across most vertebrates, it is labelled as highly conserved. If it is conserved across

most mammals but usually not beyond placental mammals, it is labelled as conserved and if it is not in any of these categories, then it is considered as poorly conserved.

The final step of the TargetScan algorithm is to look for the match score of the 3' region of miRNA and the candidate binding sites. Those candidate sites which do not form a strong binding between the rest of the miRNA and the candidate sites on the mRNA are discarded. This step is mainly governed by RNAfold (Hofacker, Fontana, Stadler, Bonhoeffer, Tacker & Schuster 1994). RNAfold calculates the thermodynamic stability score resulting from interaction between two strands of RNA. The candidate binding sites are then scored according to the thermodynamic cut-off value of each site and finally a list of potential target mRNAs is generated and sent to the output.

DIANA-microT miRNAs binding to the target mRNA is usually governed by seed region which is located at the 5' end of miRNA. In some cases, the miRNA does not form a strong binding on its 5' region, but it has a strong base pairing on its 3' region.

DIANA-microT is a human miRNA target prediction tool with the aim of addressing the necessity for a strong binding at 3' end of miRNA when 5' seed pairing is not strong. This approach also considers target sites which have only one binding site and it is an advantage compared to the previous works, because it is independent of strong miRNA seed region base pairing. It exploits the experimental deduction of rules governing miRNA-mRNA binding site. This method considers both conserved and non-conserved binding sites.

This algorithm is dependent on a set of five conditions which should be satisfied in order to have a binding accepted as a potential target:

- if three consecutive Watson-Crick (WC) matches exist.
- if the free energy is lower than a user defined threshold. The normal cut off value for free energy is 20.

- if from z1 to z10 (position one to ten on the miRNA sequence), there are more than seven WC matches or G-U matches.
- if from z8 to z15, there exists at least one loop or bulge and it should be either two to five nucleotides long if on the miRNA side or six to nine nucleotides long if on the mRNA side.
- if from z15 to z22, there are more than five WC or G-U matches and there exists at most a single-nucleotide or dinucleotide bulge, and provided that it is surrounded by two or three base-paring, respectively.

Algorithms Using Expression Level Data

Algorithms in this category are highly dependent on expression profiles of miRNAs and mRNAs. Gene expression profiling refers to the process of measuring the activity level of a selected group of genes in a cell or a tissue (Black, Falzon & Aronson 2012).

In this category, GenMiR and GenMiR++ are worth mentioning (Huang et al. 2007). GenMiR++ is a Bayesian belief network algorithm which uses expression profile of mRNA and miRNA. In conjunction to a gene expression profile, this method requires a candidate binding site predicted by a target prediction algorithm using sequence analysis, such as TargetScan.

The method then filters the given prediction list and predicts those miRNA::mRNA interactions which are closer to reality according to each miRNA expression profile and the expression level of the predicted mRNA by the given input, i.e. the TargetScan list of prediction.

This method considers the fact that mRNAs share a common background expression level in a given tissue. Then it assumes that the regulation level of mRNAs in that tissue is the linear combinatory effect of regulatory miRNAs. GenMiR++ formulates these assumptions using a Gaussian likelihood function which aims to score the given input predictions.

This method was applied to the expression profile data of 151 human miRNAs and 16,063 mRNAs of 88 tissue samples. The input was a list of

114 miRNAs and 890 mRNAs predicted by TargetScanS.

This method identified 1,597 target pairs for given miRNAs on a given list of prediction by TargetScanS with high confidence. One of the miRNAs in the list of 104 miRNAs was a well known miRNA, let-7b which is a well known miRNA. TargetScanS predicts 34 mRNAs to be putative targets of let-7b. GenMiR++ predicted 12 mRNA out of that 34 mRNAs to be let-7b targets with high confidence.

Experimental validation of the 12 prediction showed that five mRNAs are actually putative targets of let-7b and among the other 22 TargetScanS predictions, only two of them were true targets of let-7b. This method demonstrates an increase in prediction specificity, with a small decrease of sensitivity. The main problem with this method is the fact that its prediction list is limited to the input list which comes from other prediction methods.

Data Driven Algorithms

Most currently available miRNA target prediction tools are rule based. Almost all of these methods are dependent on either the presence of the seed region, high rate of conservation in the target site or the accessibility of the binding site based on the predicted structure of target mRNA. Recently however, it has been shown that this level of constraint can lead to a substantial number of missed potential targets. In contrast, data driven tools do not suffer from this rigid filtering of potential targets. In the following pages, machine learning based methods for miRNA target prediction will be reviewed.

PicTar Krek et al. (Krek et al. 2005) presented PicTar, a computational method for identifying common targets of miRNAs. PicTar is based on sequence conservation and seed region in the framework of a Hidden Markov Model (HMM).

The PicTar algorithm starts by looking for the perfect seed region match of the given miRNA in all conserved 3' region sequences, using RNAHybrid,

the algorithm calculates the minimal free energy (MFE) of the candidate binding sites. If the cut-off value is less than a given value, that binding site is removed from the list of candidate binding sites. If the number of binding sites on each candidate 3' UTR is less than a specific value, that 3' UTR is removed from the list of candidate 3' UTRs.

The PicTar algorithm then proceeds by modelling the fact that the same 3' UTR can be targetted by multiple miRNAs. This is done by building the HMM. The log ratio of HMM probability in the process of this modeling is finally used to score the 3' UTRs. These 3' sequences which have a score of more than a given cut-off are introduced as a prediction list of PicTar for the given miRNA.

For statistical tests using genome-wide alignments of eight vertebrate genomes, PicTar is able to specifically recover published miRNA targets, and experimental validation of seven predicted targets. PicTar has an excellent success rate in predicting targets for single miRNAs and for combinations of miRNAs.

miRDB miRDB (`http://mirdb.org/miRDB/`) is another popular online database for miRNA target prediction and functional annotations. MirTarget2 is an SVM based miRNA target prediction which is based on microarray data taken from two different cell lines. All the targets are predicted by a bioinformatics tool MirTarget2, which was developed by analysing thousands of genes impacted by miRNAs with an SVM learning machine (Wang 2008).

Common features associated with miRNA target binding have been identified and used to predict miRNA targets. miRDB hosts predicted miRNA targets in five species: human, mouse, rat, dog and chicken (Wang & El Naqa 2008).

In generating the training set, a gene is defined as a positive target if the expression fold change is reduced by at least 40% with a value of less than 0.001 in both cell lines, or it is a negative target if it has a fold change of 95 to 120% with a value bigger than 0.3 in both cell lines. The feature vector

used in this method is built of 113 features for a miRNA and target site.

The conservation ratio of seed region of miRNA is also considered as one of the features. Human mRNAs orthologs in dog, chicken, rat and mouse are analysed to identify the seed match conservation ratio.

Seven features are derived from binding site accessibility and the location of the binding site and finally, six seed match type features are defined as a binding site type. For those mRNAs which have multiple binding sites, a scoring system is defined to assign a score to transcripts with multiple binding locations. The following formula calculates the score for each transcript: $s = 100 * (1 - \prod_{i=1}^{n} P_i)$ where n represents the number of candidate binding sites, and P_i is the statistical significant P-value for each candidate site estimated by the support vector machine (SVM). Applying this method reveals that more than half of the binding sites available in the literature are not conserved in other species.

The main problem with this method is ignoring the possibility of the expression of candidate targets in other cell types since the backbone of this method disregards the expression profile of other types of cells. Given the fact that not all genes are expressed in all cells, the identified targets of this method might not be functional since they might not be available in other cell lines.

2.2.2 Disadvantages of miRNA Target Prediction

It has been shown that 70% of computationally predicted miRNA targets are false positives. Current miRNA target predictions are also based on seed match region, evolutionary conversation, thermodynamic features or a combination of these components.

As reviewed, some of miRNA target prediction methods are based on a combination of different miRNA target prediction methods. This combinatory prediction is also limited to all limitations of those tools which form the main method. Those prediction tools which take other prediction algorithms as their input are also limited to the constraints of their input.

To summarise, the main disadvantages of the reviewed miRNA target prediction tools are shown as follows:

- high numbers of false positive and false negative predictions.
- lack of prediction tools for target mRNAs which are expressed in specific cells
- lack of prediction tools for miRNAs which are expressed in specific cells
- rigid filtering rules which result in discarding potential targets
- seed region condition dependency
- across-species conservation dependency
- lack of predictions for not well known miRNAs
- lack of considering for biological information.

2.2.3 Computational Methods for Discovering miRNA-mRNA Regulation

At the initial stages of identifying miRNA targets, the near-perfect complementarity was adopted to predict miRNA targets for plant model species in which

the genome sequences had been encoded, such as in Arabidopsis (Rhoades, Reinhart, Lim, Burge, Bartel & Bartel 2002).However, after more genome sequences became available and a better understanding of the pairing requirements between miRNAs and their targets was acquired, it was found that some criteria in the prediction of miRNA targets can be relaxed without sacrifice of specificity (Kong & Han 2005).

Later, numerous computational software programs were developed to predict miRNA targets in other animal and plant species (Lewis et al. 2003, Lewis et al. 2005*a*, John et al. 2004, Rehmsmeier et al. 2004, Kiriakidou et al. 2004). All these approaches have been successfully used to identify miRNA-mRNA regulatory modules in various human diseases.

Analogous to transcriptional regulation, most miRNAs fine tune the expression of hundreds of genes in a combinatorial manner (Bartel 2004*a*). This combinatorial regulation manifests in at least two layers below:

- Several miRNAs have been found to regulate a single mRNA target through targetting the same mRNA transcript 3' UTR or even in combination with targetting in coding sequence (He & Hannon 2004).

- A cluster of miRNAs, which often co-expressed, could regulate functionally related proteins (Dews, Homayouni, Yu, Murphy, Sevignani, Wentzel, Furth, Lee, Enders, Mendell et al. 2006). Xu et al. (Xu & Wong 2008) provided experimental evidence that one miRNA cluster, targets three genes located in the insulin signaling pathway.

Causality Discovery-Based Methods

All of the above experimental and computational evidence demonstrated that coordinate regulation by miRNAs is a flexible and efficient strategy to regulate cellular processes in a conditional or tissue-specific manner (Bartel 2004*b*). Thus, it is important to develop novel computational methods that explicitly capture miRNA-mRNA regulatory modules.

To investigate the influence of miRNAs on transcript levels, Lim et

al. (Lim, Lau, Garrett-Engele, Grimson, Schelter, Castle, Bartel, Linsley & Johnson 2005) transfected miRNAs into human cells and used microarrays to examine changes in the mRNA profile. They found that conveying miR-124 caused the expression profile to shift towards that of the brain, the organ in which miR-124 was preferentially expressed, whereas delivering miR-1 shifted the profile towards that of the muscle, where miR-1 was preferentially expressed. In each case, about 100 messages were down-regulated after 12h. The 3' untranslated regions of these messages had a significant propensity to pair to the 5' region of the miRNA, as expected if many of these messages were the direct targets of the miRNAs. Their results suggested that metazoan miRNAs can reduce the levels of many of their target transcripts, not just the amount of protein deriving from these transcripts. Moreover, miR-1 and miR-124, and presumably other tissue-specific miRNAs, seemed to down-regulate a far greater number of targets than previously appreciated, thereby helping to define tissue-specific gene expression in humans.

An increasing number of researchers proposed some computational methods on this issue. Joung et al. (Joung, Hwang, Nam, Kim & Zhang 2007) developed a computational method based on probabilistic learning to detect the miRNA-mRNA modules from paired miRNAs and mRNAs expression profiles and binding information. Their results provided a primary source of miRNA and target sets presumed to constitute closely related parts of gene regulatory pathways.

Bayesian networks (Pearl 1988) have also been adopted by many research groups (Liu, Liu, Tsykin, Goodall, Green, Zhu, Kim & Li 2010, Peng, Li, Walters, Rosenzweig, Lederer, Aicher, Proll & Katze 2009) to detect novel miRNA-mRNA modules. For instance, the Bayesian network learning algorithm was used to search all possible networks and a scoring function based on observational data was used to score each graph. In this work, they assumed that there was a bipartite of interactions between the group of miRNAs and the group of mRNAs. The novelty factor in terms of this study is that the authors used target information to restrict the search space for

the computational expensive Bayesian network learning algorithm.

Liu et al. (Liu, Liu, Tsykin, Goodall, Green, Zhu, Kim & Li 2010) presented the correspondence latent dirichlet allocation graphic model to discover functional miRNA regulatory modules at potential biological levels by integrating heterogeneous data sets, including expression profiles of miRNAs and mRNAs, with or without the prior target binding information. They applied this model to a mouse mammary data set and captured several biological process specific modules involving miRNAs and their target mRNAs. Their results showed that expression profiles were crucial for both target identification and discovery of regulatory modules.

Another causality discovery-based method was presented by Le et al. (Le, Liu, Tsykin, Goodall, Liu, Sun & Li 2013) to uncover the causal regulatory relationship between miRNAs and mRNAs without the previous target binding information. This method firstly used *do-calculus* (Pearl 2003) to estimate the causal effects of a variable on the other variables based on observational data. The estimated causal effects simulated the effects of randomised controlled experiments. The method tackled two drawbacks of current miRNA regulatory relationships research. Firstly, the method discovered causal relationships between miRNAs and mRNAs, not just the statistical relationships. Secondly, the method assumed that miRNAs and mRNAs interact with each other in a complex system; for instance, a miRNA can causally regulate mRNAs as well as other miRNAs. This assumption is more reasonable than the assumption from commonly used approaches that considers only the bipartite of interactions between miRNAs and mRNAs.

Differential Analyses

To understand the causes of a disease, it often requires analysing the differences between normal and disease samples. For example, differentially co-expressed genes differ significantly between disease and control samples. It is very important to identify the differences in the gene regulatory networks between diseases and healthy conditions.

For this purpose, researchers discovered miRNA activity changes in two biological conditions. Highlights in this direction, are miReduce (Sood, Krek, Zavolan, Macino & Rajewsky 2006), DIANA-mirExTra (Alexiou, Maragkakis, Papadopoulos, Simmosis, Zhang & Hatzigeorgiou 2010), Sylamer (van Dongen, Abreu-Goodger & Enright 2008) and MIR (Cheng, Li et al. 2008). These methods firstly inferred the differences in gene expression levels in the two biological conditions, then correlated those alterations with the miRNA binding motifs that are predicted based on sequence data.

A method named mirAct (Liang, Zhou, He, Zheng & Wu 2011) was used to explore the miRNA activity in a sample and then to analyse the overall behaviour of miRNA activity in samples with different biological conditions. Another method called DICER was proposed to detect differential co-expression in disease and control samples (Amar, Safer & Shamir 2013). They hypothesised that changes in co-expression may be the result of changes in regulatory patterns, and thus the discovered differential co-expression may be the target of specific miRNA families. To test the approach, they identified miRNA families whose targets are enriched in the gene groups detected by their method, and tested whether those miRNAs are associated with some diseases.

Integrating Heterogeneous Data Sources

In order to know the modular organisation of the regulatory networks, researchers try to search for a set of miRNAs and their co-regulated genes by integrating heterogeneous data sources, such as sequence data, protein-protein interaction and DNA-protein interaction networks.

Zhang et al. (Zhang, Li, Liu & Zhou 2011) proposed an effective computational framework to identify the miRNA-gene regulatory modules by integrating miRNA target predictions based on sequence data, miRNA and gene expression profiles, protein-protein interaction and DNA-protein interaction networks. In their work, sequence-based miRNA target predictions were considered as a static prior network, and expression profiles were

subsequently used to identify the active miRNA-gene interactions. These active interactions were further refined by the gene-gene interaction networks (protein-protein and DNA-protein networks). Applied to the human ovarian cancer samples from The Cancer Genome Atlas (TCGA), the method discovered several miRNA gene regulatory modules. The results were then validated against the miRNA cluster from miRBase (http:// www.mirbase.org/), and the mRNAs were validated using gene functional enrichment analysis.

Similarly, Le et al. (Le & Bar-Joseph 2013) developed the Protein Interaction-based miRNA Modules (PIMiM), a regression-based probabilistic method to integrate sequence, expression and interaction data for detecting modules of mRNAs controlled by small sets of miRNAs for a specific condition. The authors firstly used a regression model to connect the express data of miRNAs and mRNAs and assumed that the expression level of an mRNA can be represented as a linear function of expression profiles of all predicted miRNAs. The predicted miRNA regulators for an mRNA were taken from sequence-based prediction databases. To interact miRNAs and mRNAs into a module, they designed a new target function to measure the strength of the predicted miRNAmRNA interactions based on the information from miRNA target information and protein-protein interactions. Specifically, the assigned function was based on the logistic-sigmoid function with parameters to adjust the contributions of the two types of interaction data. The higher the probabilities of interactions are, the more chances the interacting entities are assigned into the same module. The method was applied to a number of different types of cancer data sets from TCGA to explore the regulators (miRNAs) that are common for all cancer types and the specific active regulators for each cancer type. The results were then validated against knowledge from literature and by gene functional enrichment analysis.

Li et al. (Li, Zhang, Liu & Zhou 2012) developed a method called sparse Multi-Block Partial Least Squares (MBPLS) regression method to integrate multiple data sources, including copy number variation (CNV),

DNA methylation (DM), gene expression (GE) and miRNA expression (ME) for detecting multi-layer gene regulatory modules. This method was employed to identify multi-dimensional regulatory modules from the data. The assumption was that CNV, DM and ME all regulated the gene expression. The method projected each data type into a summary vector, and maximised the covariance between the summary vectors of source data (CNV, DM,ME) and the response data (GE). Finally, it used the weighted sum of the summary vectors of source data to represent the unique input source data, and again maximised the covariance between the input data and the response data. The method was tested on simulated data as well as the ovarian cancer data sets from TCGA and was capable of identifying the modules that have significant functional and transcriptional enrichments. The results predicted from this method proved to be better than the results from those methods that only use one type of data.

Peng et al. (Peng et al. 2009) simultaneously profiled the expression of cellular miRNAs and mRNAs across 30 HCV positive or negative human liver biopsy samples using microarray technology. They constructed a miRNA-mRNA regulatory network, and using a graph theoretical approach, identified 38 miRNA-mRNA regulatory modules in the network that were associated with HCV infection. They evaluated the direct miRNA regulation of the mRNA levels of targets in regulatory modules using previously published miRNA transfection data, and analysed the functional roles of individual modules at the systems level by integrating a large-scale protein interaction network. Finally, they found that various biological processes, including some HCV infection related canonical pathways, were regulated at the miRNA level during HCV infection. Their results provide new insights into post-transcriptional gene regulation at the miRNA level in complex human diseases.

2.3 Co-regulation miRNA Network

miRNAs are widely believed to regulate complementary mRNA targets. Co-regulation analysis of multiple miRNAs is useful for understanding complex post-transcriptional regulations in humans (Baumjohann & Ansel 2013, Guo, Zhao, Yang, Zhang & Chen 2014). Complex diseases are associated with several miRNAs rather than a single miRNA. It is still a challenging work to discover co-regulation miRNAs at a system level.

In 2004, Lai et al. (Lai, Wiel & Rubin 2004) found that miRNAs regulated non-mRNA targets, namely other miRNAs, by conducting a systematic assessment of the nearly complete catalogs of animal miRNAs. One of the earliest studies on miRNA pair co-regulation is conducted by Enright et al. (Enright, John, Gaul, Tuschl, Sander, Marks et al. 2004) for understanding the co-regulation between lin-4 and let-7 in Drosophila.

With the huge amount of expression data publicly available, newer methods have been proposed to investigate the problems of co-regulating miRNAs (Migliore & Giordano 2009, Boross, Orosz & Farkas 2009, An, Choi, Wells & Chen 2010). For example, Boross (Boross et al. 2009) proposed to construct a miRNA co-regulation network by computing the correlations between the gene silencing scores of individual miRNAs.

Since most of these studies take only the expression data of miRNAs and mRNAs without biological function analysis (Guo, Ingolia, Weissman & Bartel 2010), some true targets of these miRNAs may be ignored and some false targets may be included. One possible reason is that those miRNAs have been shown to reduce protein levels without the concomitant change in mRNA levels (Lee, Samaco, Gatchel, Thaller, Orr & Zoghbi 2008), and the mRNAs may be regulated at tissue specific levels (Guo, Maki, Ding, Yang, Xiong et al. 2014).

Yoon et al. (Yoon & De Micheli 2005*a*) considered GO information and proposed a biclique-based method to detect co-regulating groups of miRNAs and mRNAs. However, the heuristic nature of that method can lead to those miRNAs or genes being missed even when they have a high probability

of co-regulation. Recently, it has been found that interacting proteins are often regulated by similar miRNA types (Yuan, Liu, Yang, He, Liao, Kang & Zhao 2009, Liang & Li 2007). This suggests that clustered miRNAs can jointly regulate the proteins which are close to each other within a protein interaction network (Hsu, Juan & Huang 2008).

The cumulative hypergeometric statistical test devised by Shalgi et al. (Shalgi, Lieber, Oren & Pilpel 2007) was used to identify miRNA pairs that showed a high rate of co-occurrence in 3'-UTRs of the same target genes.

Zhou et al. (Zhou, Ferguson, Chang & Kluger 2007) used statistical association (interaction) measures to quantify the significance and size of the overlap between the sets of predicted targets of miRNA pairs. They used the p-values and q-values from Fisher's Exact Test to evaluate the significance of the overlap and found miRNA pairs were substantially abundant.

Signal-to-noise ratio was used by An et al. (An et al. 2010) to get high accurate regulating miRNAs for all genes. A sequence of statistical tests was then used to identify highly co-regulating miRNAs and the corresponding co-regulated gene groups.

DIANA-mirPath was developed to identify molecular pathways potentially altered by the expression of single or multiple miRNAs. This method considers the combinatorial effect of co-expressed miRNAs in the modulation of a given pathway (Papadopoulos, Alexiou, Maragkakis, Reczko & Hatzigeorgiou 2009).

Xiao et al. (Xiao, Ma, Zhu, Sun, Yin & Feng 2015) constructed the miRNA-miRNA co-regulated network to identify miRNA or target genes involved in cerebral injury caused by stroke, and to search out the associated biological processes, especially inflammation.

All these studies have demonstrated the importance of the miRNA-miRNA co-regulating network.

2.4 miRNA-TF Co-regulatory Networks

At the transcription level, TFs are believed to be the main gene regulators. TFs are fundamental players of gene expression regulation: they activate the transcription of both coding and non-coding genes (Vaquerizas, Kummerfeld, Teichmann & Luscombe 2009). Many experimental and computational methods have been developed to discover the regulatory mechanisms of these types of regulators.

On the other hand, miRNAs have been known to be the main gene regulators at the post-transcriptional level. They degrade target mRNAs or more often inhibit their translation, and function by binding to the RNA (He & Hannon 2004).

miRNAs are quite short in length, and the gene regulatory region of a miRNA is small in size, compared with a TF. Both TFs and miRNAs can regulate multiple target genes simultaneously, and target genes can be regulated by both multiple TFs and miRNAs cooperatively.

Because the predicted target of miRNAs contains many transcription factors (Hobert 2008*a*), linking the TF target genes (including miRNAs, other TFs, or other genes) with the miRNA target genes (including TFs and other genes) can provide a global insight into the gene regulation network. In addition, a unified picture of the regulatory relationships of the two main regulators and target genes can provide useful insights into the causes of diseases.

Analysis of the properties of these networks can elucidate the designing principle and provide an understanding of regulatory networks (Milo, Shen-Orr, Itzkovitz, Kashtan, Chklovskii & Alon 2002). The combined regulations of miRNAs and TFs are important but difficult to explore, as miRNAs and TFs can regulate each other in addition to regulating target genes.

Recently, some studies constructed the gene regulatory networks with the presence of both miRNAs and TFs. They employed sequence data and expression data for learning the complex regulatory network.

2.4.1 Sequence-Based Methods

A common framework for exploring miRNA-TF co-regulatory relationships is to integrate the putative target information of both TFs and miRNAs to obtain an interaction network with the three components, miRNAs, TFs and mRNAs. Researchers inferred gene regulation knowledge from the combined network by using statistical tests, network inference algorithms or gene functional enrichment analyses (Fazi, Rosa, Fatica, Gelmetti, De Marchis, Nervi & Bozzoni 2005, Guo, Xie, Fei & Chua 2005, Le Béchec, Portales-Casamar, Vetter, Moes, Zindy, Saumet, Arenillas, Theillet, Wasserman, Lecellier et al. 2011, Yang, Li, Jiang, Zhou & Qu 2013).

Predicted Databases

Some predicted databases are provided for identifying miRNA and TF shared targets, and network motifs involving miRNAs and TFs and known pathways. Fazi *et al.* discovered that human granulocytic differentiation was controlled by a regulatory circuitry involving miR-223 and two transcriptional factors, NFI-A and C/EBP (Fazi et al. 2005). The target genes of Oncogenic miR-27a were able to regulate specifical protein transcription factors and the G2-M checkpoint in MDA-MB-231 breast cancer cells (Mertens-Talcott, Chintharlapalli, Li & Safe 2007). The auxin induction of miR164 has been proved to provide a homeostatic mechanism to clear NAC1 mRNA to downregulate auxin signals (Guo et al. 2005).

ChIPBase Yang et al. (Yang, Li, Jiang, Zhou & Qu 2013) developed a web tool called ChIPBase for miRNA-TF co-regulation analysis. The web tool is available at `http://deepbase.sysu.edu.cn/chipbase/tfmiRtargetNetworks.php`. They developed ChIPBase to facilitate the integrative and interactive display, as well as the comprehensive annotation and discovery of TF-miRNA interaction maps from ChIP-Seq data that were generated from diverse tissues and cell lines from six organisms: human, mouse, dog, chicken,*Drosophila melanogaster* and *Caenorhabditis elegans*. ChIPBase

contains tens of thousands of TF-miRNA regulatory relationships. Users can select a miRNA target to see TF-miRNA, miRNA-target and TF-target networks.

MIR@NT@N Le et al. (Le Béchec et al. 2011) constructed a regulatory network by integrating available target prediction databases for both TFs and miRNAs. They have provided a web resource called MIR@NT@N for facilitating the retrieval of regulatory relationships and network motifs. MIR@NT@N is a user-friendly web resource freely available at `http://?mironton.?uni.?lu`. Users can explore the shared targets of miRNAs and TFs, and query a list of Feed-Forward Loops (FFLs) and Feed-Back Loops (FBLs) that involve miRNAs and TFs.

Depending on Shared Downstream Targets

Shalgi et al. (Shalgi et al. 2007) built the network by involving miRNAs, TFs and mRNAs using sequence data. They used evolutionary conserved binding sites of miRNA targets to construct the interactions between miRNAs and genes (including TFs). Meanwhile, conserved binding sites of TFs in promoters were used to uncover the interactions between TFs and mRNAs and the interactions between TFs and miRNAs. The combined network was then analysed to identify the shared targets of the regulators. It was found that the hub of interactions is usually TFs and it was discovered that some network motifs that involve miRNAs, TFs and mRNAs.

Zhou et al. (Zhou et al. 2007) used PicTar as the miRNA targets and Transfac as TF targets to build the network of miRNAs, mRNAs and TFs. They then used Fisher's Exact Test to measure the significance of the shared targets between the regulators, and to remove the insignificant co-regulating interactions that occurred by chance. They found that the shared targets of TF pairs and miRNA pairs were much more abundant than those of TF-miRNA pairs, and that the shared targets in feed-forward loops with TF playing as a master regulator were more statistically significant than other

types and feed-forward loops.

Martinez *et al.* used a yeast one-hybrid system to identify TFs that were bound to miRNA promoters in *C. elegans* (Martinez, Ow, Barrasa, Hammell, Sequerra, Doucette-Stamm, Roth, Ambros & Walhout 2008). They found a total of 347 high-confidence interactions between 63 miRNA promoters and 116 proteins. They then combined this regulation relationship with the computationally predicted miRNA→TF interactions. They found a total of 23 miRNA↔TF composite feedback loops in which a TF and a miRNA were mutually regulated.

Yu et al. (Yu, Lin, Zack, Mendell & Qian 2008) found that one specific regulated feedback loop (two TFs regulate each other and one miRNA regulates both of the two TFs), feed-forward loop motifs (one miRNA targets both a TF and the regulating gene of that TF), and significant pairs (one TF and one miRNA target the same gene) were the top three significantly over-represented network motifs.

A rule-based method was proposed by Tran et al. (Tran, Satou, Ho & Pham 2010) to discover the gene regulatory modules that consist of miRNAs, TFs and their target genes based on the available predicted target binding information. The authors analysed the regulatory associations among the sets of predicted miRNAs and sets of TFs on the sets of regulated genes produced by them in the human genome, and validated these modules with the GO and the literature. The results showed that their method allows them to detect functionally-correlated gene regulatory modules involved in specific biological processes.

Chen et al. (Chen, Chen, Fuh, Juan & Huang 2011) proposed a novel framework utilising gene functional enrichment analysis to identify the significant co-regulatory relationships. They also used target information to construct the co-regulation network as the first step. The authors then applied GO for gene functional enrichment analysis of the shared target genes to find the functional profiles for these co-regulation pairs. To calculate the significant levels of the shared targets, they compare their method with the

randomly pick method and use the hypergeometric distribution to calculate the P-values of the findings. It was found that some biological processes emerged only in co-regulation and that the disruption of co-regulation might be closely related to cancers, suggesting the importance of the co-regulation of miRNAs and TFs in many biological processes.

The most prominent feature in terms of the above methods is their employment of sequence-based putative target information. However, the networks constructed from sequence-based methods involve a high rate of false negative and false positive. Therefore these methods only begin to explore the complex relationships between the three components, miRNAs, TFs and mRNAs. It would be ideal if expression data or other data could be incorporated to refine the discoveries.

2.4.2 Methods Using Expression Data and Other Data

Many modules are developed to identify and characterise TF-mRNA-miRNA networks that incorporate GE into the studies (Chen & Rajewsky 2007, Sun, Gong, Purow & Zhao 2012, Le, Liu, Liu, Tsykin, Goodall, Satou & Li 2013). These methods firstly use sequence-based target prediction to initialise the network. Then, the expression data are used to refine the findings.

Chen & Rajewsky (Chen & Rajewsky 2007) compared the evolution of transcriptional regulation and post-transcriptional regulation that was mediated by miRNAs, focusing on the evolution of the individual regulators and their binding sites. As an initial step towards integrating these mechanisms into a unified framework, they proposed a simple model that describes the transcriptional regulation of new miRNA genes.

Gene regulatory factors that control the expression of genomic information come in a variety of flavours, with transcription factors and miRNAs representing the most numerous gene regulatory factors in multicellular genomes.

Sun et al. (Sun et al. 2012) proposed a network-based approach to uncover miRNA and TF regulatory networks in Glioblastoma (GBM). They firstly

filtered the miRNAs, TFs and genes related to GBM based on the literature. Then, they integrated the target prediction of miRNA and TF based on sequence data; after that, the gene expression data was used to infer gene-gene interactions by assuming that the interaction occurs when interacting genes are co-expressed. The authors then inferred 3-node FFL and 4-node motifs, which involved statistically significant miRNA-TF interactions. These motifs were integrated into a GBM-specific miRNA-TF mediated regulatory network. The authors then conducted signalling pathway and gene functional enrichment analyses to validate the results.

Le et al. (Le, Liu, Liu, Tsykin, Goodall, Satou & Li 2013) designed a framework of Bayesian network structure learning to construct gene regulatory networks involving both TFs and miRNAs from multiple sources of data, including gene expression profiles of miRNAs, TFs and mRNAs, target information based on sequence data, and sample categories. They then searched the discovered networks to identify the interplay and applied a network motif finding algorithm to further infer the network. They produced compact and meaningful gene regulatory networks that were highly relevant to the biological conditions of the data sets. The results revealed the complex gene regulatory relationships.

In another direction, researchers used target information to build the TF, miRNA and mRNA regulatory networks as the first step. The expression data was then used to identify active pathways that involved miRNAs and TFs, or to identify active regulators in different biological conditions of the data sets. For instances, Jiang et al. (Jiang, Zhang, Meng, Lian, Chen, Yu, Dai, Wang, Liu, Li et al. 2013) proposed a method to identify active TF and miRNA regulatory pathways in Alzheimer's disease (AD) by analysing AD-related mRNA and miRNA expression profiles as well as curated TF and miRNA regulation databases, including TransmiR (Wang, Lu, Qiu & Cui 2010), TRANSFAC (Matys, Fricke, Geffers, Gößling, Haubrock, Hehl, Hornischer, Karas, Kel, Kel-Margoulis et al. 2003), miRecords (Xiao, Zuo, Cai, Kang, Gao & Li 2009), TarBase (Li, Liu, Zhou, Qu & Yang 2014)

and miRTarBase (Hsu, Tseng, Shrestha, Lin, Khaleel, Chou, Chu, Huang, Lin, Ho et al. 2014). They firstly created the miRNA-gene-TF network by these databases. Then they integrated miRNA and GE from different sources and identified the differentially expressed genes (DEG) and miRNAs between normal and disease samples. The results were validated by using gene functional enrichment analysis.

2.5 Limitations of Existing Methods

Recent studies have often focused on the statistical and biological significance of single miRNAs by identifying differentially expressed individual miRNAs as biomarkers (Zhang et al. 2007*a*, Raponi et al. 2009). None of the single miRNA have good sensitivity. This is probably because target mRNAs are actually affected simultaneously by multiple miRNAs synergistically or possibly several miRNA-regulated pathways are involved in the progression of the diseases (Minna et al. 2002). In fact, the wet-lab experiments are very expensive and time-consuming. The problem is that single-miRNA rules are insufficient for accurate diagnosis. In addition, miRNAs can regulate miRNAs (Transcriptional 2006). For example, cardiac-expressed miRNAs can regulate expression of other cardiac miRNAs (Matkovich, Hu & Dorn 2013). Most of the research concerning the understanding of disease has neglected this function.

There are still unsolved problems in the detection of miRNA-mRNA regulatory modules. The key idea taken by all of these studies is the inverse expression relationship between miRNAs and their target mRNAs. An inverse expression relationship means that when the expression level of the miRNA is high (up-regulated), then the target mRNA should be down-regulated based on the principle that miRNAs deregulate the expression of targetted mRNAs (Peng et al. 2009). However, up-to-date evidence show that the inverse relationship is not always true. In fact, a miRNA can induce gene expression by binding to the gene's promoter or enhancer sequence. For

example, miR-373 can induce the expression of E-Cadherin or *CSDC2* when binding to the genes' promoters (Place, Li, Pookot, Noonan & Dahiya 2008). Recent investigation also shows that the interaction of miR-10a with RP mRNAs (those mRNAs encoding ribosomal proteins) binding at the 5' UTR region can promote these mRNAs' translational enhancement instead of repression (ørom, Nielsen & Lund 2008). It can be seen that positively regulated modules of miRNAs and mRNAs do exist, but they are widely overlooked by prior computational approaches. Therefore, the identified miRNA-mRNA interactions based purely on inverse correlation may be very incomplete in certain biological contexts (Rijlaarsdam, Rijlaarsdam, Gillis, Dorssers & Looijenga 2013).

Co-regulation analysis of multiple miRNAs is useful for understanding complex post-transcriptional regulations (Baumjohann & Ansel 2013, Guo, Zhao, Yang, Zhang & Chen 2014). One of the earliest studies on miRNA pair co-regulation was conducted by Enright et al. (Enright et al. 2004) for understanding the co-regulation between lin-4 and let-7 in Drosophila. Since most of these studies take only expression data of miRNAs and messenger RNAs (mRNAs) without biological function analysis (Guo et al. 2010), some true targets of these miRNAs may be ignored and some false targets may be included.

This suggests that biological functional analysis is a necessary assessment for the detection of complete and reliable targets of miRNAs. Gene Ontology (GO) contains comprehensive information of biological processes and functions (Ashburner, Ball, Blake, Botstein, Butler, Cherry, Davis, Dolinski, Dwight, Eppig et al. 2000). In particular, the coherence score of the GO terms annotated to a gene group can be used to compute the p-value of a co-regulated gene group which is actually a statistical measurement to judge whether or not the co-regulated gene group are reliable targets of a miRNA.

Also, it is still poorly understood how miRNAs themselves are regulated. This is partly due to the difficulty of predicting promoters from short

conserved sequence features without producing a high number of false positive and partly due to the heterogeneity of the miRNA biogenesis pathways. Many important biological processes are actually controlled by miRNAs which play the role of master regulators. Little studies have been conducted for the self-regulation miRNAs.

2.6 Summary

Since much evidence has suggested the importance role of miRNAs in the development of several diseases, studying miRNA functions will provide further insight into the causes of fatal diseases such as cancers. The huge amount of data available in different types provides opportunities and challenges for computational approaches to detect miRNA functions, which will assist in the design of wet-lab experiments. In this review, we have discussed different computational methods to infer miRNA functions, including biomarkers, co-regulation and self-regulation. The approaches are usually based on sequence data, gene expression data or integrating multiple sources of data. They provide different views on how to elucidate the complex regulatory mechanism of miRNAs.

With more and more data available, it is still challenging to design computational methods to infer miRNA functions for the purposes of assisting experimental design. It poses interesting implications for future work as these methods can help elucidate the complex gene regulatory relationships and the causes of disease.

Chapter 3

Research Methodology

The purpose of this work is to apply rule mining methods to study the miRNA expression profiles for human disease understanding, including the discovery of the miRNA biomarker, positive and negative miRNA-mRNA regulation modules, miRNA co-regulation network and miRNA self-regulation network.

The four purposes of this chapter are to (1) describe the basic knowledge used in this research, (2) explain the computational methods used to compare our rule discovery method, (3) describe some bioinformatics methods used to analyse the data, and (4) provide an explanation of performance measurement used to evaluate the methods.

3.1 Definitions for Information Gain Ratio, Euclidean Distance, 10-Fold Cross Validation and Pearson's Correlation Coefficient

3.1.1 Information Gain Ratio

In this study, we prioritise and rank all the miRNAs in the data set based on their gain ratios (Han & Kamber 2006*a*) over the whole samples' expression profiles. Gain ratio measures the collective difference of every single miRNA's expressions between the two classes. A high gain ratio indicates that the

miRNA is a high-potential biomarker differentially expressed over the two classes. In terms of the second step, we project wet-lab confirmed and intensively studied miRNAs onto this rank list. Using this step, we can recommend those highly ranked miRNAs that have not been studied in wet-labs in the past for rule discovery and potentially for fresh biological study.

Let *Attr* be the set of all attributes and *Ex* the set of all training examples, value(x, a) defines the value of a specific example x for attribute a, where $x \in E_x$,and $x \in Attr$, and the entropy specifies $H(x) = E[\log(2, 1/p(x_i))] = -\sum p(x_i)\log(2, p(x_i))(i = 1, 2, ...n)$.

The information gain for attribute $x \in Attr$ is defined as follows:

$$IG(Ex, a) = H(Ex) - \sum_{v \in values(a)} \frac{|\{x \in Ex | value(x, a) = v\}|}{|Ex|}.H(\{x \in Ex | value(x, a) = v\}) \tag{3.1}$$

The information gain is equal to the total entropy for an attribute if for each of the attribute values a unique classification can be made for the result attribute. In this case the relative entropies subtracted from the total entropy are 0. The intrinsic value for a test is defined as follows:

$$IV(Ex, a) = -\sum_{v \in value(a)} \frac{|\{x \in Ex | value(x, a) = v\}|}{|Ex|} * \log_2(\frac{|\{x \in Ex | value(x, a) = v\}|}{|Ex|}) \tag{3.2}$$

The information gain ratio is just the ratio between the information gain and the intrinsic value:

$$IGR(Ex, a) = IG/IV \tag{3.3}$$

Information gain ratio biases the decision tree against considering attributes with a large number of distinct values. So it solves the drawback of information gain, namely that, information gain applied to attributes may take on a large number of distinct values and might learn the training set too well.

3.1.2 Euclidean Distance

The Euclidean distance or Euclidean metric is the"ordinary" distance between two points that one can measure with a ruler, and is given by the Pythagorean formula. The Euclidean distance between point p and q is the length of the line segment connecting them.

In Cartesian coordinates, if $p = (p_1, p_2...p_n)$ and $q = (q_1, q_2...q_n)$ are two points in Euclidean n-space, then the distance from p to q, or from q to p is given by:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + ... + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}. \tag{3.4}$$

In this work, we just use two dimensions to calculate the distance, so p= (p1, p2) and q= (q1, q2) then the distance is given by

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}. \tag{3.5}$$

3.1.3 10-Fold Cross Validation

Ten-fold cross validation is often used to examine the performance of various classification models. In 10-fold cross validation, the data set is equally and randomly divided into ten portions. Each portion is used as testing data, and the samples in the remaining nine portions comprise the training data set. Each sample is tested once because each portion is tested once. Compared with the Jackknife test, a 10-fold cross-validation test is more efficient and provides similar results for a given data set. Thus, it has been adopted herein to examine the classification model.

3.1.4 Pearson's Correlation Coefficient

Pearson's correlation coefficient (PCC) is the covariance of the two variables divided by the product of their standard deviations. We can obtain a formula for r by substituting estimates of the covariances and variances.

If we have one data set $x_1, ..., x_n$ containing n values and another data set $y_1, ..., y_n$ containing n values then that formula for r is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{3.6}$$

This formula suggests a convenient single-pass algorithm for calculating sample correlation.

3.2 Data Mining Methods

3.2.1 A committee of decision trees

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to leaf represents classification rules.

A decision tree can be linearised into decision rules, where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause. In general, the rules have the form:

if condition 1 and condition 2 and condition 3 then outcome.

The algorithm of constructing a committee of decision trees are implemented in the R package and the source code is shown in the appendix B.1.

3.2.2 Naive Bayes Classifier

Native Bayes (NB) is a one of the most efficient and effective learning algorithms in machine learning and data mining (Rish 2001). In this study, we compared the performance between NB classifier and our discovered rules.

According to Bayes rule, the probability of an example $E = (x_1, x_2, ..., X_n)$ representing n features, it assigns to be each of K possible class C_k is

$$p(C_k|E) = \frac{p(E|C_k)p(C_k)}{P(E)} \tag{3.7}$$

E is classified as the class C_k=+ if and only if

$$f(E) = \frac{p(C_k = +|E)}{P(C_k = -|E)} \geq 1 \tag{3.8}$$

where $f(E)$ is called a Bayes classifier.

NB classifier assumes that each feature is conditionally independent of every other feature; that is

$$p(E|C_k) = p(x_1, x_2, \cdots, x_n|C_k) = \prod_{i=1}^{n} p(x_i|C_k) \tag{3.9}$$

Thus, the NB classifier is:

$$f(E) = \frac{p(C_k = +)}{p(C_k = -)} \prod_{i=1}^{n} \frac{p(x_i|C_k = +)}{p(x_i|C_k = -)} \tag{3.10}$$

3.2.3 K-nearest Neighbors Algorithm

The k-nearest neighbour approach is a powerful nonparametric technique for classification (Liao & Vemuri 2002). If k=1 or the nearest neighbour rule, then a case x is simply assigned to the class of its nearest neighbor. In order to find the point closest to x, let it be y. Now the nearest neighbour rule asks to assign the label of y to x. Distance functions include:

$$EuclideanDistance : EuD = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{3.11}$$

$$ManhattanDistance : MaD = \sum_{i=1}^{k} |x_i - y_i| \tag{3.12}$$

$$MinkowskiDistance : MiD = (\sum_{i=1}^{k} |x_i - y_i|^q)^{1/q} \tag{3.13}$$

$$HammingDistance : HaD = \sum_{i=1}^{k} |x_i - y_i| \tag{3.14}$$

It should also be noted that all the first three distance measures are only valid for continuous variables. In the instance of categorical variables

the Hamming distance must be used. It also brings up the issue of standardisation of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the data set.

Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent data set to validate the K value. Historically, the optimal K for most data sets has been between 3-10. That produces much better results than 1NN.

3.3 Our Proposed Rule Mining Methods

3.3.1 Rule Discovery

A rule discovery method was used for this study. We discover simple rules in the form:

$$a_1 \leq x_1 \leq b_1 \cap a_2 \leq x_2 \leq b_2 \tag{3.15}$$

where x_1 and x_2 represent two miRNAs, $[a_1, b_1]$ is the expression range of x_1, and $[a_2, b_2]$ is the expression range of x_2 (a_1 and a_2 can be $-\infty$; b_1 and b_2 can be $+\infty$). If every cancer sample's expression profile satisfies (falls into) the two specific expression ranges, but none of the normal sample profiles satisfies, then we say it is a 100%-frequency rule to differentiate the cancer samples from the normal samples. The complete form of this rule is denoted by

$$a_1 \leq x_1 \leq b_1 \cap a_2 \leq x_2 \leq b_2 \rightarrow cancer(100\%) \tag{3.16}$$

This work focuses on 2-miRNA or 3-miRNA 100%-frequency rules as biomarkers for the diagnosis. We do not identify 100%-frequency rules with 4 or 4+ miRNAs. Our rule discovery method is based on decision trees which usually generate rules combining 2 or 3 miRNAs with their specific expression ranges. Decision tree is a classical idea to induce a set of exclusive rules covering the training data only once, and thus the rules are sensitive

to a slight change of training data. Due to this constraint, using a single decision tree usually loses some prediction accuracy (Ho 1995).

It can be suggested that if the expression of x_1 is between a_1 and b_1 for a test normal sample, and the expression of x_2 is between a_2 and b_2, then this test sample is very likely to be a cancer cell. Similarly in this work, we also define 100%-frequency rules to differentiate normal samples from cancer samples. Such strong rules can be easily visualised in 2D spaces to facilitate biological interpretation of the computational results.

This method has two innovative parts. One is a novel idea to generate a committee of decision trees to discover 100%-frequency rules; the other is a simple projection method (gain ratio) to narrow down important miRNAs from the original miRNAs list for the induction of the decision tree ensemble.

Our feature ranking and projection method is good to select important miRNAs to derive 100%-frequency rules. However, some bias may occur as our list of "extensively studied miRNAs in the literature" may be far from complete. To ensure there is less bias, we search the whole feature space to find strong rules.

3.3.2 Strong Discriminatory Rules

Given a data set containing two classes of samples (positive and negative), we discover strong rules in the form:

$$\bigcap_{i=1}^{k} a_i \leq x_i \leq b_i, \tag{3.17}$$

where x_i represents a miRNA or a mRNA, $[a_i, b_i]$ is the expression range of x_i. If every positive sample's expression profile satisfies (falls into) the k specific expression ranges, but none of the negative sample profiles satisfies, then we say it is a 100%-frequency rule to differentiate the positive samples from the negative samples. The complete form of this rule is denoted by $\bigcap_{i=1}^{k} a_i \leq x_i \leq b_i \rightarrow positive(100\%)$. The key ideas are proposed by my chief supervisor Assoc. Prof. Jinyan Li.

This suggests that if the expression of every x_i is between a_i and b_i for a HCV test sample, then this test sample is very likely to be a positive sample. Similarly in this work, we also define the 100%-frequency rule to differentiate negative samples from positive samples. This study identifies simple 2-miRNA 100%-frequency rules (i.e., $k = 2$) to capture differentially expressed miRNAs and the miRNA expression changes in cancer and normal samples. We do not identify 3-miRNA 100%-frequency rules or the rules involving more than 3 miRNAs (i.e., $k > 3$). The stringent 100%-frequency may be unnecessary for other data sets as such a distinction may not exist. Therefore, this frequency requirement can be relaxed for other studies.

Therefore, our method is restricted to combine all possible 2- and 3-miRNAs and all possible valid expression ranges of these miRNAs to see whether the combined ranges satisfy every cancer sample's expression profile. If this is true, we then examine whether the combined ranges do not satisfy any of the normal samples. If this is true as well, then the combined expression ranges, together with the miRNAs, form a 100%-frequency rule to distinguish all of the cancer samples from all of the normal samples in 2D or 3D spaces. Similarly, we detect such rules to distinguish 100% of the normal samples from the cancer samples. We also use the distance separation technique to identify more reliable rules.

3.4 Bioinformatics Tools

3.4.1 GO Term Enrichment Analysis

Gene Ontology (GO) term enrichment is a technique for interpreting sets of genes making use of the Gene Ontology system of classification. For example, given a set of genes that are up-regulated (down-regulated) under certain conditions, an enrichment analysis will find which Gene Ontology terms are over-represented (or under-represented) using annotations for that gene set.

GO enrichment indicates the associations between genes and GO terms. For each gene set g and each GO term GO_j, a score is generated, which is

typically referred to as the gene ontology enrichment score and defined as the $-log_{10}$ of the hypergeometric test p value for a gene set G consisting of g's direct neighbours in STRING and the GO term GO_j, that can be calculated as follows:

$$S_{GO}(g, GO_j) = -\log_{10}(\sum_{m}^{n} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}})$$

where N denotes the overall number of proteins in humans, M denotes the number of proteins annotated in the gene ontology term GO_j, n denotes the number of proteins in G, and m denotes the number of proteins in G that are annotated in the gene ontology term GO_j. If the score is large for one gene set and one GO term, the gene set is associated with the GO term.

For every miRNA pair, a GO enrichment analysis (Biological Process subtype) is performed on their predicted targets to classify their functions. Only those GO terms which contain more than three genes with a significance level ($p < 1.0e-4$) are captured. Specifically, for a given miRNA pair (miRNA A and miRNA B), we use their intersecting target subsets which they co-regulate (i.e., subsets of $T(A) \cap T(B)$) to identify the biological processes under the hypergeometric distribution. Here $T(X)$ stands for the set of predicted targets of miRNA X. The analysis is preformed by the R software (GOstats and GO.db).

3.4.2 KEGG Pathway Enrichment Analysis

Kyoto Encyclopedia of Genes and Genomes (KEGG) stores a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for metabolism, genetic information processing, and environmental information.

Similarly, for each gene g and each KEGG pathway P_j, the KEGG enrichment score is defined as the $-log_{10}$ of the hypergeometric test P value for a gene set G that consists of g's direct neighbours in STRING and the

KEGG pathway P_j, which can be computed by:

$$S_{KEGG}(g, P_j) = -\log_{10}(\sum_{m}^{n} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}) \tag{3.18}$$

where N denotes the overall number of proteins in humans, M denotes the number of proteins annotated in the KEGG pathway P_j, n denotes the number of proteins in G, and m denotes the number of proteins in G that are annotated in the KEGG pathway P_j. Additionally, a higher KEGG enrichment score between g and P_j indicates a stronger relationship.

3.4.3 PPI Network Construction

The PPI network of a gene subset $T(A) \cap T(B)$ is represented by a graph, in which the proteins are represented by nodes and the interactions among them are represented by undirected edges. Using this gene subset as seed proteins, the construction of its PPI network is through the tool named UniHI (`http://193.136.227.168/UniHI/pages/unihiSearch.jsf`), which provides both experimentally determined and predicted interactions. The number of edges inserted between two seed proteins determines the network distance of the seed proteins. As found by the literature (Liang & Li 2007, Yuan et al. 2009), proteins interacting with cancer-related proteins are generally close to each other and interact more frequently compared to non-interacting proteins in the PPI networks. Therefore, we consider only those PPI networks with a primary distance no larger than 3. A primary distance between any two proteins in a PPI network is measured by the minimum number of edges required to connect them.

3.5 Performance Measurement

The prediction results for the biomarker problem can be represented by a confusion matrix consisting of four entries: a true positive (TP), a true negative (TN), a false positive (FP) and a false negative (FN). Accordingly, the prediction accuracy (ACC), specificity (SP), sensitivity (SN) and F1 score (F1) can be computed as follows:

- $ACC = \frac{TP+TN}{TP+TN+FP+FN}$
- $SP = \frac{TN}{TN+FP}$
- $SN = \frac{TP}{TP+FN}$
- $F1 = \frac{2TP}{(2TP+FP+FN)}$

In addition, the area under the receiver operating characteristic (ROC) cure (AUC) is used as a performance measure. Given a threshold parameter T, the instance is classified as "positive" if $X > T$, and "negative" otherwise. X follows a probability density $f_1(x)$ if the instance actually belongs to class "positive", and $f_0(x)$ if otherwise. Therefore, the true positive rate is given by $TPR(T) = \int_T^{\infty} f_1(x)dx$ and the false positive rate is given by $FPR(T) = \int_T^{\infty} f_0(x)dx$. The ROC curve plots TPR(T) versus FPR(T) with T as the varying parameter. It can be seen as follows:

$$\begin{aligned} A &= \int_{\infty}^{-\infty} TPR(T)FPR^{'}(T)dT \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(T^{'} > T)f_1(T^{'})f_0(T)dT^{'}dT = P(X_1 > X_0) \end{aligned} \tag{3.19}$$

where X_1 is the score for a positive instance and X_0 is the score for a negative instance.

Chapter 4

Rule Discovery and Distance Separation to Detect Reliable miRNA Biomarkers for the Diagnosis of Lung Squamous Cell Carcinoma

4.1 Introduction

As explained in the related work (Section 2.1), aberrant miRNA expressions have been linked to many diseases, and have been intensively investigated to discover miRNA biomarkers for the diagnosis of diseases including lung cancer (Raponi et al. 2009, Shen et al. 2010, Edmonston, Kushnir, Aharonov, Yanai, Benjamin, Bibbo, Thurm, Horowitz, Huang, Gilad et al. 2010). The inherent stability of miRNAs in serum and the reliability and reproducibility of expression analysis (Alevizos et al. 2011, Mraz et al. 2009, Li, Li, Zhou, Wen, Geng, Yang & Cui 2013) make them ideal candidates for biomarkers (Zeng, Cui, Wu & Lu 2014).

This work developed a novel method to find small numbers of miRNAs

that are able to separate healthy samples from Squamous Cell Carcinoma (SCC) samples with a clear and wide margin in 2D or 3D spaces. Our method was tested on the SCC miRNA expression data set from (Raponi et al. 2009). Many 2- and 3-miRNA groups (together with their specific expression ranges) were discovered as clear linear discriminant rules for the diagnosis of SCC. The basic idea of our method is the construction of an innovative committee of decision trees by using the C4.5 algorithm (Quinlan 1993*a*) iteratively. The preprocess of the data involves a prioritisation method to rank the whole number of miRNAs and then to focus on potential candidates by projecting wet-lab confirmed plasma and tissue miRNA biomarkers onto this ranked list of miRNAs ordered by miRNAs' gain ratio (Han & Kamber 2006*a*). This feature selection method is capable of recommending those highly ranked miRNAs not yet studied by wet-labs in the past for rule discovery, and capable of suggesting a good mapping between lung tissue-specific and plasma-specific miRNA biomarkers useful for a minimally invasive diagnosis. For the discovery of the most reliable rules, a distance separation technique is used to determine the Max-Min distance between the normal and cancer classes separated by each rule, and the widest distance is then taken to recommend the best rules. In addition, we also considered a computationally heavy method to detect rules from the whole feature space. We further demonstrated the reliability of these biomarkers by comparing the performance of the most reliable 2-miRNA (3-miRNA) rules with those of 1000 randomly selected 2 miRNAs (3 miRNAs) with C4.5 decision tree classifier and 10-fold cross validation, and performing a resampling test by disordering the class labels.

For all of the miRNAs involved in our 2-miRNA rules, we examined their chromosomal locations and their common target genes. We also established links between the diseases and chromosomal locus with the common target genes to show that most of the chromosomal loci have a high frequency of genomic alteration in lung cancer and that two sets of our biomarkers have confirmed associations with lung cancer.

This chapter, describing **Contribution 1**, is an extended description of my publication (Song, Liu, Hutvagner, Nguyen, Ramamohanarao, Wong & Li 2014).

4.2 Materials and Methods

4.2.1 Data Sets of miRNA Expressions in SCC Patients

Two data sets are used by this work. Data set 1 is a collection of miRNA expressions in SCC tissues which have been studied by Raponi et al. (Raponi et al. 2009) for comparative analysis of differentially expressed miRNAs between normal and SCC tissues. Here, it is used for rule discovery. In this data set, there are 61 SCC tissue samples and 10 matched adjacent normal lung tissue samples for the miRNA expression profiling. These samples were collected from patients in the University of Michigan Hospital between October 1991 and July 2002 with patient consent and institutional review board approval. Total RNAs of these 71 samples were preprocessed and then profiled on MirVan miRNA Bioarray (version 2, Ambion) which contains 328 human miRNA probes. Accordingly, this data set is a 71 x 328 relational table with each row associated with a class label "cancer" or "normal". The original miRNA expression data was normalised by the quantile and log2 methods, and it was stored at the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE16025.

Data set 2 (Tan, Qin, Zhang, Hang, Li, Zhang, Wan, Zhou, Shao & Sun 2011*b*) is used as an independent data set to assess the importance of our rules. Data set 2 comprises 187 cancer tissues and 174 adjacent normal tissue from patients described by the expression levels of 549 miRNAs. The expression levels in this data set were processed by subtracting the background as average values of the replicate spots of each miRNA and filtering out the expression signal of faint spots below 600. This data set can be downloaded from the Gene Expression Omnibus under the accession

number GSE15008. Since it is impossible to confirm the 34 paired cancerous and adjacent normal samples described by Tan et al. (Tan, Qin, Zhang, Hang, Li, Zhang, Wan, Zhou, Shao & Sun 2011*b*) from all the published studies, we are unable to choose this large sample size as the training set.

4.2.2 Rule Discovery within Top-Ranked miRNAs

The rule discovery method is described in the Section 3.3.1. This work focuses on 2-miRNA or 3-miRNA 100%-frequency rules as biomarkers for the diagnosis of SCC. We do not identify 100%-frequency rules with 4 or 4+ miRNAs. Our rule discovery method is based on decision trees which usually generate rules combining 2 or 3 miRNAs with their specific expression ranges. Decision tree is a classical idea to induce a set of exclusive rules covering the training data only once, and thus the rules are sensitive to a slight change of training data. Due to this constraint, using a single decision tree usually loses some prediction accuracy (Ho 1995).

Our method has two innovative parts. One is a novel idea to generate a committee of decision trees to discover 100%-frequency rules; the other is a simple projection method to narrow down important miRNAs from the 328 miRNAs for the induction of the decision tree ensemble.

As the first step of the projection method, we prioritise and rank the 328 miRNAs in the data set based on their gain ratios over the 71 samples' expression profiles. Gain ratio (Han & Kamber 2006*a*) measures a collective difference of every single miRNA's expressions between the two classes. A high gain ratio indicates that the miRNA is a high-potential biomarker differentially expressed over the two classes. As the second step, we project wet-lab confirmed and intensively studied miRNAs onto this rank list. Using this step, we can recommend those highly ranked miRNAs that have not been studied in wet-labs in the past for rule discovery and potentially for fresh biological study.

In this work, we use 5 plasma biomarkers (miR-486, miR-126, miR-182, miR-210 and miR-21) identified in 28 NSCLC patients including 14

adenocarcinoma and 14 SCC patients (Shen et al. 2010) for the above rank list projection. All of these miRNAs are confirmed as key biomarkers in early lung cancer diagnosis. These miRNAs in plasma are also a subset of 12 previously identified tissue biomarkers validated by paired SCC tissues and noncancerous tissues associated with early-stage lung cancer (Yu et al. 2010). So these 5 miRNAs can serve as a guideline for the next step of tissue-specific biomarkers identification.

The projection of the 5 plasma biomarkers against the list of prioritised 328 miRNAs is shown in Table 4.1. The 5 confirmed miRNAs are mapped to positions 1, 3, 5, 13 and 19. However, none of these 19 individual miRNAs is a good biomarker to separate the two classes of data as shown in Figure 4.1. So, we concentrate on the entire expression data of these 19 miRNAs to derive groups of miRNAs for 100%-frequency rules. The remaining data (i.e., excluding the 19 miRNAs) is used for comparison to examine the effectiveness of our rule discovery method.

Table 4.1: Projection of 5 important miRNAs onto a prioritised list of 328 miRNAs, resulting in 19 miRNAs ranked as high as these 5 miRNAs.

miRNA	Rank	GE	P-value	miRNA	Rank	GE	P-value
miR-486	1	Down	3.12e-05	miR-125a	11	Down	8.857e-02
miR-98	2	Down	4.631e-07	miR-93	12	Up	6.401e-06
miR-126	3	Down	1.14e-02	**miR-210**	13	Up	5.548e-12
miR-205	4	Up	3.678e-07	miR-224	14	Up	2.866e-14
miR-182	5	Up	2.2e-16	miR-17-5p	15	Up	3.646e-11
miR-106b	6	Up	1.224e-09	miR-373-AS	16	Down	3.647e-03
miR-133a	7	Down	4.208e-03	miR-483	17	Down	4.11e-02
miR-513	8	Down	2.263e-02	miR-139	18	Down	3.812e-03
miR-451	9	Down	2.713e-05	**miR-21**	19	Up	1.293e-04
miR-331	10	Up	4.124e-02				

To construct a committee of decision trees for the discovery of multiple 100%-frequency rules, we induce the first decision tree from the 19-miRNA

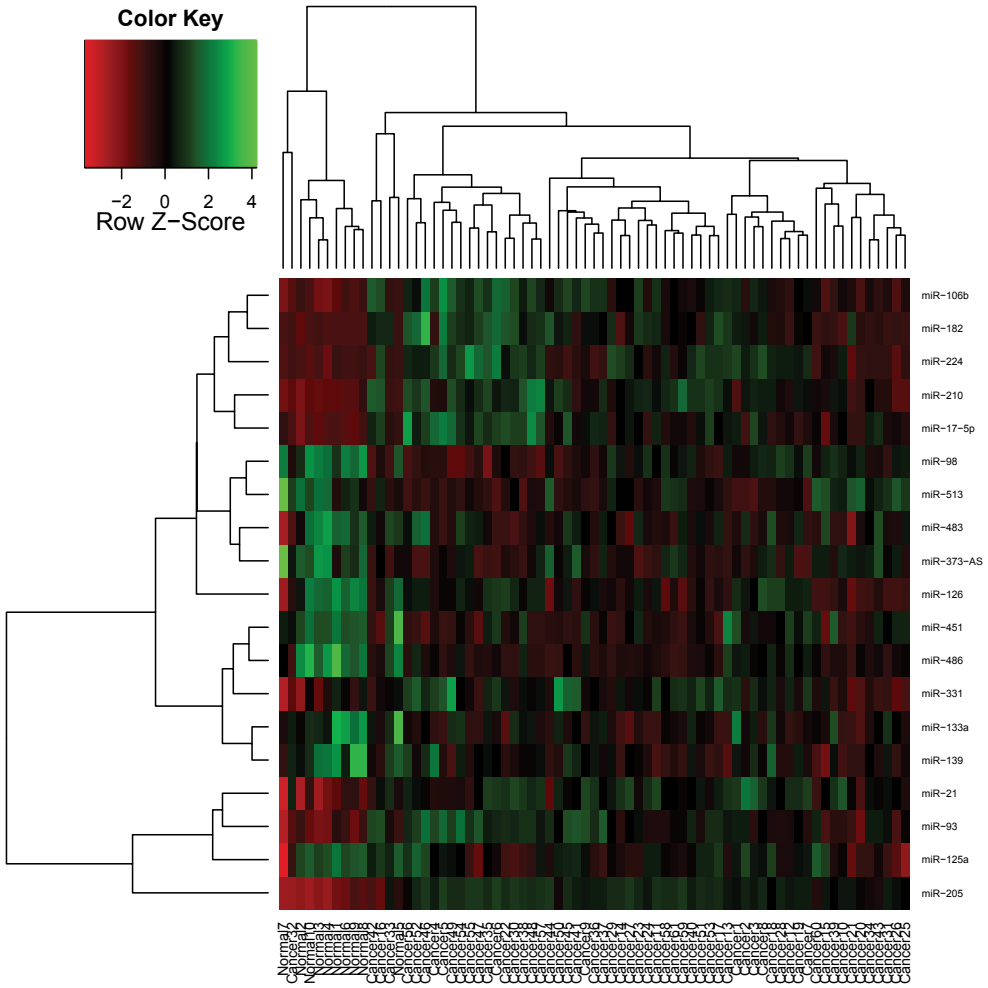


Figure 4.1: **Heatmap representation of the expression levels of the 19 miRNAs.** A single miRNA is unable to distinguish cancer samples from normal samples, while the combination of 2 or 3 miRNAs can faultlessly identify cancer (or normal) samples from normal (or cancer) samples.

data set. To induce the second tree, we remove the field (attribute values from the data) of the root node miRNA of the first tree from the data set. Iteratively, we construct a subsequent decision tree by removing the data of the root node miRNA of the current tree. This process continues until there are only two miRNAs left in the data set. We use the R software

package (Team 2013) and its C4.5 implementation to construct each decision tree (The source code of the algorithm constructing a committee of decision trees is described in the Appendix: Algorithm of Prim Code).

Every 100%-frequency rule with two or three miRNAs can separate the cancer samples clearly from the normal samples in 2D or 3D spaces. As a wider separation suggests a more reliable biomarker rule (Figure 4.2), we measure the separation extent by using the shortest pair-wise Euclidean distance between the cancer and normal samples. When multiple 100%-frequency rules are generated, further data analysis is on those with a wider separation distance (i.e., the Max-Min distance).

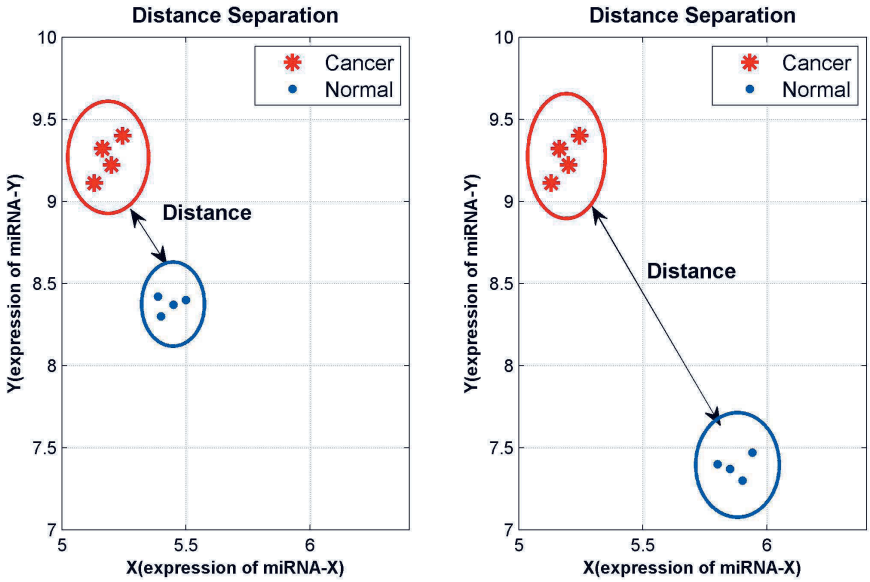


Figure 4.2: **Distance separation by 100%-frequency rules in 2D space.** The left panel shows a shorter distance separation between the cancer and normal samples than the separation shown in the right panel.

The entire work flow of our rule discovery method with feature space projection and distance separation is summarised in Figure 4.3.

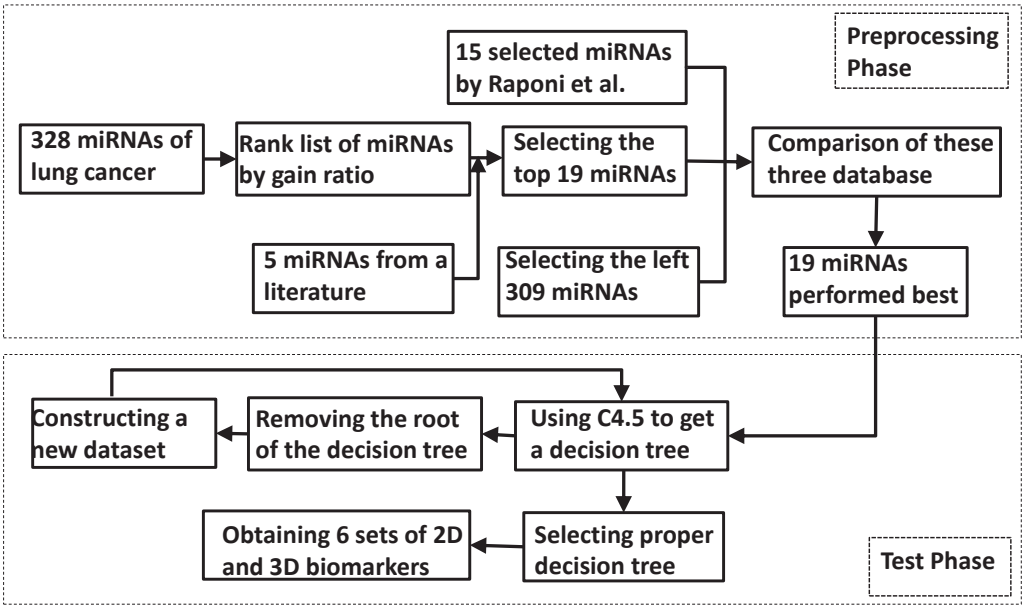


Figure 4.3: **The procedure of rule discovery with 19 miRNAs.** The up panel is the dataset processing phase, and 19 miRNAs are obtained. The down panel is the discovery phase to get biomarkers.

4.2.3 Rule Discovery across the Whole Feature Space

Our feature ranking and projection method is good to select important miRNAs to derive 100%-frequency rules. However, some bias may occur as our list of "extensively studied miRNAs in the literature" may be far from complete. To ensure there is less bias, we search the whole feature space, namely across all of the 328 miRNAs, to find strong rules. However, the exploration of every possible combination of these 328 miRNAs leads to exponentially computational cost.

Therefore, our method is restricted to combine all possible 2- and 3-miRNAs and all possible valid expression ranges of these miRNAs to see whether the combined ranges satisfy every cancer sample's expression profile. If this is true, we then examine whether the combined ranges do not satisfy any of the normal samples. If this comes true as well, then the combined expression ranges, together with the miRNAs, form a 100%-frequency rule

to distinguish all of the cancer samples from all of the normal samples in 2D or 3D spaces. Similarly, we detect such rules to distinguish 100% of the normal samples from the cancer samples. We also use the distance separation technique to identify more reliable rules.

4.3 Results

Our results are presented in five parts. The first part reports 2-miRNA and 3-miRNA rules and classification performance. The second part is related to distance separation of the rules in 2D or 3D spaces. The third part illustrates the reliability of the identified best miRNA rules. The fourth part presents the chromosomal locations of the miRNAs, and the last part is related to association studies between miRNA biomarkers and disease genes.

4.3.1 Prediction Performance by Rules

Comparison with Literature Methods

To show the effectiveness of our feature projection method on prediction accuracy, we compared the prediction performance of three commonly used classifiers on four data sets. One is the data set prepared by Raponi et al. (Raponi et al. 2009) which consists of 15 differentially expressed miRNAs extracted from the initial 328 miRNAs. The second data set contains only the 5 plasma miRNAs (Shen et al. 2010) which we used to project out our top-ranked 19 miRNAs. The third data set is our data set consisting of the 19 top-ranked miRNAs (Table 4.1). The fourth data set contains all the data after the removal of the third data set (the 19-miRNA data set) from the 328-miRNA data set. Note that there is not much miRNA overlapping between the first and third data set (only 6 miRNAs in common). We used the k-nearest neighbour classifier (KNN, k=1), Naive Bayes (NB), and the C4.5 decision tree (C4.5) classifier to conduct the prediction under a 10-fold cross-validation scheme.

Table 4.2 shows the prediction performance (specificity, sensitivity, F1 measure and receiver operating characteristic (ROC) area) of the three classifiers on these four data sets. It can be seen that the three classifiers all performed better on the 5-plasma miRNAs data set and on our 19-miRNA data set than on the other two data sets. This indicates that the 5 plasma biomarkers are indeed good biomarkers, and the 19 prioritised and projected miRNAs are indeed good potential candidates for rule discovery and biomarker identification.

Table 4.2: Comparisons of three classifiers on four data sets

Data sets	Algorithms	Specificity	Sensitivity	F-Measure	ROC Area
15 miRNAs (Raponi et al. 2009)	KNN	0.9833	0.8182	0.975	0.934
	NB	0.9833	0.8182	0.975	0.934
	C4.5	0.9516	0.7778	0.959	0.827
5 miRNAs (Shen et al. 2010)	KNN	0.9839	1.0000	**0.992**	0.944
	NB	0.9839	1.0000	**0.992**	0.989
	C4.5	0.9672	0.8000	**0.967**	0.84
19 miRNAs (top ranked)	KNN	0.9839	1.0000	**0.992**	0.944
	NB	0.9836	0.9000	**0.984**	0.946
	C4.5	0.9524	0.8750	**0.968**	0.798
309 miRNAs (lower ranked)	KNN	0.9833	0.8182	0.975	0.926
	NB	0.8413	0.6250	0.935	0.779
	C4.5	0.8413	0.3846	0.891	0.666

Multiple Rules Derived from the Top-Ranked 19 miRNAs

We applied C4.5 to our 19 top-ranked miRNAs data set to construct the first decision tree (denoted by DT1). As described in the Method section, we then removed the root node miRNA of DT1 from the data set to construct the second tree (denoted by DT2). By iteration, we constructed a total of 18 decision trees. Interestingly, DT1 does not contain any 100%-frequency rules. In fact, only 6 of the 18 decision trees (DT2, DT3, DT4, DT9, DT10,

and DT15) contain rules consisting of 2 or 3 miRNAs. Table 4.3 shows the details of the rules and expression ranges of these 2D and 3D biomarkers.

Table 4.3: Multiple 100%-frequency rules derived from the 19-miRNA data set through our iterative decision tree method.

Tree ID	miRNAs, their expression ranges and the rules
DT2	7.356 $\leq$miR-98$\leq$ 8.123 $\cap$ 5.105 $\leq$miR-205 $\leq$ 9.601 $\rightarrow$Normal(100%)
DT3	6.145 $\leq$miR-126$\leq$ 8.825 $\cap$ 5.105 $\leq$miR-205 $\leq$ 9.601 $\cap$5.551 $\leq$miR-182$\leq$ 8.966 $\rightarrow$Cancer(100%)
DT4	6.148 $\leq$miR-451$\leq$ 8.054 $\cap$ 5.105 $\leq$miR-205 $\leq$ 9.601 $\rightarrow$Normal(100%)
DT9	4.760 $\leq$miR-133a$\leq$ 5.493 $\cap$ 5.745 $\leq$miR-210$\leq$ 9.780 $\cap$4.662 $\leq$miR-373-AS$\leq$ 5.731 $\rightarrow$Cancer(100%)
DT10	5.014 $\leq$miR-224$\leq$ 9.417 $\cap$ 4.662 $\leq$ miR-373-AS$\leq$ 5.731 $\cap$4.760 $\leq$miR-133a$\leq$ 5.493 $\rightarrow$Cancer (100%)
DT15	4.2032 $\leq$miR-139$\leq$ 5.858 $\cap$ (4.760 $\leq$miR-133a$\leq$ 5.400 $\cap$6.129 $\leq$ miR-513$\leq$ 7.853 $\cup$ (5.400 $\leq$miR-133a$\leq$ 5.4927 $\cap$7.507 $\leq$miR-513$\leq$ 7.853)) $\rightarrow$Cancer (100%)

As an example, Figure 4.4 displays the tree structures of DT2 and DT4. Both of them contain only two miRNAs. The 100%-frequency rules derived from these two trees separate the cancer and normal samples in a way as shown in Figure 4.5 where the x-y axis of the 2D planes represents the expression ranges of these miRNAs. We obtained the left and right bounds of the two rules from the rectangles in the planes.

Classification Performance under 5-fold Training-Test Experiments

The derived rules above can separate the two classes of samples clearly without any mistake. However, they are derived from the top-ranked miRNAs based on all of the 71 samples. To demonstrate the generalisation ability of the rules induced by our method, we conducted C4.5's 5-fold

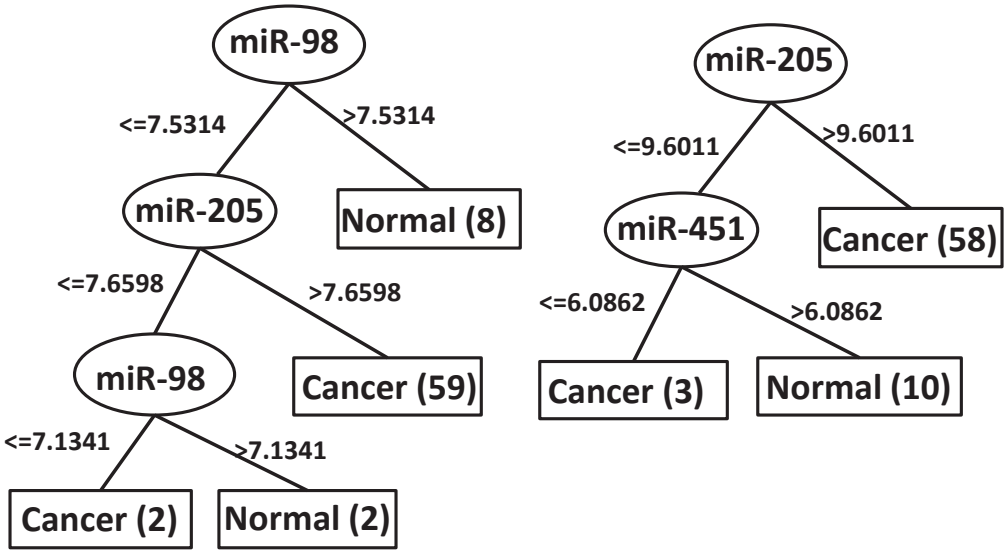


Figure 4.4: **Decision trees.** The left panel is a decision tree made of miR-205 and miR-98. The right panel is a decision tree made of miR-205 and miR-451.

training-test experiments. The initial 10 normal samples and 61 cancerous samples are randomly divided into 5 parts. Four parts of the data set were used as a training data set, and 5 training data sets were constructed (TrS1, TrS2, TrS3, TrS4 with 57 samples, and TrS5 with 56 samples). Correspondingly, the remaining part was reserved as a test data set, and 5 test data sets were constructed (TeS1, TeS2, TeS3, TeS4 with 14 samples, and TeS5 with 15 samples, each containing two normal samples). By our method, the gain ratio and the 5 plasma miRNAs projection method were applied to select miRNAs from the 5 training sets. Actually we obtained 27, 21, 14, 32, and 20 top-ranked miRNAs respectively. Then the rules were derived within these top-ranked miRNAs and the Max-min distance step was applied to determine the most reliable rule. The TrS1, TrS2, TrS4, and TrS5 training data sets have the same best rule (made from miR-205 and miR-451), while the TrS3 has the rule made from miR-205 and miR-21. Finally, we applied these reliable rules to the corresponding test sets, and all achieved an

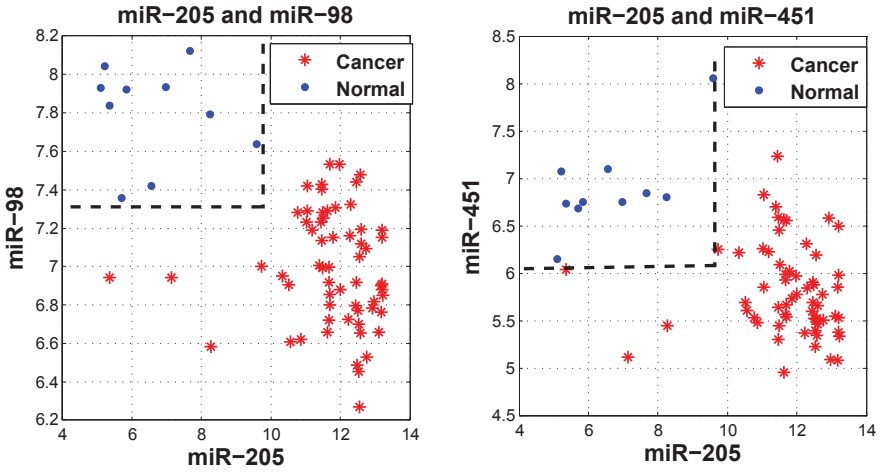


Figure 4.5: **Expression data on 2D planes.** The left panel is the plane co-ordinated by miR-205 and miR-98. The right panel is coordinated by miR-205 and miR-451. The blue rectangles indicate the expression ranges of all of the normal samples.

accuracy of 100%, except TeS4 with 92.86% (1 cancer sample misclassified). The details are described in Attached file 1.

Assessing the Importance of MiRNA Biomarkers by Using an Independent Data Set

Data set 2 (Tan, Qin, Zhang, Hang, Li, Zhang, Wan, Zhou, Shao & Sun 2011*b*) contains miRNA expression data of 187 cancer tissues and 174 adjacent normal tissue from patients. The platform for generating data set 2 (the National Engineering Research Center mammalian microRNA microarray with 549 human miRNAs) is different from the platform of data set 1 (MirVan miRNA Bioarray, version 2). The two data sets are preprocessed by different methods as well. Because of these differences, it

is not reasonable to directly test the miRNA expression ranges on data set 2 for a rule derived from data set 1. However, the miRNAs in a rule of data set 1 can be still validated on the data set 2 by testing whether these miRNAs are able to classify the samples in data set 2 with a high accuracy. A high classification performance would suggest that these miRNAs are robust across different data sets and thus they are worth further investigation. We note that the miRNAs in a rule from data set 1 are detected independently from data set 2.

To test whether the miRNA biomarkers discovered from data set 1 have a good generalisation ability, we carried out 10-fold cross-validation on the expression data of only these miRNAs of data set 2 (the independent data set) to see the classification performance in C4.5. We compared the sensitivity, specificity, accuracy, ROC area and F-measure for three data sets: data set 2 of 549 miRNAs, the data set of top-ranked 158 miRNAs, and the data set of 3 miRNAs (miR-126, miR-205 and miR-182) which are from the best rule from data set 1 (with the largest distance 0.7799). The classification performance on these three data sets are shown in Table 4.4. We can see that the classification using just the 3 miRNAs from the best rule of data set 1 achieved an accuracy of 84.49%, sensitivity of 91.40% and specificity of 77.14%.

Table 4.4: The performance comparison of three datasets.

Data sets	Sensitivity	Specificity	Accuracy	ROC area	F-measure
549 miRNAs	0.8441	0.8343	0.8393	0.817	0.844
158 miRNAs	0.8656	0.8111	0.8393	0.845	0.847
3 miRNAs	0.9140	0.7714	0.8449	0.853	0.859

This performance is better than the classification performance by using all miRNAs in data set 2. Although the specificity decreases, the cost in real-life diagnosis would be lower using the just 3 miRNAs, because the cost of misclassifying 'normal' as 'cancer' is much smaller than misclassifying 'cancer' as 'normal'.

These results demonstrate that the miRNA biomarkers identified from data set 1 are also biomarkers to separate the two classes of samples in the independent data set 2 with a high accuracy. This implies that our miRNA biomarkers have a good generalisation ability in classification.

Rules Derived by Using the Whole Feature Space

On the whole feature space, our rule mining method detected a total of 14 new 100%-frequency rules each of which combines only two or three miRNAs, in addition to the 6 rules identified by the decision tree committee. Two of them are displayed in Figure 4.6. The rules are: $7.970 \leq let-7a \leq 11.989 \cap 5.105 \leq miR-205 \leq 9.601 \rightarrow Normal(100\%); 7.755 \leq miR-103 \leq 9.879 \cap 6.145 \leq miR-126 \leq 8.825 \rightarrow Cancer(100\%)$. Again, it can be seen that these two sets of biomarkers are able to distinguish the 71 cancer and normal samples with no mistake. Examples of 3-miRNA 100%-frequency rules are shown in Figure 4.7. The rules are: $4.760 \leq miR-133a \leq 5.844 \cap 7.381 \leq miR-21 \leq 11.014 \cap 4.324 \leq miR-520a-AS \leq 5.229 \rightarrow Cancer(100\%); 5.165 \leq miR-100 \leq 8.706 \cap 5.518 \leq miR-199a \leq 7.091 \cap 5.814 \leq miR-200c \leq 9.890 \rightarrow Normal(100\%)$.

4.3.2 Distance Separation in 2D and 3D Spaces to Identify Reliable Biomarkers

We calculated the Euclidean distance for the rules discovered from the whole data set 1 (i.e., the 71 samples), and used the shortest pair-wise distance and the Max-Min technique to identify the best miRNA biomarkers (Table 4.5). For 2D spaces, we used the distance cut-off threshold 0.20 to focus our further biological analysis, and cut-off the threshold 0.45 in 3D spaces.

From Table 4.5, it can be seen that miR-205 and miR-98 constitute our best 2D rule that

$$7.356 \leq miR-98 \leq 8.123 \cap 5.105 \leq miR-205 \leq 9.601 \rightarrow Normal(100\%) \tag{4.1}$$

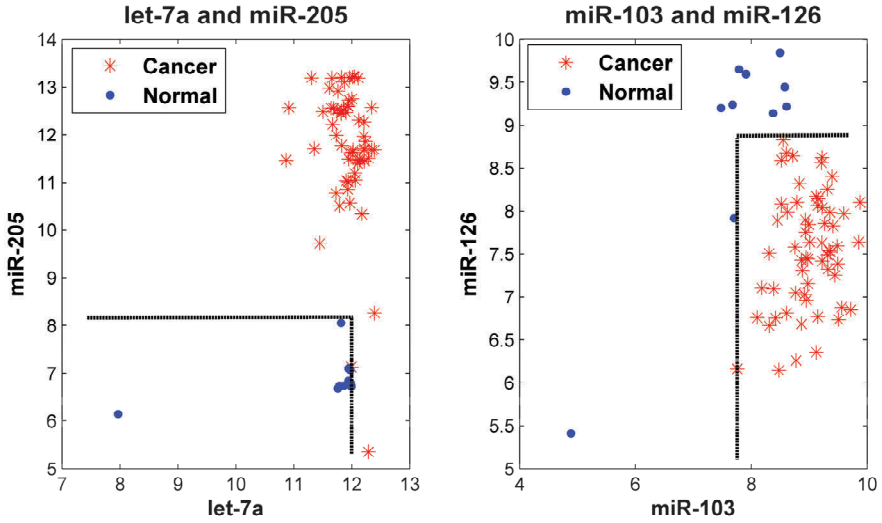


Figure 4.6: **Examples of 2D rules.** The left panel describes two miRNAs whose class-label is related to normal. The right panel shows two miRNAs whose class-label is related to cancer.

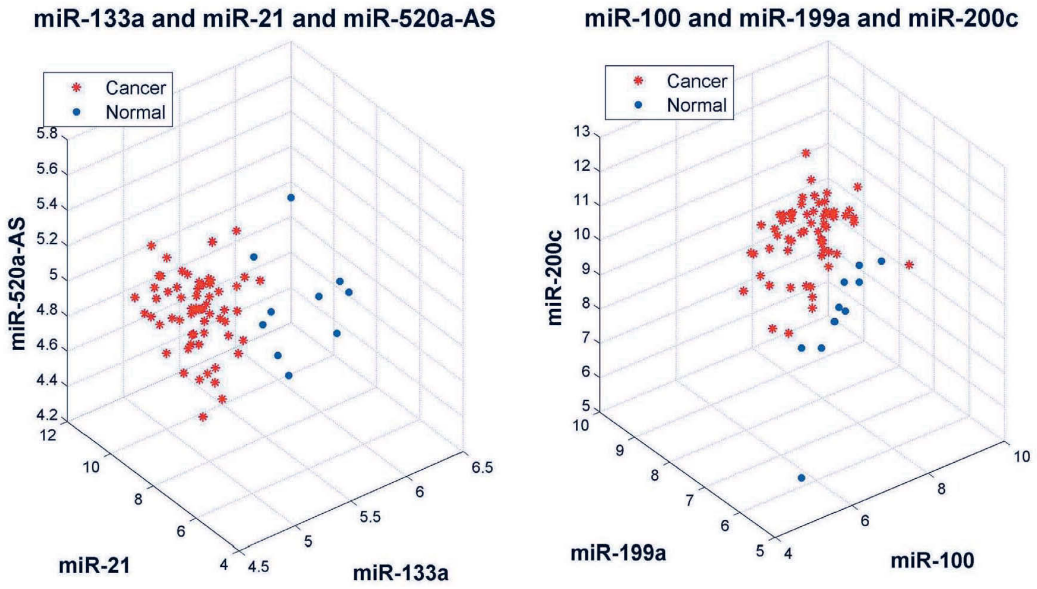


Figure 4.7: **Examples of 3D rules.** The left panel contains miR-100, miR-199a and miR-200c. The right panel contains miR-133a, miR-21 and miR-520a-AS.

Table 4.5: Shortest pair-wise Euclidean distance between the cancer and normal samples in 2D and 3D biomarker spaces.

Methods	miRNAs in the Rules	Shortest Distance	Rank
Rule discovery within the 19 top-ranked miRNAs	**miR-205 and miR-98**	**0.5421**	**2D.1**
	miR-205 and miR-451	**0.4311**	**2D.2**
	miR-126, miR-205 and miR-182	**0.7799**	**3D.1**
	miR-210, miR-373-AS and miR-133a	0.1068	3D.9
	miR-224, miR-373-AS and miR-133a	0.1786	3D.5
	miR-133a, miR-513 and miR-139	0.1238	3D.8
Rule discovery across the whole feature space (328 miRNAs)	**Let-7a and miR-205**	**0.2496**	**2D.4**
	miR-103 and miR-126	**0.3591**	**2D.3**
	Let-7b and miR-486	0.0835	2D.11
	miR-106b and miR-29b	0.1498	2D.7
	miR-137 and miR-98	0.1660	2D.6
	miR-149 and miR-182	0.0941	2D.9
	miR-133a, miR-21 and miR-520a-AS	**0.4515**	**3D.3**
	miR-210 and miR-98	0.1892	2D.5
	miR-133b, miR-139 and miR-210	0.2459	3D.4
	Let-7i, miR-130a and miR-224	0.1231	3D.7
	miR-324-3p and miR-43	0.0879	2D.10
	miR-17-5p and miR-451	0.1398	2D.8
	miR-1, miR-106a and miR-203	0.1589	3D.6
	miR-100, miR-199a and miR-200c	**0.7275**	**3D.2**

for the diagnosis of lung cancer. In fact, this rule separates the normal and cancer classes by a distance of at least 0.5421 in 2D space. Their chromosomal locations, common target genes, and associations with disease genes are presented in a later part.

Classification performance on the data of only these two miRNAs was also evaluated. The performance (F1 Measure: KNN-1.000, NB-0.984, C4.5-0.976) is higher than that on the 19-miRNA data set, or on the 15-miRNA

data set (Table 4.2).

The other three important 2D rules are formed by miR-205 and miR-451, by miR-103 and miR-126, or by Let-7a and miR-205. The best 3D rule is formed by miR-126, miR-205 and miR-182; the second best is by miR-100, miR-199a and miR-200c; and the third best is by miR-133a, miR-21 and miR-520a-AS.

4.3.3 The Reliability of Identified Best 2D and 3D Biomarkers

We applied the 10-fold cross-validation test on the best 2D (miR-205 and miR-98) and 3D rules (miR-126, miR-205 and miR-182) to see the classification performance by C4.5 (R package RWeka). We further performed a randomisation test to see whether the best 2D (or 3D) miRNAs are better predictors than randomly selected 2 miRNAs (or 3 miRNAs). The random selection was repeated 1000 times. All the area under ROC curves (AUCs) were calculated and compared. The best 2D rule had an average AUC=1.0 in the 10-fold cross-validation, and the best 3D rule had an average AUC=0.9975. For the randomly selected 2 miRNAs, only a probability of 0.007 could produce an AUC$\geq$0.999 for the 1000 repeated tests. For the randomly selected 3 miRNAs, only a probability of 0.012 could produce an AUC$\geq$0.9975. The probabilities in different AUC scales are shown in Table 4.6. These results indicate that our miRNA biomarkers are significant and reliable, instead of random. We further performed a resampling test by disordering the class labels, and no rules were found using our method.

4.3.4 The Genomic Location of Biomarker miRNAs

Many known human miRNAs reside in particular genomic regions that are prone to alteration in cancer cells. For example, the main chromosomal alteration loci of miR-15 and miR-16 are identified at 13q14 with down-regulation, which is the first association study between miRNA genes and

Table 4.6: The probability of different AUC values in the 1000 randomization tests.

2-miRNA AUCs	Probability	3-miRNA AUCs	Probability
≥ 0.9	0.177	≥ 0.9	0.328
≥ 0.95	0.089	≥ 0.95	0.19
≥ 0.98	0.035	≥ 0.98	0.091
≥ 0.99	0.025	≥ 0.99	0.062
≥ 0.998	0.009	≥ 0.9975	0.02
≥ 0.999	0.007	≥ 0.999	0.012

cancer (Breu, Gil, Kirkpatrick & Werman 1995, Dostie, Mourelatos, Yang, Sharma & Dreyfuss 2003). We obtained the chromosomal locations of all of the 13 miRNAs in the 100%-frequency rules of a wide separation in 2D and 3D spaces (the 7 top-ranked rules in Table 4.5). This location information was obtained through a keyword search from the miRNAMap database (`mirnamap.mbc.nctu.edu.tw`) and miRBase database (`www.mirbase.org`) (Griffiths-Jones 2006, Hsu, Huang, Hsu, Lin, Tsou, Tseng, Stadler, Washietl & Hofacker 2006). For the miRNAs let-7a, miR-133a and miR-199a, we obtained three loci for each of them. Details are presented in Table 4.7.

Table 4.7: The chromosomal location of the 13 miRNAs in our 2D and 3D biomarker rules

miRNAs	Chr location	miRNAs	Chr location
Let-7a-1,-2,-3	9q22.2,11q24.1,	miR-199a-1,-2	19p13.2, 1q23.2
	22q13.3	miR-133a-1,-2	18q11.1,20q13.3
miR-21	17q23.2	miR-200c	12p13.31
miR-98	**Xp11.2**	**miR-205**	**1q32.2**
miR-100	11q24.1	miR-451	17q11.2
miR-126	9q34	miR-520a-AS	19q13.42

It has been previously reported that there are many chromosomal arms

having frequent loss of heterozygosity (Calin, Sevignani, Dumitru, Hyslop, Noch, Yendamuri, Shimizu, Rattan, Bullrich & Negrini 2004), such as 1p, 3p, 4p, 4q, 5q, 8p, 9p (p16), 9q, 10p, 10q, 13q (Rb), 15q, 17p (p53), 18q, 19p, Xp, and Xq, in frequency order for lung cancer (Alevizos et al. 2011, Girard, Zchbauer-Mller, Virmani, Gazdar & Minna 2000, Griffiths-Jones 2006, Hsu et al. 2006). In this study, we identified some new chromosomal arms such as 11q, 22q, 17q, 20q, 1q and 12p. In particular, the best 2D rule biomarkers miR-98 and miR-205 are located at Xp11.2 and the new arm 1q32.2. In fact, these two arms have been studied before for various purposes. It was reported by Prot et al. (Prot, Boccon-Gibod, Bouvier, Doz, Fournet, Frneaux, Vieillefond & Couturier 2003) that there are 5 cases of renal cell carcinoma with translocation involving Xp11.2 in children. It was found by Gregory et al. (Gregory, Bert, Paterson, Barry, Tsykin, Farshid, Vadas, Khew-Goodall & Goodall 2008) that chromosome 1q32.2, based on an alignment of the mature miR-205, controlled epithelial-to-mesenchymal transition. It was also claimed by Meyer et al. (Meyer, Clark, Flanigan & Picken 2007) that renal cell carcinomas are associated with Xp11.2 translocation in five adult patients. Sham et al. (Sham, Tang, Fang, Sun, Qin, Wu, Xie & Guan 2002) identified several nonrandom chromosomal changes in 31 primary ovarian carcinomas in Chinese women, including gains of 1q (10 cases, 32%), and that the losses of 1q32.2 were observed as alterations in comparative genomic hybridisation studies. These results showing the alterations of these two locations in cancers support our suggestion that combining miR-98 and miR-205 is a good approach to lung cancer study.

4.3.5 Target Genes of Biomarker miRNAs and Their Associated Diseases

For each 100%-frequency rule containing 2 or 3 miRNAs, we detected target mRNAs of these miRNAs. Then we identified their common targets. From these common targets, we also linked to the OMIM disease database to examine disease gene information.

The target genes of the miRNAs in the 4 top-ranked 2D rules (Table 4.5) were extracted from the TargetscanHuman database (`www.targetscan.org`) (Dweep, Sticht, Pandey & Gretz 2011). All of them have many target genes. For example, miR-451, -126, -98, -205, -103 and let-7a have 20, 25, 46, 415, 531 and 84 target genes respectively. Then we looked at the common target genes of the miRNAs involved in one rule. Interestingly, the common targets are not many. For example, miR-98 and miR-205 have only two common targets FZD3 and RPS6KA3. Details are shown in Table 4.8.

Table 4.8: The targets and associated disease of our biomarkers

Biomarkers	Common targets	OMIM gene/ disorder	Relate to lung cancer or carcinoma
miR-98 and miR-205	FZD3 RPS6KA3	606143/- 300075/303600	**carcinoma** **squamous cell carcinoma**
miR-451 and miR-205	AEBP2	–/–	irrelevant
Let-7a and miR-205	PARD6B NKD1 MAP3K2 RBMS2 EPB41	608975/– 607851/– 609487/– 602387/– 130500/61804	irrelevant **lung cancer** irrelevant irrelevant **lung cancer**
miR-103 and miR-126	AKAP13	604686/–	irrelevant

The first and third top-ranked miRNA pairs (Table 4.8) have an opposite change of expression in normal samples compared to the disease samples. These pairs of miRNA may affect different complementary pathways. It is possible that the down regulated miRNA inhibited a transcription factor that regulates the other miRNA. On the other hand, the common targets of the pairs of miRNAs are sensible only when (i) down-regulation of their common targets cause cancers, and (ii) their common targets have normal or high expression in normal tissues. For example, NKD1, FZD2 and EPB41

fit the biological behaviour expected above. Especially, down regulation of NKD1 (common target of let-7a and miR-205) increases the invasive potential of NSCLC (Zhang, Wang, Dai & Wang 2011). FZD3 works the same way ("The proliferation and invasion ability of SACC-M cells were enhanced when the expressions of FZD2 and FZD3 genes were inhibited in SACC-M cells" http://mt.china-papers.com/7/?p=6645). EPB41 (common target of let-7a and miR-205) is another example that works this way. It is absent in most NSCLC cancer. Its presence suppresses these lung cancer cells' growth (`www.wikigenes.org/e/gene/e/2035.html`) (Zheng, Qi, Gao, Wang, Qi, Shi & An 2009).

From these target genes, we further conducted disease gene analysis. First, we obtained the common target genes' Online Mendelian Inheritance in Man (OMIM) information and their associated diseases from Human Disease Gene List (http://www.genecards.org/) with the target genes' name. To this end, we compared the associated diseases of these biomarkers. It was found that: (i) the two miRNAs (miR-98 and miR-205) involved in our best rule have both been confirmed as being associated with carcinoma; and (ii) let-7a and miR-205 (in the second best rule) have been confirmed to be directly associated with lung cancer. On the other hand, we did not find evidence in the literature to show the pair miR-451 and miR-205, or the pair miR-103 and miR-126 linked to lung cancer in any way (Table 4.8).

4.4 Discussion

As described, this work applied a new rule discovery and distance separation technique to discover 2D and 3D 100%-frequency rules for lung SCC diagnosis. We constructed a data set consisting of 19 important miRNAs by projecting 5 plasma miRNA biomarkers onto the whole list of 328 miRNAs ordered by gain ratio. Classification performance on this data set is better than on other data sets. This study also provides knowledge so that we can develop potential non-invasive or minimally invasive diagnostic biomarkers

for early lung cancer diagnosis. Of the 5 previously intensively studied plasma miRNAs, three of them (miR-21, miR-126 and miR-182) have been employed to form our diagnostic rules for lung tissue diagnosis. So, these 2D and 3D rules and the corresponding miRNAs identified from the tumour tissues may be good plasma miRNA biomarkers as well.

The present study suggests that a minimal 2-miRNA or 3-miRNA rule can distinguish lung SCC tissues from normal tissues. These rules are entirely new, because complex diseases are often affected by various miRNAs rather than a single miRNA, and single-miRNA rules are insufficient for accurate diagnosis.

The advantage of the method presented here can be extended to the study of biomarkers identification in lung cancer prognosis. Also, we can validate the prognostic utility of these identified diagnostic biomarkers in early lung cancer. In addition, the discovered rules and distance separation technique can potentially be applied to further investigate biomarkers in other cancer diagnosis and prognosis, including breast cancer, pancreatic cancer, etc.

4.5 Conclusion

Rule discovery followed by distance separation is a powerful computational method for reliable identification of miRNA biomarkers. The visualization of the rules and the clear separation between the normal and cancer samples by our rules will help biology experts for their analysis and biological interpretation.

This chapter addresses **Contribution 1** of this thesis as listed in Section 1.3 by proposing a rule mining method to detect 2- and 3-miRNA biomarkers for the diagnosis of SCC. In the proposed method, this work has illustrated the computational difficulties of multi-miRNA analysis of expression data, and presented our effective approach to 2D or 3D biomarker discovery for lung SCC diagnosis. We have proposed a novel method to construct a committee of decision trees which may subsequently be used to derive 100%-

frequency rules containing 2 or 3 miRNAs. To detect more reliable rules, we have applied a Max-Min distance separation technique to look for the clear boundaries between the normal and cancer sample groups. The chromosomal loci of the miRNAs in these rules are identified, and the target genes of these biomarker miRNAs are also obtained from databases to determine the common mRNAs. These common target genes are then linked to diseases.

Chapter 5

Connecting rules from paired miRNA and mRNA expression data sets of HCV patients to detect both inverse and positive regulatory relationships

5.1 Introduction

As is explained in the related work (Section 2.2), miRNAs affect the stability and translational efficiency of target mRNAs by binding to their 3' UTRs to inhibit expression (Lewis et al. 2005*a*). A miRNA can have many target mRNAs and a mRNA can be regulated by multiple miRNAs, forming complicated many-to-many regulatory modules between miRNAs and mRNAs.

The identification of miRNA-mRNA regulatory modules has proven to be important for understanding complex cellular systems (Lewis et al. 2005*a*). It is also useful for understanding the infection process of various human diseases (Yoon & De Micheli 2005*b*). A recent computational method

based on probabilistic learning has been specially designed which uses the paired expression profiles and binding information of miRNAs and mRNAs on human cancer samples to discover miRNA-mRNA modules (Joung et al. 2007). Bayesian networks have also been adopted by many research groups (Friedman, Linial, Nachman & Pe'er 2000, Liu, Li & Tsykin 2009*a*) to detect novel miRNA-mRNA modules.

The key idea in all of these studies is the inverse expression relationship between miRNAs and their target mRNAs. An inverse expression relationship means that when the expression level of the miRNA is high (up-regulated), the target mRNA should be down-regulated based on the principle that miRNAs deregulate the expression of targeted mRNAs (Lim et al. 2005). However, up-to-date evidence shows that the inverse relationship does not always hold. First, a miRNA can induce gene expression by binding to the gene's promoter or enhancer sequence. For example, miR-373 can induce the expression of E-Cadherin or *CSDC2* when binding to these genes' promoters (Place et al. 2008). Second, recent investigation also shows that the interaction of miR-10a with RP mRNAs (those mRNAs encoding ribosomal proteins) binding at their 5' UTRs can promote the translational enhancement of these mRNAs instead of repression (ørom et al. 2008). Third, some positively regulated modules of miRNAs and mRNAs have been studied by wet-labs. For example, Van *et al.* have used quantitative PCR technology to evaluate the expression of miRNAs in the inflammatory breast cancer (IBC) and 50 non-IBC samples. Their results showed 7012 negative correlated miRNA-mRNA pairs and 10283 positive correlated miRNA-mRNA pairs (Van der Auwera, Limame, Van Dam, Vermeulen, Dirix & Van Laere 2010). Enerly *et al.* have reported strong positive correlations between miRNA clusters and their target genes of distinct biological processes in primary human breast tumours (Enerly, Steinfeld, Kleivi, Leivonen, Aure, Russnes, Rønneberg, Johnsen, Navon, Rødland et al. 2011). Nazarov *et al.* have identified several interactions in the form of negative or positive correlations between miRNAs and mRNAs, and subsequently identified

positively correlated miRNA-mRNA interaction networks in the frontal cortex of mice by differential expression analysis and weighted gene co-expression network analysis (Nazarov, Reinsbach, Muller, Nicot, Philippidou, Vallar & Kreis 2013). Therefore, the identified miRNA-mRNA interactions based purely on the inverse regulatory relationship are only an incomplete part of the modules in a certain biological context. Nunez *et al.* have firstly reported the positively correlated miRNA-mRNA networks in an animal model, which they proposed as an adaptive mechanism to reinstate cellular homeostasis (Nunez, Truitt, Gorini, Ponomareva, Blednov, Harris & Mayfield 2013).

This work focuses on the detection of both inverse and positive regulatory relationships in the paired miRNA and mRNA expression data of HCV-affected tissue samples. Paired miRNA and mRNA expression profiling provides an excellent platform for capturing those miRNA expression changes between two classes of samples that lead, positively or negatively, to the changes in mRNA expressions between the two classes of samples. We present a novel two-step sequential method to capture such "changes-to-changes". Our method derives discriminatory rules from miRNA expression data as the first step, and derives discriminatory rules from mRNA expression data as the second step. These rules are then combined to discover miRNA-mRNA regulatory modules.

The first step works on the miRNA data of the HCV negative and positive tissue samples to derive differentially expressed miRNAs and discriminatory rules (i.e., the miRNA expression changes between the two classes of samples). For each of these rules, we search for the predicted mRNA targets of every miRNA from the public miRNA target database TargetScan (Friedman, Farh, Burge & Bartel 2009*b*). We then narrow the search findings to a selected mRNA data set by removing the expression data of those mRNAs which do not belong to the predicted target mRNAs from the original mRNA data set. Discriminatory rules are derived from this selected and relevant data set of mRNA expression to concentrate on

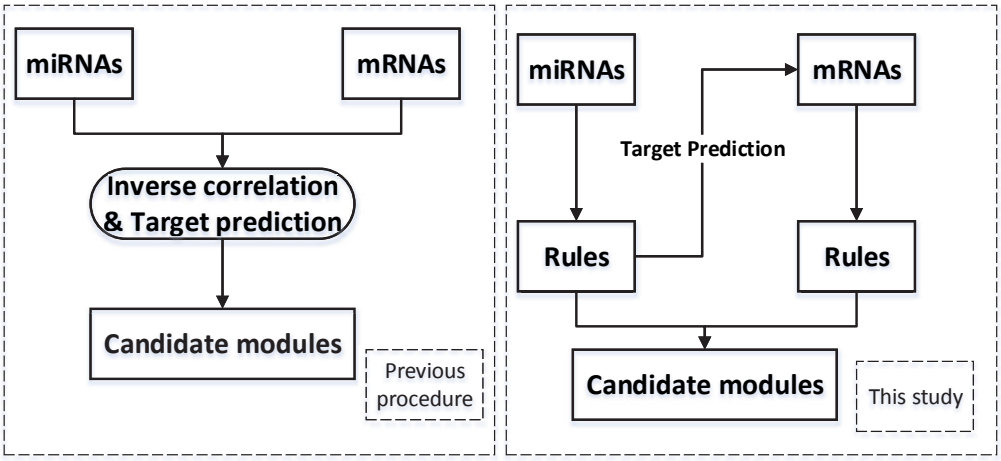


Figure 5.1: **Our approach in comparison to previous approaches.** We construct miRNA-mRNA regulatory modules using rule-based methods as shown in the right panel where the mRNA data set is narrowed down by the identified miRNA rules which are derived at the first step.

gene expression patterns that show significant differences between HCV positive and negative tissue samples (i.e., the mRNA expression changes led by those miRNA expression changes detected in the first step). Then, all the miRNAs in a rule and the mRNAs involved in the mRNA rules are combined to form a potential miRNA-mRNA regulatory module which is subsequently analysed using Pearson's correlation coefficients and biological literature results. Our approach does not use expression similarity networks or gene clusters to connect the two expression data fundamentally from the traditional approaches (Yoon & De Micheli 2005*b*, Joung et al. 2007, Peng et al. 2009, Jayaswal, Lutherborrow, Ma & Yang 2011) (see Figure 5.1 for detailed description).

This chapter, describing **Contribution 2**, is an extended description of my publication (Song, Liu, Liu & Li 2015).

5.2 Methods

5.2.1 miRNA and mRNA Expression Data Sets

The HCV data set from Peng *et al. (2009)* is used in this study (downloaded from the NCBI Gene Expression Omnibus database under the SuperSeries accession number GSE15387). This data set contains 36 tissue samples (24 HCV positive/+ and 12 HCV negative/-) described by the expression levels of 470 human miRNAs and 22575 mRNAs. The miRNA and mRNA data sets were both preprocessed using the Agilent Feature Extraction v9.5.3 under the default miRNA or mRNA parameters. Each miRNA value is the total gene signal, while each mRNA value is the log (REDsignal/GREENsignal) per feature (processed signals used, base 10). Of the 36 samples, 30 (24 HCV+ and 6 HCV-) samples have paired miRNA and mRNA expression profiles. Experiments were conducted on all samples using four technical replicates with the exception of sample28, sample33 and sample35, for which only three replicates were used. It is very costly for wet-lab experiments to obtain such a paired miRNA and mRNA expression data set. To our best of our knowledge, the paired data set used in this work is the largest microarray paired data set in the existing literature.

5.2.2 Rule-based Identification of miRNA-mRNA Regulatory Modules

We take the following steps to detect miRNA-mRNA regulatory modules. The first step is to use rule discovery to identify differentially expressed miRNA rules of 100%-frequency. Then for every miRNA in each rule, we obtain its predicted mRNA targets by searching for a public database (Friedman et al. 2009*b*). We then construct a selected mRNA data set for each rule consisting of all the samples but only those predicted target mRNAs presented in the original mRNA data set. We subsequently detect mRNA rules of 100% frequency from this selected mRNA data set. The mRNA

rules and their miRNA rule are then combined to form a miRNA-mRNA regulatory module. Our method is summarised in Figure 5.2 and detailed in the following subsections.

Rule Discovery from miRNA Expression Data

We rank all of the 470 miRNAs using the gain ratio criteria (Quinlan 1993*b*, Han & Kamber 2006*b*) through the Weka 3.6 software package (website: http://www.cs.waikato.ac.nz/ml/weka/). The top-ranked miRNAs (the most significant miRNAs) are then extracted to construct a new data set. We take a committee tree approach to detect 100%-frequency rules from this new data set, and to generate a committee of decision trees. We use the implementation of the C4.5 algorithm (Quinlan 1993*b*) in the R software package(RWEKA) to construct the tree committee. The first tree is derived based on the above miRNA data set. To derive the second tree, we change the data set by removing the root node of the first decision tree. This process is repeated until the data set has only two miRNAs left. If all of the training samples can be correctly classified by a rule in one of these trees, then this rule is a 100%-frequency rule. As mentioned, this work focuses on only 2-miRNA 100%-frequency rules as differentially expressed miRNAs for the simple diagnosis of HCV infection.

All the 100%-frequency rules are evaluated by Euclidean distance (Breu et al. 1995) and the average area under receiver operating characteristic (ROC) curves (AUCs) in the 10-fold decision tree cross-validation to determine their significance. Euclidean distance of a 100%-frequency rule indicates the separation extent between the HCV+ and HCV- samples. The separation extent is measured by the shortest pair-wise Euclidean distance of the HCV+ and HCV- samples (i.e., the Max-Min distance). The wider the separation is, the more reliable is the rule. 100%-frequency rules with a wide separation distance are of our interest for further investigation.

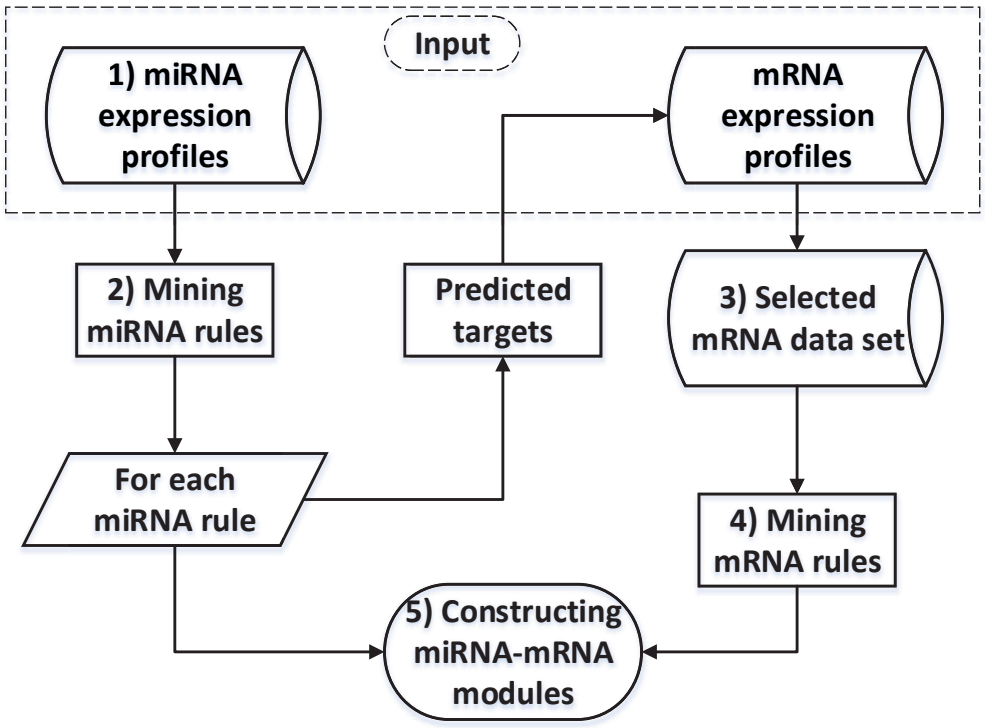


Figure 5.2: **Computational steps for the identification of miRNA-mRNA regulatory modules.** 1) Collection of miRNA expression profile data set. 2) Discovery of discriminatory rules from the miRNA expression data set using our rule discovery algorithm. 3) Construction of a selected and relevant mRNA expression data set. 4) Discovery of discriminatory rules from the relevant mRNA data set. 5) Identification of candidate miRNA-mRNA regulatory modules by combining the miRNAs and mRNAs in the discovered rules.

Rule Discovery from the mRNA Data Set

The systematic function of a miRNA is ultimately defined by its interaction with its target mRNAs or genes. We thus investigate the co-expressed miRNAs in each rule and their corresponding target mRNAs that are corporately involved in HCV infections. For each rule, we obtain computationally predicted target mRNAs through the Targetscan database (Friedman et al. 2009*b*). Using these predicted target mRNAs and their corresponding expression profiles from the original mRNA data, we apply the data mining techniques below to discover the rules of mRNA targets.

Given a dataset D with the class label set C (e.g., positive and negative), we detect rules for each $c \in C$. There may be more than one rule for each $c \in C$, we use several rounds of rule analysis to detect the rules. In each round, we detect a rule for each $c \in C$. We enumerate every attribute x_i to get its expression range a_i and b_i, and calculate the compactness $p = N_c/N$ where N or N_c is the number of all samples or c's samples in the expression range. Then, $a_i \leq x_i \leq b_i$ with the highest compactness is added to the rule. This process is repeated until (i) $p = 100\%$ (a rule for c is detected), or (ii) p cannot be improved but is still below 100% (there is no rule for c). In the selected mRNA expression dataset used in this work, in each round of rule analysis, we detect two rules: one for HCV+ and the other for HCV-. For the next round, all mRNAs in the discovered rules beforehand are not considered, and the rule analysis is performed again to detect more rules. The whole process is continued until none of the classes has a rule. This computationally heavy method is used at this step, because it is hard to identify the significant mRNAs with high gain ratio among tens of thousands of mRNAs by the tree-based analysis, but this computational heavy method can work out many 100%-frequency rules from the mRNA data sets.

Rule-Based miRNA-mRNA Regulatory Modules

We group all of its mRNA rules of 100%-frequency for each miRNA rule. A miRNA-mRNA regulatory module is formed by using a bipartite graph

representation (West et al. 2001), in which all the mRNAs in these rules comprise the mRNA partite, while the miRNAs are placed at the miRNA partite. To show the significant part of the modules and to assess the modules in the validation, we focus on the top four mRNA rules: two rules for classifying HCV+ samples and two rules for HCV-. We refer to these miRNAs and those mRNAs in the top four rules as significant components of the miRNA-mRNA regulatory module.

We also review the existing empirical literature to assess the biological importance of the regulatory modules. Furthermore, Pearson's correlation coefficient is calculated to detect the relationships (positive or negative correlation) between the miRNAs and mRNAs.

5.3 Results

5.3.1 2-miRNA Discriminatory Rules from the miRNA Expression Data

On the original miRNA data set of the 36 samples and 470 miRNAs, the gain ratio method selects 21 top-ranked miRNAs as the most significant miRNAs for the distinction between the HCV+ and HCV- samples. Each of the 21 miRNAs has a gain ratio > 0.5. The other miRNAs have a gain ratio ≤ 0.5 and thus are not considered here. Statistical analysis is also carried out using the two-sided student's t-test and the statistical significance is set as $P<0.05$ (Table 5.1).

On the data set of the above 21 miRNAs and all of the 36 samples, a total of nine 100%-frequency rules covering 10 miRNAs are derived through our committee tree approach. Each of these rules can classify the 36 samples into HCV+ or HCV- without any misclassification. An example of these miRNA rules is related to miR-557 and miR-214. The rule is: every HCV+ sample's expression profile satisfies the two miRNAs' specific expression ranges ($8.94 \leq miR-557 \leq 43.53 \cap 95.54 \leq miR-214 \leq 1057.51$), but none of the

Table 5.1: The top-ranked miRNAs with a gain ratio larger than 0.5

miRNA	Rank	p-value	miRNA	Rank	p-value
miR-202	1	2.072e-08	miR-519e*	12	4.222e-04
miR-601	2	1.060e-05	miR-526b	13	8.246e-04
miR-498	3	6.196e-09	miR-345	14	9.802e-05
miR-557	4	1.148e-05	miR-17-3p	15	2.927e-05
miR-34a	5	1.767e-02	miR-520a	16	0.276
miR-493-3p	6	3.127e-06	miR-452	17	4.170e-05
miR-214	7	4.629e-03	miR-501	18	4.328e-07
miR-184	8	1.470e-06	miR-130a	19	7.261e-04
miR-129	9	3.752e-03	miR-34b	20	1.278e-02
miR-765	10	1.243e-08	miR-221	21	4.622e-02
miR-210	11	1.668e-08			

HCV- samples satisfies these two expression ranges. The minimum Euclidean distance separating the two classes of samples for each rule and the average AUCs in the 10-fold decision tree cross-validation are also calculated. As shown in Table 5.2, the rule consisting of miR-557 and miR-214 has the maximum distance and maximum AUC.

5.3.2 Rules from the mRNA Expression Data

For each miRNA rule, the predicted mRNA targets from TargetScan are used to narrow down the original mRNA data set to a relevant mRNA data set for mRNA rule discovery. As shown in Table 5.3, some of the predicted targets (mRNAs) of a miRNA are not in the list of the probes used in the original mRNA expression data set (Table 5.2). Therefore, the mRNA expression profiles of only those targets (mRNAs) of the miRNAs in the probe list are used for the rule discovery (fourth column of Table 5.2). We note that the miRNAs involved in each rule may have common targets. For example, miR-557 and miR-214 have two common targets. On the 9 new mRNA data sets each for one miRNA rule, many 100%-frequency rules were mined by our

Table 5.2: The target mRNAs and their rules for each miRNA rule.

Rule ID	Euclidean distance	Average AUC	#mRNA in dataset	Class[1]	#rules[2] in HCV+	#rules[3] in HCV-	#mRNAs in all rules	#mRNAs[4] in top rules
R1	4.8946	0.9323	300	HCV+	2	14	110	15
R2	3.2888	0.9323	517	HCV+	2	21	159	12
R3	2.5160	0.9323	329	HCV+	2	14	85	12
R4	2.3360	0.8889	247	HCV+	2	11	75	12
R5	0.2256	0.8681	184	HCV+	2	5	41	13
R6	1.6425	0.9115	650	HCV-	8	34	269	10
R7	1.2757	0.8750	398	HCV-	7	28	227	10
R8	2.6420	0.9028	186	HCV-	2	6	55	20
R9	0.9806	0.8958	289	HCV-	1	12	97	11

R1: miR-557 and miR-214; R2: miR-34a and miR-214; R3: miR-493-3p and miR-214; R4: miR-214 and miR184; R5: miR-184 and miR-210; R6: miR-129 and miR-765; R7: miR-765 and miR-210; R8: miR-210 and miR-452; R9: miR-452 and miR-17-3p. [1]: the miRNA rule defines a region covering all samples of a class.

[2]([3]): the number of mRNA rules, each of which defines a region covering all samples of HCV+(HCV-).

[4]: the number of mRNAs in the top four mRNA rules: the top two mRNA rules in HCV+ (one rule is used if it is the only mRNA rule in HCV+) and another top two rules in HCV-.

Table 5.3: The number of predicted mRNA targets in the TargetScan database and those targets common in our mRNA data set

miRNA	mRNA targets	
	Predicted by TargetScan	in our used data set
miR-557	97	78
miR-214	301	224
miR-34a	387	293
miR-493-3p	131	105
miR-184	28	23
miR-129	320	236
miR-765	1105	414
miR-452	32	25
miR-210	218	161
miR-17-3p	353	264

proposed rule mining method (Table 5.2). In detail, we identified 28 mRNA rules for HCV+ and 145 mRNA rules for HCV- covering a total of 1118 mRNAs for all of the 9 miRNA rules (Table 5.2). Lastly, the top 4 rules, 2 from HCV+ (one rule is used if it is only one mRNA rule in HCV+) and 2 from HCV- are chosen as differentially expressed mRNAs for the subsequent miRNA-mRNA regulatory module study.

5.3.3 A miRNA-mRNA Regulatory Interaction Network

The above detected miRNA rules and significant mRNA rules are merged to form 9 miRNA-mRNA regulatory modules. These 9 miRNA-mRNA regulatory modules are then integrated to form a bigger miRNA-mRNA regulatory network (Figure 5.3). Many miRNAs and mRNAs in bold in these modules are related to diseases especially Hepatocellular carcinoma as supported by literature work (Table 5.4).

The numbers of mRNAs in these significant modules are shown in the last column of Table 5.2. Figure 5.4 and Figure 5.5 show two examples of

Table 5.4: All target mRNAs of miRNAs in HCV+ and HCV- modules.

miRNAs	Targetted mRNAs	modules
miR-557	ADRA1D, **ACVR1C**,DNAJA3,FAM120A	HCV+
miR-214	ASB16, GALNTL4,CBX5,BNC2,PDLIM2,**RAB43**, SPCS2,NKTR,ASXL1, ACLY,C6orf192, ING4,GLG1,SHOC2	HCV+
miR-34a	CPLX2, **FNDC5**	HCV+
miR-493-3p	**WDR33**	HCV+
miR-184	EPB41L5,**ALDH4A1**	HCV+
miR-129	CBLB, OCRL, COMT,DENND2C	HCV-
miR-765	ABCC5, BRD3, **ANKRD12**, AUTS2, PCID2, NMD3, NUP43	HCV-/+
miR-210	FGD4, HDAC4, CACNA2D2, OAZ2, ADAMTS5, AK3, CDKN1B, EPM2AIP1, PPP1R12B, PRPF4B, STAM2, EZ6L, SAMD4A, PISD, **KCTD9**, FAM118A, CHD2, KIT, TCF4	HCV-
miR-452	**ARMC1**, ZNF462, EFNA3, SMG5, FAM73B	HCV-
miR-17-3p	BNC2, **DICER1**, GFRA1, KIAA1804, ENPP1, ZNF558, ERO1L, SNX27, ZNF718	HCV-

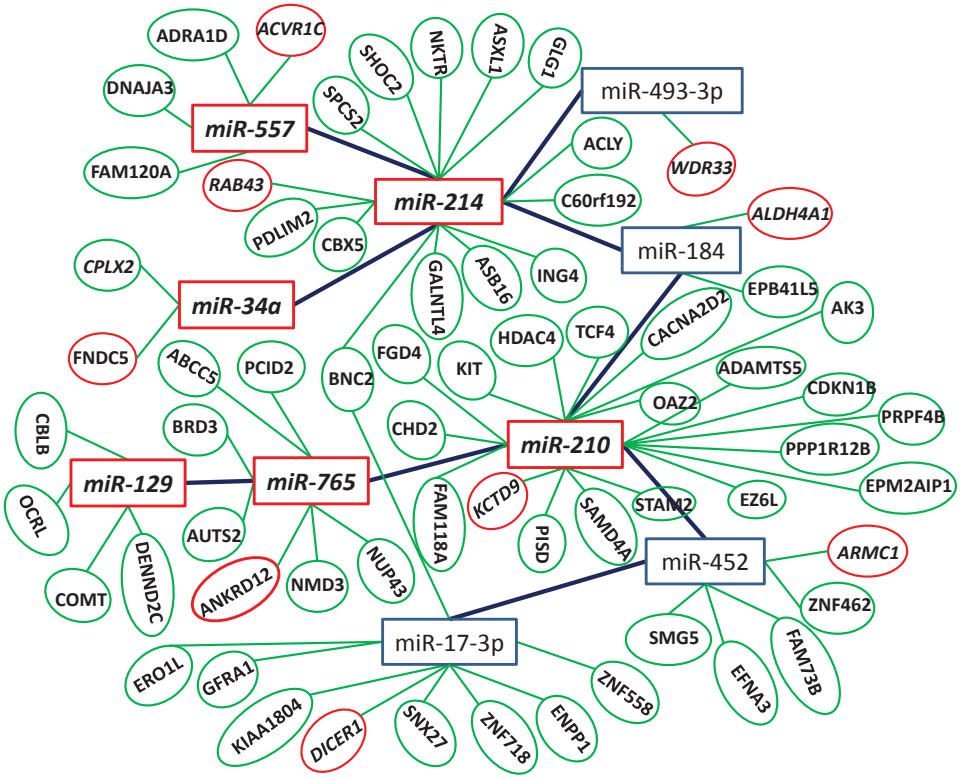


Figure 5.3: **A miRNA-mRNA regulatory interaction network.** There is an edge between two miRNAs if they are components of a miRNA rule. The edge between a miRNA and a mRNA represents a regulation of the miRNA for its target. Six miRNAs (miR-214, miR-34a, miR-129, miR-765 and miR-210) and 9 mRNAs (ACVR1C, RAB43, FNDC5, WDR33, ALDH4A1, ANKRD12, KCTD9, ARMC1 and DICER1) all in red are confirmed by literature work.

these significant modules, and all the miRNAs in bold and the mRNAs with underlining and italics can be confirmed by the literature. The validation results are presented below:

- In the module for miR-557 and miR-214, miR-557 has been reported as a novel candidate biomarker for hepatocellular carcinoma (Katayama, Maeda, Miyaguchi, Nemoto, Yasen, Tanaka, Mizushima, Fukuoka, Arii & Tanaka 2012), and miR-214-5p has been shown to up-regulate in human and mouse livers in a fibrosis progression-dependent manner. The expression of miR-214-5p increased during the culture-dependent activation of mouse primary stellate cells and was significantly higher in stellate cells than in hepatocytes (Iizuka, Ogawa, Enomoto, Motoyama, Yoshizato, Ikeda & Kawada 2012). As miR-214 and miR-557 expression patterns in hepatocellular carcinoma are tissue specific, they can both serve as novel biomarkers for chronic liver diseases. Meanwhile, a target mRNA *ACVR1C* of miR-557 is also associated with a reduction in HCV-infected cells (Zhang, Daucher, Armistead, Russell & Kottilil 2013), while the target mRNA *RAB43* of miR-214 is a key RAB to maintain a functional Golgi complex in human cells (Fukuda 2011), has been found to interact with HCV NS5A proteins (Sklan, Staschke, Oakes, Elazar, Winters, Aroeti, Danieli & Glenn 2007) and can also mediate the replication of HCV (Fukuda 2011).

- In the module of miR-34a and miR-214, besides the confirmed miR-214 and its mRNA *RAB43*, miR-34a has been reported to up-regulate in both liver fibrosis and hepatocellular carcinoma, and the serum levels of miR-34a are significantly higher in chronic hepatitis C infection patients than in controls (Cermelli, Ruggieri, Marrero, Ioannou & Beretta 2011). In addition, its target mRNA Fibronectin (*FNDC5*) was down-regulated and associated with hepatic fibrosis (Clark 2012).

- In the module of miR-493-3p and miR-214, the sole target mRNA *WDR33* of miR-493-3p has been found to result in increased viral

infection with two or more siRNAs (Brass, Huang, Benita, John, Krishnan, Feeley, Ryan, Weyer, van der Weyden & Fikrig 2009).

- In the module of miR-184 and miR-214, a mRNA *ALDH4A1* of miR-184 was believed to contribute to HBV- or HCV- induced liver (Xie, Cheng, Xing, Wang, Su, Wei, Zhou & Zheng 2011).

- In the module of miR-129 and miR-765, miR-129 has been strongly believed to be involved in the significant dysregulation in hepatocellular carcinogenesis (Katayama et al. 2012, Lu, Lin, Tien, Wu, Uen & Tseng 2013), and miR-765 is one of the promising candidate miRNA biomarkers to detect hepatocellular carcinoma among hepatitis C virus patients (Abdalla & Haj-Ahmad 2012). Meanwhile, mRNA *ANKRD12* of miR-765 is involved in one of the important roles of th host miRNAs in regulating the liver-specific HCV (Liu, Wang, Wakita & Yang 2010).

- In the module of miR-765 and miR-210, besides the validation of miR-765 and its target mRNAs above, miR-210 was up-regulated in HBV-producing HepG2.2.15 cells compared to parental HepG2 cells, and identified to suppress the hepatitis B virus (Zhang, Li, Zheng, Liu, Li & Tang 2010). In addition, a target mRNA (*KCTD9*) of miR-210 has been found to contribute to liver injury (Chen, Zhu, Zhou, Pi, Liu, Deng, Zhang, Wang, Wu & Han 2013).

Other mRNAs in these modules are also confirmed to be involved in hepatocellular diseases. For example, a mRNA of miR-452, *ARMC1* is up-regulated and frequently amplified in human hepatocellular carcinoma (Lee, Ho, Roy, Kosinski, Patil, Tward, Fridlyand & Chen 2008). Target mRNA *DICER1* of miR-17-3p, a component of the RNAi machinery, can markedly reduce HCV production and intracellular HCV RNA levels (Lupberger, Brino & Baumert 2008).

All these validation results suggest that the identified modules are closely related to hepatocellular carcinoma and are important for understanding the

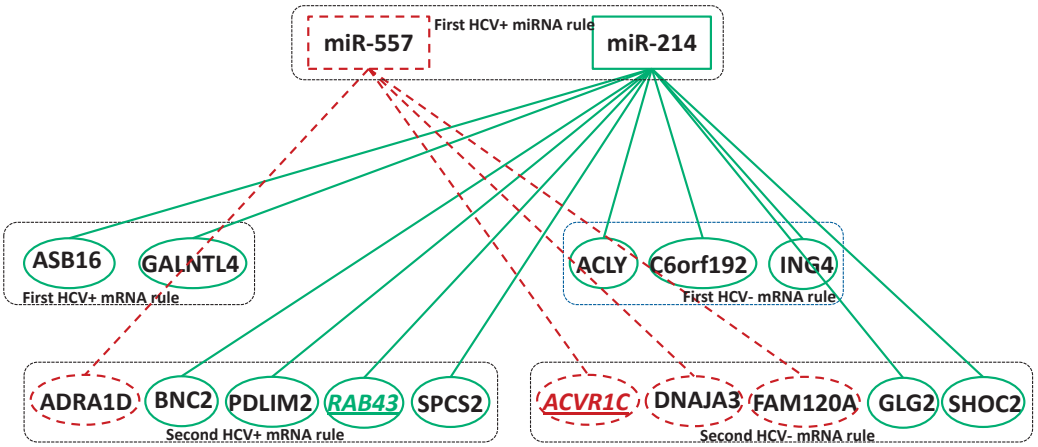


Figure 5.4: **The regulatory module inferred from the first miRNA rule and its corresponding mRNAs.** miR-557 and miR-214, the miRNAs of the first HCV+ miRNA rule are placed in the up panel. Four mRNA rules are identified and their mRNAs are placed in the middle and bottom panels. The edges linking miR-214 and its mRNA targets are in solid lines, while the edges linking miR-557 and its mRNA targets are in dashed lines. The confirmed target mRNAs are also highlighted with an underline.

miRNA-mRNA regulation in the host responses and pathogenesis of HCV infection.

5.3.4 Many-to-Many miRNA-mRNA Regulatory Modules

The big regulatory module (Figure 5.3) is a miRNA-mRNA interaction network integrated from the 9 simple regulatory modules corresponding to the 9 miRNA rules. A many-to-many miRNA-mRNA regulatory module usually consists of a cohort of miRNAs and a set of their target mRNAs, in which a target mRNA is regulated by multiple miRNAs, and a miRNA has multiple mRNAs as its target. We especially examined those many-to-many miRNA-mRNA regulatory modules in which mRNAs are targetted by at least 3 miRNAs. Figure 5.6 shows such an example.

This regulatory module contains 8 miRNAs and 6 target mRNAs. The

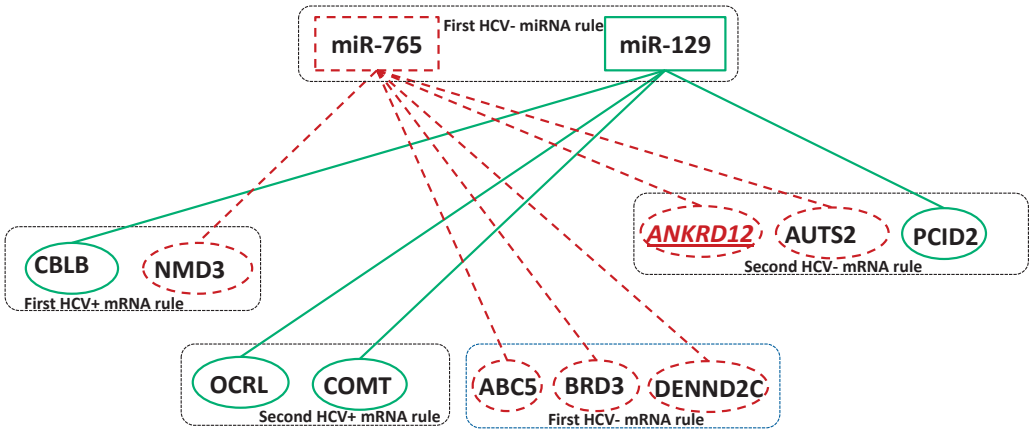


Figure 5.5: **The regulatory module inferred from the first HCV- rule consisting of miR-129 and miR-765.** In this module, miR-765 targets 6 mRNAs and miR-129 regulates 4 mRNAs. ANKRD12, a target of miR-765, is validated to be associated with chronic liver disease by existing works.

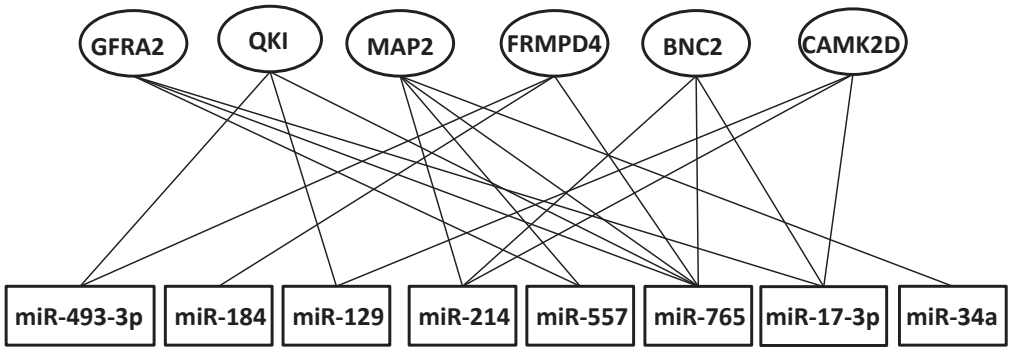


Figure 5.6: **The many-to-many relationship between some mRNAs and miRNAs identified in our modules** (i.e., one mRNA is targeted by many miRNAs and one miRNA can regulate many mRNAs).

literature shows that the miRNAs in this regulatory module are causally connected to human hepatocellular carcinoma or related diseases. For example, miR-129, miR-214 and miR-34a are found to associate with human hepatocellular carcinoma (Katayama et al. 2012, Lu et al. 2013, Xia, Ooi & Hui 2012, Cermelli et al. 2011). Mature miR-184 of over-expression can act as an oncogene in the antiapoptotic and proliferative processes of tongue Squamous Cell Carcinoma (Wong, Liu, Wong, Ng, Yuen & Wei 2008). miR-129 can regulate multiple tumour cell lines and primary tumours including medulloblastoma, undifferentiated gastric cancers, lung adenocarcinoma, endometrial cancer and colorectal carcinoma through down-regulating *CDK6* expression (Wu, Qian, Li, Kwok, Cheng, Liu, Perdomo, Kotton, Vaziri & Anderlind 2010). miR-34a can act as a tumour suppressor gene in a broad range of tumours including breast cancer, lung cancer, colon cancer, kidney cancer, bladder cancer and pancreatic carcinoma cell lines (Lodygin, Tarasov, Epanchintsev, Berking, Knyazeva, Korner, Knyazev, Diebold & Hermeking 2008). The mRNAs targeted by the miRNAs in this regulatory module are also engaged with cancer. Tumour suppressor *QKI* (the common target of miR-493-3p, miR-129 and miR-765) is expressed at significantly low levels in most of the gastric cancer tissues (Bian, Wang, Lu, Yang, Zhang, Fu, Lu, Wei, Sun & Zhao 2012). *MAP2* has been reported to be involved with malignant oral cancer tissues by playing important roles in neuronal and non-neuronal development (Liu, Chen, Tseng, Hung, Chiang, Chen, Shieh, Chen, Jou & Chen 2008).

5.3.5 Negatively and Positively Regulated mRNAs by Multiple miRNAs

The miRNA-mRNA expression relationships in the above many-to-many regulatory module were further assessed by analysing the Pearson's correlation coefficients of the 19 paired miRNA and mRNA expression levels of the 30 patients (i.e., the 19 edges in Figure 5.6). These coefficients are shown in Table 5.5. As expected, most of these relationships are negative. For

Table 5.5: Pearson's correlation coefficients between the miRNAs and mRNAs in the many-to-many regulatory module.'-' indicates the mRNA (in a column) is not the target of the miRNA (in a row).

	GFRA2	QKI	MAP2	FRMPD4	BNC2	CAMK2D
miR-493-3p	-	-0.68	-	0.12	-	-
miR-184	-	-	-	-0.05	-	-
miR-129	-	-0.71	-	-	-	0.01
miR-214	-	-	-0.15	-	-0.01	0.03
miR-557	0.18	-	-0.01	-	-	-
miR-765	0.21	-0.44	-0.02	-0.04	-0.10	-
miR-17-3p	0.26	-	-	-	-0.13	-0.13
miR-34a	-	-	-0.05	-	-	-

example, *QKI*, *MAP2* and *BNC2* have an inverse expression relationship with all of their regulator miRNAs. *FRMPD4* is also negatively correlated with their regulators except miR-493-3p. *CAMK2D* has a random correlation with miR-129 and miR-214, but it is negatively correlated with miR-17-3p.

One of our novel findings is a positive regulatory relationship between a mRNA and multiple miRNAs. As can be seen from Table 5.5, *GFRA2* has a clear positive relationship with the expression of all of miR-557, miR-765 and miR-17-3p with Pearson's correlation coefficients 0.18, 0.21, and 0.26 respectively. Figure 5.7 details the positively regulated expression levels of *GFRA2* in the 30 patients in comparison with the expression levels of the three miRNAs. If $|PCC| < 0.1$, we define the relationship should be a random or uncertain relationship. As indicated by the gain ratios shown in Table 5.1, the expression levels of each of these three miRNAs are able to separate these HCV+ and HCV- samples (also seen from the horizontal lines in Figure 5.7). It is the expression change of these three miRNAs that leads to a positive expression change of *GFRA2* between the two classes of patients.

The sequence matching between these miRNAs and *GFRA2* was also

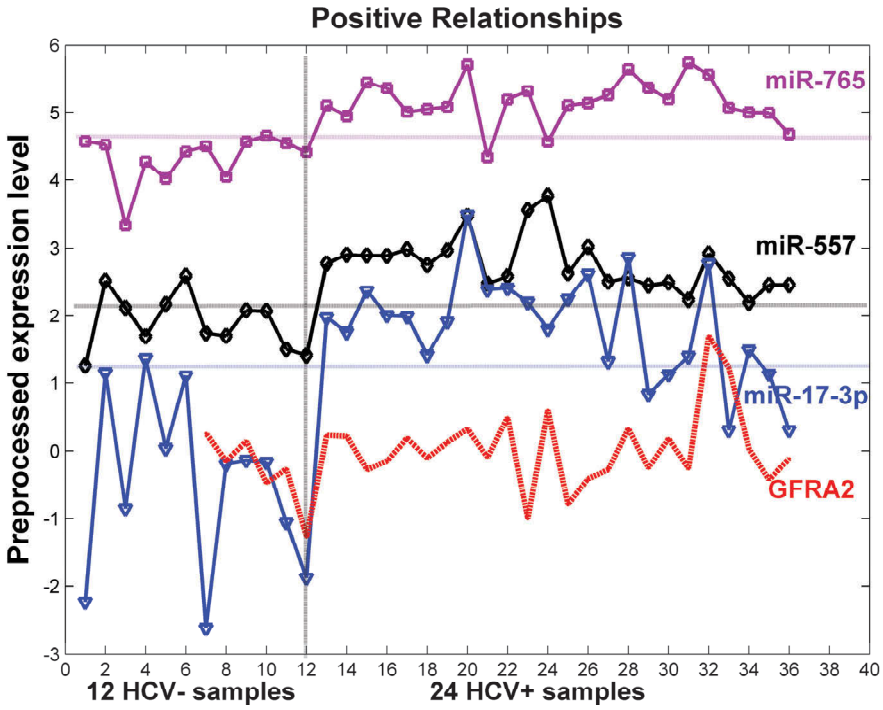


Figure 5.7: **The positive expression relationship between *GFRA2* mRNA and miR-557, miR-765, and miR-17-3p.** The expression levels of the three miRNAs are preprocessed in the log scale, and the expression levels of *GFRA2* are expanded by 10 times. The three miRNAs all have a high gain ratio, separating the HCV+ and HCV- samples very well.

studied. ørom *et al.* reported that the binding of miR-10a at the 5' UTRs of ribosomal protein (RP) mRNAs can promote their translational enhancement instead of repression (ørom et al. 2008). We attempted to verify whether the 5' UTR of *GFRA2* has a full or partial complementary sequence pairing with the seed region of miR-557, miR-765 or miR-17-3p. The fact is that the seed region of these three miRNAs is complementary to the 5' UTR of *GFRA2* with just one mismatched pair. In detail, the seed region of miR-557 matches the positions from 132 to 138 of *GFRA2* 5' UTRs, the seed region of miR-17-3p matches from 225 to 231, and the seed region of miR-765 matches from 495 to 502 (Figure 5.8). Therefore, it is likely that these three miRNAs bind at the 5' UTR end of *GFRA2* mRNA to enhance its translation for a positive regulation.

The statistical significance of this sequence complementarity in the defined manner (which also includes a mismatch) was also analysed using a Markov Model (MM) (Chung 1967). Based on the first-order Markov model (Marín & Vaníček 2011), the complementary significance was assessed by computing a probability (P) for each miRNA-5' UTR pair. It is an approximate probability that a complementary to the miRNA seed is found in the corresponding 5' UTR. The lower the P is, the higher the chances that the 5' UTR is a functional target. The length of the 5' UTR of *GRFA2* mRNA is 675, being composed of 151 purine bases adenine (A), 182 guanine (G), 161 the pyrimidine bases uracil (U), and 181 cytosine (C). The number of nucleotide in the miRNA seed region is 7. The transition matrix is shown as in Table 5.6. The complementary probability of the sequence matching between the seed region of the three miRNAs (miR-557, miR-765 and miR-17-3p) and 5' UTRs of *GFRA2* are 1.337e-05, 1.488e-04, and 1.133e-04 respectively which all imply a strong indication of a functional target.

To the best of our knowledge, the expression relationship between *GFRA2* and any of the three miRNAs has not been studied before in spite of intensive research into this field. *GFRA2* is a member of the *GDNF* receptor family encoding *GDNF* family receptor alpha-2 protein. *GFRA2*

Table 5.6: Transition probability of two adjacent bases in the 5' UTRs of *GFRA2*.

	A	G	U	C	Sum
A	33 (0.219)	50 (0.331)	32 (0.212)	36 (0.238)	151 (1.000)
G	60 (0.330)	60 (0.330)	20 (0.110)	42 (0.230)	182 (1.000)
U	22 (0.137)	33 (0.205)	62 (0.385)	44 (0.273)	161 (1.000)
C	36 (0.199)	38 (0.210)	47 (0.260)	60 (0.331)	181 (1.000)
Sum	151 (0.885)	181 (1.076)	161 (0.967)	182 (1.072)	675 (4.000)

Figure 5.8: **The partial complementary sequence pairing between the 5' UTRs of *GFRA2* and the seed sites of miR-557, miR-765 and miR-17-3p.** The mismatched base pairs are shown in smaller font.

is also a glycosylphosphatidylinositol (GPI)-linked cell surface receptor for both the Glial cell line-derived neurotrophic factor (GDNF) and neurturin (NTN) (Airaksinen & Saarma 2002), and it can affect the activation of the RET tyrosine kinase receptor (Buj-Bello, Adu, Pinon, Horton, Thompson, Rosenthal, Chinchetru, Buchman & Davies 1997). *GFRA2* is a candidate gene for RET-associated diseases. The brain-derived neurotrophic factor in patients has been found to be related to chronic Hepatitis C (Fábregas, de Miranda, Barbosa, Moura, Carmo & Teixeira 2012). Independent of the research on *GFRA2*, miR-557 (Katayama et al. 2012), miR-765 (Abdalla & Haj-Ahmad 2012) and miR-17-3p (Shan, Fang, Shatseva, Rutnam, Yang, Du, Lu, Xuan, Deng & Yang 2013) have all been reported to be associated with hepatocellular carcinoma. This suggests that the binding and interaction of mRNA *GFRA2* with miR-557, miR-765, or miR-17-3p, or with their combinations is a new research area, worth of comprehensive investigation by wet-lab experiments.

We also closely examined a strong negative regulatory relationship, shown in Figure 5.6. This regulatory relationship is between *QKI* mRNA and multiple miRNAs being miR-493-3p, miR-129 and miR-765 (see Figure 5.9). The seed matching sequence of miR-129 is located within the 3' UTRs end of *QKI*. But, the 5' UTRs of *QKI* mRNA does not contain the miR-129 complementary seed site. It is believed that miR-129 binds at the 3' UTRs end of *QKI* mRNA to down-regulate its translation.

Pearson's correlation coefficients were similarly examined for the literature-confirmed miRNAs and their corresponding mRNAs in Figure 5.3. It was found that miR-17-3p and *DICER1* mRNA have a strong negative regulatory relationship (Pearson's correlation coefficient: -0.53). A protein possessing an RNA helicase motif can be encoded by *DICER1* gene. The encoded protein functions as a ribonuclease and is required to produce the active small RNA component that represses gene expression, which may affect the biogenesis of miRNA (Liu, An, Liu, Wen, Zhai, Liu, Pan, Jiang, Wen, Liu et al. 2013).

As found by this work, the strongest positively regulated relationship is

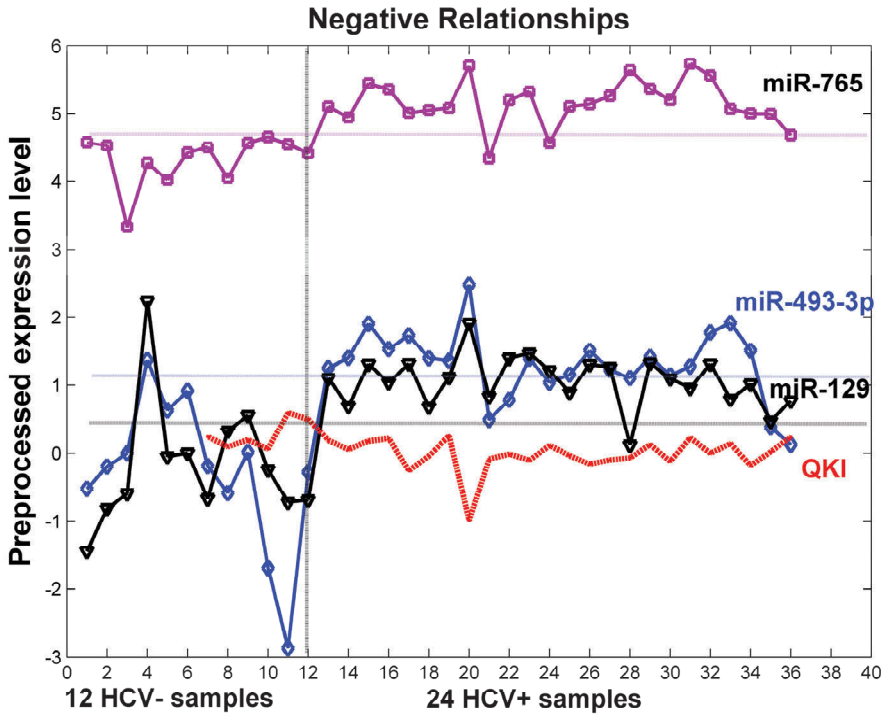


Figure 5.9: **The negative expression relationship between the *QKI* mRNA and miR-493-3p, miR-129, and miR-765.** The expression levels of the three miRNAs are preprocessed in the log scale. The three miRNAs all have a good gain ratio, separating the HCV+ and HCV- samples very well.

between miR-184 and *ALDH4A1* mRNA. Its Pearson's correlation coefficient is 0.43 (Figure 5.10). The *ALDH4A1* mRNA is up-regulated in late HCV cirrhosis (Mas, Maluf, Stravitz, Dumur, Clark, Rodgers, Ferreira-Gonzalez & Fisher 2004) and HBV pathogenesis (Xie et al. 2011). In Drosophila, a luciferase reporter assay has shown that miR-184 can target some of mRNAs in the protein coding region (Easow, Teleman & Cohen 2007). We found that the seed region of miR-184 is complementary to the coding region or to the 5' UTR of *ALDH4A1* with just one mismatched pair. The seed region of miR-184 matches the positions from 705 to 711 of *ALDH4A1*'s coding region or with the positions from 257 to 263 at *ALDH4A1* 5' UTR. Based on this evidence, the miRNA-184 target sites in 5' UTRs or the coding region may make a significant contribution to miR-184 mediated regulation. The functionality of miR-184 when binding at the 5' UTR or the coding region of *ALDH4A1* deserves thorough investigation to expand the current research on the 3' UTR.

In addition, we also checked our discovery results in the starBase database (http://starbase.sysu.edu.cn/) (Li et al. 2014), and an interaction is confirmed between hsa-miR-129-5p (previous ID: hsa-miR-129) and QKI, with the highest Pearson's Correlation Coefficient.

5.4 Conclusion

The literature review of miRNA-mRNA relationships detection shows that for most computational methods, the key idea in all of these studies is the inverse expression relationship between miRNAs and their target mRNAs. However, up-to-date evidence shows that the inverse relationship does not always hold.

This chapter addresses **Contribution 2** of this thesis as listed in Section 1.3 by connecting rule mining methods to detect both inverse and positive miRNA-mRNA relationships in HCV patients. In this work, we have proposed rule-based methods for the discovery of miRNA-mRNA regulatory

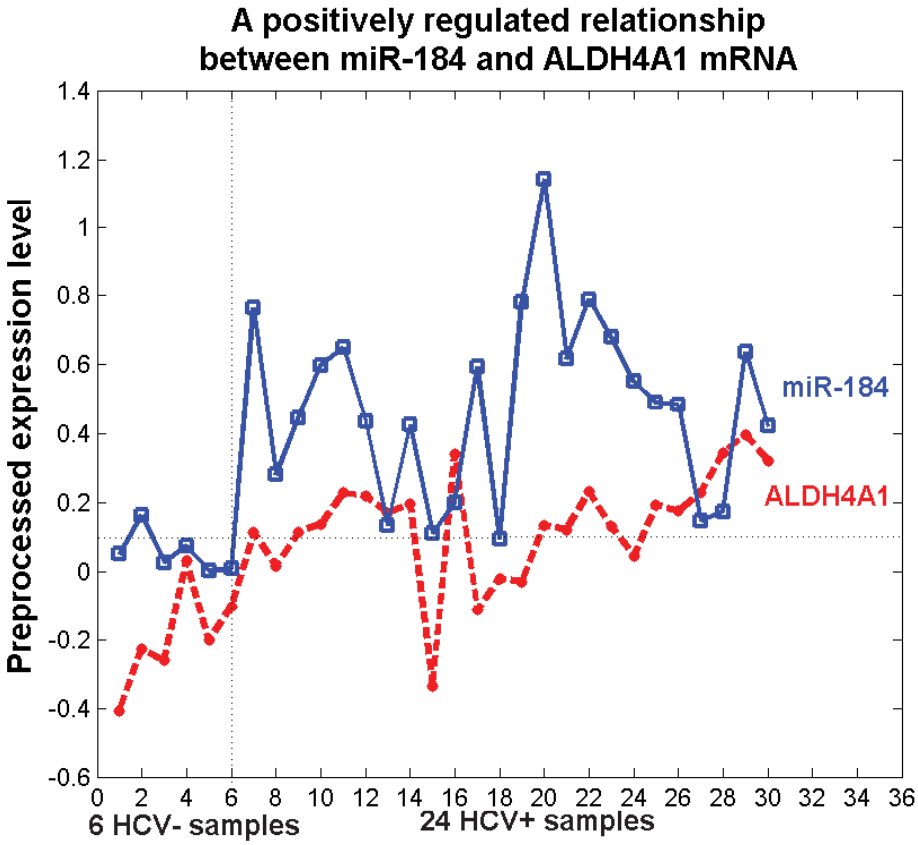


Figure 5.10: **The positive expression relationship between the *ALDH4A1* mRNA and miR-184.** The expression levels of miR-184 are preprocessed by dividing 10 and this has a good gain ratio, classifying the HCV+ and HCV- samples very well.

modules in HCV infection. We followed the biological principle that inverse expression relationships and positively regulated miRNA-mRNA pairs can both exist in many-to-many regulatory modules. We detected 100%-frequency rules from the most differentially expressed miRNAs and then mined 100%-frequency rules from the relevant target mRNAs expression data for each miRNA rule. We integrated the miRNA rules and their mRNA rules to construct miRNA-mRNA regulatory modules. Many detected miRNAs and mRNAs can be supported by recent work in the literature. We also detected novel positive and inverse regulatory relationships. For example, mRNA *GFRA2* is positively regulated by multiple miRNAs miR-557, miR-765 and miR-17-3p which all likely bind at the 5' UTR end of *GFRA2*. The detected miRNA-mRNA regulatory modules will provide new insights into the regulation of host responses and the pathogenesis of HCV infection. We conclude that our rule discovery method is useful for integrating binding information and the expression profile for identifying HCV miRNA-mRNA regulatory modules and can be applied to the study of the expression profiles of other complex human diseases.

Chapter 6

Identification of Lung Cancer miRNA-miRNA Co-regulation Networks Through a Progressive Data Refining Approach

6.1 Introduction

As is explained in the related work (Section 2.3), with these biological observations, research interests have been focusing on the regulation relationships between a miRNA and its target genes for many years (Hashimoto, Akiyama & Yuasa 2013, Suzuki, Mihira, Watabe, Sugimoto & Miyazono 2013, Yang, Sun, Hu, Zheng, Ji, Pecot, Zhao, Reynolds, Cheng, Rupaimoole et al. 2013). However, the co-regulation relationship among the miRNAs themselves, for example among a miRNA cluster, has not been intensively studied though it was first reported in 2005 and 2006 (He, Thomson, Hemann, Hernando-Monge, Mu, Goodson, Powers, Cordon-Cardo, Lowe, Hannon et al. 2005, Cui, Yu, Purisima & Wang 2006).

Co-regulation analysis of multiple miRNAs is useful for understanding complex post-transcriptional regulations (Baumjohann & Ansel 2013, Guo, Zhao, Yang, Zhang & Chen 2014). One of the earliest studies on miRNA pair co-regulation is by Enright et al. (Enright et al. 2004) for understanding the co-regulation between lin-4 and let-7 in Drosophila. With the huge amount of expression data publicly available, newer methods have been proposed to investigate the problems of co-regulating miRNAs (Migliore & Giordano 2009, Boross et al. 2009, An et al. 2010). For example, Boross (Boross et al. 2009) proposed to construct a miRNA co-regulation network by computing the correlations between the gene silencing scores of individual miRNAs. Since most of these studies take only expression data of miRNAs and messenger RNAs (mRNAs) without biological function analysis (Guo et al. 2010), some true targets of these miRNAs may be ignored and some false targets may be included. One possible reason for this can be explained by examples of those miRNAs being demonstrated to reduce protein levels without the concomitant change in mRNA levels (Lee, Samaco, Gatchel, Thaller, Orr & Zoghbi 2008), which may be regulated at tissue specific levels (Guo, Maki, Ding, Yang, Xiong et al. 2014).

This suggests that biological functional analysis is a necessary assessment for the detection of complete and reliable targets of miRNAs. Gene Ontology (GO) contains comprehensive information of biological processes and functions (Ashburner et al. 2000). In particular, the coherence score of the GO terms annotated to a gene group can be used to compute the p-value of a co-regulated gene group which actually is a statistical measurement to judge whether or not the co-regulated gene group are reliable targets of a miRNA. Yoon and De Micheli (Yoon & De Micheli 2005*a*) had considered GO information and proposed a biclique-based method to detect co-regulating groups of miRNAs and mRNAs. However, the heuristic nature of that method can lead to those miRNAs or genes missing each other even when they have a high probability of co-regulation. Recently, it has been found that interacting proteins are often regulated by similar miRNA

types (Yuan et al. 2009, Liang & Li 2007). This suggests that clustered miRNAs can jointly regulate those proteins which are close to each other within a protein interaction network (Hsu et al. 2008). In this work, we take a novel approach to the discovery of miRNA-miRNA co-regulation networks by sequentially and progressively integrating expression correlations, GO function knowledge, and protein interaction information.

The co-regulation networks of miRNAs in lung cancer have not been well investigated (Vincent 2013), despite that lung cancer is the leading cause of cancer-related deaths worldwide. Our integrative computational method is applied to identify a miRNA-miRNA co-regulation network common to three lung cancer miRNA expression data sets of different subtypes. Our method has three main steps. Firstly, all the common miRNAs to the three data sets are selected to get relevant miRNA expression data. At this step, Pearson's correlation coefficient (PCC) and Targetscan database are used to discover highly correlated miRNA pairs and the pairs' common targets for each of the three processed data sets. Secondly, some of these miRNA pairs are filtered by performing a GO functional enrichment and a protein interaction analysis on their common targets. If a subset of target genes has a significant functional enrichment in the GO analysis and has a close proximity in the protein interaction network, then this subset of target genes is defined as a functional module. Thirdly, the analysis focuses on candidate co-regulating miRNA pairs targetting the same functional modules. Those miRNAs which are detected at least twice from the three data sets are finally selected and then assembled to construct the common miRNA-miRNA co-regulation network of lung cancer.

Important databases such as the KEGG pathway database, miR2Disease and OMIM, and graph theories have been employed to interrogate the validity of the miRNA co-regulation network. We found that this network is a scale free network with a power law distribution, indicating it is far from a random network. This network also contains some intensively-studied co-regulating miRNAs (e.g., miR-221/222, miR-15b/16 and let-7a/b/c/d/f/g).

On the other hand, some co-regulating miRNA pairs are novel. For example, miR-18a/b from the same family is found to function together to co-regulate their common targets. Furthermore, we discovered that lung cancer related miRNAs have more synergism than lung cancer un-related miRNAs, suggesting that they have more influence at the post-transcriptional stage and on the fundamental cellular processes.

This chapter, describing **Contribution 3**, is an extended description of my publication (Song, Catchpoole, Kennedy & Li 2015).

6.2 Materials

Three data sets of different lung cancer subtypes are used in this work. Dataset1 contains small cell lung cancer, large cell neuroendocrine cancer, squamous cell carcinoma and adenocarcinoma samples; Dataset2 contains non-small cell lung cancer samples; and Dataset3 contains only squamous cell carcinoma samples. These data sets are available at the Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information (NCBI) database (under accession IDs GSE19945, GSE29250 and GSE15008).

Dataset1 (GSE19945) has a panel of 600 human miRNAs for 63 distinct tissues (55 lung cancer samples and 8 normal samples). These miRNA expression profiles were measured by the Agilent Human 0.6K miRNA Microarray G4471A platform. The raw data was processed with the GeneSpring GX10 software (Agilent) by the original author: Raw data of intensities < 1.0 were transformed to 1.0, and then log2 transformed. The signal intensities of each sample were then normalised to its 75 percentile intensity by the GeneSpring normalisation option. Those features having raw intensities < 5.0 in all samples were excluded. Our work uses the normalised data only.

Dataset2 (GSE29250) consists of 859 human miRNAs for 12 distinct samples (6 lung cancer samples and 6 normal samples) (Ma, Huang, Zhu, Zhou, Zhou, Zeng, Liu, Zhang & Yu 2011). These miRNA expression profiles

were extracted by the Illumina Human v2 MicroRNA expression beadchip platform. Our work uses the normalised data set from GEO, which was normalised via an Illumina Genomestudio software.

Dataset3 (GSE15008) contains 549 human miRNAs for 361 lung tissue samples (Tan, Qin, Zhang, Hang, Li, Zhang, Wan, Zhou, Shao, Sun et al. 2011). The expression profiles were measured by the National Engineering Research Center mammalian microRNA microarray platform. The miRNA expression data in these tissues were derived after the average values of the replicate spots of each miRNA were background subtracted and the faint spots were filtered out when the expression signal was lower than 800 in all samples. Our work uses all the 187 cancer tissue samples and the related 174 adjacent normal tissue samples.

The Targetscan database (Grimson, Farh, Johnston, Garrett-Engele, Lim & Bartel 2007) provides predicted targets of miRNAs. The GO files on Biological Process (BP) are from the GO consortium. Our protein interaction data sets are from Unified Human Interactome, a big database containing several large sets of protein-protein interactions. These data sets and their work flow are depicted as a diagram in Figure 6.1. The details of the computational steps are described in the subsequent sections.

6.3 Methods

6.3.1 Preprocessing of miRNA Expression Data

The three data sets are denoted by DS1, DS2, and DS3 in Figure 6.1. As our goal is to detect a miRNA-miRNA co-regulation network common to lung cancer, we concentrate on only those common miRNAs of these data sets. In fact, there are 401 common miRNAs, obtained by mapping the probe sets of these three data sets to the miRBase database (Griffiths-Jones, Saini, van Dongen & Enright 2008). Each of these common miRNAs has a unique ID number. We note that specially for DS3, the miRNA expression values of these probe sets are actually averaged because these probe replicates

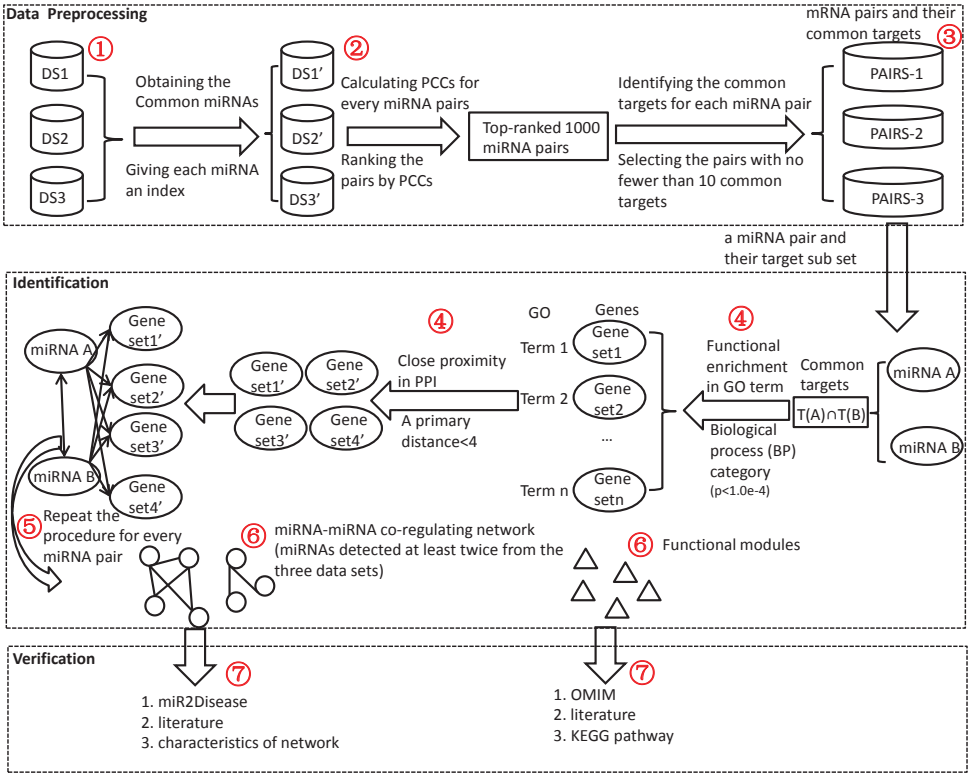


Figure 6.1: **The flowchart of constructing a miRNA-miRNA co-regulation network starting from three lung cancer data sets (①: DS1, DS2 and DS3). Preprocessing:** ②: using DS1′, DS2′ and DS3′ to represent the common miRNAs of DS1, DS2 and DS3; ③ : using data sets PAIRS-1, PAIRS-2 and PAIRS-3 for storing miRNA pairs and their common targets by selecting highly correlated miRNA pairs containing no fewer than 10 common targets. **Identification:** ④: Identifying a miRNA pair co-regulating the same function modules by performing GO function and protein interaction analyses; ⑤: repeating the procedure for every miRNA pair in PAIRS-1, PAIRS-2 or PAIRS-3; ⑥: identifying the functional modules and constructing a common miRNA-miRNA co-regulation network by assembling all the miRNA pairs with miRNAs which are detected at least twice from the three data sets. **Verification:** ⑦: using existing databases (KEGG pathway, miR2Disease and OMIM) and graph theoretical methods to validate this co-regulation network and functional modules.

correspond to the same miRNA. Such processed DS1, DS2, and DS3 are denoted by DS1′, DS2′, and DS3′ in Figure 6.1.

With the data from 401 miRNAs across our cohorts, we decided that the data volume was sufficiently reduced to warrant the analysis of all miRNA pairs and no further filtering based on differential expression, noise or other quality measures would be addressed. Hence, for each miRNA pair in DS1′, DS2′ and DS3′, we assess their expression relation via Pearson's absolute correlation coefficient (PCC). We also rank all of the miRNA pairs in each data set with regard to their PCCs. In this work, we focus on the 1000-top ranked miRNA pairs and their absolute PCC values.

When a pair of miRNAs have a high PCC, they should have a strong potential to synergistically co-regulate their targets. PCC is not a sufficient condition to establish a reliable causal relationship, since it may produce indirect interactions and the elimination of indirect interactions is very important (Barzel & Barabási 2013, Feizi, Marbach, Médard & Kellis 2013). Thus, further refinement is needed.

Given a pair of miRNAs having a high absolute PCC, they are more likely to be co-regulating if they are predicted to regulate a large number of common targets. So, their common predicted targets are checked at the Targetscan database (http://www.targetscan.org/). If one miRNA pair contains 10 or more common targets, then this miRNA pair and their common targets are stored at PAIRS-1, PAIRS-2, or PAIRS-3 (Figure 6.1) for our next analysis.

6.3.2 Network Construction for Co-regulating miRNAs

GO enrichment analysis For every miRNA pair from PAIRS-1, PAIRS-2, and PAIRS-3, a GO enrichment analysis (Biological Process subtype) is performed on their predicted targets to classify their functions. Only those GO terms which contain more than three genes with a significance level ($p < 1.0e-4$) are captured. Specifically, for a given miRNA pair (miRNA A and miRNA B), we use their intersecting target subsets which they co-regulate (i.e., subsets of $T(A) \cap T(B)$) to identify the biological processes

under the hypergeometric distribution. Here $T(X)$ stands for the set of predicted targets of miRNA X. The analysis is preformed by the R software (GOstats and GO.db).

PPI Network Construction The PPI network of a gene subset of $T(A) \cap T(B)$ is represented by a graph, in which the proteins are represented by nodes and the interactions among them are represented by undirected edges. Using this gene subset as seed proteins, the construction of its PPI network is through the tool named UniHI (`http://193.136.227.168/UniHI/pages/unihiSearch.jsf`), which provides both experimentally determined and predicted interactions. The number of edges inserted between two seed proteins determines the network distance of the seed proteins. As found by literature (Liang & Li 2007, Yuan et al. 2009), proteins interacting with cancer-related proteins are generally close to each other and interact more frequently compared to non-interacting proteins in the PPI networks. Therefore, we consider only those PPI networks with a primary distance no larger than 3. A primary distance between any two proteins in a PPI network is measured by the minimum number of edges required to connect them.

Combining Co-regulating miRNA Pairs from the Three Data Sets If the miRNA pair of A and B contains target subsets having significant GO enrichment and having at least one network of close distance, then A and B are defined to co-regulate the corresponding target genes. The procedure is repeated for every miRNA pair from PAIRS-1, PAIRS-2, and PAIRS-3, and store only those miRNA pairs which contain miRNAs presented at least twice from the three data sets. Then, these stored miRNA pairs are integrated to generate a miRNA-miRNA co-regulation network. A node stands for a miRNA, and two nodes are connected if the corresponding miRNA pair shows a co-regulation relationship. The corresponding regulated gene sets (the PPI networks) are defined as functional modules.

6.3.3 Validation of the Co-regulating miRNAs

KEGG pathway enrichment analysis The KEGG pathway database is a collection of manually drawn pathway maps representing knowledge of molecular interactions and reaction networks. To determine significant changes of the target genes in signaling pathways, KEGG pathway analysis is conducted through the R software (org.Hs.eg.db) to study pathway terms that contain more than two genes with $p < 1.0e-4$. We also look at KEGG pathways of microRNAs related to cancer, especially lung cancer.

Literature-Based Verification miR2Disease, a manually curated database (`http://www.mir2disease.org/`), provides a comprehensive resource of miRNA deregulation information for various human diseases. The miRNAs in our identified co-regulation network are matched with this database to see whether they have been found to be associated with lung cancer. Also, the Online Mendelian Inheritance in Man (OMIM: `http://www.omim.org/`) database is used to understand whether genes involved in the functional modules are related to lung cancer.

Topological Analysis on Hub Proteins Most proteins interact with only a few other proteins, while a small number of proteins may have many interaction partners in the PPI networks. Hubs are proteins with a large number of interactions in a protein-protein interaction network. They are the principal agents in the interaction network and affect its function and stability. Therefore, we calculate the hub degree of the interaction network formed by the gene groups and detect hub proteins to discover their relationships with lung cancer, then we see whether the genes have potential as targets for lung cancer treatment.

Characteristics of the miRNA-miRNA Co-regulation Network Co-regulation miRNA networks and random networks are compared to examine whether miRNA-miRNA co-regulation networks are scale-free or

random. A scale-free network is a network whose degree distribution follows a power law. That is, the fraction $P(k)$ of nodes in the network having k connections to other nodes goes for large values of k as $P(k) \sim k^{-\gamma}$, where γ is a parameter whose value is typically in the range $2 < \gamma < 3$.

Analysis on Transcription Factors Those transcription factors (TFs) are compared which are located within the promoter regions of the miRNAs of our miRNA-miRNA co-regulation network. TFs regulate the transcription of miRNAs in a pol II dependent manner similar to that of protein-coding genes; that is, by binding to the conventional transcription factor binding site sequences located in or near the promoter regions upstream of the miRNAs. The ChIPBase database (Yang, Li, Jiang, Zhou & Qu 2013) is used to construct TF-miRNA and TF-miRNA-mRNA regulatory networks. The cooperativity of miRNAs is evaluated by examining their shared transcription factors.

6.4 Results

Our results are presented in six parts. The first part reports newly discovered co-regulating miRNA pairs and their interacting network. The second part presents the topological characteristics of the co-regulation network. The third part describes how the lung cancer related miRNAs can regulate more functional modules and have more functional synergism than un-related miRNAs. The fourth part highlights the lung cancer related miRNAs and genes in the discovered co-regulation network and functional modules. The last two parts present KEGG pathway analysis results and TF-miRNA and TF-miRNA-mRNA regulatory networks.

6.4.1 Co-regulating miRNA Pairs and Their Big Network

Pearson's correlation coefficients were computed for all of the possible miRNA pairs for each of the three data sets (DS1′, DS2′, and DS3′, having

401 common miRNAs). Most of these pairs have a PCC>0. For DS1′, PCCs range from -0.5714 to 0.9997; of the top 1000 miRNA pairs, the absolute PCCs range from 0.7990 to 0.9997. For DS2′, PCCs range from -0.9309 to 0.9996; of the top 1000 miRNA pairs, the absolute PCCs range from 0.8254 to 0.9996. Similarly, PCCs from DS3′ range from -0.5714 to 0.9850; of the top 1000 miRNA pairs, the absolute PCCs range from 0.8483 to 0.9850. Most of these top-ranked miRNA pairs have a small number of common targets. But, there still exist 182, 237 and 132 miRNA pairs in DS1′, DS2′, and DS3′ respectively which have common targets of at least 10.

The significance level was set at the threshold p<1.0e-4 for GO functional analysis on the target subset of each of these miRNA pairs. There are 99, 47 and 28 miRNA pairs from DS1′, DS2′, and DS3′ respectively satisfying this biological process enrichment requirement.

The target subsets enriched by the GO term were further filtered by their PPI network distance properties. A total of 31 gene networks (functional modules) were constructed to satisfy the topological distance condition of PPI networks (the primary distance no larger than 3). Under this condition, there are only 36, 15 and 3 miRNA pairs for the three data sets. As there are some identical miRNA pairs, there are actually only 41 unique miRNA pairs containing 43 unique miRNAs. If every miRNA is required to be present at least twice at the three data sets, then only 30, 11 and 3 miRNA pairs are left for these three data sets. In particular there exist three overlapping miRNA pairs (let-7b and let7c, miR-18a and miR-18b, and miR-302c and miR-373). Table 6.1 summarises the change of these numbers of the miRNA pairs when our analysis and requirements were progressed and refined.

These co-regulating miRNA pairs satisfying both the functional enrichment and PPI network requirement are assembled to construct a miRNA-miRNA co-regulation network (see Figure 6.2). Every node in this network represents a miRNA; two nodes are connected if the corresponding miRNA pair has a co-regulation relationship.

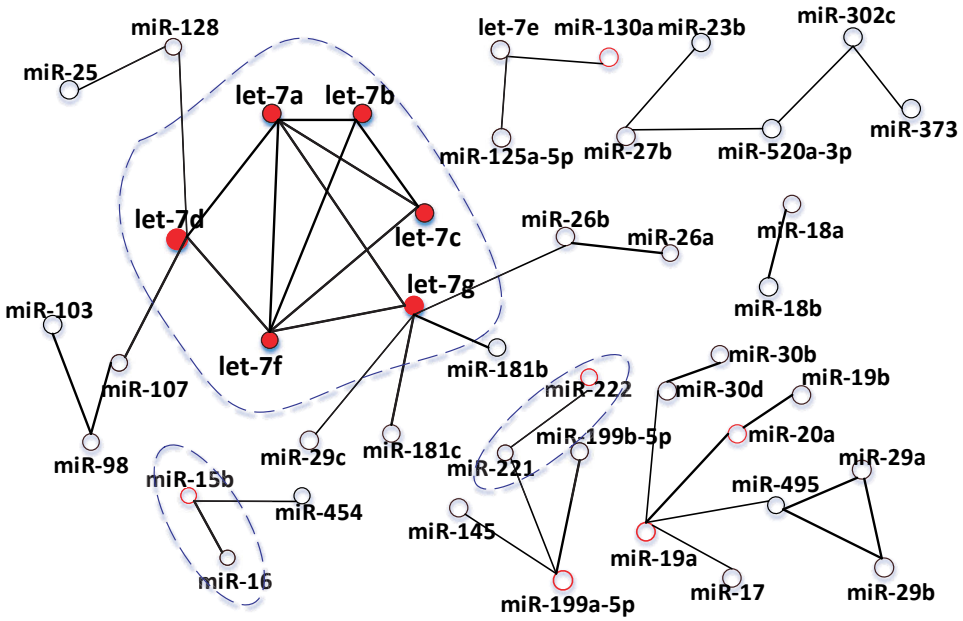


Figure 6.2: **A miRNA-miRNA co-regulation network. There are 41 connections and 43 nodes in this network.** A node stands for a miRNA, and an edge connecting two nodes represents a co-regulation. The miRNAs with red circles (points) are confirmed to be associated with lung cancer from the miR2Disease database. The verified co-regulating miRNA pairs are highlighted in the blue dashed circles and let-7a/b/c/d/f/g are in red.

Table 6.1: Progress of the miRNA pairs numbers when our analysis and requirements were getting refined.

Dataset	#Stage 1	#Stage 2	#Stage 3	#Stage 4	#Stage 5
Dataset1	1000	182	99	36	30
Dataset2	1000	273	47	15	11
Dataset3	1000	132	28	3	3

#Stage 1 indicates the number of miRNA pairs by selecting the 1000 miRNA pairs with high Pearson's correlation coefficients. #Stage 2 is the number of miRNA pairs after detecting their common targets no fewer than 10. #Stage 3 represents the number of miRNA pairs after performing the GO functional analysis for their targets from stage 2 (P<1.0e-4). #Stage 4 shows the number of miRNA pairs after constructing the PPI networks for the targets from stage 3 (the primary distance < 4). #Stage 5 stands for the number of miRNA pairs in which the miRNAs exist at least twice from the three data sets (three overlapping miRNA pairs in the three data sets).

6.4.2 Topological Characteristics of the Co-regulation Network and the Functional Modules

From Figure 6.2, we can see that some miRNAs can correlate with a relatively large number of miRNA partners, while the majority of miRNAs have just one or two co-regulating partners. The degree distribution of this network follows a power law (Figure 6.3, $R^2 = 0.9011$), indicating that this network is scale free instead of random. The five miRNAs having a degree of at least 4 are let-7a, d, f, g and miR-19a.

This work also found that miRNAs from the same family tend to have similar functions—24 of the 41 edges are directly or indirectly shared by the same family. For example, both miR-15b and miR-16 are located at 3q25.33 and play important roles for apoptosis by targetting BCL2 in human diseases (e.g., chronic lymphocytic leukemia (Cimmino, Calin, Fabbri, Iorio, Ferracin, Shimizu, Wojcik, Aqeilan, Zupo, Dono et al. 2005) and gastric cancer (Xia, Zhang, Du, Pan, Zhao, Sun, Hong, Liu & Fan 2008)). The connected miR-221/222 are involved in the same functional modules, and they had been both found to act as oncogenes or tumour suppressors in tumour development (Garofalo, Quintavalle, Romano, Croce & Condorelli

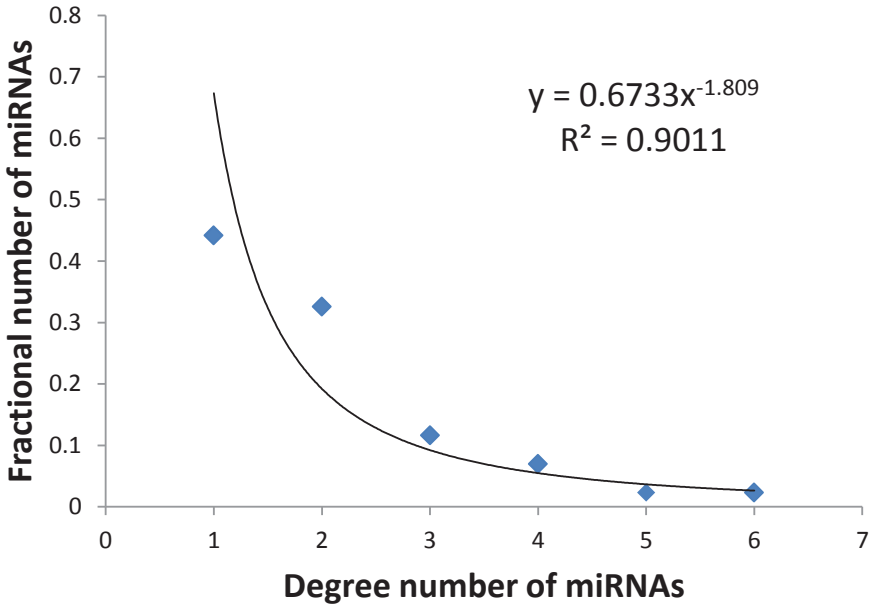


Figure 6.3: **Degree distribution of the miRNAs in the co-regulation network.** The X axis stands for the degrees of each miRNA and the Y axis represents the proportion of each degree category in the miRNA-miRNA co-regulation network. There are 19 nodes having a degree of 1, while there is only one miRNA with a degree of 6.

2012). As another example, the co-regulation between miR-16/15b is also re-discovered in this network.

Special interest was paid to the 'hub' proteins in the functional module of every miRNA pair of this co-regulation network. A hub protein in a functional module is a protein having a far bigger number of connections than the other proteins. We calculated the hub degree of the 31 corresponding interaction networks formed by the target gene products and found that Tumour Protein p63 (TP63) has the highest hub degree, forming a module with six other genes (e.g., PRKD2, FOXP1, TIPARP, TSHZ3, PKD2 and MYLK). TP63 may be involved in oncogenesis in a broader range of tumours

Table 6.2: GO functional analysis of a functional module with seven genes in the TP63 interaction network (p-value <1.0e-04).

GOBPID	P-value	Count	Size	Term
GO:0048745	1.8997e-17	6	20	smooth muscle tissue development
GO:0060537	7.1130e-10	6	322	muscle tissue development
GO:0009888	3.7576e-08	7	1285	tissue development
GO:0009887	1.6968e-07	6	802	organ morphogenesis
GO:0072001	2.5428e-06	4	246	renal system development
GO:0048513	3.1266e-06	7	2414	organ development
GO:0001655	4.5669e-06	4	285	urogenital system development
GO:0035295	2.4799e-05	4	437	tube development
GO:0048731	3.4820e-05	7	3405	system development
GO:0044767	3.8408e-05	7	3453	single-organism developmental process
GO:0009653	4.6691e-05	6	2069	anatomical structure morphogenesis
GO:0048856	9.5426e-05	7	3932	anatomical structure development

including lung tumours (Au, Gown, Cheang, Huntsman, Yorida, Elliott, Flint, English, Gilks & Grimes 2004). Miki (Miki, Kubo, Takahashi, Yoon, Kim, Lee, Zo, Lee, Hosono, Morizono et al. 2010) reported that genetic variation in TP63 may influence susceptibility to lung adenocarcinoma in Japanese and Korean populations. TP63, also known as transformation-related protein 63, is a member of the p53 family of transcription factors, that are essential for the prevention of cancer formation.

The GO functional analysis on the genes in this module was performed. Table 6.2 shows that smooth muscle tissue development-related genes are the most significant. Smooth muscle plays a critical role in pulmonary function by regulating air flow in the lungs, and smooth muscle function can often be compromised as a result of lung disease (Low & White 1998).

6.4.3 Lung Cancer Related miRNAs Have More Functional Synergism

As shown above, the degree information of nodes in a network is one of the most important topological measurements of the network as it can indicate a local centrality of the nodes in the network (Wei, Deng, Zhang, Deng & Mahadevan 2013). The greater the degree is, the more important is the node for the stabilisation of the network. For the miRNA-miRNA co-regulation network (Figure 6.2), we divided all its miRNAs into two categories to understand the difference between the subnetwork of lung cancer-related miRNAs and the subnetwork of the other miRNAs (classified according to the miR2Disease database).

The total degree of the 33 lung cancer miRNAs is 69 and that of the lung cancer un-related miRNAs is 14. The median and average degree of the lung cancer-related miRNAs are 3 and 2.0909±1.2836 respectively, while those of un-related miRNAs are 2 and 1.4±0.6992. This indicates a difference in the functional complexity of these two subnetworks of miRNAs. The functional complexity of miRNAs can be also understood by looking at the number of their regulation modules. The 33 lung cancer-related miRNAs are observed to regulate more functional modules and have more functional synergism than the un-related miRNAs. Therefore, our results can suggest that the dysregulation of those miRNAs co-regulating more biological processes is more likely to cause lung cancer.

6.4.4 Lung Cancer Related miRNAs and Their Functional Modules

From the miR2Disease database, we can understand that 33 of the 43 miRNAs are related to lung cancer (i.e., disease miRNAs). Indeed, the let-7 family members (let-7a, b, c, d, e, f, and g) can especially synergistically regulate the same functional gene set (ACVR2B, ACVR1B, ACVR2A and SMAD2). In our work, ACVR2B, ACVR1B, ACVR2A and SMAD2 all

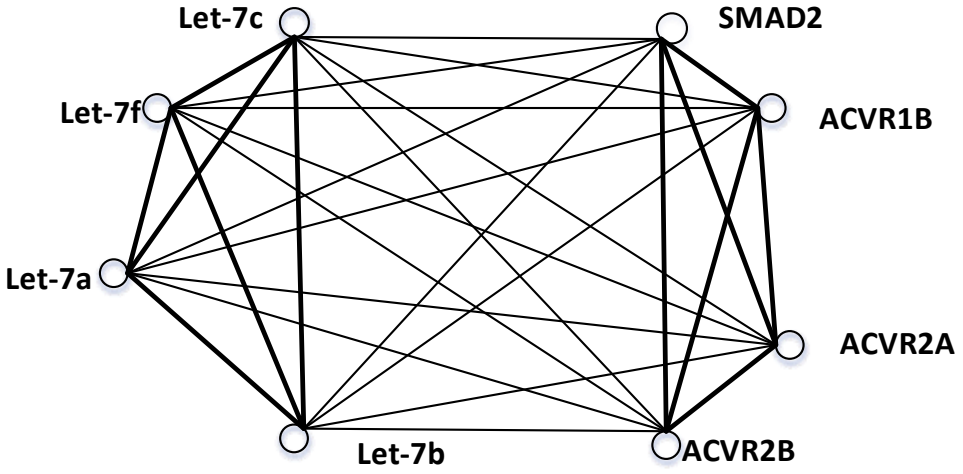


Figure 6.4: **An example of co-regulation between miRNAs (let-7a, b, c and f) and one of their functional modules (SMAD2, ACVR1B, ACVR2A and ACVR2B).** There is a co-regulation between let-7c/f, let-7c/a, let-7c/b, let-7a/b, let-7a/f, and let-7a/b. SMAD2, ACVR1B, ACVR2B and ACVR2A define a functional module, and they are directly connected to each other in the protein-protein interaction network.

together define a functional module of interacting proteins enriched in the "TGF-beta signaling pathway". Some mutations of these proteins in the functional module can affect the development of many cancers (Orton, Sturm, Vyshemirsky, Calder, Gilbert & Kolch 2005). Furthermore, let-7 is a very attractive potential therapeutic that can prevent tumour genesis and angiogenesis for lung cancer patients. let-7 has several key oncogenic mutations including P53, RAS and MYC, some of which may directly correlate with the reduced expression of let-7 and be repressed by introduction of let-7. Figure 6.4 shows an example of the co-regulating miRNAs and their regulated functional modules, in which the five miRNA pairs can regulate the four targets synergistically.

There are a total of 124 unique targets in these 31 different functional modules. About 54.8% of these target genes are associated with lung

cancer and 13.7% of them exist in at least 3 functional modules (Table 6.3). According to the records in AceView (`http://www.ncbi.nlm.nih.gov/`, a well maintained, comprehensive and non-redundant sequence representation of all public mRNA sequences), about 95.2% of these target genes are expressed at high level and their sequences are defined by many GeneBank accessions from cDNA clones (some from lung).

6.4.5 KEGG Pathway Analysis Results

A KEGG pathway analysis was conducted on the genes of the functional modules to determine significant changes in signalling pathways. Those pathway terms containing more than two genes with p<1.0e-4 are shown in Figure 6.5 (the pathways with number > 3). It can be seen that these genes are more often be involved in the mTOR signaling pathway, chronic myloid leukemia pathway and the pancreatic cancer pathway.

mTOR pathway is an intracellular signalling pathway important in regulating the cell cycle. It is directly related to cellular quiescence, proliferation, cancer and longevity. mTOR at the top of the list is consistent with our current knowledge of this pathway in lung cancer. There are many known studies showing that the dysregulation of mTOR signalling frequently happens in a wide variety of cancers including lung cancer (Ekman, Wynes & Hirsch 2012, Fumarola, Bonelli, Petronini & Alfieri 2014) and some miRNAs can function as a tumour suppressor in NSCLC metastasis by inactivating the mTOR signalling pathway (Yu, Li, Yan, Liu, Lin, Zhao, Sun, Zhang, Cui, Zhang, He & Yao 2015).

The TFG-beta signalling pathway participates in various biological processes and plays a critical role in lung cancer as the mTOR signaling pathway. TFG-beta signalling can inhibit tumour growth in early-stage tumours and contribute to lung cancer progression (Jeon & Jen 2010, Jakubowska, Naumnik, Niklińska & Chyczewska 2015). All these existing studies provide strong support for such a high rate of hits in mTOR and TFG-beta signalling pathways in Figure 6.5.

Table 6.3: Target genes in the functional modules

ID	Name	Lung cancer	no.
EIF2C1	eukaryotic translation initiation factor 2C,1	relevant	6
SMAD2	SMAD family member 2	irrelevant	6
HIPK2	homeodomain interacting protein kinase 2	irrelevant	6
TNRC6B	trinucleotide repeat containing 6B	irrelevant	5
DICER1	dicer 1, ribonuclease type III	relevant	5
ACVR2A	activin A receptor, type IIA	irrelevant	4
LRP6	low density lipoprotein receptor-related protein 6	irrelevant	4
PTEN	phosphatase and tensin homolog	relevant	4
ACVR1B	activin A receptor, type IB	irrelevant	3
TNRC6A	trinucleotide repeat containing 6A	irrelevant	3
VEGFA	vascular endothelial growth factor A	relevant	3
EIF2C4	eukaryotic translation initiation factor 2C, 4	irrelevant	3
FXR1	fragile X mental retardation, autosomal homolog 1	relevant	3
CDKN1C	cyclin-dependent kinase inhibitor 1C	relevant	3
RPS6KA3	ribosomal protein S6 kinase, 90kDa, polypeptide 3	relevant	3
TSC1	tuberous sclerosis 1	relevant	3
ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4	relevant	3

The column labelled "Lung cancer" means whether the gene is relevant to lung cancer in the OMIM database. The column of "no." indicates the existing times of the genes in all the functional modules. This table only focuses on the "no." of at least 3.

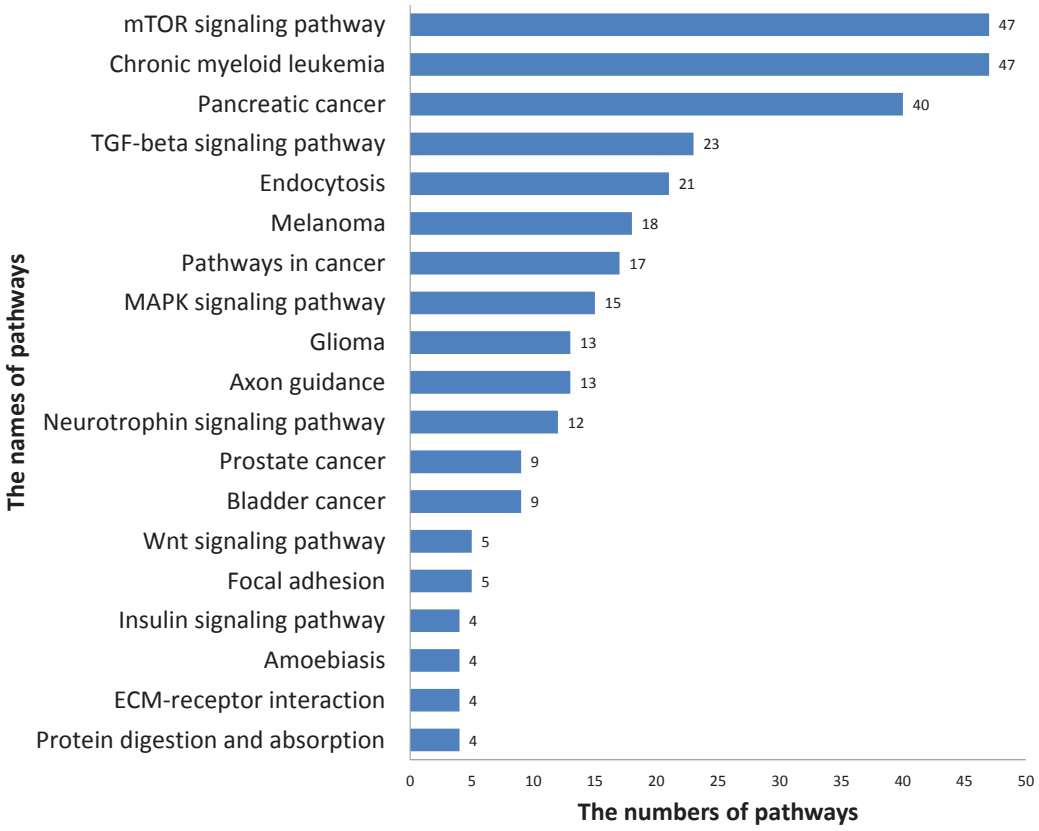


Figure 6.5: **KEGG pathway enrichment analysis for the target subsets of each miRNA pair in the co-regulation miRNA network.** The X axis shows the existing numbers of the corresponding pathways and the Y axis describes the pathways' names (P-value $<$ 1.0e-4) in the three data sets.

In addition, a co-regulating miRNA pair miR-221/222 (Garofalo et al. 2012, Dentelli, Traversa, Rosso, Togliatto, Olgasi, Marchiò, Provero, Lembo, Bon, Annaratone et al. 2014) is also verified by the pathway of "MicroRNAs in Cancer" in the KEGG pathway database, suggesting that miR-221/222 co-ordinate to up-regulate the common targets (PTEN, p27 and TMP3) and are involved in the lung epithelial cell, tumorigenesis, survival, angiogenesis and invasion/metastasis.

6.4.6 Transcription Factors Related to Lung Cancer

Genes and their mRNAs are controlled not by a single, but by a combination of TFs or miRNAs. Cooperative regulation therefore can provide the mechanistic basis for reading out combinatorial expression patterns for both TFs and miRNAs. From the ChIPBase database (Yang, Li, Jiang, Zhou & Qu 2013), we obtained the target genes' transcription factors of 42 of the 43 miRNAs (nodes) in the co-regulation network (except for miR-103, details shown in **Supplementary file**). We can see that miRNAs involved in an co-regulatory module can be confirmed to share common TFs.

From the shared TFs, we found that a transcription factor named caudal type homeobox 2 (*CDX2*) is shared by 31 miRNAs in the miRNA-miRNA co-regulation network. Bai (Bai, Miyake, Iwai & Yuasa 2003) reported that the CDX2 homeobox transcription factor can unregulate transcription of the p21/WAF1/CIP1 gene, and that p21 plays key roles in differentiation and tumour suppression. Many other studies also indicated that CDX2 is a potential tumour suppressor gene in colon and gastric cancer (Bonhomme, Duluc, Martin, Chawengsaksophak, Chenard, Kedinger, Beck, Freund & Domon-Dell 2003, Gross, Duluc, Benameur, Calon, Martin, Brabletz, Kedinger, Domon-Dell & Freund 2007, Do Youn Park, Kim, Mino-Kenudson, Deshpande, Zukerberg, Am Song, Lauwers et al. 2009). Furthermore, Liu (Liu, Zhang, Zhan, Brock, Herman & Guo 2012) have demonstrated that CDX2 is frequently methylated in lung cancer, and that the expression of CDX2 is regulated by a promoter region hypermethylation in lung cancer.

Therefore, CDX2 may serve as a tumour suppressor in lung cancer and a lung cancer detection marker, inhibiting lung cancer cell proliferation by suppressing Wnt signalling. These evidences show that the CDX2 homeobox transcription factor shared by most miRNAs in the miRNA-miRNA co-regulation network may play critical roles in lung cancer.

6.5 Summary and Conclusion

As described, this work applied an integrative computational method to discover a common miRNA-miRNA co-regulation network from three lung cancer miRNA data sets, supporting **Contribution 3** of the thesis as listed in Section 1.3 . As the first step, the three data sets are processed to find common miRNAs. Then, miRNA pairs are ranked using Pearson's correlation coefficient and their common targets provided by the Targetscan database. We observed that co-regulating miRNAs always show a high correlation in their expression profiles. A GO functional enrichment and a protein interaction analysis on the common targets have been further used to filter some of these miRNA pairs. GO functional enrichment is another factor that matters when miRNAs regulate mRNAs without changing their expression levels. Protein interaction analysis has advantages to avoid the incompleteness of GO functional enrichment (Thomas, Wood, Mungall, Lewis, Blake, Consortium et al. 2012), allowing us to analyse miRNAs' functionality according to the feature of their protein products (Yuan et al. 2009). Proteins usually fulfill certain functions by means of interaction. The closer these proteins are in the PPI network, the more likely the targetting miRNAs are located in the same cluster (Liang & Li 2007). So, integrating different types of data from various sources is potentially more successful than any single database, which can help to decrease the false positive results and understand the results from many biological perspectives (Le & Bar-Joseph 2013).

The present study suggests that the newly discovered miRNA-miRNA

co-regulation network is scale free and its degree distribution follows a power law. These lung cancer related miRNAs have more synergistic influence; and miRNAs from the same family tend to have similar functions and high correlation (Gong, Kakrana, Arikit, Meyers & Wendel 2013). Some miRNA interactions have been identified by previous work, including miR-15b/16 (Cimmino et al. 2005, Xia et al. 2008), miR-221/222 (Dentelli et al. 2014) and let-7a/b/c/d/g/f (Johnson, Grosshans, Shingara, Byrom, Jarvis, Cheng, Labourier, Reinert, Brown & Slack 2005).

We also confirm that known lung cancer related miRNAs have more synergism than lung cancer un-related miRNAs. KEGG pathway enrichment analysis and transcription factor analysis have all demonstrated the biological relevance of the miRNA-miRNA co-regulation network to lung cancer. This study discovered that potential co-regulating miRNAs and potential signalling pathways may lend insight into lung cancer. Our analysis can help scientists to look at these significant relationships. The proposed method can also be applied to other diseases data sets for constructing their respective miRNA-miRNA co-regulation networks.

Chapter 7

A Novel Framework for Inferring Self-regulation miRNAs

7.1 Introduction

As is explained in the related work (Section 2.4), TFs and miRNAs are known to positively or negatively regulate transcription. Gene expression is an important mechanism to shape the cell-specific gene regulatory system. miRNAs have their own characteristics, making it difficult or impossible to apply the experimental and computational methods used for other gene regulation (Hobert 2008*b*).

miRNAs are mainly located in intergenic regions or in the introns of protein coding genes (Kim & Nam 2006). A promoter region is located around the transcription start site of a transcript and is regulated by proteins that bind to this region. Evidence suggests that binding sites for transcription factors are similarly distributed within the promoters of both protein coding genes and miRNA transcripts (Hobert 2008*b*).

A miRNA gene is controlled by several TFs whose binding sites (TFBS) are located near the TSS of this gene. When transcribed, the miRNA gene

produces a long pri-miRNA molecule. The pri-miRNA molecule is cleaved by Drosha and yields part of the hairpin along with producing the miRNA-miRNA* duplex (Krol, Sobczak, Wilczynska, Drath, Jasinska, Kaczynska & Krzyzosiak 2004). One chain of the miRNA duplex is incorporated into the RISC complex and can regulate miRNA translation by binding in a sequence specific manner to the 3' UTR of mRNAs (Bernstein, Caudy, Hammond & Hannon 2001).

Most of the research over the past decade have concentrated on elucidating the mechanisms of miRNA-mediated post-transcriptional regulation in cancer and other diseases, and on the potential clinical applications of this knowledge (Winter, Jung, Keller, Gregory & Diederichs 2009, Chekulaeva & Filipowicz 2009).

Many important biological processes are actually controlled by miRNAs which act as the role of master regulators. Liu et al. (Liu, Roth, Yu, Morris, Bersani, Rivera, Lu, Shioda, Vasudevan, Ramaswamy et al. 2013) found that miR-483-5p, which is located in an intron of IGF2, was up-regulated to the transcription of IGF2 active. Ectopic expression of miR-483-5p in IGF2-dependent sarcoma cells increased tumour size in mice, strengthening the function of this microRNA and positive feedback regulation of its host gene in tumorigenesis. This is the case for instance in the miRNA-mediated Feed Forward Loop (FFL) or the miRNA mediated self-loop, in which the miRNA plays the role of master regulator.

It is still poorly understood how miRNAs themselves are regulated. This is partly due to the difficulty of predicting promoters from short conserved sequence features without producing a high number of false positive and partly due to the heterogeneity of the miRNA biogenesis pathways.

In this work, we design a novel framework (called SRmiR) to integrate multiple data types for exploring self-regulated miRNAs for understanding their mechanisms. Particularly, SRmiR is aimed at discovering the self-regulated miRNAs specific to humans, by using heterogeneous data. We define a self-regulated miRNA if the miRNA regulates a TF and together with

it one or more target genes and if there is a TF-target interaction. Firstly, we collected a total of 1881 human miRNAs from the miRBase and obtained the promoter regions of all the miRNA primary transcripts. Secondly, we collected 690 ChIP-seq datasets representing a total of 161 unique regulatory transcription factors from the ENCODE project (Consortium et al. 2004). Thirdly, we discovered the potential miRNA-TF relationships between TFs and miRNAs by comparing the miRNAs' promoter regions and transcription factor binding sites (TFBS). After that, we obtained the miRNA-target relationships between miRNAs and genes, and the TF-gene relationships between TFs and genes. Finally, 13 genes (BACH1, BRCA1, CTBP2, EBF1, HDAC2, HNF4G, IRF1, MEF2C, MTA3, NFIC, SMC3, TAL1 and TCF7L2) are shown to have self-regulations, and we discussed the FFL involving these genes as Transcription Factors and targets. This addresses **Contribution 4** of the thesis.

7.2 Materials

7.2.1 Construction of TF-miRNA Relationships and miRNA-target Relationships in the Post-transcriptional Regulatory Network:

As miRNAs located within protein coding genes tend to be co-regulated with their host genes, we focus on identifying TFs that regulate intergenic miRNAs in this work. We downloaded the TFBS data set from the TFs database ENCODE (`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/`). The data set contains 690 ChIP-seq data sets representing 161 unique regulatory factors, which span 91 human cell types and some are in various treatment conditions. These data sets were generated by the five ENCODE TFBS ChIP-seq production groups: Broad, Stanford/Yale,UC-Davis/Harvard, HudsonAlpha Institute, University of Texas-Austin, University of Washington, and University of

Chicago. All ChIP-seq experiments were performed at least in duplicate, and were scored against an appropriate control designated by the production groups.

Transcription of protein coding genes as well as some small RNAs, such as miRNAs, is carried out by Pol II. While Pol II binds to the DNA at the transcription initiation point, it is not capable of directly recognising its target. A complex of proteins in a region known as the core promoter binds to the DNA whereupon they recruit Pol II to the transcription start site (TSS). Other proteins, called TFs, then bind to the proximal prompter or enhancer regions to either initiate or block the activation of Pol II. The core promoter region typically consists of hundreds to thousands of base pairs surrounding the TSS of a gene.

A total of 1881 human miRNA were obtained from the miRBase (`ftp://mirbase.org/pub/mirbase/CURRENT/genomes/hsa.gff3`). Since the majority of primary transcripts of intergenic miRNAs are shorter than protein-coding transcripts, with TSSs located within 2,000 bp upstream and poly(A) signals located within 2,000 bp downstream of the pre-miRNAs and the promoters of most miRNA genes are found within 500-bp upstream of the TSS, potential promoter regions (from 2,000 bp upstream up to 500 bp downstream to the expected TSSs) were obtained to predict the potential TF binding sites.

We mapped TFBSs to the individual promoter regions of miRNA genes using the BEDTools (Quinlan & Hall 2010) for the comparison of sequence alignments between miRNA promoters and TFBSs. BEDTools are a new software suite for the comparison, manipulation and annotation of genomic features in Browser Extensible Data (BED), Sequence Alignment/Map (SAM) and General Feature Format (GFF) format (Quinlan & Hall 2010).

We used *intersectBed* to extract overlapping features between BED-files of TFBSs and miRNA promoters. If there were overlapping features with predicted TFBSs in the promoters of miRNA genes, this suggests that these TFs are involved in regulating pri-miRNAs transcription, at a level similar

to protein coding genes.

As potential targets of miRNAs we selected only transcripts corresponding to protein-coding genes completely annotated in Ensemble 83, for a total of 102450 known transcripts. To define miRNA targets, we used the four most commonly cited prediction algorithms: DIANA((Alexiou et al. 2010)), Miranda ((John et al. 2004)), PicTar ((Krek et al. 2005)) and TargetScan ((Lewis et al. 2003)). Integrating the four databases, we annotated how many databases confirm the target genes with these miRNAs involving the miRNA-TF relationships. Then, out of these interactions we selected those targets involving at least two databases.

7.2.2 Construction of TF-target Relationships in the Transcriptional Regulatory network:

Transcription factors (TFs) regulating a miRNA often regulates its target genes and TFs are the main regulators of gene transcription. 690 ChIP-seq studies cover 161 transcription factors. We used a package named *tftargets* of R software and chose the ENCODE data set to obtain the 161 putative human transcription factor targets based on ChIP-seq data from ENCODE (source: `http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegTfbsClustered/`).

7.2.3 Identification of Important Network Motifs: miRNA-mediated Feed Forward Loops

We constructed the list of putative miRNA-TF, miRNA-Target and TF-Target links obtained above. In the miRNA-mediated FFL circuit, a transcription factor TF (X_1) regulates a miRNA (X_2), and they both regulate a target mRNA (X_3). Further, three models based on ordinary differential equations are examined to describe the miRNA and target mRNA expression kinetics. All models consider X_1 as a forcing function and describe the rate of change of X_2 and X_3 as the balance between their synthesis (S_i) and

degradation (D_i) with the basal expression level (X_ib) as the initial condition; the topological model is shown in Figure 7.1. Thus, for i=2,3, the differential equation describing the variables is

$$X_i(t) = S_i(t) - D_i(t)$$

$$X_i(0) = X_{ib}$$

The synthesis is expressed as the sum of a basal term (S_{ib}), plus a positive (activation) or negative (repression) term (ΔS_i) encoding the effect of the specific TF on the transcription of miRNA and target mRNA.

The list of miRNA FFLs can be found from the links by using a tool named FANMOD (Wernicke & Rasche 2006). In order to reduce the number of false positives, we selected only the FFLs with both miRNA regulatory links confirmed by all four databases. The specific parameters used in FANMOD were as follows:

- Network Number of nodes: 6379 number of genes Number of edges: 44647 (44576 single, 71 bidirectional) number of interactions between genes
- Algorithm Size of subgraphs: 3 nodes Algorithm: enumeration
- Random Networks Number of Random Networks: 1000 Edge exchange parameters: 4 per edge, 4 tries per exchange
- Computation Run time: 30.26 hours System: PC running Windows 7 64-bit, Intel Core2 Duo CPU E8600 @ 3.33GHz, and 4GB RAM

FANMOD outputs all 3-node subgraphs (with gene IDs), grouped by topological equivalency. For each of the 13 subgraphs (there are only 13 possible permutations for 13 the interactions in a 3-node subgraph), we calculated all phase locking indices from 1:1 to 4:3 for each possible gene pairing.

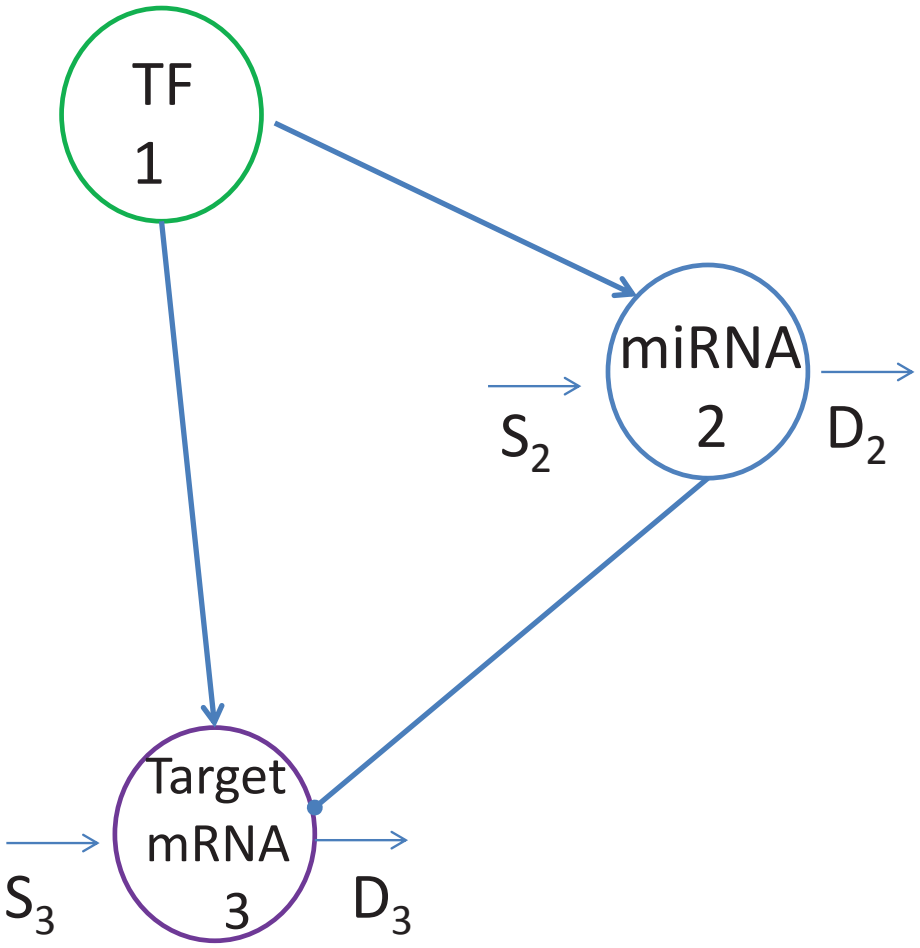


Figure 7.1: **The topological model of miRNA-mediate FFLs.** A TF regulates a miRNA, and they both regulate the target mRNA, and miRNA regulation of the target gene is negative. S and D represents synthesis and degradation respectively.

7.2.4 Identification of Experimentally Validated Regulatory Interactions

The list of miRNA FFLs with experimentally validated regulatory interactions was obtained combining information collected from several databases. For the miRNA-Target and the miRNA-TF interactions we used the last versions of miRTarBase V 3.5 (updated November, 2012), miRecords V.3 (updated

on November, 2010) and miR2Disease (updated on Jun, 2010). We obtained in this way a list of experimentally validated miRNA-T interactions. For TF-Target interactions we used data from ENCODE and the last version of Tfact(v.2). Tfact contains genes responsive to transcription factors which focus on humans, according to experimental evidence reported in the literature. It reports two data sets: (i) a sign sensitive catalogue that indicates the type (up or down) of TF regulation exerted on its targets and (ii) a signless catalogue that includes all regulatory interactions contained in sign sensitive one plus further interactions without the specific type of regulation.

7.3 Results

Our results are presented in five parts. The first part reports self-regulated miRNAs and their interacting network. The second part presents self-regulated transcription factors and their interacting network. The third part illustrates discovery of miRNA-mediate feed-forward loops. The fourth part describes the validation of miRNAs self-regulations.

7.3.1 Self-Regulated miRNAs in the Human Regulatory Network

A detailed description of our procedure is reported in the Materials and Methods section. Accordingly, we only report here the main steps. Briefly, we constructed a list of putative miRNA FFLs combining miRNA-TF, miRNA-Target and TF-Target regulatory interactions which were obtained as follows: for the miRNA-TF side, we integrated information obtained from the miRNAs sequence data and TFBSs ChIP-seq data provided by miRBase and ENCODE respectively. We selected potential regulations if there were overlaps between miRNA promoter regions and TFBSs. For the miRNA-Target side we selected the miRNAs contained in the miRNA-TF regulations.

Then, we used information obtained from four freely available databases of miRNA-Target interactions: DIANA (Alexiou et al. 2010),Miranda (John et al. 2004), PicTar (Krek et al. 2005) and TargetScan (Lewis et al. 2003). We selected as potential targets only transcripts corresponding to protein-coding genes completely annotated in Ensemble 83 and occurring in at least three databases. For the TF-T side we selected the TFs contained in the miRNA-TF regulations. Then we obtained the putative human transcription factor targets based on ChIP-seq data from ENCODE. In fact with the ENCODE list, based on ChIP-seq experiments, we expected to have a smaller rate of false positives results with respect to a purely bioinformatic approach. At the same time, using only the ENCODE list we were able to induce a statistical bias in the results due to the fact that ChIP-seq experiments were performed only for a small subset of TFs which were selected for their particular biological relevance.

7.3.2 Identification of Self-regulated Transcription Factors

In total, we obtained a number of 436,644 relationships among 159 TFs, 688 miRNAs, and 14,464 genes. 13 genes (BACH1, BRCA1, CTBP2, EBF1, HDAC2, HNF4G, IRF1, MEF2C, MTA3, NFIC, SMC3, TAL1 and TCF7L2) are involved in the self regulations, which indicate that they can target some miRNAs' host genes and can also be regulated by these miRNAs. A total of 402 miRNAs were associated with these self-regulated TFs, and eight miRNAs targetted at least 8 self-regulated TFs (Figure 7.2 and Figure 7.3).

In addition, 16 miRNAs involved in seven self-edge TFs (MEF2C, BACH1, HDAC2, TCF7L2 (Karginov & Hannon 2013), TAL1, SMC3 and EBF1 (Tavazoie, Alarcón, Oskarsson, Padua, Wang, Bos, Gerald & Massagué 2008)) have the same relationships in both the miRNA-target and miRNA-TF regulations. The relationships are shown in Figure 7.4.

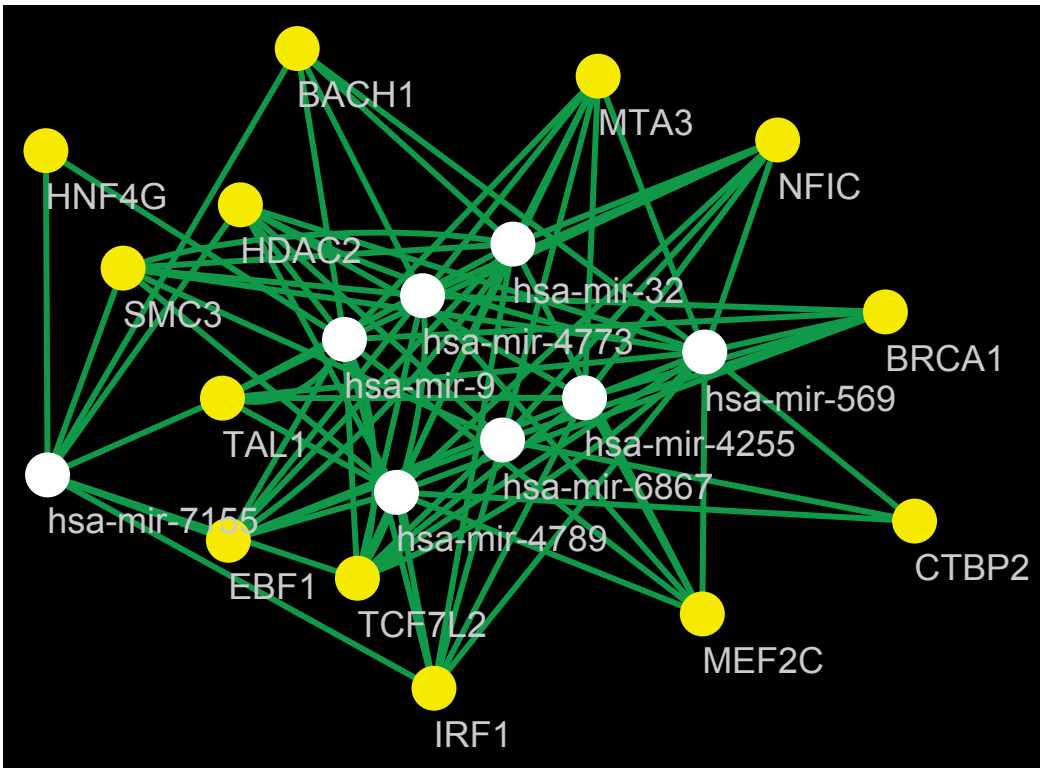


Figure 7.2: **The relationships between miRNAs and the self-edge TFs.** The number of TFs is no less than 8.

	HNF4G	BACH1	MTA3	NFIC	BRCA1	CTBP2	MEF2C	IRF1	TCF7L2	EBF1	TAL1	SMC3	HDAC2
hsa-mir-32													
hsa-mir-9													
hsa-mir-4255													
hsa-mir-569													
hsa-mir-4773													
hsa-mir-7155													
hsa-mir-6867													
hsa-mir-4789													

Figure 7.3: **The relationships between miRNAs and the self-edge TFs.** The number of TFs is no less than 8.

7.3.3 Identification of miRNA-mediate Feed-Forward Loops

miRNAs are known to be involved in feed-forward loops where a TF regulates a miRNA and they both regulate a target gene. FANMOD provides

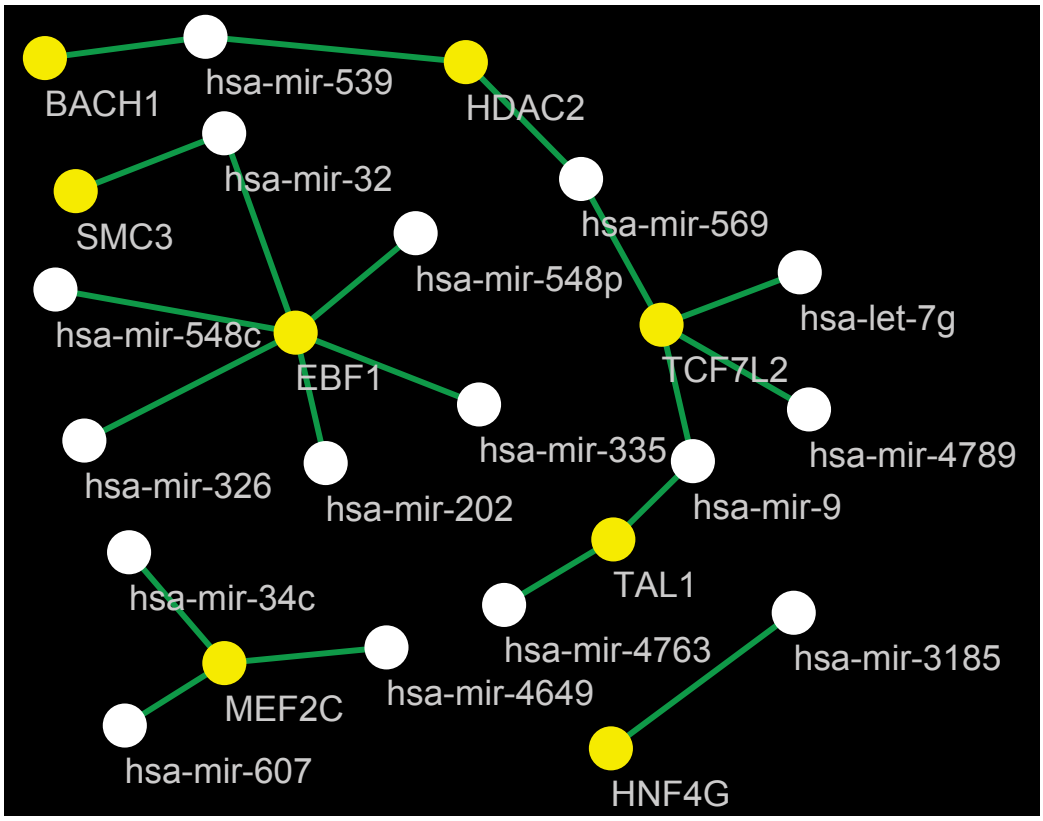


Figure 7.4: **Self-regulated miRNAs involved in self-edge TFs.** These relationships exist in both the miNRA-target and miRNA-TF regulations.

several statistical values alongside significant network motifs (details in the footnotes). The FANMOD tool was run with a subgraph size of 3, using the full enumeration algorithm option and generating 100 random networks for determining subgraph significance. The FANMOD reported a feed-forward loop motif with an ID of 38, as well as a single-input module with and ID of 6. The following table shows the results for full enumeration of the network, enumerating subgraphs of size three. All of the graphs are ordered by descending Z-Score, so that the most significant network motifs are listed first.

Table 7.1: All of the identified motifs using a FANMOD trial with a subgraph size of 3.

ID	Adj	Frequency[a] [Original]	Mean-Freq [Random]	Standard-Dev [Random]	Z-Score[b]	P-Value[c]
6	**000 000 110**	**96.014%**	**96.083%**	**2.9631e-005**	**-23.301**	**1**
12	000 001 100	1.7567%	1.6883%	2.9426e-005	23.251	0
36	000 100 100	1.7402%	1.6698%	2.9095e-005	24.204	0
38	**000 100 110**	**0.23967%**	**0.31121%**	**3.048e-005**	**-23.473**	**1**
14	000 001 110	0.22325%	0.21996%	1.521e-005	2.1616	0.018
46	000 101 110	0.020082%	0.021881%	7.6272e-006	-2.3587	0.986
164	010 100 100	0.0046682%	0.0042945%	3.0896e-007	12.094	0
140	010 001 100	0.00044733%	0.00036873%	1.9796e-007	3.9706	0
102	001 100	0.00032164%	0.00035179%	1.4998e-007	-2.0103	0.977

166	110 010 100	0.000298%	0.00045117%	1.4503e-007	-10.561	1
78	110 001 001	0.00026729%	0.00023289%	6.0374e-008	5.6983	0
174	110 010 101	7.82e-005%	0.00012363%	6.0488e-008	-7.5112	1
238	110 011 101	3.5345e-006%	3.1194e-008%	1.9159e-009	18.286	0

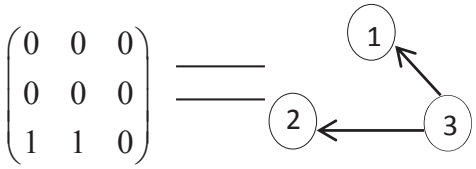
[1]

The adjacency matrix from FANMOD representing co-regulated genes. Each row corresponds to the regulator and each column corresponds to the gene that is regulated. For example, 3^{nd} row has a 1 for column 1 and 2, thus gene 3 regulated genes 1 and 2.

[1a] The frequency denotes the frequency with which a motif occurred in the original network.

[b] The Z-score is one way of determining how significant a network motif is. The FANMOD documentation and manual describes how the Z-score is calculated, "The Z-Score is the original frequency minus the random frequency divided by the standard deviation." Motifs with the highest Z-scores are the most significant, so the following tables of motifs are organized in order of decreasing Z-score.

[c] P-Values range from zero to one; smaller p-Values indicate more significant motifs because a smaller p-value indicates that the motif occurs more often in the network than would occur by random chance. The p-Value is calculated in the following way, "The p-Value of a motif is the number of random networks in which it occurred more often than in the original network, divided by the total number of random networks."



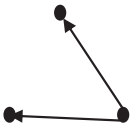
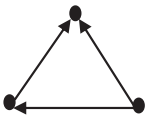
ID	Adj	Frequency [Original]	Mean-Freq [Random]	Standard-Dev [Random	Z-Score	P-Value
6		96. 014%	96.083%	2.9631e-005	-23.301	1
38		0.23967%	0.31121%	3.048e-005	-23.473	1

Figure 7.5: **Results of FANMOD trial with a subgraph size of 3.** only the ID of 38 and 6 are shown in this figure.

Figure 7.5 reported only the feed-forward loop motif with an ID of 38 or 6.

7.3.4 Validation of miRNAs Self-Regulations

We employed four manually curated databases (miRTarbase (Chou, Chang, Shrestha, Hsu, Lin, Lee, Yang, Hong, Wei, Tu et al. 2015), Tarbase (Vlachos, Paraskevopoulou, Karagkouni, Georgakilas, Vergoulis, Kanellos, Anastasopoulos, Maniou, Karathanou, Kalfakakou et al. 2014), miRecords (Xiao et al. 2009) and miRWalk (Dweep et al. 2011, Dweep & Gretz 2015)) to show the evidence related to the regulatory effect of these self-regulated miRNAs over its TF targets. miR-335 was verified to show an experimentally validated regulatory relationship with Early B-cell Factor 1 (EBF1) (Tavazoie et al. 2008) by the microarrays method. (Tavazoie et al. 2008) identified that the expression of miR-335 is lost in the majority of primary breast tumours from patients

who relapse, and the loss of expression of either microRNA is related to poor distal metastasis-free survival. Thus, miR-335 is discovered as a metastasis suppressor microRNA in human breast cancer. In addition, Let-7g has been verified to have an experimentally validated regulatory relationship with Transcription Factor 7-Like 2 (TCF7L2) (Karginov & Hannon 2013) by the immunoprecipitation method.

In addition to the existing databases, we also looked into the literature work and found that miR-9 is expressed specifically in neurogenic areas of the brain and may be involved in neural stem cell self-renewal and differentiation (Zhao, Sun, Li & Shi 2009).

7.4 Discussion

As described, this work applied an integrative computational method to discover the common miRNA-miRNA co-regulating network in three lung cancer data sets. The three data sets are used to find the common miRNAs for improving miRNAs' reliability and robustness. We identified the miRNA pairs by using Pearson's correlation coefficient and their common targets provided by Targetscan database. We observed that co-regulating miRNAs always show high correlation in their expression levels. The higher the correlation is in the expression data, the more promising the miRNA pair is co-regulated. A GO functional enrichment and a protein interaction analysis on the common targets have been used to filter some of the miRNA pairs. GO functional enrichment is another factor that matters when the miRNAs regulate mRNAs without changing their expression levels. Protein interaction analysis has advantages to avoid the incompleteness of GO functional enrichment (Thomas et al. 2012), allowing us to analyze miRNAs' functionality according to the feature of their protein products (Yuan et al. 2009). Proteins usually fulfill certain functions by means of interaction. The closer these proteins are in the PPI network, the more likely the targeting miRNAs are located in the same cluster (Liang & Li 2007). So, integrating

different types of data from various sources (e.g., PPI network) is potentially more successful than any single database, which can help to decrease the false positive results and understand many biological perspectives (Le & Bar-Joseph 2013).

The present study suggests that a miRNA-miRNA co-regulating network is scale free and its degree distribution follows a power law. These lung cancer related miRNAs have more synergistic influence and miRNAs from the same family tend to have similar functions and high correlation (Gong et al. 2013). Some miRNA interactions have been identified in previous work, including miR-15b/16 (Cimmino et al. 2005, Xia et al. 2008), miR-221/222 (Dentelli et al. 2014) and let-7a/b/c/d/g/f (Johnson et al. 2005), and new co-regulating miRNAs especially the miRNAs from a same family (e.g., miR-18a/b) will allow expansion of our understanding of lung cancer. Further validation is still required for the results since our analysis was based on some imbalanced data sets.

As described, this work applied a novel framework to discover self-regulation miRNAs in humans. We constructed the miRNA promoter region information. Then, the miRNA-TF relationships, miRNA-target relationships and TF-target relationships were constructed to discover the feed-forward loops and to detect the self-regulation miRNAs. Experimentally validated miRNA-gene databases were employed to verify the results. miR-335, let-7 and miR-9 are shown to involved in the self regulation.

The present study suggests that a self-regulation miRNA can regulate the transcription factor target. These rules are entirely new, because complex diseases are often affected by various miRNAs rather than a single miRNA, and single-miRNA rules are insufficient for accurate diagnosis.

The advantage of the method presented here s that we can study all the human miRNAs and use the ChIP-Seq data to discover the important self-regulation miRNAs. In addition, the discovered self-regulation miRNAs can potentially be applied to further investigation of therapeutic targets in various human disease.

7.5 Conclusion

Our results provide strong evidence that coordinated transcriptional and post-transcriptional regulation via miRNAs is a recurrent motif to enhance the robustness of gene regulation in human genomes. As suggested by our findings, self-regulation miRNAs tend to play important roles in various human disease and the miRNA-mediate repression will provide a comprehensive view on how gene expression is regulated at the systems level.

This chapter addresses **Contribution 4** as listed in section 1.3 by proposing a novel framework to integrate multiple data types for exploring miRNA self-regulations. We defined a self-regulated miRNA if the miRNA regulated a TF and one or more target genes and if there was a TF-target interaction. We collected human miRNAs from miRBase and obtained the promoter regions of all the miRNA primary transcripts. ChIP-seq datasets representing unique regulatory transcription factors were collected from ENCODE at UCSC. The potential miRNA-TF relationships were discovered between TFs and miRNAs by comparing the miRNAs' promoter regions and transcription factor binding sites (TFBS). The miRNA-target relationships were detected between miRNAs and genes, and TF-gene relationships are detected between TFs and genes based on the miRNA-TF relationships.

Chapter 8

Summary and Conclusion

The study set out to design rule mining methods for analysing miRNA expression profiles in human disease, and identified the miRNA biomarkers in lung cancer, co-regulation miRNAs in lung cancer, miRNA-mRNA regulations in HCV infection and miRNA self-regulations in humans. The general literature on this subject is inconclusive in relation to several vital questions. The study sought to answer two of these questions:
1. How do we identify the significant miRNA biomarkers associated with prognosis, diagnosis and progression in cancers?
2. How do we identify the uncovered systematical functions of miRNAs in human disease?

This chapter summarises the research findings on rule discovery to detect reliable miRNA biomarkers and miRNA-mRNA relationships. It also summarises the findings from miRNA-miRNA regulations and miRNA self-regulations. At the end of this chapter, we propose future directions for this research as well as further research opportunities.

8.1 Main Achievements

Rule mining is a method for discovering interesting relations between variables in large data sets. The advancement in the knowledge of miRNA

functions and the large amounts of miRNA expression profiles have created a great need for the understanding of how these small non-coding RNAs are regulated in various human diseases. This work is intended to identify strong rules discovered in the data sets between miRNAs, mRNAs and TFs.

Previous studies have demonstrated the importance of miRNA biomarkers, co-regulation miRNAs, miRNA-mRNA regulations and miRNA self-regulations in human disease. This allows for rule mining methods to be applied in the study of miRNA functions.

Based on the premise of this research, I have conducted a comprehensive investigation to study the effect of miRNAs on human disease by rule mining methods. The main findings are chapter specific and were summarised within the respective chapters (Figure 8.1). This section will synthesise the findings to answer the study's two research questions.

1. How do we identify the significant miRNA biomarkers associated with prognosis, diagnosis and progression in cancers?

- A novel rule method was proposed to discover reliable miRNA biomarkers that can be used to distinguish between healthy and cancer tissue samples. The method can be broadly useful for the study of diagnosis and prognosis of different kinds of diseases including lung cancer, HCV infection, and leukaemia. This study (Song et al. 2014) was presented in the International Conference on Bioinformatics 2014 and published in BMC Genomic (covered in Chapter 4).

2. How do we identify the uncovered systematical functions of miRNAs in human disease?

- A "change to change" method was proposed to derive discriminatory rules for detecting both inverse and positive regulatory relationships. Specifically, rules from the paired miRNA and mRNA expression data of human disease samples and controls are connected to identify the many-to-many miRNA-mRNA regulatory modules involved in cancers. The rule discovery method is useful to integrate binding information

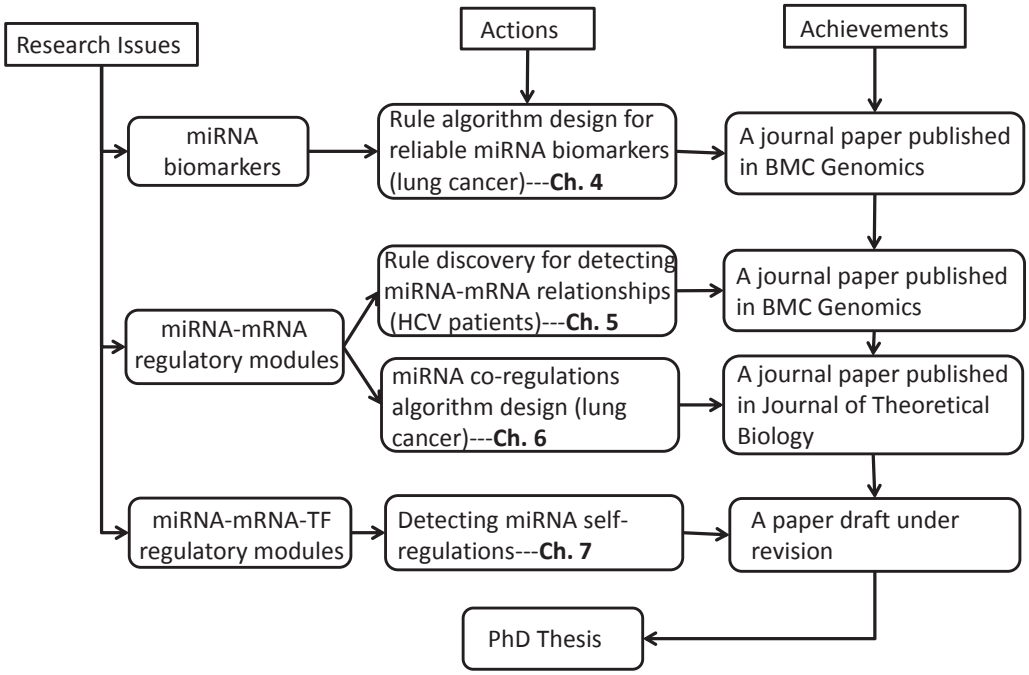


Figure 8.1: **Main achievements of my PhD study**

and the expression profile for identifying miRNA-mRNA regulatory modules and can be applied to the study of other complex human disease expression profiles. This study (Song, Liu, Liu & Li 2015) was presented in the Asia Pacific Bioinformatics Conference 2015 and published in BMC Genomic (covered in Chapter 5).

- An integrative computational method was designed to identify a miRNA-miRNA co-regulation network common to the three lung cancer miRNA expression data sets of different subtypes. The newly discovered miRNA-miRNA co-regulation network is scale free and its degree distribution follows a power law. These lung cancer related miRNAs have more synergistic influence; and miRNAs from the same family tend to have similar functions and a high correlation. This study (Song, Catchpoole, Kennedy & Li 2015) was published in the Journal of Theoretical Biology (covered in Chapter 6).

- A robust methodology was designed to mine big regulatory modules especially the self-regulation miRNAs in the pre-and post-transcriptional level from paired miRNAs, mRNAs and TFs sequence data. The advantage of the method presented here is that we can study all the human miRNAs and use the ChIP-Seq data to discover the important self-regulation miRNAs. In addition, the discovered self-regulation miRNAs can potentially be applied to further investigation of therapeutic targets in human diseases. This paper draft is under revision (covered in Chapter 7).

Conclusions

This section concludes the results and findings that have been achieved in this study. The studies described in this dissertation have helped advance the rule mining on miRNA expression profiles for human disease understanding. The study began in Chapter 2 in which we were able to demonstrate

the advantages of miRNA biomarkers and the disadvantages of existing methods. This led us to use a rule mining method to identify the 2D or 3D biomarker for lung SCC diagnosis. It is the first study to date to apply a rule mining method to identify 2-miRNA and 3-miRNA biomarkers. Analysis of miRNA-mRNA regulatory relationships in Chapter 5 with a rule mining method demonstrated that miRNAs share positive and negative relationships with mRNAs. That analysis also suggested that the inverse relationship is not the only regulatory relationship between miRNAs and mRNAs, and some miRNAs can positively regulate some mRNAs. The proposed "change to change" method is able to discover both the positive and negative relationships at the same time.

In Chapter 6 we further investigated the co-regulations of miRNAs by using a novel integrative approach. The method was able to discover a miRNA-miRNA co-regulation network and co-regulating functional modules common in lung cancer. An example of these functional modules consists of genes SMAD2, ACVR1B, ACVR2A and ACVR2B. This module is synergistically regulated by let-7a/b/c/f, enriched in the same GO category, and has a close proximity in protein interaction network. The similarity in promoters between miRNA and protein coding genes provided us with the incentive in Chapter 7 to search for TFBS resulting in the identification of self-regulation miRNAs.

8.2 Direction for Future Research

Following the above discussion, there is no doubt that rule mining is an ideal research method for miRNA studies. Therefore, due to the success of this study, I do encourage all potential researchers in the bioinformatics field to consider, and hopefully adopt rule mining as their research method.

In addition, miRNAs play a key role in diverse biological processes in eukaryotes, and aberrantly expressed miRNAs play key roles in the development of human disease. It is still a necessary but challenging field of

work in cancer research. I also encourage all potential researchers to study miRNAs and their functions.

The study has offered a rule mining method to study miRNA expression profiles in human diseases. As a direct consequence of the materials and methods, the study encountered a number of limitations, which need to be considered. The first is the small sample size. Further larger studies are thereby required to confirm these results. The second one is the lack of wet-lab experiments, and all the discovered results can only be evaluated by the existing knowledge. The wet-lab experiments are quite useful for verifying the preferred candidate results. Finally, the study mainly considered the expression profiles. It would be a high throughput analysis if we took advantage of the next-generation sequence data.

While the studies in this dissertation have provided a good first step into understanding the regulation and regulatory networks of miRNAs by rule mining methods applied on miRNA expression profiles, there is still much left to be discovered. The advent of next generation sequencing (NGS) technologies makes it possible to get a comprehensive miRNA landscape for data analysis. The next important step in the identification of regulatory networks and new pathways involving miRNAs is a high throughput analysis method for analysing NGS data, which brings greater understanding of the mechanisms of diseases, leading to rational drug design. Consequently, further research is needed to achieve the objectives as described in the following sections:

miRNAs' Systematic Function Analysis Using High Throughput Sequencing Data Further research may be conducted to investigate the systematic function of miRNAs by using high throughput sequencing data, including next-generation sequencing data (e.g., RNA-seq). This study will offer several advantages. First, next-generation sequencing platforms have produced huge amounts of sequence data, and sequencing data is more accurate than array-based methods for determining miRNA expression

levels. Second, potential novel miRNAs can be detected by using various computational methods for characterising miRNAs. Third, sequencing data can be used to identify the miRNAs' systematic function with high accuracy.

miRNA Study Based on Multi-Layer Hierarchical MapReduce Framework Further research could be conducted in order to process more than one million miRNA sequences in acceptable time by using the multi-layer hierarchical MapReduce framework, which can gather computational resources from different clusters and run MapReduce jobs across them. In detail, MapReduce is a programming model well suited to processing large data sets using high-throughput parallelism running on a large number of computational resources (Dean & Ghemawat 2008).

A MapReduce job divides a large data set into independent chunks and organises them into key and value pairs for parallel processing. A key-value pair is a set of two linked data items: a key, which is a unique identifier for some items of data, and the value, which is either the data that is identified or a pointer to the location of that data. The mapping and reducing functions receive not just values, but (key, value) pairs. This parallel processing improves the speed and reliability of the cluster, returning solutions more quickly and with greater reliability. Every MapReduce job consists of at-least three parts: The driver, Mapper and Reducer.

The key feature of the MapReduce framework is the parallelism of the analysis process, so that the execution time for a single miRNA can be accelerated as desired by allocating more resources.

Mapping Phase The first phase of a MapReduce program is called mapping. A list of data elements are provided, one at a time, to a function called the Mapper, which transforms each element individually to an output data element. The Map function divides the input into ranges by the InputFormat and creates a map task for each range in the input. The JobTracker distributes those tasks to the worker nodes. The output of each map task is partitioned into a group of key-value pairs for each reduction.

Reducing Phase Reducing lets you aggregate values together. A reducer function receives an iterator of input values from an input list. It then combines these values together, returning a single output value. The reducer function then collects the various results and combines them to answer the larger problem that the master node needs to solve. Each reduce pulls the relevant partition from the machines where the maps are executed, and then writes its output back into HDFS. Thus, the reducer function is able to collect the data from all of the maps for the keys and combine them to solve the problem.

8.3 Closing Summary

The scope of this research was exclusively aimed and strongly focused on rule mining on miRNA expression profiles for human disease understanding. A large segment of this research concentrated on studying the miRNA expression profiles for important roles in human disease. This research was instigated to contribute more knowledge to computational biology in general and to the miRNAs' functions in particular. Since my first introduction to rule mining and miRNAs, I have always had a deep interest in the discipline of rule mining on miRNA study.

Appendix A

Appendix: Long Table

Table A.1: Categorisation of miRNA target prediction tools (Categorisation was taken from a survey of computational algorithms for miRNA target prediction)

Database	Regions scanned	Method	Species	Implementation	Reference
TargetScan	8&7mer sites, reading frames	Rule-based	Human, mouse, rat, dog and chicken	PerlScript	(Agarwal et al. 2015)
miRDB	3'-UTR,CDS, 5'-UTR	Data-driven	Human, mouse, rat, dog and chicken	Web-driven	(Wong & Wang 2014, Wang 2016)
PicTar	3' region	Data-driven	Vertebrate, mouse, flies and nematode	Web-driven	(Krek et al. 2005)
TargetScanS	3' region	Rule-based	Human, mouse, rat, dog and chicken	Web-driven	(Lewis et al. 2005*b*)
miRanda	3' region	Rule-based	Human, mouse, rat, fruit fly and nematode	Predictions available	(Betel et al. 2008)
RNAHybrid	3' region	Rule-based	Any	Prediction available	(Rehmsmeier et al. 2004)
miRNAMap	3' region	Rule-based	12 species	Web-driven	(Griffiths-Jones et al. 2006)
DIANA-microT	3' region	Rule-based	Human and mouse	Web-driven	(Maragkakis et al. 2009)
PITA	3' region and CDs	Rule-based	Human, mouse, worm and fly	PerlScript	(Kertesz et al. 2007)
GenMiR++	3' region	GE data	Any	code	(Huang et al. 2007)
RNA22	Unspecific	Data-driven	Human	PerlScript	(Miranda et al. 2006)
SVMicro	3'-UTR	Data-driven	Human	PerlScript	(Liu, Yue, Chen, Gao & Huang 2010)
TargetSpy	3' region	Data-driven	Human, mouse, rat, chicken and flies	Web-driven	(Sturm et al. 2010)
NBmiRTar	3' region	Data-driven	Any	Web-driven application	(Yousef et al. 2007)
MirMap	3' region	Combinatory	8 species	Web-driven	(Vejnar & Zdobnov 2012)
miRTarPRi	3' region	Data-driven	Human	Web-driven	(Wang et al. 2013)

Appendix B

Appendix: Algorithm of Prim Code

Algorithm B.1 Algorithm of constructing a committee of decision trees

1: Input: data.txt, a 71x329 relational table.

2: Output: DTc, a committee of 2- or 3- miRNAs decision trees with 100% accuracy; nrule, the number of selected decision trees; myoutput.txt, all the contents printed in the screen; mygraphs.pdf, pictures of all the selected decision trees.

```
sink("myoutput.txt",append=TRUE, split=TRUE); ♯ save the contents of the screen
pdf("mygraphs.pdf"); ♯ save the output pictures
rm(list=ls());
library(RWeka); ♯load RWeka
library(stringr); ♯load String
library(gplots); ♯load Plot
d<-read.table("data.txt",sep="‵"); ♯ load data from a file
g<-GainRatioAttributeEval(V329 .,data=d); ♯ rank the miRNAs by gain ratio
t=19; ♯ choose the top-ranked 19 miRNAs after mapping the 5 plasma biomarkers
i<-order(g,decreasing=T)[1:t];
i<-c(i,length(d)) ♯ add the column of class label to the new dataset
n<-d[,i]; ♯ obtain a new dataset
nrule<-0; ♯ number of rules
q<-length(i)-1; ♯ the times of construction decisions
for (c in 1:q) ♯ the procedure continues until only two miRNAs are left
DTc<-J48(V329 .,data=n); ♯ use C4.5 to construct a decision tree
bc<-summary(DTc)$details["pctCorrect"][[1]]; ♯ the accuracy of the decision tree
♯ JUSTIFY THE ACCURACY OF THE DECISION TREE
if(bc==100) ♯ if the accuracy equals 100%, then print and draw the decision tree
♯ JUSTIFY THE NUMBER OF NODES IN THE DECISION TREE
str<-DTc$classifier$toString(); ♯ the string structure of the decision tree
str1<-strsplit(str," <"); ♯ split the string, str1 is a list
str2<-unlist(str1); ♯ transfer a list to a character
l<-length(str2); ♯ the length of the character
id<-seq(1,328,1); id=0;
for (i in 1:(l-1))
st<-str2[i];
ll<-nchar(st,type="chars",allowNA=FALSE); ♯ the length of a character
nst<-substr(st,ll-3,ll); ♯ choose the last four characters
nst1<-strsplit(nst,"V"); ♯ separate the character in V
num<-nst1[[1]][2];
num<-as.numeric(num);
id[i]=num; ♯ the ID of node in the decision tree
node<-length(unique(id)); ♯ the number of nodes in the decision tree
if (node<4) plot(DTc);
nrule<-nrule+1;
dtcstr<-DTc$classifier$toString(); ♯ select the root of a decision tree
s1<-strsplit(dtcstr,"J48 pruned tree");
s2<-strsplit(s1[[1]][2]," <",);
s3<-s2[[1]][1];
if(is.na(s3)=="TRUE") n[,-1];
else
n[,eval(s3)]<-NULL; ♯ remove the column of the root
print(sprintf('The number of selected decision trees is %d', nrule));
sink();
dev.off();
```

Appendix C

Appendix: List of Symbols

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

$miRNA$	microRNA
$mRNA$	Messenger RNA
TF	Transcription Factor
HCV	Hepatitis C Virus
$UTRs$	Untranslated Regions
WHO	World Health Organization
$NSCLC$	Non-Small Cell Lung Cancer
IFN	Interferon
$NS3$	Non-structural Protein 3
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myelogenous Leukemia
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myelogenous Leukemia

SCC	Squamous Cell Carcinoma
GO	Gene Ontology
$KEGG$	Kyoto Encyclopedia of Genes and Genomes
$pri-miRNA$	Primary miRNA
Pol	Polymerase
$pre-miRNA$	Precursor miRNA
$qRT-PCR$	Quantitative Reverse Transcription-Polymerase Chain Reaction
PCR	Polymerase Chain Reaction
$ORFs$	Open Reading Frames
WC	Watson-Crick
SVM	Support Vector Machine
$TCGA$	The Cancer Genome Atlas
$PIMiM$	Protein Interaction-based miRNA Modules
$MBPLS$	Multi-Block Partial Least Squares
CNV	Copy Number Variation
$D-M$	DNA Methylation
GE	Gene Expression
ME	miRNA Expression
MFE	Minimal Free Energy
HMM	Hidden Markov Model

$FFLs$	Feed-Forward Loops
$FBLs$	Feed-Back Loops
GBM	Glioblastoma
AD	Alzheimer's Disease
DEG	Differentially Expressed Gene
GEO	Gene Expression Omnibus
$NCBI$	National Center for Biotechnology Information
KNN	K-Nearest Neighbour
NB	Naive Bayes
EuD	Euclidean Distance
MaD	Manhattan Distance
MiD	Minkowski Distance
HaD	Hamming Distance
$C4.5$	C4.5 decision tree
ROC	Receiver Operating Characteristic
AUC	Area Under ROC curves
DT	Decision Tree
TrS	Training Set
TeS	Testing Set
$OMIM$	Online Mendelian Inheritance in Man
$FZD3$	Frizzled Class Receptor 3

$RPS6KA3$	Ribosomal Protein S6 Kinase, 90kDa, Polypeptide 3
$AEBP2$	AE Binding Protein 2
$PARD6B$	Par-6 Family Cell Polarity Regulator Beta
$NKD1$	Naked Cuticle 1 Homolog
$MAP3K2$	Mitogen-Activated Protein Kinase Kinase Kinase 2
$RBMS2$	RNA Binding Motif, Single Stranded Interacting Protein 2
$EPB41$	Erythrocyte Membrane Protein Band 4.1
$AKAP13$	A-Kinase Anchoring Protein 13
$CSDC2$	Cold Shock Domain Containing C2, RNA Binding
$ACVR1C$	Activin A Receptor type IC
$RAB43$	RAB43, member RAS oncogene family
$FNDC5$	Fibronectin type III Domain Containing 5
$WDR33$	WD Repeat Domain 33
$ALDH4A1$	Aldehyde Dehydrogenase 4 Family member A1
$ANKRD12$	Ankyrin Repeat Domain 12
$KCTD9$	Potassium Channel Tetramerization Domain Containing 9
$ARMC1$	Armadillo Repeat Containing 1
$DICER1$	Dicer 1 ribonuclease III
$ASB16$	Ankyrin Repeat and SOCS Box Containing 16
$GALNTL4$	polypeptide N-acetylgalactosaminyltransferase 18

$ADRA1D$	Adrenoceptor Alpha 1D
$BNC2$	Basonuclin 2
$PDLIM2$	PDZ and LIM domain 2
$SPCS2$	Signal Peptidase Complex Subunit 2
$ACLY$	ATP Citrate Lyase
$C6orf192$	Chromosome 6 Open Reading Frame 192
$ING4$	Inhibitor of Growth Family member 4
$DNAJA3$	DnaJ heat shock protein family (Hsp40) member A3
$FAM120A$	Family with Sequence Similarity 120A
$GLG2$	Glycogenin 2
$SHOC2$	SHOC2 leucine-rich repeat scaffold protein
$CBLB$	Cbl proto-oncogene B, E3 ubiquitin protein ligase
$NMD3$	NMD3 ribosome export adaptor
$OCRL$	Oculocerebrorenal syndrome of Lowe
$COMT$	Catechol-O-methyltransferase
$BRD3$	Bromodomain containing 3
$DENND2C$	DENN/MADD domain containing 2C
$AUTS2$	Autism Susceptibility candidate 2
$PCID2$	PCI Domain containing 2
$GFRA2$	GDNF Family Receptor Alpha 2
QKI	QKI, KH domain containing, RNA binding

$MAP2$	Microtubule Associated Protein 2
$FRMPD4$	FERM and PDZ Domain containing 4
$CAMK2D$	Calcium/Calmodulin-dependent protein kinase II delta
$CDK6$	Cyclin-Dependent Kinase 6
GPI	Glycosylphosphatidylinositol
$GNDF$	Glial Cell Line-Derived Neurotrophic Factor
NTN	neurturin
$ZNF718$	Zinc Finger protein 718
$SNX27$	Sorting Nexin family member 27
$ERO1L$	ERO1-like
$ZNF558$	Zinc Finger protein 558
$ENPP1$	Ectonucleotide Pyrophosphatase/Phosphodiesterase 1
$KIAA1804$	Mixed Lineage Kinase 4
$GFRA1$	GDNF family receptor alpha 1
$FAM73B$	Family with sequence similarity 73 member B
$SMG5$	SMG5 nonsense mediated mRNA decay factor
$EFNA3$	Ephrin-A3
$ZNF462$	Zinc Finger protein 462
$TCF4$	Transcription Factor 4
KIT	KIT proto-oncogene receptor tyrosine kinase
$CHD2$	Chromodomain Helicase DNA binding protein 2

$FAM118A$	Family with sequence similarity 118 member A
$KCTD9$	Potassium Channel Tetramerization Domain containing 9
$PISD$	Phosphatidylserine Decarboxylase
$SAMD4A$	Sterile Alpha Motif Domain containing 4A
$STAM2$	Signal Transducing Adaptor Molecule 2
$PRPF4B$	Pre-mRNA Processing Factor 4B
$PPP1R12B$	Protein Phosphatase 1 Regulatory Subunit 12B
$EPM2AIP1$	EPM2A (laforin) Interacting Protein 1
$CDKN1B$	Cyclin-Dependent Kinase Inhibitor 1B (p27, Kip1)
$AK3$	Adenylate Kinase 3
$ADAMTS5$	ADAM Metallopeptidase with Thrombospondin type 1 motif 5
$OAZ2$	Ornithine Decarboxylase Antizyme 2
$CACNA2D2$	Calcium Channel, voltage-dependent, alpha 2/delta subunit 2
$HDAC4$	Histone Deacetylase 4
$FGD4$	FYVE, RhoGEF and PH domain containing 4
$NUP43$	Nucleoporin 43kDa
$NMD3$	Ribosome-binding protein NMD3
$PCID2$	PCI domain containing 2
$AUTS2$	Autism susceptibility candidate 2

$ANKRD12$	Ankyrin repeat domain 12
$BRD3$	Bromodomain containing 3
$ABCC5$	ATP binding cassette subfamily C member 5
$DENND2C$	DENN/MADD domain containing 2C
$OCRL$	Oculocerebrorenal syndrome of Lowe
$EPB41L5$	Erythrocyte membrane protein band 4.1 like 5
$CPLX2$	Complexin 2
$GLG1$	Golgi Glycoprotein 1
$ING4$	Inhibitor of growth family member 4
$ASXL1$	Additional sex combs like 1, transcriptional regulator
$NKTR$	Natural Killer Cell Triggering Receptor
$SPCS2$	Signal Peptidase Complex Subunit 2
$PDLIM2$	PDZ and LIM domain 2 (mystique)
$CBX5$	Chromobox 5
PCC	Pearson's Correlation Coefficient
PPI	Protein Protein Interaction
$BCL2$	B-cell CLL/lymphoma 2
$TP63$	Tumor Protein p63
$PRKD2$	Protein Kinase D2
$FOXP1$	Forkhead box P1
$TIPARP$	TCDD-inducible poly(ADP-ribose) polymerase

$TSHZ3$	Teashirt zinc finger homeobox 3
$PKD2$	Polycystic Kidney Disease-2
$MYLK$	Myosin light chain kinase
$SMAD2$	SMAD family member 2
$ACVR1B$	Activin A receptor type IB
$ACVR2A$	Activin A receptor type IIA
$ACVR2B$	Activin A receptor type IIB
MYC	V-myc avian myelocytomatosis viral oncogene homolog
$EIF2C1$	Eukaryotic Initiation Factor 2C1
$HIPK2$	Homeodomain Interacting Protein Kinase 2
$TNRC6B$	Trinucleotide Repeat Containing 6B
$LRP6$	LDL Receptor Related Protein 6
$PTEN$	Phosphatase and tensin homolog
$TNRC6A$	Trinucleotide repeat containing 6A
$VEGFA$	Vascular endothelial growth factor A
$EIF2C4$	Eukaryotic Initiation Factor 2C4
$FXR1$	Fragile X mental retardation, autosomal homolog 1
$CDKN1C$	Cyclin-Dependent Kinase inhibitor 1C (p57, Kip2)
$TSC1$	Tuberous Sclerosis 1
$ERBB4$	Erb-b2 Receptor tyrosine kinase 4
$TMP3$	Tropomyosin 3

$CDX2$	Caudal type homeobox 2
TSS	Transcription Start Site
BED	Browser Extensible Data
GFF	General Feature Format
SAM	Sequence Alignment/Map
$TFBS$	Transcription Factor Binding Site
$IGF2$	Insulin-like Growth Factor 2
FFL	Feed Forward Loop
$SRmiR$	Self-regulation miRNA
$BACH1$	BTB and CNC homology 1,basic leucine zipper transcription factor 1
$BRCA1$	Breast Cancer 1
$CTBP2$	C-terminal Binding Protein 2
$EBF1$	Early B-cell Factor 1
$HDAC2$	Histone Deacetylase 2
$HNF4G$	Hepatocyte Nuclear Factor 4 Gamma
$IRF1$	Interferon Regulatory Factor 1
$MEF2C$	Myocyte Enhancer Factor 2C
$MTA3$	Metastasis Associated 1 family member 3
$NFIC$	Nuclear Factor I/C (CCAAT-binding transcription factor)
$SMC3$	Structural Maintenance of Chromosomes 3

$TAL1$	T-cell Acute Lymphocytic leukemia 1
$TCF7L2$	Transcription Factor 7 Like 2

Bibliography

Abdalla, M. A. & Haj-Ahmad, Y. (2012), 'Promising candidate urinary microrna biomarkers for the early detection of hepatocellular carcinoma among high-risk hepatitis c virus egyptian patients', *Journal of Cancer* **3**, 19.

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

Ai, J., Zhang, R., Li, Y., Pu, J., Lu, Y., Jiao, J., Li, K., Yu, B., Li, Z., Wang, R. et al. (2010), 'Circulating microRNA-1 as a potential novel biomarker for acute myocardial infarction', *Biochemical and Biophysical Research Communications* **391**(1), 73–77.

Airaksinen, M. S. & Saarma, M. (2002), 'The gdnf family: signalling, biological functions and therapeutic value', *Nature Reviews Neuroscience* **3**(5), 383–394.

Alevizos, I., Alexander, S., Turner, R. J. & Illei, G. G. (2011), 'MicroRNA expression profiles as biomarkers of minor salivary gland inflammation and dysfunction in Sjögren's syndrome', *Arthritis Rheum* **63**(2), 535–544.

Alexiou, P., Maragkakis, M., Papadopoulos, G. L., Simmosis, V. A., Zhang, L. & Hatzigeorgiou, A. G. (2010), 'The DIANA-mirExTra web server: from gene expression data to microRNA function', *PloS one* **5**(2), e9171.

Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M. J., Tuschl, T. & Margalit, H. (2005), 'Clustering and conservation patterns of human microRNAs', *Nucleic Acids Research* **33**(8), 2697–2706.

Amar, D., Safer, H. & Shamir, R. (2013), 'Dissection of regulatory networks that are altered in disease via differential co-expression'.

An, J., Choi, K. P., Wells, C. A. & Chen, Y.-P. P. (2010), 'Identifying co-regulating microRNA groups', *Journal of Bioinformatics and Computational Biology* **8**(01), 99–115.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al. (2000), 'Gene Ontology: tool for the unification of biology', *Nature Genetics* **25**(1), 25–29.

Au, N., Gown, A., Cheang, M., Huntsman, D., Yorida, E., Elliott, W., Flint, J., English, J., Gilks, C. & Grimes, H. (2004), 'P63 expression in lung carcinoma: a tissue microarray study of 408 cases', *Applied Immunohistochemistry & Molecular Morphology* **12**(3), 240–247.

Bai, Y.-Q., Miyake, S., Iwai, T. & Yuasa, Y. (2003), 'CDX2, a homeobox transcription factor, upregulates transcription of the p21/WAF1/CIP1 gene', *Oncogene* **22**(39), 7942–7949.

Bartel, D. P. (2004*a*), 'Microrna: genomics, biogenesis, mechanism, and function', *Cell* **116**(2), 281–297.

Bartel, D. P. (2004*b*), 'Microrna: genomics, biogenesis, mechanism, and function', *Cell* **116**(2), 281–297.

Bartels, C. L. & Tsongalis, G. J. (2009), 'MicroRNAs: novel biomarkers for human cancer', *Clin Chem* **55**(4), 623–631.

Barzel, B. & Barabási, A.-L. (2013), 'Network link prediction by global silencing of indirect correlations', *Nature Biotechnology* .

Baumjohann, D. & Ansel, K. M. (2013), 'MicroRNA-mediated regulation of T helper cell differentiation and plasticity', *Nature Reviews Immunology* **13**(9), 666–678.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R. et al. (2008), 'Accurate whole human genome sequencing using reversible terminator chemistry', *nature* **456**(7218), 53–59.

Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. (2001), 'Role for a bidentate ribonuclease in the initiation step of RNA interference', *Nature* **409**(6818), 363–366.

Betel, D., Wilson, M., Gabow, A., Marks, D. S. & Sander, C. (2008), 'The microRNA. org resource: targets and expression', *Nucleic Acids Research* **36**(suppl 1), D149–D153.

Bian, Y., Wang, L., Lu, H., Yang, G., Zhang, Z., Fu, H., Lu, X., Wei, M., Sun, J. & Zhao, Q. (2012), 'Downregulation of tumor suppressor QKI in gastric cancer and its implication in cancer prognosis', *Biochemical and Biophysical Research Communications* **422**(1), 187–193.

Black, E. R., Falzon, L. & Aronson, N. (2012), 'Gene expression profiling for predicting outcomes in stage II colon cancer'.

Bonhomme, C., Duluc, I., Martin, E., Chawengsaksophak, K., Chenard, M., Kedinger, M., Beck, F., Freund, J. & Domon-Dell, C. (2003), 'The Cdx2 homeobox gene has a tumour suppressor function in the distal colon in addition to a homeotic role during gut development', *Gut* **52**(10), 1465–1471.

Boross, G., Orosz, K. & Farkas, I. J. (2009), 'Human microRNAs co-silence in well-separated groups and have different predicted essentialities', *Bioinformatics* **25**(8), 1063–1069.

Brass, A. L., Huang, I., Benita, Y., John, S. P., Krishnan, M. N., Feeley, E. M., Ryan, B. J., Weyer, J. L., van der Weyden, L. & Fikrig, E. (2009), 'The ifitm proteins mediate cellular resistance to influenza a h1n1 virus, west nile virus, and dengue virus', *Cell* **139**(7), 1243–1254.

Breu, H., Gil, J., Kirkpatrick, D. & Werman, M. (1995), 'Linear time Euclidean distance transform algorithms', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **17**(5), 529–533.

Brin, S., Motwani, R., Ullman, J. D. & Tsur, S. (1997), Dynamic itemset counting and implication rules for market basket data, *in* 'ACM SIGMOD Record', Vol. 26, ACM, pp. 255–264.

Buj-Bello, A., Adu, J., Pinon, L., Horton, A., Thompson, J., Rosenthal, A., Chinchetru, M., Buchman, V. L. & Davies, A. M. (1997), 'Neurturin responsiveness requires a gpi-linked receptor and the ret receptor tyrosine kinase', *Nature* **387**(6634), 721.

Calin, G. A. & Croce, C. M. (2006), 'MicroRNA signatures in human cancers', *Nat Rev Cancer* **6**(11), 857–866.

Calin, G. A., Dumitru, C. D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K. et al. (2002), 'Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia', *Proceedings of the National Academy of Sciences* **99**(24), 15524–15529.

Calin, G. A., Sevignani, C., Dumitru, C. D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F. & Negrini, M. (2004), 'Human microRNA genes are frequently located at fragile sites

and genomic regions involved in cancers', *Proc Natl Acad Sci U S A* **101**(9), 2999–3004.

Carrington, J. C. & Ambros, V. (2003), 'Role of microRNAs in plant and animal development', *Science* **301**(5631), 336–338.

Cermelli, S., Ruggieri, A., Marrero, J. A., Ioannou, G. N. & Beretta, L. (2011), 'Circulating micrornas in patients with chronic hepatitis c and non-alcoholic fatty liver disease', *PLoS One* **6**(8), e23937.

Chekulaeva, M. & Filipowicz, W. (2009), 'Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells', *Current opinion in cell biology* **21**(3), 452–460.

Chen, C.-Y., Chen, S.-T., Fuh, C.-S., Juan, H.-F. & Huang, H.-C. (2011), 'Coregulation of transcription factors and microRNAs in human transcriptional regulatory network', *BMC bioinformatics* **12**(Suppl 1), S41.

Chen, K. & Rajewsky, N. (2007), 'The evolution of gene regulation by transcription factors and microRNAs', *Nature Reviews Genetics* **8**(2), 93–103.

Chen, T., Zhu, L., Zhou, Y., Pi, B., Liu, X., Deng, G., Zhang, R., Wang, Y., Wu, Z. & Han, M. (2013), 'Kctd9 contributes to liver injury through nk cell activation during hepatitis b virus-induced acute-on-chronic liver failure', *Clinical Immunology* **146**(3), 207–216.

Chen, X., Ba, Y., Ma, L., Cai, X., Yin, Y., Wang, K., Guo, J., Zhang, Y., Chen, J., Guo, X. et al. (2008), 'Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases', *Cell Research* **18**(10), 997–1006.

Cheng, C., Li, L. M. et al. (2008), 'Inferring microRNA activities by combining gene expression with microRNA target prediction', *PloS one* **3**(4), e1989.

Chou, C.-H., Chang, N.-W., Shrestha, S., Hsu, S.-D., Lin, Y.-L., Lee, W.-H., Yang, C.-D., Hong, H.-C., Wei, T.-Y., Tu, S.-J. et al. (2015), 'miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database', *Nucleic acids research* p. gkv1258.

Chung, K. L. (1967), *Markov chains*, Springer.

Cimmino, A., Calin, G. A., Fabbri, M., Iorio, M. V., Ferracin, M., Shimizu, M., Wojcik, S. E., Aqeilan, R. I., Zupo, S., Dono, M. et al. (2005), 'miR-15 and miR-16 induce apoptosis by targeting BCL2', *Proceedings of the National Academy of Sciences of the United States of America* **102**(39), 13944–13949.

Clark, P. J. (2012), Translational genomics, transcriptomics and metabolomics analyses of the metabolic effects of chronic hepatitis C infection and their clinical implications, PhD thesis, The University of New South Wales.

Consortium, E. P. et al. (2004), 'The ENCODE (ENCyclopedia of DNA elements) project', *Science* **306**(5696), 636–640.

Cui, Q., Yu, Z., Purisima, E. O. & Wang, E. (2006), 'Principles of microRNA regulation of a human cellular signaling network', *Molecular systems biology* **2**(1).

Dahiya, N., Sherman-Baust, C. A., Wang, T.-L., Davidson, B., Shih, I.-M., Zhang, Y., Wood III, W., Becker, K. G. & Morin, P. J. (2008), 'MicroRNA expression and identification of putative miRNA targets in ovarian cancer', *PloS One* **3**(6), e2436.

Dean, J. & Ghemawat, S. (2008), 'MapReduce: simplified data processing on large clusters', *Communications of the ACM* **51**(1), 107–113.

Dentelli, P., Traversa, M., Rosso, A., Togliatto, G., Olgasi, C., Marchiò, C., Provero, P., Lembo, A., Bon, G., Annaratone, L. et al. (2014), 'miR-

221/222 control luminal breast cancer tumor progression by regulating different targets', *Cell Cycle* **13**(11), 0–1.

Dews, M., Homayouni, A., Yu, D., Murphy, D., Sevignani, C., Wentzel, E., Furth, E. E., Lee, W. M., Enders, G. H., Mendell, J. T. et al. (2006), 'Augmentation of tumor angiogenesis by a Myc-activated microRNA cluster', *Nature Genetics* **38**(9), 1060–1065.

Dimri, G. P., Lee, X., Basile, G., Acosta, M., Scott, G., Roskelley, C., Medrano, E. E., Linskens, M., Rubelj, I. & Pereira-Smith, O. (1995), 'A biomarker that identifies senescent human cells in culture and in aging skin in vivo', *Proceedings of the National Academy of Sciences* **92**(20), 9363–9367.

Do Youn Park, A. S., Kim, G. H., Mino-Kenudson, M., Deshpande, V., Zukerberg, L. R., Am Song, G., Lauwers, G. Y. et al. (2009), 'CDX2 expression in the intestinal-type gastric epithelial neoplasia: frequency and significance', *Modern Pathology* **23**(1), 54–61.

Dostie, J., Mourelatos, Z., Yang, M., Sharma, A. & Dreyfuss, G. (2003), 'Numerous microRNPs in neuronal cells containing novel microRNAs', *RNA* **9**(2), 180–186.

Dreyfuss, G., Kim, V. N. & Kataoka, N. (2002), 'Messenger-RNA-binding proteins and the messages they carry', *Nature Reviews Molecular Cell Biology* **3**(3), 195–205.

Dweep, H. & Gretz, N. (2015), 'miRWalk2. 0: a comprehensive atlas of microRNA-target interactions', *Nature methods* **12**(8), 697–697.

Dweep, H., Sticht, C., Pandey, P. & Gretz, N. (2011), 'miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes', *Journal of biomedical informatics* **44**(5), 839–847.

Easow, G., Teleman, A. A. & Cohen, S. M. (2007), 'Isolation of microRNA targets by miRNP immunopurification', *RNA* **13**(8), 1198–1204.

Edmonston, T. B., Kushnir, M., Aharonov, R., Yanai, G. L., Benjamin, H., Bibbo, M., Thurm, C., Horowitz, L., Huang, Y., Gilad, S. et al. (2010), 'microRNAs as clinical biomarkers for lung cancer classification', *Proceedings of the American Association for Cancer Research* **2010**(1_Molecular_Diagnostics_Meeting), B8.

Ekman, S., Wynes, M. W. & Hirsch, F. R. (2012), 'The mTOR pathway in lung cancer and implications for therapy and biomarker analysis', *Journal of Thoracic Oncology* **7**(6), 947–953.

Enerly, E., Steinfeld, I., Kleivi, K., Leivonen, S.-K., Aure, M. R., Russnes, H. G., Rønneberg, J. A., Johnsen, H., Navon, R., Rødland, E. et al. (2011), 'miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors', *PloS One* **6**(2), e16915.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., Marks, D. S. et al. (2004), 'MicroRNA targets in Drosophila', *Genome Biology* **5**(1), R1–R1.

Fábregas, B. C., de Miranda, A. S., Barbosa, I. G., Moura, A. S., Carmo, R. A. & Teixeira, A. L. (2012), 'Brain-derived neurotrophic factor in patients with chronic hepatitis c: beyond neurotrophic support', *Biological Psychiatry* **72**(4), e13–e14.

Fazi, F., Rosa, A., Fatica, A., Gelmetti, V., De Marchis, M. L., Nervi, C. & Bozzoni, I. (2005), 'A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPα regulates human granulopoiesis', *Cell* **123**(5), 819–831.

Feizi, S., Marbach, D., Médard, M. & Kellis, M. (2013), 'Network deconvolution as a general method to distinguish direct dependencies in networks', *Nature Biotechnology* .

Filipowicz, W., Bhattacharyya, S. N. & Sonenberg, N. (2008), 'Mechanisms of post-transcriptional regulation by microrna: are the answers in sight?', *Nature Reviews Genetics* **9**(2), 102–114.

Fontana, R., Sanderson, D., Taylor, W., Woolner, L., Miller, W., Muhm, J. & Uhlenhopp, M. (1984), 'Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Mayo Clinic study', *Am J Respir Crit Care Med* **130**(4), 561.

Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000), 'Using Bayesian networks to analyze expression data', *Journal of Computational Biology* **7**(3-4), 601–620.

Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. (2009*a*), 'Most mammalian mRNAs are conserved targets of microRNAs', *Genome Research* **19**(1), 92–105.

Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. (2009*b*), 'Most mammalian mRNAs are conserved targets of microRNAs.', *Genome Research,* **19**(1), 92–105.

Frost, J., Ball Jr, W., Levin, M., Tockman, M., Baker, R., Carter, D., Eggleston, J., Erozan, Y., Gupta, P. & Khouri, N. (1984), 'Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Johns Hopkins study', *Am J Respir Crit Care Med* **130**(4), 549.

Fukuda, M. (2011), 'Tbc proteins: Gaps for mammalian small gtpase rab?', *Bioscience Reports* **31**, 159–168.

Fumarola, C., Bonelli, M. A., Petronini, P. G. & Alfieri, R. R. (2014), 'Targeting PI3K/AKT/mTOR pathway in non small cell lung cancer', *Biochemical pharmacology* **90**(3), 197–207.

Garofalo, M., Quintavalle, C., Romano, G., Croce, C. & Condorelli, G. (2012), 'miR221/222 in cancer: their role in tumor progression and response to therapy', *Current Molecular Medicine* **12**(1), 27.

Gilad, S., Meiri, E., Yogev, Y., Benjamin, S., Lebanony, D., Yerushalmi, N., Benjamin, H., Kushnir, M., Cholakh, H. & Melamed, N. (2008), 'Serum microRNAs are promising novel biomarkers', *PLoS One* **3**(9), e3148.

Girard, L., Zchbauer-Mller, S., Virmani, A. K., Gazdar, A. F. & Minna, J. D. (2000), 'Genome-wide allelotyping of lung cancer identifies new regions of allelic loss, differences between small cell lung cancer and non-small cell lung cancer, and loci clustering', *Cancer Res* **60**(17), 4894–4906.

Gong, L., Kakrana, A., Arikit, S., Meyers, B. C. & Wendel, J. F. (2013), 'Composition and expression of conserved microRNA genes in diploid cotton (Gossypium) species', *Genome Biology and Evolution* **5**(12), 2449–2459.

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

Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A. & Enright, A. J. (2006), 'miRBase: microRNA sequences, targets and gene nomenclature', *Nucleic Acids Research* **34**(suppl 1), D140–D144.

Griffiths-Jones, S. M. (2006), 'The microRNA sequence database', *Methods Mol Biol* **342**, 129–38.

Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. (2008), 'miRBase: tools for microRNA genomics', *Nucleic Acids Research* **36**(suppl 1), D154–D158.

Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P. & Bartel, D. P. (2007), 'MicroRNA targeting specificity in mammals: determinants beyond seed pairing', *Molecular Cell* **27**(1), 91–105.

Gross, I., Duluc, I., Benameur, T., Calon, A., Martin, E., Brabletz, T., Kedinger, M., Domon-Dell, C. & Freund, J. (2007), 'The intestine-specific homeobox gene Cdx2 decreases mobility and antagonizes dissemination of colon cancer cells', *Oncogene* **27**(1), 107–115.

Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. (2010), 'Mammalian microRNAs predominantly act to decrease target mRNA levels', *Nature* **466**(7308), 835–840.

Guo, H.-S., Xie, Q., Fei, J.-F. & Chua, N.-H. (2005), 'MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for Arabidopsis lateral root development', *The Plant Cell Online* **17**(5), 1376–1386.

Guo, L., Zhao, Y., Yang, S., Zhang, H. & Chen, F. (2014), 'Integrative Analysis of miRNA-mRNA and miRNA-miRNA Interactions', *BioMed Research International* **2014**.

Guo, Z., Maki, M., Ding, R., Yang, Y., Xiong, L. et al. (2014), 'Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues', *Scientific Reports* **4**.

Habbe, N., Koorstra, J.-B. M., Mendell, J. T., Offerhaus, G. J., Ryu, J. K., Feldmann, G., Mullendore, M. E., Goggins, M. G., Hong, S.-M. & Maitra, A. (2009), 'MicroRNA miR-155 is a biomarker of early pancreatic neoplasia', *Cancer Biology & Therapy* **8**(4), 340–346.

Han, J. & Kamber, M. (2006*a*), *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, California.

Han, J. & Kamber, M. (2006*b*), *Data mining: concepts and techniques*, Morgan Kaufmann.

Hashimoto, Y., Akiyama, Y. & Yuasa, Y. (2013), 'Multiple-to-multiple relationships between microRNAs and target genes in gastric cancer', *PloS One* **8**(5), e62589.

He, L. & Hannon, G. J. (2004), 'MicroRNAs: small RNAs with a big role in gene regulation', *Nature Reviews Genetics* **5**(7), 522–531.

He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S. W., Hannon, G. J. et al. (2005), 'A microRNA polycistron as a potential human oncogene', *Nature* **435**(7043), 828–833.

He, Y., Tan, S.-L., Tareen, S. U., Vijaysri, S., Langland, J. O., Jacobs, B. L. & Katze, M. G. (2001), 'Regulation of mrna translation and cellular signaling by hepatitis c virus nonstructural protein ns5a', *Journal of Virology* **75**(11), 5090–5098.

Heneghan, H., Miller, N., Lowery, A., Sweeney, K. & Kerin, M. (2009), 'MicroRNAs as novel biomarkers for breast cancer', *Journal of oncology* **2010**, 1–7.

Hennessey, P. T., Sanford, T., Choudhary, A., Mydlarz, W. W., Brown, D., Adai, A. T., Ochs, M. F., Ahrendt, S. A., Mambo, E. & Califano, J. A. (2012), 'Serum microRNA biomarkers for detection of non-small cell lung cancer', *PloS One* **7**(2), e32307.

Ho, T. K. (1995), Random decision forests, *in* 'Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on', Vol. 1, pp. 278–282 vol.1.

Hobert, O. (2008*a*), 'Gene regulation by transcription factors and microRNAs', *Science* **319**(5871), 1785–1786.

Hobert, O. (2008*b*), 'Gene regulation by transcription factors and microRNAs', *Science* **319**(5871), 1785–1786.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. & Schuster, P. (1994), 'Fast folding and comparison of RNA secondary structures', *Monatshefte für Chemie/Chemical Monthly* **125**(2), 167–188.

Hsu, C.-W., Juan, H.-F. & Huang, H.-C. (2008), 'Characterization of microRNA-regulated protein-protein interaction network', *Proteomics* **8**(10), 1975–1979.

Hsu, P. W., Huang, H.-D., Hsu, S.-D., Lin, L.-Z., Tsou, A.-P., Tseng, C.-P., Stadler, P. F., Washietl, S. & Hofacker, I. L. (2006), 'miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes', *Nucleic Acids Res* **34**(suppl 1), D135–D139.

Hsu, S.-D., Tseng, Y.-T., Shrestha, S., Lin, Y.-L., Khaleel, A., Chou, C.-H., Chu, C.-F., Huang, H.-Y., Lin, C.-M., Ho, S.-Y. et al. (2014), 'miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions', *Nucleic Acids Research* **42**(D1), D78–D85.

Huang, J. C., Babak, T., Corson, T. W., Chua, G., Khan, S., Gallie, B. L., Hughes, T. R., Blencowe, B. J., Frey, B. J. & Morris, Q. D. (2007), 'Using expression profiling data to identify human microRNA targets', *Nature methods* **4**(12), 1045–1049.

Iizuka, M., Ogawa, T., Enomoto, M., Motoyama, H., Yoshizato, K., Ikeda, K. & Kawada, N. (2012), 'Induction of microRNA-214-5p in human and rodent liver fibrosis', *Fibrogenesis Tissue Repair* **5**, 12.

Iorio, M. V., Ferracin, M., Liu, C.-G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M. et al. (2005), 'MicroRNA gene expression deregulation in human breast cancer', *Cancer Research* **65**(16), 7065–7070.

Iorio, M. V., Visone, R., Di Leva, G., Donati, V., Petrocca, F., Casalini, P., Taccioli, C., Volinia, S., Liu, C.-G., Alder, H. et al. (2007), 'MicroRNA signatures in human ovarian cancer', *Cancer Research* **67**(18), 8699–8707.

Issaq, H. J., Waybright, T. J. & Veenstra, T. D. (2011), 'Cancer biomarker discovery: opportunities and pitfalls in analytical methods', *Electrophoresis* **32**(9), 967–975.

Jakubowska, K., Naumnik, W., Niklińska, W. & Chyczewska, E. (2015), 'Clinical Significance of HMGB-1 and TGF-β Level in Serum and BALF of Advanced Non-Small Cell Lung Cancer', *Advances in Experimental Medicine and Biology* .

Jayaswal, V., Lutherborrow, M., Ma, D. & Yang, Y. (2011), 'Identification of microrna-mrna modules using microarray data', *BMC Genomics* **12**(1), 138.

Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., Smigal, C. & Thun, M. J. (2006), 'Cancer statistics, 2006', *CA Cancer J Clin* **56**(2), 106–130.

Jeon, H.-S. & Jen, J. (2010), 'TGF-β signaling and the role of inhibitory smads in non-small cell lung cancer', *Journal of Thoracic Oncology: official publication of the International Association for the Study of Lung Cancer* **5**(4), 417.

Jiang, W., Zhang, Y., Meng, F., Lian, B., Chen, X., Yu, X., Dai, E., Wang, S., Liu, X., Li, X. et al. (2013), 'Identification of active transcription factor and miRNA regulatory pathways in Alzheimers disease', *Bioinformatics* **29**(20), 2596–2602.

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

Johnson, S. M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K. L., Brown, D. & Slack, F. J. (2005), '*RAS* Is Regulated by the *let-7* MicroRNA Family', *Cell* **120**(5), 635–647.

Jopling, C. L., Yi, M., Lancaster, A. M., Lemon, S. M. & Sarnow, P. (2005), 'Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA', *Science* **309**(5740), 1577–1581.

Joung, J.-G., Hwang, K.-B., Nam, J.-W., Kim, S.-J. & Zhang, B.-T. (2007), 'Discovery of microRNA-mRNA modules via population-based probabilistic learning', *Bioinformatics* **23**(9), 1141–1147.

Judice, A. & Geetha, K. P. (2013), 'A Novel Assessment of Various Bio-Imaging Methods for Lung Tumor Detection and Treatment by using 4-D and 2-D CT Images', *International Journal of Biomedical Science: IJBS* **9**(2), 54.

Karginov, F. V. & Hannon, G. J. (2013), 'Remodeling of Ago2–mRNA interactions upon cellular stress reflects miRNA complementarity and correlates with altered translation rates', *Genes & development* **27**(14), 1624–1632.

Katayama, Y., Maeda, M., Miyaguchi, K., Nemoto, S., Yasen, M., Tanaka, S., Mizushima, H., Fukuoka, Y., Arii, S. & Tanaka, H. (2012), 'Identification of pathogenesis-related micrornas in hepatocellular carcinoma by expression profiling', *Oncology Letters* **4**(4), 817.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. (2007), 'The role of site accessibility in microRNA target recognition', *Nature genetics* **39**(10), 1278–1284.

Kim, V. N. & Nam, J.-W. (2006), 'Genomics of microRNA', *TRENDS in Genetics* **22**(3), 165–173.

Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z. & Hatzigeorgiou, A. (2004), 'A combined computational-experimental approach predicts human microRNA targets', *Genes & Development* **18**(10), 1165–1178.

Kong, Y. & Han, J.-H. (2005), 'MicroRNA: biological and computational perspective', *Genomics Proteomics Bioinformatics* **3**(2), 62–72.

Kosaka, N., Iguchi, H. & Ochiya, T. (2010), 'Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis', *Cancer Science* **101**(10), 2087–2092.

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M. et al. (2005), 'Combinatorial microRNA target predictions', *Nature Genetics* **37**(5), 495–500.

Krol, J., Sobczak, K., Wilczynska, U., Drath, M., Jasinska, A., Kaczynska, D. & Krzyzosiak, W. J. (2004), 'Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design', *Journal of Biological Chemistry* **279**(40), 42230–42239.

Lai, E. C., Wiel, C. & Rubin, G. M. (2004), 'Complementary miRNA pairs suggest a regulatory role for miRNA: miRNA duplexes', *RNA* **10**(2), 171–175.

Le Béchec, A., Portales-Casamar, E., Vetter, G., Moes, M., Zindy, P.-J., Saumet, A., Arenillas, D., Theillet, C., Wasserman, W. W., Lecellier, C.-H. et al. (2011), 'MIR@ NT@ N: a framework integrating transcription factors, microRNAs and their targets to identify sub-network motifs in a meta-regulation network model', *BMC bioinformatics* **12**(1), 67.

Le, H.-S. & Bar-Joseph, Z. (2013), 'Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation', *Bioinformatics* **29**(13), i89–i97.

Le, T. D., Liu, L., Liu, B., Tsykin, A., Goodall, G. J., Satou, K. & Li, J. (2013), 'Inferring microRNA and transcription factor regulatory networks in heterogeneous data', *BMC bioinformatics* **14**(1), 92.

Le, T. D., Liu, L., Tsykin, A., Goodall, G. J., Liu, B., Sun, B.-Y. & Li, J. (2013), 'Inferring microRNA–mRNA causal regulatory relationships from expression data', *Bioinformatics* **29**(6), 765–771.

Lee, R. C., Feinbaum, R. L. & Ambros, V. (1993), 'The C. elegans heterochronic gene¡ i¿ lin-4¡i¿ encodes small RNAs with antisense complementarity to¡ i¿ lin-14¡i¿', *Cell* **75**(5), 843–854.

Lee, S. A., Ho, C., Roy, R., Kosinski, C., Patil, M. A., Tward, A. D., Fridlyand, J. & Chen, X. (2008), 'Integration of genomic analysis and in vivo transfection to identify sprouty 2 as a candidate tumor suppressor in liver cancer', *Hepatology* **47**(4), 1200–1210.

Lee, Y., Jeon, K., Lee, J.-T., Kim, S. & Kim, V. N. (2002), 'MicroRNA maturation: stepwise processing and subcellular localization', *The EMBO Journal* **21**(17), 4663–4670.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H. & Kim, V. N. (2004), 'MicroRNA genes are transcribed by RNA polymerase II', *The EMBO Journal* **23**(20), 4051–4060.

Lee, Y., Samaco, R. C., Gatchel, J. R., Thaller, C., Orr, H. T. & Zoghbi, H. Y. (2008), 'miR-19, miR-101 and miR-130 co-regulate ATXN1 levels to potentially modulate SCA1 pathogenesis', *Nature Neuroscience* **11**(10), 1137–1139.

Lewis, B. P., Burge, C. B. & Bartel, D. P. (2005*a*), 'Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets', *Cell* **120**(1), 15–20.

Lewis, B. P., Burge, C. B. & Bartel, D. P. (2005*b*), 'Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets', *Cell* **120**(1), 15–20.

Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. (2003), 'Prediction of mammalian microRNA targets', *Cell* **115**(7), 787–798.

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H. & Yang, J.-H. (2014), '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, W., Zhang, S., Liu, C.-C. & Zhou, X. J. (2012), 'Identifying multi-layer gene regulatory modules from multi-dimensional genomic data', *Bioinformatics* **28**(19), 2458–2466.

Li, Y., Li, Z., Zhou, S., Wen, J., Geng, B., Yang, J. & Cui, Q. (2013), 'Genome-Wide Analysis of Human MicroRNA Stability', *BioMed Research International* **2013**.

Liang, H. & Li, W.-H. (2007), 'MicroRNA regulation of human protein–protein interaction network', *RNA* **13**(9), 1402–1408.

Liang, Z., Zhou, H., He, Z., Zheng, H. & Wu, J. (2011), 'mirAct: a web tool for evaluating microRNA activity based on gene expression data', *Nucleic acids research* **39**(suppl 2), W139–W144.

Liao, Y. & Vemuri, V. R. (2002), 'Use of k-nearest neighbor classifier for intrusion detection', *Computers & Security* **21**(5), 439–448.

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S. & Johnson, J. M. (2005), 'Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs', *Nature* **433**(7027), 769–773.

Liu, B., Li, J. & Tsykin, A. (2009*a*), 'Discovery of functional miRNA-mRNA regulatory modules with computational methods', *Journal of Biomedical Informatics* **42**(4), 685.

Liu, B., Li, J. & Tsykin, A. (2009*b*), 'Discovery of functional miRNAmRNA regulatory modules with computational methods', *J Biomed Inform* **42**(4), 685.

Liu, B., Liu, L., Tsykin, A., Goodall, G. J., Green, J. E., Zhu, M., Kim, C. H. & Li, J. (2010), 'Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation', *Bioinformatics* **26**(24), 3105–3111.

Liu, H., Yue, D., Chen, Y., Gao, S.-J. & Huang, Y. (2010), 'Improving performance of mammalian microRNA target prediction', *BMC bioinformatics* **11**(1), 476.

Liu, L., An, J., Liu, J., Wen, J., Zhai, X., Liu, Y., Pan, S., Jiang, J., Wen, Y., Liu, Z. et al. (2013), 'Potentially functional genetic variants in microRNA processing genes and risk of HBV-related hepatocellular carcinoma', *Molecular Carcinogenesis* **52**(S1), 148–154.

Liu, M., Roth, A., Yu, M., Morris, R., Bersani, F., Rivera, M. N., Lu, J., Shioda, T., Vasudevan, S., Ramaswamy, S. et al. (2013), 'The IGF2 intronic miR-483 selectively enhances transcription from IGF2 fetal promoters and enhances tumorigenesis', *Genes & development* **27**(23), 2543–2548.

Liu, S.-Y., Chen, Y.-T., Tseng, M.-Y., Hung, C.-C., Chiang, W.-F., Chen, H.-R., Shieh, T.-Y., Chen, C.-H., Jou, Y.-S. & Chen, J. Y.-F. (2008),

'Involvement of microtubule-associated protein 2 (MAP2) in oral cancer cell motility: a novel biological function of MAP2 in non-neuronal cells', *Biochemical and Biophysical Research Communications* **366**(2), 520–525.

Liu, X., Wang, T., Wakita, T. & Yang, W. (2010), 'Systematic identification of microrna and messenger rna profiles in hepatitis c virus-infected human hepatoma cells', *Virology* **398**(1), 57–67.

Liu, X., Zhang, X., Zhan, Q., Brock, M. V., Herman, J. G. & Guo, M. (2012), 'CDX2 serves as a Wnt signaling inhibitor and is frequently methylated in lung cancer', *Cancer Biology & Therapy* **13**(12), 1152.

Lodygin, D., Tarasov, V., Epanchintsev, A., Berking, C., Knyazeva, T., Korner, H., Knyazev, P., Diebold, J. & Hermeking, H. (2008), 'Inactivation of miR-34a by aberrant CpG methylation in multiple types of cancer', *Cell Cycle* **7**(16), 2591–2600.

Low, R. B. & White, S. L. (1998), 'Lung smooth muscle differentiation', *The International Journal of Biochemistry & Cell Biology* **30**(8), 869–883.

Lowery, A. J., Miller, N., McNeill, R. E. & Kerin, M. J. (2008), 'MicroRNAs as prognostic indicators and therapeutic targets: potential effect on breast cancer management', *Clinical Cancer Research* **14**(2), 360–365.

Lu, C.-Y., Lin, K.-Y., Tien, M.-T., Wu, C.-T., Uen, Y.-H. & Tseng, T.-L. (2013), 'Frequent dna methylation of mir-129-2 and its potential clinical implication in hepatocellular carcinoma', *Genes, Chromosomes and Cancer* **42**(8), 1273–1281.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H. & Ferrando, A. A. (2005), 'MicroRNA expression profiles classify human cancers', *Nature* **435**(7043), 834–838.

Ludwig, J. A. & Weinstein, J. N. (2005), 'Biomarkers in cancer staging, prognosis and treatment selection', *Nat Rev Cancer* **5**(11), 845–856.

Lupberger, J., Brino, L. & Baumert, T. F. (2008), 'Rnai-a powerful tool to unravel hepatitis c virus-host interactions within the infectious life cycle', *Journal of Hepatology* **48**(3), 523–525.

Ma, L., Huang, Y., Zhu, W., Zhou, S., Zhou, J., Zeng, F., Liu, X., Zhang, Y. & Yu, J. (2011), 'An integrated analysis of miRNA and mRNA expressions in non-small cell lung cancers', *PLoS One* **6**(10), e26502.

Maragkakis, M., Reczko, M., Simossis, V. A., Alexiou, P., Papadopoulos, G. L., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K. et al. (2009), 'DIANA-microT web server: elucidating microRNA functions through target prediction', *Nucleic acids research* p. gkp292.

Marín, R. M. & Vaníček, J. (2011), 'Efficient use of accessibility in microRNA target prediction', *Nucleic Acids Research* **39**(1), 19–29.

Martinez, N. J., Ow, M. C., Barrasa, M. I., Hammell, M., Sequerra, R., Doucette-Stamm, L., Roth, F. P., Ambros, V. R. & Walhout, A. J. (2008), 'A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity', *Genes & Development* **22**(18), 2535–2549.

Mas, V. R., Maluf, D. G., Stravitz, R., Dumur, C. I., Clark, B., Rodgers, C., Ferreira-Gonzalez, A. & Fisher, R. A. (2004), 'Hepatocellular carcinoma in HCV-infected patients awaiting liver transplantation: Genes involved in tumor progression', *Liver Transplantation* **10**(5), 607–620.

Matkovich, S. J., Hu, Y. & Dorn, G. W. (2013), 'Regulation of Cardiac MicroRNAs by Cardiac MicroRNAs', *Circulation Research* **113**(1), 62–71.

Mattie, M. D., Benz, C. C., Bowers, J., Sensinger, K., Wong, L., Scott, G. K., Fedele, V., Ginzinger, D., Getts, R. & Haqq, C. (2006), 'Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies', *Molecular Cancer* **5**(1), 24.

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V. et al. (2003), 'TRANSFAC®: transcriptional regulation, from patterns to profiles', *Nucleic acids research* **31**(1), 374–378.

Mertens-Talcott, S. U., Chintharlapalli, S., Li, X. & Safe, S. (2007), 'The oncogenic microRNA-27a targets genes that regulate specificity protein transcription factors and the G2-M checkpoint in MDA-MB-231 breast cancer cells', *Cancer Research* **67**(22), 11001–11011.

Meyer, P. N., Clark, J. I., Flanigan, R. C. & Picken, M. M. (2007), 'Xp11. 2 translocation renal cell carcinoma with very aggressive course in five adults', *Am J Clin Pathol* **128**(1), 70–79.

Migliore, C. & Giordano, S. (2009), 'MiRNAs as new master players', *Cell Cycle* **8**(14), 2185–2186.

Miki, D., Kubo, M., Takahashi, A., Yoon, K.-A., Kim, J., Lee, G. K., Zo, J. I., Lee, J. S., Hosono, N., Morizono, T. et al. (2010), 'Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations', *Nature Genetics* **42**(10), 893–896.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002), 'Network motifs: simple building blocks of complex networks', *Science* **298**(5594), 824–827.

Minna, J. D., Roth, J. A. & Gazdar, A. F. (2002), 'Focus on lung cancer', *Cancer Cell* **1**(1), 49.

Miranda, K. C., Huynh, T., Tay, Y., Ang, Y.-S., Tam, W.-L., Thomson, A. M., Lim, B. & Rigoutsos, I. (2006), 'A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes', *Cell* **126**(6), 1203–1217.

Miska, E. A., Alvarez-Saavedra, E., Townsend, M., Yoshii, A., Šestan, N., Rakic, P., Constantine-Paton, M. & Horvitz, H. R. (2004), 'Microarray analysis of microRNA expression in the developing mammalian brain', *Genome biology* **5**(9), R68.

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., Peterson, A., Noteboom, J., O'Briant, K. C. & Allen, A. (2008), 'Circulating microRNAs as stable blood-based markers for cancer detection', *Proc Natl Acad Sci USA* **105**(30), 10513–10518.

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., Peterson, A., Noteboom, J., O'Briant, K. C., Allen, A. et al. (2008), 'Circulating microRNAs as stable blood-based markers for cancer detection', *Proceedings of the National Academy of Sciences* **105**(30), 10513–10518.

Mraz, M., Malinova, K., Mayer, J. & Pospisilova, S. (2009), 'MicroRNA isolation and stability in stored RNA samples', *Biochem Biophys Res Commun* **390**(1), 1–4.

Murakami, Y., Aly, H. H., Tajima, A., Inoue, I. & Shimotohno, K. (2009), 'Regulation of the hepatitis c virus genome replication by mir-199a*', *Journal of Hepatology* **50**(3), 453.

Nakasa, T., Miyaki, S., Okubo, A., Hashimoto, M., Nishida, K., Ochi, M. & Asahara, H. (2008), 'Expression of microRNA-146 in rheumatoid arthritis synovial tissue', *Arthritis & Rheumatism* **58**(5), 1284–1292.

Nazarov, P. V., Reinsbach, S. E., Muller, A., Nicot, N., Philippidou, D., Vallar, L. & Kreis, S. (2013), 'Interplay of microRNAs, transcription factors and target genes: linking dynamic expression changes to function', *Nucleic Acids Research* **41**(5), 2817–2831.

Nunez, Y. O., Truitt, J. M., Gorini, G., Ponomareva, O. N., Blednov, Y. A., Harris, R. A. & Mayfield, R. D. (2013), 'Positively correlated miRNA-mRNA regulatory networks in mouse frontal cortex during early stages of alcohol dependence', *BMC Genomics* **14**(1), 725.

ørom, U. A., Nielsen, F. C. & Lund, A. H. (2008), 'MicroRNA-10a binds the 5' UTR of ribosomal protein mRNAs and enhances their translation', *Molecular Cell* **30**(4), 460–471.

Orton, R. x., Sturm, O. x., Vyshemirsky, V., Calder, M., Gilbert, D. x. & Kolch, W. (2005), 'Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway', *Biochem. J* **392**, 249–261.

Papadopoulos, G. L., Alexiou, P., Maragkakis, M., Reczko, M. & Hatzigeorgiou, A. G. (2009), 'DIANA-mirPath: Integrating human and mouse microRNAs in pathways', *Bioinformatics* **25**(15), 1991–1993.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Pub.

Pearl, J. (2003), 'Causality: models, reasoning, and inference', *Econometric Theory* **19**, 675–685.

Peng, X., Li, Y., Walters, K.-A., Rosenzweig, E. R., Lederer, S. L., Aicher, L. D., Proll, S. & Katze, M. G. (2009), 'Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers', *BMC Genomics* **10**(1), 373.

Place, R. F., Li, L.-C., Pookot, D., Noonan, E. J. & Dahiya, R. (2008), 'Microrna-373 induces expression of genes with complementary

promoter sequences', *Proceedings of the National Academy of Sciences* **105**(5), 1608–1613.

Prot, C., Boccon-Gibod, L., Bouvier, R., Doz, F., Fournet, J.-C., Frneaux, P., Vieillefond, A. & Couturier, J. (2003), 'Five new cases of juvenile renal cell carcinoma with translocations involving Xp11. 2: a cytogenetic and morphologic study', *Cancer Genet Cytogenet* **143**(2), 93–99.

Quinlan, A. R. & Hall, I. M. (2010), 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics* **26**(6), 841–842.

Quinlan, J. R. (1993*a*), *C4. 5: programs for machine learning*, Vol. 1, Morgan Kaufmann, San Francisco, California.

Quinlan, J. R. (1993*b*), *C4. 5: programs for machine learning*, Vol. 1, Morgan kaufmann.

Raponi, M., Dossey, L., Jatkoe, T., Wu, X., Chen, G., Fan, H. & Beer, D. G. (2009), 'MicroRNA classifiers for predicting prognosis of squamous cell lung cancer', *Cancer Res* **69**(14), 5776–5783.

Rehmsmeier, M., Steffen, P., Höchsmann, M. & Giegerich, R. (2004), 'Fast and effective prediction of microRNA/target duplexes', *Rna* **10**(10), 1507–1517.

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B. & Bartel, D. P. (2002), 'Prediction of plant microRNA targets', *Cell* **110**(4), 513–520.

Rijlaarsdam, M. A., Rijlaarsdam, D. J., Gillis, A. J., Dorssers, L. C. & Looijenga, L. H. (2013), 'mimsg: a target enrichment algorithm for predicted mir-mrna interactions based on relative ranking of matched expression data', *Bioinformatics* **29**(13), 1638–1646.

Rish, I. (2001), An empirical study of the naive Bayes classifier, *in* 'IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence', Vol. 3, pp. 41–46.

Shalgi, R., Lieber, D., Oren, M. & Pilpel, Y. (2007), 'Global and local architecture of the mammalian microRNA-transcription factor regulatory network', *PLoS Comput Biol* **3**(7), e131.

Sham, J. S., Tang, T. C.-M., Fang, Y., Sun, L., Qin, L.-X., Wu, Q.-L., Xie, D. & Guan, X.-Y. (2002), 'Recurrent chromosome alterations in primary ovarian carcinoma in Chinese women', *Cancer Genet Cytogenet* **133**(1), 39–44.

Shan, S. W., Fang, L., Shatseva, T., Rutnam, Z. J., Yang, X., Du, W., Lu, W.-Y., Xuan, J. W., Deng, Z. & Yang, B. B. (2013), 'Mature mir-17-5p and passenger mir-17-3p induce hepatocellular carcinoma by targeting pten, galnt7 and vimentin in different signal pathways', *Journal of Cell Science* **126**(6), 1517–1530.

Shen, J., Todd, N. W., Zhang, H., Yu, L., Lingxiao, X., Mei, Y., Guarnera, M., Liao, J., Chou, A. & Lu, C. L. (2010), 'Plasma microRNAs as potential biomarkers for non-small-cell lung cancer', *Lab Invest* **91**(4), 579–587.

Sklan, E. H., Staschke, K., Oakes, T. M., Elazar, M., Winters, M., Aroeti, B., Danieli, T. & Glenn, J. S. (2007), 'A rab-gap tbc domain protein binds hepatitis c virus ns5a and mediates viral replication', *Journal of Virology* **81**(20), 11096–11105.

Song, R., Catchpoole, D. R., Kennedy, P. J. & Li, J. (2015), 'Identification of lung cancer miRNA–miRNA co-regulation networks through a progressive data refining approach', *Journal of Theoretical Biology* **380**, 271–279.

Song, R., Liu, Q., Hutvagner, G., Nguyen, H., Ramamohanarao, K., Wong, L. & Li, J. (2014), 'Rule discovery and distance separation to detect reliable miRNA biomarkers for the diagnosis of lung squamous cell carcinoma', *BMC Genomics* **15**(9), 1.

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**(Suppl 2), S11.

Sood, P., Krek, A., Zavolan, M., Macino, G. & Rajewsky, N. (2006), 'Cell-type-specific signatures of microRNAs on target mRNA expression', *Proceedings of the National Academy of Sciences of the United States of America* **103**(8), 2746–2751.

Sturm, M., Hackenberg, M., Langenberger, D. & Frishman, D. (2010), 'TargetSpy: a supervised machine learning approach for microRNA target prediction', *BMC bioinformatics* **11**(1), 292.

Su, A. I., Pezacki, J. P., Wodicka, L., Brideau, A. D., Supekova, L., Thimme, R., Wieland, S., Bukh, J., Purcell, R. H. & Schultz, P. G. (2002), 'Genomic analysis of the host response to hepatitis c virus infection', *Proceedings of the National Academy of Sciences* **99**(24), 15669–15674.

Sun, J., Gong, X., Purow, B. & Zhao, Z. (2012), 'Uncovering microRNA and transcription factor mediated regulatory networks in glioblastoma', *PLoS Comput Biol* **8**(7), e1002488.

Suzuki, H. I., Mihira, H., Watabe, T., Sugimoto, K. & Miyazono, K. (2013), 'Widespread inference of weighted microRNA-mediated gene regulation in cancer transcriptome analysis', *Nucleic Acids Research* **41**(5), e62–e62.

Tan, X., Qin, W., Zhang, L., Hang, J., Li, B., Zhang, C., Wan, J., Zhou, F., Shao, K. & Sun, Y. (2011*a*), 'A 5-microRNA signature for

lung squamous cell carcinoma diagnosis and hsa-mir-31 for prognosis', *Clinical Cancer Research* **17**(21), 6802–6811.

Tan, X., Qin, W., Zhang, L., Hang, J., Li, B., Zhang, C., Wan, J., Zhou, F., Shao, K. & Sun, Y. (2011*b*), 'A 5-microRNA signature for lung squamous cell carcinoma diagnosis and hsa-miR-31 for prognosis', *Clin Cancer Res* **17**(21), 6802–6811.

Tan, X., Qin, W., Zhang, L., Hang, J., Li, B., Zhang, C., Wan, J., Zhou, F., Shao, K., Sun, Y. et al. (2011), 'A 5-microRNA signature for lung squamous cell carcinoma diagnosis and hsa-miR-31 for prognosis', *Clinical Cancer Research* **17**(21), 6802–6811.

Tavazoie, S. F., Alarcón, C., Oskarsson, T., Padua, D., Wang, Q., Bos, P. D., Gerald, W. L. & Massagué, J. (2008), 'Endogenous human microRNAs that suppress breast cancer metastasis', *nature* **451**(7175), 147–152.

Taylor, D. D. & Gercel-Taylor, C. (2008), 'MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer', *Gynecologic Oncology* **110**(1), 13–21.

Team, R. C. (2013), 'R: A Language and Environment for Statistical Computing'.
URL: *http://www.R-project.org/*

Thomas, P. D., Wood, V., Mungall, C. J., Lewis, S. E., Blake, J. A., Consortium, G. O. et al. (2012), 'On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: a short report', *PLoS Computational Biology* **8**(2), e1002386.

Tijsen, A. J., Creemers, E. E., Moerland, P. D., de Windt, L. J., van der Wal, A. C., Kok, W. E. & Pinto, Y. M. (2010), 'MiR423-5p as a circulating biomarker for heart failure', *Circulation Research* **106**(6), 1035–1039.

Tran, D. H., Satou, K., Ho, T. B. & Pham, T. H. (2010), 'Computational discovery of miR-TF regulatory modules in human genome', *Bioinformation* **4**(8), 371.

Transcriptional, C. (2006), 'miRNAs Regulate miRNAs', *Cell Cycle* **5**(21), 2473–2476.

Van der Auwera, I., Limame, R., Van Dam, P., Vermeulen, P., Dirix, L. & Van Laere, S. (2010), 'Integrated miRNA and mRNA expression profiling of the inflammatory breast cancer subtype', *British journal of cancer* **103**(4), 532–541.

van Dongen, S., Abreu-Goodger, C. & Enright, A. J. (2008), 'Detecting microRNA binding and siRNA off-target effects from expression data', *Nature methods* **5**(12), 1023–1025.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. (2009), 'A census of human transcription factors: function, expression and evolution', *Nature Reviews Genetics* **10**(4), 252–263.

Vejnar, C. E. & Zdobnov, E. M. (2012), 'miRmap: Comprehensive prediction of microRNA target repression strength', *Nucleic acids research* **40**(22), 11673–11683.

Vincent, M. (2013), 'Conference Scene: Australia Lung Cancer Conference 2012: focus on systemic treatment of advanced non-small-cell lung cancer', *Lung Cancer* **2**(1), 27–30.

Vlachos, I. S., Paraskevopoulou, M. D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I.-L., Maniou, S., Karathanou, K., Kalfakakou, D. et al. (2014), 'DIANA-TarBase v7. 0: indexing more than half a million experimentally supported miRNA: mRNA interactions', *Nucleic acids research* p. gku1215.

Wang, G.-K., Zhu, J.-Q., Zhang, J.-T., Li, Q., Li, Y., He, J., Qin, Y.-W. & Jing, Q. (2010), 'Circulating microRNA: a novel potential biomarker

for early diagnosis of acute myocardial infarction in humans', *European Heart Journal* **31**(6), 659–666.

Wang, J., Lu, M., Qiu, C. & Cui, Q. (2010), 'Transmir: a transcription factor–microrna regulation database', *Nucleic acids research* **38**(suppl 1), D119–D122.

Wang, P., Ning, S., Wang, Q., Li, R., Ye, J., Zhao, Z., Li, Y., Huang, T. & Li, X. (2013), 'mirTarPri: improved prioritization of microRNA targets through incorporation of functional genomics data', *PloS one* **8**(1).

Wang, X. (2008), 'miRDB: a microRNA target prediction and functional annotation database with a wiki interface', *RNA* **14**(6), 1012–1017.

Wang, X. (2016), 'Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-Ligation studies', *Bioinformatics* p. btw002.

Wang, X. & El Naqa, I. M. (2008), 'Prediction of both conserved and nonconserved microRNA targets in animals', *Bioinformatics* **24**(3), 325–332.

Wei, D., Deng, X., Zhang, X., Deng, Y. & Mahadevan, S. (2013), 'Identifying influential nodes in weighted networks based on evidence theory', *Physica A: Statistical Mechanics and its Applications* **392**(10), 2564–2575.

Wernicke, S. & Rasche, F. (2006), 'FANMOD: a tool for fast network motif detection', *Bioinformatics* **22**(9), 1152–1153.

West, D. B. et al. (2001), *Introduction to graph theory*, Vol. 2, Prentice hall Englewood Cliffs.

Wilby, K. J., Partovi, N., Ford, J.-A. E., Greanya, E. D. & Yoshida, E. M. (2012), 'Review of boceprevir and telaprevir for the treatment of chronic hepatitis c', *Canadian Journal of Gastroenterology* **26**(4), 205.

Winter, J., Jung, S., Keller, S., Gregory, R. I. & Diederichs, S. (2009), 'Many roads to maturity: microRNA biogenesis pathways and their regulation', *Nature Cell Biology* **11**(3), 228–234.

Wong, N. & Wang, X. (2014), 'miRDB: an online resource for microRNA target prediction and functional annotations', *Nucleic acids research* p. gku1104.

Wong, T.-S., Liu, X.-B., Wong, B. Y.-H., Ng, R. W.-M., Yuen, A. P.-W. & Wei, W. I. (2008), 'Mature miR-184 as potential oncogenic microRNA of squamous cell carcinoma of tongue', *Clinical Cancer Research* **14**(9), 2588–2592.

Wu, J., Qian, J., Li, C., Kwok, L., Cheng, F., Liu, P., Perdomo, C., Kotton, D., Vaziri, C. & Anderlind, C. (2010), 'miR-129 regulates cell proliferation by downregulating Cdk6 expression', *Cell Cycle* **9**(9), 1809–1818.

Xia, H., Ooi, L. L. P. & Hui, K. M. (2012), 'mirna-214 targets β-catenin pathway to suppress invasion, stem-like traits and recurrence of human hepatocellular carcinoma', *PLoS One* **7**(9), e44206.

Xia, L., Zhang, D., Du, R., Pan, Y., Zhao, L., Sun, S., Hong, L., Liu, J. & Fan, D. (2008), 'miR-15b and miR-16 modulate multidrug resistance by targeting BCL2 in human gastric cancer cells', *International Journal of Cancer* **123**(2), 372–379.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X. & Li, T. (2009), 'miRecords: an integrated resource for microRNA–target interactions', *Nucleic acids research* **37**(suppl 1), D105–D110.

Xiao, S., Ma, Y., Zhu, H., Sun, H., Yin, Y. & Feng, G. (2015), 'miRNA functional synergistic network analysis of mice with ischemic stroke', *Neurological Sciences* **36**(1), 143–148.

Xie, H.-Y., Cheng, J., Xing, C.-Y., Wang, J.-J., Su, R., Wei, X.-Y., Zhou, L. & Zheng, S.-S. (2011), 'Evaluation of hepatitis b viral replication and proteomic analysis of hepg2. 2.15 cell line after knockdown of hbx', *Hepatobiliary & Pancreatic Diseases International* **10**(3), 295–302.

Xie, Y., Todd, N. W., Liu, Z., Zhan, M., Fang, H., Peng, H., Alattar, M., Deepak, J., Stass, S. A. & Jiang, F. (2010), 'Altered miRNA expression in sputum for diagnosis of non-small cell lung cancer', *Lung Cancer* **67**(2), 170–176.

Xu, J. & Wong, C. (2008), 'A computational screen for mouse signaling pathways targeted by microRNA clusters', *RNA* **14**(7), 1276–1283.

Yang, D., Sun, Y., Hu, L., Zheng, H., Ji, P., Pecot, C. V., Zhao, Y., Reynolds, S., Cheng, H., Rupaimoole, R. et al. (2013), 'Integrated analyses identify a master microRNA regulatory network for the mesenchymal subtype in serous ovarian cancer', *Cancer Cell* **23**(2), 186–199.

Yang, J.-H., Li, J.-H., Jiang, S., Zhou, H. & Qu, L.-H. (2013), 'ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data', *Nucleic Acids Research* **41**(D1), D177–D187.

Yang, N., Kaur, S., Volinia, S., Greshock, J., Lassus, H., Hasegawa, K., Liang, S., Leminen, A., Deng, S., Smith, L. et al. (2008), 'MicroRNA microarray identifies Let-7i as a novel biomarker and therapeutic target in human epithelial ovarian cancer', *Cancer Research* **68**(24), 10307–10314.

Yang, Y., Li, X., Yang, Q., Wang, X., Zhou, Y., Jiang, T., Ma, Q. & Wang, Y.-J. (2010), 'The role of microRNA in human lung squamous cell carcinoma', *Cancer Genet Cytogenet* **200**(2), 127–133.

Yoon, S. & De Micheli, G. (2005*a*), 'Prediction of regulatory modules comprising microRNAs and target genes', *Bioinformatics* **21**(suppl 2), ii93–ii100.

Yoon, S. & De Micheli, G. (2005*b*), 'Prediction of regulatory modules comprising microRNAs and target genes', *Bioinformatics* **21**(suppl 2), ii93–ii100.

Yousef, M., Jung, S., Kossenkov, A. V., Showe, L. C. & Showe, M. K. (2007), 'Naïve Bayes for microRNA target predictionsmachine learning for microRNA targets', *Bioinformatics* **23**(22), 2987–2992.

Yu, L., Todd, N. W., Xing, L., Xie, Y., Zhang, H., Liu, Z., Fang, H., Zhang, J., Katz, R. L. & Jiang, F. (2010), 'Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers', *Int J Cancer* **127**(12), 2870–2878.

Yu, T., Li, J., Yan, M., Liu, L., Lin, H., Zhao, F., Sun, L., Zhang, Y., Cui, Y., Zhang, F., He, X. & Yao, M. (2015), 'MicroRNA-193a-3p and -5p suppress the metastasis of human non-small-cell lung cancer by downregulating the ERBB4/PIK3R3/mTOR/S6K2 signaling pathway', *Oncogene* **34**(4), 413–423. Supplementary information available for this article at http://www.nature.com/onc/journal/v34/n4/suppinfo/onc2013574s1.html.

Yu, X., Lin, J., Zack, D. J., Mendell, J. T. & Qian, J. (2008), 'Analysis of regulatory network topology reveals functionally distinct classes of microRNAs', *Nucleic Acids Research* **36**(20), 6494–6503.

Yuan, X., Liu, C., Yang, P., He, S., Liao, Q., Kang, S. & Zhao, Y. (2009), 'Clustered microRNAs' coordination in regulating protein-protein interaction network', *BMC Systems Biology* **3**(1), 65.

Zeng, L., Cui, J., Wu, H. & Lu, Q. (2014), 'The emerging role of circulating microRNAs as biomarkers in autoimmune diseases', *Autoimmunity* **2014**(0), 1–11.

Zhang, B., Pan, X., Cobb, G. P. & Anderson, T. A. (2007*a*), 'microRNAs as oncogenes and tumor suppressors', *Dev Biol* **302**(1), 1–12.

Zhang, B., Pan, X., Cobb, G. P. & Anderson, T. A. (2007*b*), 'microRNAs as oncogenes and tumor suppressors', *Developmental biology* **302**(1), 1–12.

Zhang, G.-l., Li, Y.-x., Zheng, S.-q., Liu, M., Li, X. & Tang, H. (2010), 'Suppression of hepatitis b virus replication by microrna-199a-3p and microrna-210', *Antiviral Research* **88**(2), 169–175.

Zhang, S., Li, Q., Liu, J. & Zhou, X. J. (2011), 'A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules', *Bioinformatics* **27**(13), i401–i409.

Zhang, S., Wang, Y., Dai, S.-D. & Wang, E.-H. (2011), 'Down-regulation of NKD1 increases the invasive potential of non-small-cell lung cancer and correlates with a poor prognosis', *BMC Cancer* **11**(1), 186.

Zhang, X., Daucher, M., Armistead, D., Russell, R. & Kottilil, S. (2013), 'Microrna expression profiling in hcv-infected human hepatoma cells identifies potential anti-viral targets induced by interferon-α', *PLoS One* **8**(2), e55733.

Zhao, C., Sun, G., Li, S. & Shi, Y. (2009), 'A feedback regulatory loop involving microRNA-9 and nuclear receptor TLX in neural stem cell fate determination', *Nature structural & molecular biology* **16**(4), 365–371.

Zheng, X., Qi, Y., Gao, Y., Wang, X., Qi, M., Shi, X. & An, X. (2009), 'Expression and significance of membrane skeleton protein 4.1 family in nonDsmall cell lung cancer', *Chinese Journal of Cancer* **28**(7).

Zhou, Y., Ferguson, J., Chang, J. T. & Kluger, Y. (2007), 'Inter-and intra-combinatorial regulation by transcription factors and microRNAs', *BMC Genomics* **8**(1), 396.