Stony Brook University

Joint Genetic and MicroRNA Study of the Human Thrombocytosis under a System Biology Scheme

A Dissertation Presented

by

Xiao Xu

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Biomedical Engineering

Stony Brook University

August 2011

Stony Brook University

The Graduate School

Xiao Xu

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Wei Zhu – Dissertation Advisor
Professor, Applied Mathematics and Statistics & Biomedical Engineering

Michael Hadjiargyrou - Chairperson of Defense
Associate Professor, Biomedical Engineering

Anil Dhundale, Research Assistant Professor
Biomedical Engineering & Long Island High Technology Incubator

Wadie Bahou, Professor
Department of Medicine/Hematology & Vice Dean Office of Scientific Affairs
Stony Brook University Health Science Center

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Joint Genetic and MicroRNA Study of the Human Thrombocytosis under a System Biology Scheme

by

Xiao Xu

Doctor of Philosophy

in

Biomedical Engineering

Stony Brook University

2011

The disease of human thrombocytosis is a type of myeloproliferative disorder (MPD) characterized by a complex alteration of genetic and epigenetic factors. Consequently, the pathological cellular environment is permeated by transcriptomic deregulation, interferomic aberrations, protein level irregularity and disrupted cellular pathways. In this dissertation, we present a novel study of the human thrombocytosis disease through a multi-layered genetic, microRNA and proteomic platelet modeling and analysis in a systems biology scheme. This includes (1) the extraction of distinctive messenger RNA and microRNA expression profiles in human platelets for two subtypes of thrombocytosis patients versus the normal control group, (2) the derivation of a classification model that differentiates between subtypes of thrombocytosis as well as between general diseased and normal subjects using classical and modern discriminant analysis methods, (3) the profiling of platelet microRNAs in thrombocytosis and controls, (4) a microRNA-mRNA correlational analysis to identify a joint genetic and microRNA biological

regulatory network in thrombocytotic platelets, and (5) The final validation of identified joint genetic and microRNA platelet regulatory pathways using preliminary proteomics data. This system biological study of the thrombocytosis sets stage for further exploration of the disease mechanism and a more comprehensive recognition of the underlying etiologies. In summary, this dissertation features the discoveries of cellular regulatory networks underlying human thrombocytosis based on the analysis of the joint platelet genetic, microRNA and proteomic data under the system biology scheme. Further expansion is expected to advance our understanding of the platelet maturation pathway under the pathological environment of human thrombocytosis.

Dedicated to my parents and my beloved grandpa

Contents:

List of Tables

List of Figures

Acknowledgments

Upon the completion of my Ph.D dissertation, I would express my sincere appreciation to my advisor, Dr. Wei Zhu, for guiding me in the field of biostatistics and helping me through my graduate research career. I am also highly grateful to Dr. Wadie F. Bahou for his sponsorship and leading role in initiating all these medical projects that have provided me with various research opportunities. In addition, I am grateful to Dr. Dmitri V. Gnatenko for his invaluable help in our daily research; I also owe my success at BME to Dr. Michael Hadjiargyrou for his excellent academic advising. I would like to thank Dr. Anil Dhundale for his work producing custom microarray platforms that enabled the generation of platelet genomics data without which this dissertation would not be written. I want to thank Ms. Jean Wainer for her devotion and expertise in performing many biological experiments for my analysis. I thank Dr. Jingfang Ju for his suggestions and collaborations in my research project especially on all things related to miRNA.. Thank all my past and current colleagues in Dr. Zhu's lab (Dr. Xiao Wu, Dr. Tianyi Zhang, Hongyan Chen, Shaonan Zhang, Hao Han, Tim, Dr. Jianping Zhang, Dr. Ling Leng, Erya Huang, etc). Finally, I would like to extend my sincere gratitude to my mother, my father and my grandpa as well as all the other family members for their love and support through all these years.

Chapter 1

Introduction to Thrombocytosis Study

Among the many types of hematopoietic disorders, essential or primary thrombocythemia (ET) is a rare chronic myeloproliferative disorder characterized by excessive production of platelets from megakaryocytes in bone marrow [1]. This disease is often associated with sustained megakaryocytes proliferation without a definitive cause [2, 3]. Although in most of the ET cases, people do not experience further complications and they can have a normal life expectancy, a subset of the ET patients develop a strong tendency for arterial or/and venous thrombosis and hemorrhage that underlies the lethality of this serious medical condition. The common clinical records of ET have shown a major predisposition to vascular occlusive events such as clots detected in cerebrovascular, coronary and peripheral circulation. The typical symptoms of ET include headache, lightheadedness, vision disturbance, erythromelalgiand, tingling and numbness or burning pain in the hands and feet. In contrast, some patients would not

experience any symptoms or abnormal sensations throughout the course of disease. The prognosis of ET varies on individual level, while most of the patients can have a normal life expectancy without severe complications, a subset of cases do report severe neurological, cardiac or peripheral artery manifestations. In rare scenarios, advanced ET disease could evolve into acute leukemia or myelodysplasia [1, 4-6]. The prevalence of ET in general population is about 30/100,000 and is more frequent in people older than 50. Women have a naturally higher susceptibility to ET than men with an approximate ratio of 2:1 based on surveys conducted among confirmed cases [2, 7]. This disease shares many phenotypic and etiological similarities with other myeloproliferative neoplasms (MPN) among which the polycythemia vera (PV) and primary myelofibrosis (PMF) resembled ET disease most in both phenotypes and clinical manifestations [5, 8]. The diagnostic criteria for ET is characterized by a consensus of various clinical factors, including platelet counts greater than 600,000/μL, megakaryocytic hyperplasia, detection of acquired pathogenic mutations, splenomegaly, lack of reactive cause and a clinical records plagued by thrombotic and/or hemorrhagic events [8-12].

The reactive or secondary thrombocythemia (RT) is a similar but different form of thrombocytosis in comparison to ET. The RT disease shares a lot of phenotypic and pathological attributes with ET but with different disease etiology. While the essential thrombocythemia is a type of connatural hematopoietic defect with a strong genetic background [8], the RT disease is usually caused by various external stimuli, including trauma/surgery, infection, inflammatory disorders malignancy, chemotherapy and blood loss in a predisposing genetic environment [13, 14]. The overall pathological development of RT is benign and self-limiting (usually disappear automatically in one week), while it may also result in hemorrhage or thrombosis in a small number of cases. Under extreme conditions, the RT disease can evolve into acute myocardial infarction, mesenteric vein thrombosis and pulmonary embolism which is life threatening [15,

16]. The treatment of RT is simpler than that of ET since the therapy only needs to focus on reducing the occurrence of blood clotting using primarily physical measures (e.g. compression stockings)； in few cases, the administration of aspirin or other blood thinning medicine is used.

1.1 Etiology of Thrombocytosis

Although the ET disease was first described in the 1930s by Epstein and Goedel [17], its pathogenic background was not well understood until in 2005 when several different research groups independently identified an acquired mutation in the JAK2 tyrosine kinase in most patients with PV, ET or PMF [18-20]. The JAK2 is a type of cytoplasmic non-receptor tyrosine kinase. It interacts with signaling from the GM-CSF receptor family (IL-3R, IL-5R and GM-CSF-R), type II cytokine receptor family, gp130 receptor family (IL6ST, IL6-beta or CD130) and other single chain receptors [21-24]. Specifically for JAK2 gene, a mutation of a guanine-to-thymidine substitution was found in the vast majority of PV patients and 40%~50% percent of ET or PMF cases. This acquired genetic mutation will entail a substitution of valine for phenylalanine at codon 617 of JAK2 (V617F), which is likely to render hematopoietic stem cells more sensitive to growth factors such as thrombopoietin and erythropoietin [25, 26]. While the ET disease shares molecular pathogenic pathways with other clonal myeloproliferative disorders (MPDs) such as Chronic myeloid leukemia (CML), Polycythemia Vera (PV), Myelofibrosis (MF) with myeloid metaplasia of the spleen (IMF), its unique genetic profiles do feature a set of independent clinical criterions in its diagnosis and management as suggested in previous studies [8, 10, 27-29]. A series of MPD diagnostic guidelines dictated by World Health Organization (WHO) was included in this chapter (**Table 1-1)** with highlights on the distinctive clinical features of ET disease [30]. A comparative histological illustration is also presented to show the

difference between ET and its close 'relative' PV using pathological criteria (**Figure 1-1**). On the other hand, despite the existing clinical manifestations (e.g. the presence of the Philadelphia (Ph)-positive chromosomal abnormality in CML) and JAK2-617F mutation status, little consensus has been reached on the intrinsic molecular mechanism underlying the pathological origin of these myeloproliferative disorders. The JAK2 mutation mechanism has clear indication for the Philadelphia-negative MPDs as it implicitly elevates the proliferative and survival response to the target cell in bone marrow by rendering them more sensitive to incoming stimulatory signals, causing clonal expansion of hematopoietic progenitors during blood cell biogenesis.

Table 1-1: Diagnostic guidelines for ET and PV diseases [30, 31]

	Essential Thrombocythemia	Polycythemia Vera
Primary Clinical Criteria	1) Sustained platelet count > 450 $\times 10^9$ /L	1) Hemoglobin > 18.5 g/dL (men) > 16.5 g/dL (women) or Hemoglobin > 17 g/dL (men), or > 15 g/dL (women)
	2) Presence of an acquired pathogenic mutation such as JAK2V617F or MPL515	2) Hematocrit > 0.51 (men) or hematocrit > 0.48 (women)
	3) Presence of large or giant platelets in a peripheral blood smear	2) Presence of JAK2V617F mutation
	4) Regular level of hemoglobin, erythrocytes, hematocrit and white blood cells count.	3) Low serum erythropoietin level
	5) Absence of the Philadelphia chromosome or any other cytogenetic fusion-gene abnormality	
	6) No reactive cause for thrombocytosis	
Secondary Clinical criteria	7) JAK2V617F homozygous clone ratio	4) Histology: Bone marrow tri-lineage myeloproliferation (**Fig 1-1)**
	8) Bone marrow trephine histology showing enlarged megakaryocytes in prominent large hyperlobulated forms; reticulin is generally not increased **(Fig 1-1)**	5) Granulocytes >10 × 109 l−1 or leukocytes >12 × 109 l−1 and/or raised LAP-score or increased PRV-1 expression in the absence of fever
		6) Endogenous erythroid colony growth

The RT disease resembles ET in both clinical manifestations and symptoms, yet has not been identified with a clear genetic cause [32]. The intrusion of external factors such as bacterial infection and physical trauma are considered as the primary causes of this disease. One mechanism has indicated an association between a high level of interleukin-6 with the development of RT [33]; Alternatively the occurrence of RT is closely related to the immunological role of platelets since they can attach and engulf the pathogens, facilitating the functional protective role of white blood cells. Upon stimulation by bacteria or thrombin, the temporary spiking platelets level may reflect the autoimmune response of body [34]. Other disease causing factors of RT include chronic blood loss-iron deficiency, acute blood loss, hemolytic anemia, vasculitis, drug use (e.g. *vinca alkaloids*, *epinephrine*), splenectomy and rebound thrombocytosis [35-37]. Currently, the lack of clinical criterion in differentiating ET to RT diagnosis has posed another major obstruction in the RT disease management. A careful medical history review and supportive clinical tests need to be done in order to distinguish RT disease from ET or other MPDs. We published a study in 2009 describing a set of 11 biomarkers that have discriminant power to delineate ET patients from RT and normal controls [32]. This study will be discussed specifically in chapter 2.

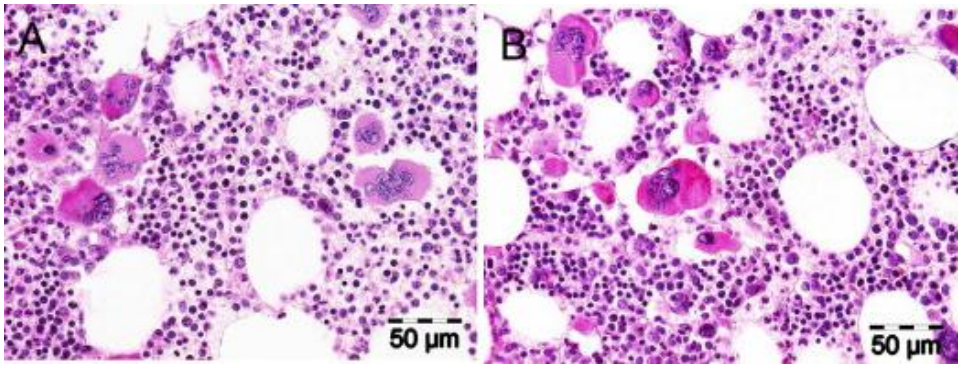

Figure 1-1: Histological bone marrow features of myeloproliferative disorders (A) Essential thrombocythemia (ET) showing loose clustering of enlarged mature megakaryocytes with hyperlobulated nuclei and moderate granulocytic proliferation. (B) Polycythemia Vera (PV) showing clustering of small to enlarged megakaryocytes with mature cytoplasm and hyperlobulated nuclei as well as higher cellularity due to increased erythropoiesis and granulopoiesis. (Image source: Michael J.J. et al [38])

1.2 Genetic Risk Factors of Thrombocytosis

JAK2 (V617F) mutation (Janus kinase 2), occurring within exon 14 of JAK2 and located on 9p24 is the top recurrent mutation in MPDs; It is found in roughly 96% PV cases and 50%~65% ET and PMF cases [39-41]. This mutation alters the auto-inhibitory sites JH2 of JAK2, which subsequently leads to constitutive activation of JAK2 and JAK/STAT signaling pathway. The JAK2 (V617F mutation) has been the most widely recognized genetic abnormality found associated with ET disease. It is also prevalent in PV or other types of MPDs [42]. Aside from the prominent JAK2 tyrosine kinase mutations, there exist other genetic aberrations detected in a fraction of ET patients and these mutations may affect the hematopoietic cell growth, differentiation and maturation in different ways. Examples of these genetic risk factors include the abnormal gain of function in myeloproliferative leukemia virus (MPL) gene [43], the deactivation of ten eleven translocation (TET2) tumor suppressor gene [44, 45]; The mutation in Additional Sex Combs-Like 1 (ASXL1) [46] and casitas B-lineage lymphoma (c-CBL) [47]. In some of the ET cases, the altered expression of thrombopoietin receptor c-MPl as well as the deficient thrombopoietin (TPO)-mediated tyrosine phosphorylation of proteins can directly lead to thrombocytosis as the TPO is the primary humeral regulator of platelet production [48]. Besides the genetic abnormalities mentioned above, there are common MPD deficiencies are far more frequently observed in other types of MPDs than ET, including IDH1/IDH2 (exon4), LNK, EZH2, NF1 and IKAROS mutations [46, 49, 50]. In both ET and PV diagnosis, an impaired expression of c-MPL protein in bone marrow megakaryocytes as well as an overexpression of PRV-1 are considered of important clinical indicators [51, 52]. The polycythemia rubra vera-1(PRV-1) is a member of uPAR receptor superfamily and increased PRV-1 mRNA expression level is another harbinger of MPDs, although it is less commonly detected in ETs than the PV cases [51, 53].

1.3 Potential microRNA Regulation in Hematological Malignancies

The microRNAs (miRNA) are a special group of short RNA species consisting of 21~24 nucleotides and interact with target mRNAs to affect their translational efficacy [54]. The miRNA regulation represents an important post-transcriptional control to the primary cellular regulatory pathway in which gene expression level dictates the final protein synthesis. Recent studies have suggested the extensive regulatory role of hematopoietic-specific miRNAs in megakaryocytic proliferation and differentiation process [55-59]. Prominent cases include the interaction between miR-10a and its direct target HOXA-1; miR-130's repression on transcription factor MAFB [58] and miR-150's role in the differentiation process of megakaryocyte–erythrocyte progenitor cells [60]. Unsurprisingly, the deregulation of some key miRNAs is extensively associated with the pathogenesis of a few hematopoietic malignancies. This includes the aberrant profiles of some hematopoietic-specific miRNAs (e.g. let-7a, miR-182, miR-143, miR-145 and miR-223) in clonal myeloproliferative disorder (MPD) reported by Bruchova et al [61, 62]; the over-expression of miR-28 in a fraction of BCR-ABL negative myeloproliferative neoplasms cases as well as the down-regulation of neutrophilic miR-133a, miR-1 in patients with PV and ET from Slezak et al [63]. These experimental findings have suggested the widespread and effective regulation of miRNAs in hematological system, which laid a firm ground for us to investigate the miRNAs' role in the origin and development of thrombocytosis.

1.4 The Model of Our Study: Human Platelets

As a key component of human blood, platelets involve in a wide variety of blood functions: They regulate and initiate the important hemostasis process while simultaneously

transporting the negatively charged phospholipids to its surface, which serves as catalytic sites for contact phase-mediated propagation of the coagulation cascade. Platelets also release a multitude of growth factors including Platelet-derived growth factor (PDGF), a potent chemotactic agent, and TGF beta, which stimulate the deposition of extracellular matrix. In some cases, protein over-expression may favor platelet activation and thrombus formation, which provides conceptual support for the presence of biomarkers that, may result in enhanced thrombosis susceptibility risk. In general, genetic or protein biomarkers are closely linked to clinical platelet phenotypes, including those in pathological settings. Human platelets are produced in thrombopoiesis by budding off from megakaryocytes and are generally less complex than nucleated cells in its transcriptome [64, 65]. Minimal translational activities are observed in quiescent platelets for some of its mRNAs are inherited from megakaryocytes. A recent study has also confirmed the existence of complete miRNA post-transcriptional regulatory pathways inside human platelets. This includes an abundant and diverse spectra of expressed human miRNAs as well as the functional Dicer and Argonaute 2 (Ago2) complexes [66]. The platelet miRNAs come from their parental megakaryocytes and therefore reflects the miRNA regulatory profiles in hematopoietic stem cells to a certain extent.

In our own study, we have conducted extensive research on the microRNA expression level in ET and RT, and compared them to normal controls. We found that a wide array of miRNAs are expressed in human platelets during thrombocytosis development and in some pathological conditions, the number of expressed miRNAs are many more than those in normal platelets (unpublished work [67]). Besides, in a comparison between ET and normal platelets we also noticed that there are roughly equal amount of up and down regulated miRNAs expressed across the two conditions. The details of platelet microRNA study will be discussed in C-3 (C refers to Chapter in this dissertation).

1.5 The Study of Thrombocytosis in a System Biology Framework

Current hematological research in thrombocytosis faces many challenges although some of the genetic mechanisms underlying these diseases have been unveiled (i.e. JAK2 mutation). Behind the many fields to be explored, the lack of understanding in the complete regulatory pathways underlying the development of the disease has been the main clue of our research. In specific terms, the genetic aberration, microRNA deregulation and the corresponding protein abnormalities associated with the pathological development of thrombocytosis and its different subtypes are three primary focuses in our research. We planned our research scheme based on the general hypothesis that there exist both critical gene features and microRNA regulators representing the biological difference between thrombocytotic patients and normal subjects. These genetic and miRNA factors also interact with each other and mediates a complete regulatory pathways that eventually express their controlling efficacy at the protein output level.

By studying the individual expression profile of three molecular-level regulatory factors as well as their interactions in human platelet model, we have formulated the first dimension of our study in thrombocytosis under the system biology frame work (**Figure 1-2**). In the second dimension, we tend to expand our research longitudinally along the CD34+ hematopoietic stem cell maturation lineage to the up-stream cellular system of megakaryocytes (MK) and further back to the hematopoietic stem cells. At current stage, we have explored the genetic profiles in MK system in a separate study (not discussed in this dissertation) under normal physiological conditions. A complete characterization on its miRNA and protein levels will be included in a future research scheme and more investigative efforts are planned in the longitudinal direction . Besides, we also tried to include additional layer of experimental conditions where thrombin-receptor-activating peptides (TRAP) stimulation is introduced to activate the G-protein coupled signaling pathway [68]. At current stage, only platelet protein level is measured in this study and

we are expected to expand our research scope to miRNA and genes later. The details of this experiment using TRAP stimulation on platelets are illustrated in C-5. Inside the genomic characterization of thrombocytotic platelets, we sought to address the current clinical challenge of lacking in genetic criteria delineating between thrombocytosis subtypes: ET vs RT. This part of research was already done by us and has been included in the general system biology framework (**Figure 1-2**).

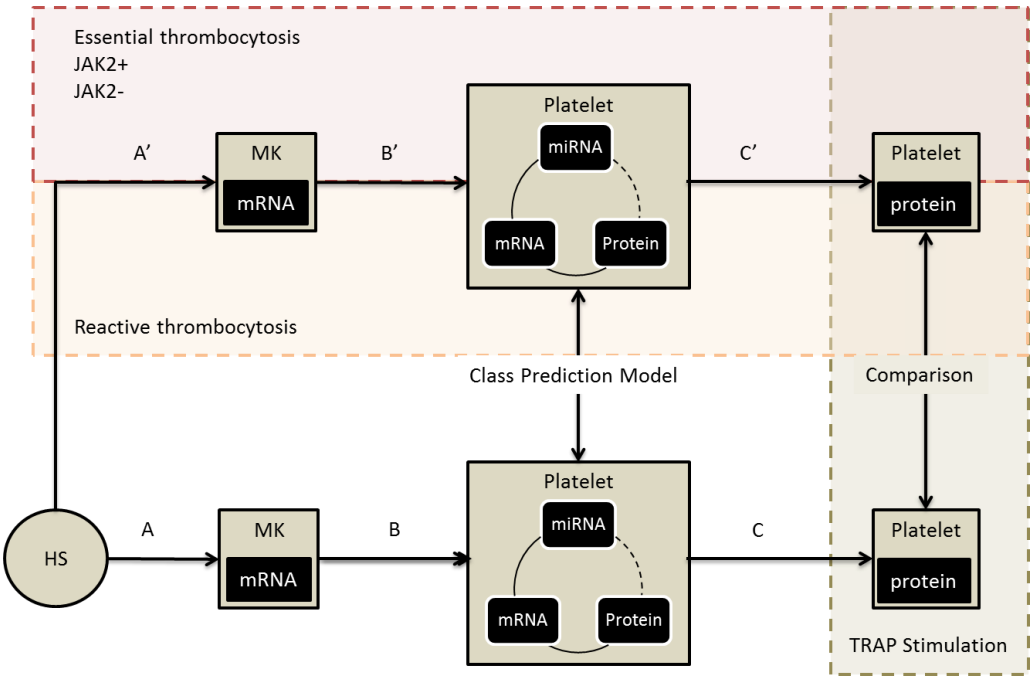

Figure 1-2 The system biology scheme of our study in the CD34+ hematopoietic stem cell maturation lineage with/ without the thrombocytosis condition. Acronym of graph: HS, hemocytoblast; MK, Megakaryocte; PLT, Platelet. We studied a complete CD 34+ hematopoietic differentiation pathway with a focus on the hematopoietic cell → megakaryocyte → platelet lineage under both normal and the pathological condition of thrombocytosis. The first leg (path A): from hemocytoblast to megakaryocyte was included in a separate study and is not discussed in details in this dissertation. The megakaryocyte matures and breaks down to give birth to platelets (path B, 1 MK cell can generate ~400 platelets [69]). The platelet has been the main focus of our study because it serves as a unique yet simplified model for us to study the molecular regulation pathways during hematopoietic stem cell differentiation process at platelet lineage (dimension I) and the interactions of the regulatory factors inside each cell system (dimension II). In general scale, this system biology scheme represents an in-depth review for the studies we have done and a prospective blueprint to guide our future research directions.

1.6 Organization of the Thesis

The review of our study in human thrombocytosis disease under system biological scheme is discussed on 5 layers in order to reflect the specific approaches and perspectives we planned in our research. Accordingly, the dissertation structure consists of 5 chapters with each section corresponding to a specified layer. The first chapter introduces the biological background and significance of the study, detailing on a series of thrombocytosis conditions caused by myeloproliferative disorders (MPDs). In specific items, I discussed the etiology of the primary and secondary thrombocytosis and compared them to the other MPDs. A further exploration into the pathogenesis of the disease is then presented from two perspectives: 1) The genetic network control and 2) microRNA regulation underlying the development of thrombocytosis. C-1 is finalized with a discussion on the difference between ET and RT, which paves the way into C-2 where I introduced our study of a class prediction model using discriminant analysis to delineate two thrombocytosis conditions: ET vs RT. we discussed in details about the statistical method we used in deriving such a genotype prediction model from a set of biomarkers and reported the classification results. Subsequently in C-3, we explored the pathogenic process of human thrombocytosis from the perspective of microRNA regulatory network. The differentially expressed microRNAs between ET, RT and normal are reported with their biological implications discussed in detail. In the ensuing C-4, we further linked our findings of important microRNA features to their potential messenger RNA targets (miRNA-mRNA correlation) using both knowledge driven and data driven validation methods. Finally in C-5, we presented another layer of validation using additional proteomics data from a preliminary thrombocytosis study: the miRNA-mRNA mediated regulatory network was further linked to the corresponding protein output. We then discuss this validation analysis under several biological schemes to infer the potential regulation pathways. In conclusion, the entire thesis is based on the general hypothesis

that there exists a set of vibrant and functional regulatory pathways involving gene level control, microRNA regulation and the final protein output inside the platelet cell lineage, which underlies the hematopoietic cell differentiation and the pathological development of thrombocytosis. Our study aims to uncover these comprehensive pathogenic pathways associated with human thrombocytosis as well as its subtypes based on joint genetic, microRNA and proteomic data, and the analysis was conducted in five inter-related layers as presented in this dissertation.

Chapter 2:

Class Prediction Models of Thrombocytosis Based on Genomic Features

The lack of genetic criterion in characterizing primary thrombocythemia (ET) and secondary thrombocythemia (RT) remained a major restriction in the current understanding of thrombocytotic pathology [32]. Although in recent years some of the notable genetic mutations were uncovered and studied for their roles in the pathogenesis of ET or other types of MPDs, there are still a substantial fraction of ET cases attributed to unknown genetic causes (e.g. those JAK2V617F negative patients) . In addition, there are reported cases when both clonal (ET) and non-clonal (RT) thrombocytosis share a similar JAK2 mutation as co-existent medical conditions could complicate the traditional delineation between ET and RT cases [70]. Even based on the most recent WHO guidelines on myeloproliferative disorders [30], constant diagnostic refinement has to be factored in to classify cases with indistinguishable genetic features [71, 72]. For all these clinical ambiguity and diagnostic confusions, we felt the urgency to derive a definitive and positive test to classify ET, RT and normal controls based on their platelet

messenger RNA expression profiles. On the other hand, the JAK617F mutation, a widely recognized gene marker for ET may also be associated with a set of complex gene control network and we are therefore highly interested in the potential genetic characteristics behind the ET JAK2 mutations, too. By identifying a subset of gene markers that predicts between the JAK2617F+ and JAK2617F- we can further our understanding in the underlying role of JAK/STAT signaling pathway and the overall the pathological mechanism of thrombocytosis. Our research hypothesis is thus formulated as the following: there exists a set of predictive gene markers which delineates between thrombocytosis and normal controls as well as the two subtypes of thrombocytosis: ET & RT; moreover, there is another set of genetic markers that are highly associated with ET JAK2 status, thereby have predication power over ET JAK2 genotypes. In specific research schemes, the enucleate human platelets represent a unique model of hematopoietic regulatory control and they play a key role in various blood function (C-1.4). The molecular defects of platelets and their association with thrombocytosis are of particular interest in our investigation. In this regard, we studied the regulatory network of genetic biomarkers inside human platelets on a cohort of thrombocytosis patients set of thrombocytosis patients (ET and RT, as well as normal controls). All the biological experiments including microarray and qRT-PCR processing completed and provided courtesy of Dr. W. Bahou.

2.1 Experimental Design

We characterized the platelet transcript profiles from 126 patients (48 healthy controls, 38 RT, 40 ET [24 contained the Jak2V617F mutant allele; 16 were homozygous normal]) using stepwise variable selection method on custom cDNA microarrays to identify a subset of platelet transcripts which are associated with disease phenotypes. We subsequently quantified identified

biomarkers using quantitative real-time PCR (qRT-PCR) platform for an independent validation. The qRT-PCR experiment, commonly regarded as the 'gold standard' for transcript profile quantifications, was conducted on a separate cohort of patients with thrombocytosis. (**Figure 2-1**) The results suggested a good concordance between the two platforms, which also serves to further validate the selected biomarkers. Among the many goals we formulated in our experiment, the differentiation of clonal (ET) from non-clonal (RT) thrombocytosis has important diagnostic and therapeutic significance as discussed above. We therefore concentrated on identifying and validating gene markers that have optimal predictive power in classifying between the ET and RT conditions. After testing on several classifiers, a linear prediction algorithm based on nonparametric LDA is established to predict disease phenotypes using a subset of genetic profiles.

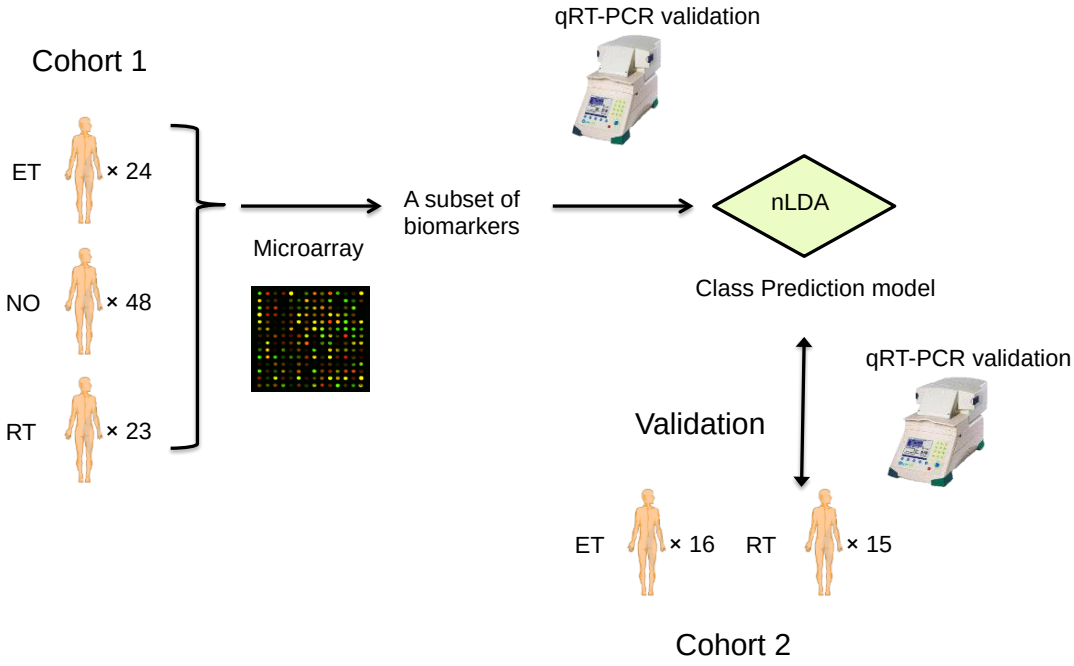

Figure 2-1: A schematic view of the experimental design of the thrombocytotic class prediction model. A total of 126 subjects were recruited into the study. The cohort 1 patient mix included 95 samples (ET [n = 24]; RT [n = 23 and healthy controls [n = 48]) and was used as the exploratory analysis. The stepwise variable selection method was applied to identify a subset of 11 gene markers for class prediction modeling (among 3 groups and between ET and RT). The result was further validated by qRT-PCR experiment using nonparametric LDA model (denoted as nLDA in figure). The second cohort consists of 16 ET and 15 RT patients. The transcripts profile of the 11 gene markers

were quantified from this group of patients and used for ET vs RT classification only. Microarray and qPCR processing completed and provided courtesy of Dr. W. Bahou.

We initially selected 432 platelet exclusive genes and hybridized them onto our custom oligonucleotide chips with Arabidopsis probes as normalization controls. The design of chip probes on our custom array platform resembles that of an Affymetrix Human genome 133A so that we can directly correlate our probes to actual transcripts using Affymetrix annotation. Besides, some of the leukocytes RNAs are also included to reflect the leukocyte genetic profiles during thrombocytosis. From the 432 gene candidates, we first narrow down to a subset of transcripts using feature selection method to differentiate among ET, RT and normal controls; Statistical model is subsequently established for supervised clustering of sample phenotypes and cross-platform consistency of platelet gene profiles between ET and RT was validated using qPCR (qRT-PCR). Another smaller subset is also selected to differentiate between JAK2V617F positive and negative among the ET samples. These genes are believed to have prediction power and are worth of exploring biologically. The general genetic marker selection and analytic pipeline is illustrated in **Figure 2-2**. A variety of platforms are involved in the above statistical analysis including: SPSS® (SPSS. Inc, Chicago, IL) software, Matlab® (The MathWorks, Inc. Natick, MA) and SAS® (SAS Institute Inc. Cary, NC).

2.2 Statistical Analysis

There are a total of 126 thrombocytosis patients and healthy controls involved in this study. This includes 48 healthy controls, 38 RTs and 40 ETs. The female to male ratio is about 3:2 in ET and RT and is roughly equal in normal controls. The samples are initially divided into two cohorts with the first one serving in biomarker identification and the second cohort to be

used in confirmatory study. The custom microarray is based on a two channel design where human reference RNA (Stratagene®, TX) are used in Cy5 and patient RNAs are deposited in Cy3. The fluorescence intensity of each probe spot was quantified in GenePix Pro® 6 Software (Molecular Devices Inc., Sunnyvale, CA) and the array quality was examined by between-sample spearman correlations. Qualified arrays are passed down to further analysis [73]. A preprocessing procedure was then carried out by taking the ratio of Cy3/Cy5 to represent the value of each transcript, followed by an array-wise median alignment using 5 house-keeping Arabidopsis genes. Besides, the genes are also filtered by their expression values as probes with fluorescence intensities <10 in more than 70% of probes were excluded from further analysis. Because the data we dealt with are actually ratios between Cy3 and Cy5, negative or extremely small values in either channel are treated and replaced with a constant cutoff. In the subsequent statistical analysis, we designed our pipeline based on a primary hypothesis consisting of several aims: In the first aim, we are interested in the gender effect and its interaction with disease phenotypes. A non-parametric two-way ANOVA (Friedman test) was performed to examine the two effects. The result indicates that at the significance level of 0.05 only genes with disease effect can be detected. We then turned to Kruskal-Wallis, the non-parametric one-way ANOVA to identify differentially expressed genes among the three cohorts. As our primary interest lies in the difference between ET and RT, we narrow down our scope to this pairwise comparison: a nonparametric Wilcoxon rank-sum test was used to examine median differences between two independent samples, including gender effects, the comparison between ET and RT subjects, as well as comparison within ET subjects by Jak2 genotypes. All statistical analysis results were evaluated at the significance level of 0.05 (two-sided).

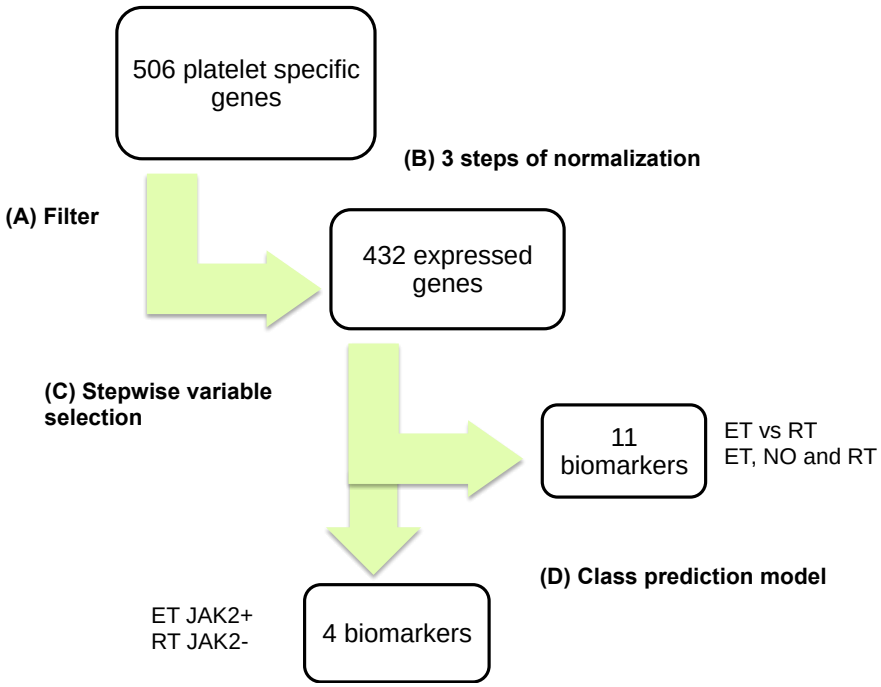

Figure 2-2: The summary of the analysis pipeline from the perspective of genomic dimension reduction process. The analysis starts with 506 platelet-restricted genes that were hybridized onto the custom microarray. (A) Preliminary filter: Genes with fluorescence intensities less than 10 (the default background threshold in GeneSpring ®) in more than 70% of arrays were excluded from further analysis. (B) Three normalizing procedures were applied: 1) per-spot 2) per-gene and 3) per-array. (C) Stepwise variable selection method was used to identify the best subset of gene features that can optimally delineate among three groups and between ET vs RT. The classification algorithm (D) is based on the classifier of nonparametric linear discriminant analysis. Both Leave-one-out and cross-platform (cohort 1 as training set while cohort 2 as testing set) validation scheme was used in the study.

To select features (biomarkers) for the establishment of a classification model, stepwise discriminant analysis was used to filter the candidate genes and the results are validated by linear discriminant analysis (LDA) with a leave-one-out setting. With all the analysis, 11 biomarkers were eventually determined as genetic features with predictive power for the following classification model. A leave-one-out cross validation procedure was used with the 11 biomarkers to delineate among ET, RT and Normal. Besides, it's also used in differentiating between ET and RT samples. In this step, we tried several classifiers and compared their performance in terms of their prediction accuracy. These included parametric LDA, nonparametric LDA, quadratic discriminant analysis (QDA), SVM, random forest and KNN.

The nonparametric LDA as a simple linear classifier was shown to outperform others on default settings with the best prediction rate. Specifically, at two cohorts level (ET vs RT) an overall classification success rate of 93.6% was achieved, and for three cohorts delineation, 86.3% of the total samples are correctly classified (ET, RT and Normal).

This classifier is based on Bayes theorem:

$$p(t \mid x) = \frac{q_t f_t(x)}{f(x)} \tag{2-1}$$

in which a prior probability of group membership is assumed. And *p(t|x)* is, the probability of x belonging to group t is evaluated by the formula above. q_t is the prior probability of membership in group t; f_t is the probability density function of group t; *f(x)* is the estimated unconditional density at x and $f_t(x)$ is the group-specific density estimate at x from group t: it is estimated by

$$f(x) = \frac{1}{n_t} \sum_y K_t(x - y) \tag{2-2}$$

In which the kernel K_t is calculated by

$$K_t(z) = \frac{1}{c_o(t)} \exp(-\frac{1}{2r^2} z^t V_t^{-1} z) \tag{2-3}$$

Where $c_o(t) = (2\pi)^{\frac{p}{2}} r^p |V_t|^{\frac{1}{2}}$ and V_t is the pooled covariance matrix. x in the above formula is a p-dimensional vector containing quantitative variables of an observation; p is the pooled covariance matrix; t is the grouping assignment and n_t is the number of training set observations in group t.

2.3 Results and Discussion

In the first step, we did not find any genes with significant gender effect using one-way ANOVA. On the class prediction model we derived, a fairly high classification rate was achieved

using 11 biomarkers we identified from stepwise variable selection process (**Table 2-1**). In addition to differentiating between ET and RT phenotypes, 4 genetic features were also identified to delineate on Jak2V617F genotypes within the ET samples: the classification rate for this discretion is quite good, too.

Table 2-1: The 11 biomarkers identified from ET vs RT comparison and the 4 biomarkers from ET JAK2617F mutation positive and negative comparison.

11 Biomarkers between ET vs RT		4 biomarkers on JAK2617F Status	
Gene Names	**Description**	**Gene Names**	**Description**
ACOT7	acyl-CoA thioesterase 7	HIST1H1A	Histone cluster1, H1a
APP	amyloid beta (A4) precursor protein	SRP72	Signal recognition particle 72-like
CLEC1B	C-type lectin domain family 1, member B	C20orf103	Chromosome 20 open reading frame 103
CTNS	cystinosis, nephropathic	CRYM	Crystalline, mu
H3F3A/H3F3B	H3 histone, family 3B (H3.3B)		
HIST1H2AG	histone cluster 1, H2ag		
LAPTM4B	lysosomal protein transmembrane 4 beta		
NGFRAP1	nerve growth factor receptor (TNFRSF16) associated protein 1		
TGFB2	transforming growth factor, beta 2		
TPM1	tropomyosin 1 (alpha)		
WASF3	WAS protein family, member 3		

Besides, the fidelity of all the microarray results are subsequently verified by a follow-up quantitative PCR experiment in which the same set of statistical procedures were carried out to validate the predictive power of the selected biomarkers on our second cohort of patients [73]. All the class prediction results are presented (**Table 2-2**) to reflect the discriminant power of selected biomarkers. The primary goal of our analysis is to derive a definitive class prediction model of different thrombocytosis phenotypes for the sake of easier clinical diagnosis and a

better understanding of the genetic mechanism behind the disease pathogenesis. Moreover, the widely reported JAK2V617F mutation may have associations or concurrence effect with genes outside of JAK-STAT signaling pathway [74, 75], which is also in our interest to explore. Back to the beginning of this study, we started our class prediction models by selecting most relevant biomarkers among the 3 disease groups and between ET vs RT. We opted to use stepwise variable selection of LDA since this is a wrapper based feature selection algorithm [76] which is closely related to the final class prediction process. After identifying the significant biomarkers, the next step is to seek appropriate classifiers for the supervised clustering analysis (classification). Popular classifiers include LDA, quadratic discriminant analysis (QDA), support vector machines (SVM), Neural networks and random forest method, etc. To maximize the prediction power of our gene list, we decided to examine our features from simple and linear classifiers such as LDA; if these methods failed us, we would like to increase the classifying power and try more advanced classifiers. Fortunately in our study, the nonparametric LDA alone already suited our needs and provided high prediction accuracy under a leave-one-out setting. In the end, we established our classification model with nonparametric LDA classifier for both ET-RT and Jak2 mutation settings. These biomarkers were later confirmed by quantitative PCR analysis on another independent cohort of patients.

Table 2-2 Class prediction results using identified biomarkers.

Prediction	NO. of biomarkers	Cohort I (Microarray)	Cohort II (Quantitative PCR)
Three Way Classification	11	86.3%	Not Applicable
Two Way Classification	11	93.6%	87.1%
Jak2 class prediction	4	87%	72%

Chapter 3:

MicroRNA regulation in Human Thrombocytosis

The microRNAs (miRNA) consist of a large family of ~22 nucleotide-long RNAs which have emerged as key post-transcriptional regulatory factors in plants and metazoans. Since the discovery of the first miRNA family, lin-4 back in 1993 [77], our understanding of genetic regulatory mechanisms has been totally revolutionized. Major functional studies showed that the miRNA regulatory network is powerful and widespread and is involved in almost every cellular process [78, 79]. The alteration or defects in certain miRNAs or their expression levels are found to be closely related to many human pathologies including hematological malignancies [78, 80]. Based on accumulated evidences of miRNAs' regulatory role in similar blood cell system (see C-1.3), we have thus proposed our hypothesis that there exists a subset of important miRNA markers differentially expressed between thrombocytosis and normal controls and their regulation mechanism is highly associated with the pathological development of thrombocytosis. The miRNAs themselves are no different than other protein coding RNAs except that they never entered translational process to produce functional proteins. Instead, they interact with other

messenger RNAs and primarily suppress the expressional activity of their target mRNAs [81]. There are two biological modes in the miRNA-to-mRNA silencing process: one is the miRNA induced cleavage and degradation of target mRNAs; the other mechanism involves target mRNAs not to be actively destroyed but rather, to be destablized by deadenylation and decapping [82, 83] which results in a total or partially translational repression effect. The miRNA induced silencing regulation occurs through the direct interaction between microRNA and its target mRNAs. This process takes place when the miRNA binds its targets by base-paring with the 3` untranslated region (3 UTRs) of the mRNA sequence. The binding are in perfect complementary pairs in plants but relies on far more complex paring mechanisms in animal cells [84, 85]. In this regard, the miRNA target prediction process, while simple and straightforward in plants [86], is highly diverse and sophisticated in animal eukaryotic cells [87, 88]. Currently, there are several popular miRNA target prediction tools available on line for public use, these include DIANA-microT [89], ElMMo [90], miRanda [91], TargetScan [92, 93] and PicTar [94], etc. These tools differ in their concentrations on the various factors involved in miRNA-mRNA interactions. However, comparative studies have suggested that inconsistency and false positives are prevelant in current miRNA target prediction algorithms [95-97]. Therefore, the common practice in identifying miRNA-mRNA regulatory pairs requires the experimental validation of mRNA targets derived computationally, which is the next step in our research plan. In this study, all the biological experiments including microarray and qRT-PCR processing completed and provided courtesy of Dr. W. Bahou.

3.1 miRNA regulation in hematopoitic functions

MicroRNAs in hematopoietic cell differentiations

The regulatory network of microRNAs in the human hematopoietic system is prevelant and diverse. It was already found that the miRNA expression patterns are very different between hematopoietic and non-hematopoietic cell lines, suggesting an active role the miRNAs may play in the functional process of hematopoietic cell system [98]. Examples of this category include miRNAs' regulatory roles in multipotent stem cell differentiation process [99-101]. Three important miRNAs: miR-142, miR-181 and miR-223 were found to be dynamically invovled in the process of early hematopoiesis and lineage commitment [57, 102]; In addition, miR-150 was proved to regulate megakaryocyte-erythrocyte progenitors during hematopoiesis lineage specification [103]. The downregulation of miR-221 and miR-222 was observed at different stagets of erythrocyte maturation and differentiation [104]. In different megakaryocyte cell lines, The miR-10a inversely interacted with HOXA1, which forms an important pathway during megakaryocytic differentiation [58]. Theoretical considerations are given to the miRNA mediated cell lineage separation mechanism, one appealing explanation states that miRNAs induced translational represssion eliminates trivial expression of unnecessary genes and hence refine unilineage gene expression during hematopoiesis. Additional studies demonstrated that miRNAs regulate not only differentiation but also activity of hematopoietic cells by targeting growth factor receptors and transcription factors in cellular responses to external stimuli [105].

MicroRNAs in hematopoietic malignancies

On the other hand, the deregulation of miRNA in different hematopoitic cell lines were also strongly associated with the pathogenisis of various blood malignancies. Examples of this include the reduced level of miR-15a and miR-16-1 in at least 75% chronic lymphocytic leukemia cases for their potential targets at 13q14 deletions [106]; The overexpression of miR-28 found in a fraction of MPN patients [107] and the oncogenic role of miR-17-92 polycistron and miR-155 within B-cell integration cluster for their involvement in malignant lymphoma [108,

109]. Other exploratory studies also showed the aberrant profiles of some hematopoetic-specific miRNAs (i.e. miR-150) in clonal myeloproliferative disorder (MPD) development [42, 43], as well as the down-regulation of neutrophilic miR-133a, miR-1 in patients with polycythemia vera (PV) and essential thrombocythosis (ET) from Slezak et al report [44]. All these studies have suggested the microRNA involvement in a variety of hematopoietic malignancies including chronic MPD and possibly reactive thrombocytosis, which are of our primary research interest.

MicroRNA control inside human platelets

The human platelet represents a unique yet important cellular structure in the hematopoitic system. They are widely involved in blood funcitonal pathways and mediates multiple thrombotic processes (discussed in Chapter 1.3). On the other hand, the molecular defects causally implicated in platelet-associated bleeding or thrombotic dysfunctionality are closely associated with the deregulation of platelet internal control. For example, some of the defects are observed resulting in the absence of glycoproteins (GP) IIb/IIIa [IIb3; Glanzmann thrombasthenia] or the GPIb-IX-V complex [Bernard-Soulier syndrome] in human platelets [110]. In other cases, Protein over-expression may favor platelet activation and thrombus formation, which provides conceptual support for the presence of biomarkers that may result in increased thrombotic susceptibility. In general, genetic or protein biomarkers are closely linked to clinical platelet phenotypes. Platelets are produced in thrombopoiesis by budding off from megakaryocytes and are generally less complex than nucleated cells in its transcriptome [64, 65]. Minimal translational activities are observed in quiescent platelets as some of its mRNAs are directly inherited from megakaryocytes. Recent studies have confirmed the existence of a few post-transcriptional regulatory networks including a set of diverse and functional microRNA pathways [66]. These platelet microRNAs interact with the remaining mRNAs and mediate platelet translational activities in a variety of normal and pathological environment [66]. Our

study is focused on the deregulation of miRNA networks in thrombocytosis environment, this include 1) identification of differentially expressed miRNAs between different pathological conditions; 2) discovery and validation of gene targets from the selected miRNAs; 3) The exploratory analysis and confirmation of complete miRNA-mRNA-protein pathways underlying the pathogenesis of thrombocytosis condition.

3.2 Experimental Design

Sample preparation

A total of 80 subjects were used in the study with comparable amount of samples assigned into 3 cohorts: ET (n = 28) RT (n = 22) and healthy controls (n = 30), based on initial clinical diagnosis. The gender and age distribution were paralleled with the patient cohort assignments by a male : female ratio of 1:1.32 and patient ages ranging from 23 to 78 years. In our research plan, we have specifically focused on the miRNA regulatory network inside human platelets, therefore the platelets we collected from peripheral blood samples of the donor are subjected to strict purification process. In the sample preparation stage, leukocytes and platelets were initially obtained and purified following a protocol described in previous studies [32, 64]. Leukocytes contamination in platelets was confined to a minimal level of 1×10-5 by cell counts. Platelet RNA were isolated using Trizol reagents following recommended protocols [111]. Subsequent quality control and quantification of platelet RNAs were conducted on Agilent 2100 bioanalyzer. The allele frequency of the JAK2V617F mutation (exon 12, 1849G3T trans-version) was quantified as described elsewhere [112]. Samples with more than 5% mutated sequences (from the DNA/RNA nucleic acids pool) detected in this method were defined as JAK2 positive in our study.

Platelet microRNA profiling

Platelet miRNA expression profiling was conducted using Agilent Human miRNA Microarray Kit V3. This Agilent miRNA assay incorporates 866 human and 89 viral microRNAs from the Sanger mirBase v 12.0. For each platelet sample, 20 ng total RNA fluorescence-labeled with Cyanine 3-pCp were hybridized with the miRNA array and further processed following the manufacturer's instructions. The slides were then scanned using an Agilent High-Resolution Microarray Scanner® with the minimum detection level of 0.05 chromophores per square micron. Subsequently, the array images were quantified and processed with Agilent Feature Extraction Software (version 10.7). The final output data sets were saved into separate Microsoft Excel files for further statistical analysis. Preliminary quality control tests in each individual sample with default software setting were also performed. Within group coefficient of variation (CV) of replicate probes were calculated at this step to evaluate intra-array reproducibility [113]; Samples resulting in CV values below Agilent default standard (CV <15%) were re-arrayed for better quality until they become qualified for the acceptable requirement [114, 115].

qRT-PCR Validation

In a subsequent validation step, a list of miRNAs selected from the microarray profiling results was also arrayed on the qRT-PCR platform to verify their expression levels in each pathological condition. To standardize the expression profiles of microRNAs in qRT-PCR analysis, two miRNAs (hsa-mir-22 and hsa-mir-720) were selected as house-keeping controls from the Agilent microarray data using minimum across-array coefficient of variances criterion (top 2 selected). The abundance of the selected microRNAs was determined from triplicate assays arrayed in parallel for each primer. The median of triplicates was used to represent the expression level of these microRNA features. Relative fold changes of each primer standardized

by controls were calculated by $_{\Delta}$Ct method and reported as $2^{-\Delta Ct}$ [116] in the following statistical analysis. Spearman correlation was calculated between the microarray and qPCR platform on a per-miRNA base and boxplots are generated to reflect the relationship between thrombocytosis groups across platforms. The entire experimental scheme is illustrated in **Figure 3-1**:

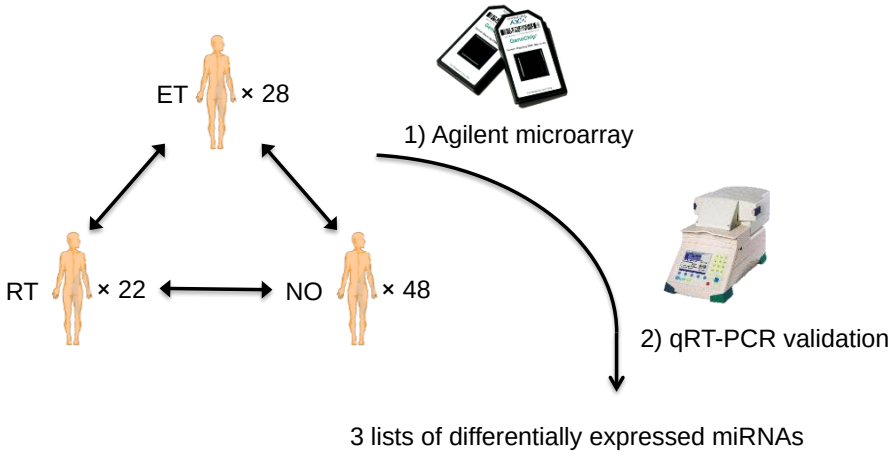

Figure 3-1: A schematic view of the experimental design in thrombocytotic microRNA profiling and comparison. A total of 80 subjects were recruited into the study including three patient groups (ET [n = 28]; RT [n = 22 and healthy controls [n = 30]). The Significant Analysis of Microarray (SAM) was applied to identify differentially expressed miRNAs using transcript profiles from Agilent microarray platform. An ensuing qRT-PCR experiment was performed to validate the results from microarray experiment. Three lists of miRNAs were obtained in this way as important miRNA markers between the three thrombocytotic phenotypes. Both microarray and qRT-PCR processing completed and provided courtesy of Dr. W. Bahou.

3.3 Data Preprocessing and Statistical Analysis

Data preprocessing and filtering

From the Agilent microarray experiment, the miRNA expression values were initially obtained from Agilent Feature Extraction Software after microarray image scanning and quantification process. The output data were then normalized and preprocessed using statistical software R 2.11 with Bioconductor® packages (http://www.bioconductor.org/). In the initial preprocessing step, a filtering procedure was applied to remove miRNAs not expressed in human platelets. This procedure was performed using the default spot-flagging information provided by

Agilent Feature Extraction Software [117]. Specifically, an absence-call standard was enforced to retain microRNA entries with more than 70% non-absent instances in any of the 3 groups (ET, Normal and RT). The other miRNA entries were removed from subsequent analysis. In addition, potential outlying values were also detected and deleted using a criterion of 3 standard deviation units from the mean for each miRNA expression. The resultant missing values were subsequently imputed with KNN algorithm using 10 neighboring seeds [118].

Normalization and preliminary analysis

After arriving at a filtered data set, a quantile normalization procedure was applied to adjust between-array variations [119, 120] in both microarray and qPCR platform. A following two-way ANOVA analysis was performed to examine both the gender effects and its interaction with phenotypic groups using a p value cutoff of 0.05. To determine the number of expressed platelet miRNA in 3 phenotypic cohorts, a simple filter similar to the aforementioned absent-call standard was used: platelet miRNA containing less than 30% absent calls in the specified group is considered a meaningful expression for that patient cohort.

Unsupervised Clustering and SAM Analysis

Subsequently, an unsupervised hierarchical clustering of average linkage with 1 – Spearman correlation was performed to generate the dendrograms for both arrays and median-centered miRNA profiles. The data was centered to mean 0 and standard deviation = 1 before clustering and a preliminary cluster dissection was determined by visual inspection. To explore the difference between 2 individual cohorts, Significance Analysis of Microarrays (SAM) (http://www-stat.stanford.edu/Btibs/SAM/) was applied comparing ET with Control, ET with RT and RT with controls sequentially with a cutoff >= 2 fold change and false discovery rate (FDR)

<= 0.05. The differentially expressed microRNA signatures between ET and control groups were selected for further confirmatory analysis and miRNA targets prediction study.

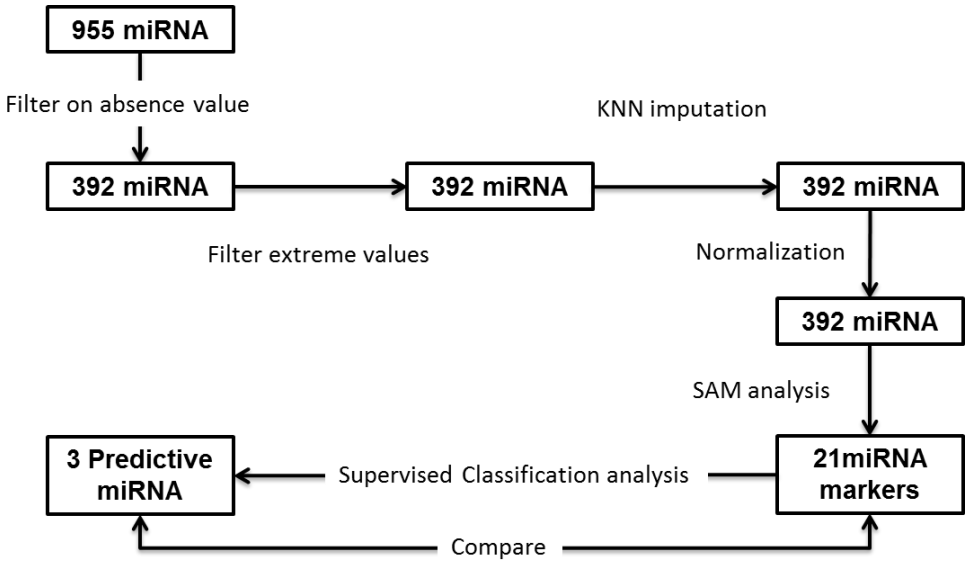

Figure 3-2: A schematic view of the preprocessing and statistical procedures for thrombocytotic microRNA data analysis. The analysis starts with 955 miRNAs based on Agilent miRNA Microarray Kit V3 platform. A preliminary filter was applied to remove miRNAs that are not expressed in the platelet system (only candidates with more than 70% non-absent instances in any of the 3 groups are retained). Extreme values (> 2 standard deviation from group mean) are also removed, followed by a KNN imputation procedure. A quantile normalization procedure was applied to adjust between array variance and the final SAM and supervised classification analysis were used to identify important miRNA features with their results compared to each other.

3.4 Results from Preliminary Analysis

General miRNA profiles and gender effects

From Agilent microarray data, 392 out of 955 microRNAs are identified as nontrivially expressed after the filter step and normalization adjustment. In addition, One ET sample (THR 127) is removed from the analysis due to its failure to comply with absence-call standard. The initial gender effect analysis identified 16 miRNAs with significant gender difference independent from group effect with a threshold of ANOVA P <= 0.05. Among these features, 3

miRNAs exhibits gender effect both within and across phenotypic groups, they are hsa-miR-143*, hsa-miR-497 and hsa-miR-193a-3p. Considering that the occurrence of chronic MPDs are highly relevant to gender difference [1, 5], it is not surprising to observe a handful of miRNAs differentially expressed between two genders (**Table 3-1**). After the preprocessing step, a total of 392 miRNAs × 79 samples (ET [n = 27]; NO [n = 30]; RT [n = 22]) entered further statistical analysis. The initial evaluation of expressed platelet miRNAs using absence-call standard in each of the 3 phenotypic groups revealed distinctive patterns across ET, Normal and RT cohorts.

Table 3-1: miRNAs with significant gender effects (ANOVA P Value)

miRNA	Gender Effect	Gender × Genotypes
hsa-miR-27a	0.005	0.159
**hsa-miR-143*	0.007	0.029
hsa-miR-574-3p	0.015	0.668
hsa-miR-150	0.016	0.340
hsa-miR-451	0.017	0.361
**hsa-miR-193a-3p	0.021	0.049
hsa-miR-361-3p	0.022	0.570
hsa-miR-145	0.024	0.457
hsa-miR-29a*	0.028	0.749
hsa-miR-33a	0.032	0.618
hsa-miR-143	0.034	0.769
hsa-let-7c	0.040	0.761
hsa-miR-486-3p	0.040	0.608
**hsa-miR-497	0.040	0.036
hsa-miR-16-2*	0.041	0.201
hsa-miR-449a	0.044	0.621

** hsa-miR-143*, hsa-miR-193a-3p and hsa-miR-497 have both significant gender effect and a non-trivial interaction between the gender and genotypes among the examined patients.

Unsupervised Clustering Analysis

The dendrogram from two-way unsupervised clustering analysis showed a general expression pattern across 3 groups (**Figure 3-3**) which implies a more distinctive delineation between ET and the other two groups (controls and RT). With a direct visual observation, 18 out of 27 ET samples (66.7%) were assigned to the same cluster while RT and controls were randomly mingled in other clusters. We also investigated ET genotypes sub-stratified by the absence (GG) or presence (GT/TT) of the $JAK2^{V617F}$ mutation. Based on the unsupervised clustering dendrogram, 13 out 19 ET samples (68.4%) with $JAK2^{V617F}$ (GT/TT) were assigned in the ET enriched cluster as compared to 5 out 8 (62.5%) ET JAK2 wild-type samples in the same cluster. Due to the limited resource available and the fact that ET disease, being a chronic MPDs, may have a much stronger genetic and epigenetic regulatory background [2, 20, 61], we decided to further explore the difference between ET patient cohort and the healthy control group.

SAM Analysis for differentially expressed miRNAs

Using SAM analysis, 21 microRNA signatures were found differentially expressed between ET and NO cohorts. Among these miRNA features, 10 miRNA are down-regulated and 11 are up-regulated in ET cohort compared to normal controls based on the Agilent microarray data. We also observed that three miRNA complementary sequences (6 features) exhibiting a fold change higher than 3 between two patient groups, including the hsa-miR-490 family, the hsa-miR-34 family and the hsa-miR-9 family. We also investigated the group differences between ET vs RT and NO vs RT on the same set of miRNAs. The general group comparison information of the identified 21 features is included in **Table 3-2** (SAM P values and fold changes are shown).

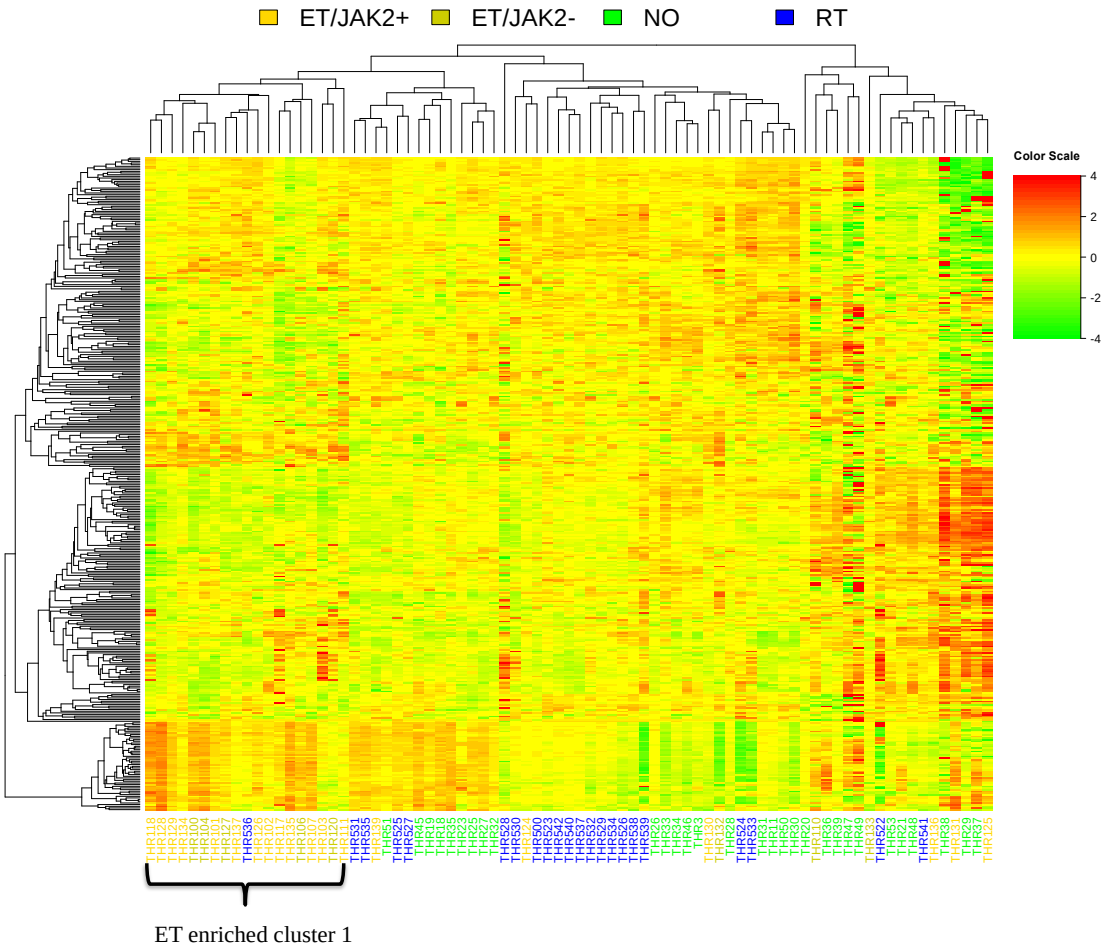

Figure 3-3: Two-way unsupervised clustering dendrogram of ET (Jak2+ and Jak2-), RT and normal controls. The two-way hierarchical clustering analysis rearranges microRNAs and patient samples into the dendrogram based on their correlation proximity (Average linkage is used). miRNA expressions were standardized to mean = 0, standard deviation = 1 prior to clustering; Besides, the data range were confined to -4 ~ +4, reflected in the heat map color gradient. Sample labels were color-coded with prior diagnosis information -- ET cohorts were substratefied into $JAK2^{V617F}$ mutation (ET/JAK2+ in figure) and JAK2 wild-type (ET/JAK2- in figure). Normal controls were labeled as NO and RT represents RT samples. It is obviously shown that the ET samples (in two yellow colors) are primarily enriched in the leftist cluster in the dendrogram. Figure is created using R software.

qRT-PCR Confirmatory Analysis

A subset of 20 microRNAs from the 21 identified biomarkers were arrayed using quantitative real time PCR platform as part of validation study to Agilent microarray data. The consistency of microarray and qPCR platforms was evaluated using Spearman correlation coefficient for each miRNA signature as well as a direct group-wise comparison on cohort median and inter-quantile ranges (**Figure 3-4 A**). A high correlation coefficient was observed between the two methods both in general (r = 0.69, p = 0.024) and at per-miRNA level (r = 0.26

~ 0.92, $p = 1 \times 10^{-7}$ ~ 0.28) (**Figure 3-4 C**). 19 out of the 20 miRNA features exhibit concordant group-medians with consistent data profiles (**Figure 3-4 A & B**). The concordant observation between microarray and qPRT-PCR data verified the fidelity of the 20 differentially expressed biomarkers. It also paved the road for our further pursuit into the subset of these features.

Table 3-2: miRNA signatures identified by SAM analysis between two cohorts**

miRNA Signatures	FC: NO/RT (P value)	FC: ET/RT (P Value)	FC: ET/NO (P Value)
hsa-miR-551b	1.069(0.45)	0.368(<0.001)	0.344(<0.001)
hsa-miR-9	0.835(0.225)	3.153(<0.001)	3.776(<0.001)
hsa-miR-29a	0.754(0.026)	1.532(0.002)	2.032(<0.001)
hsa-miR-490-5p	0.456(0.009)	4.367(<0.001)	9.587(<0.001)
hsa-miR-182	2.031(0.002)	0.705(0.027)	0.347(<0.001)
hsa-miR-1301	1.625(0.001)	0.781(0.031)	0.481(<0.001)
hsa-miR-1274a	0.729(0.037)	1.502(0.003)	2.059(<0.001)
hsa-miR-10a	1.106(0.391)	0.542(0.001)	0.49(0.001)
hsa-miR-9*	0.798(0.195)	2.716(0.001)	3.402(<0.001)
hsa-miR-424	0.408(0.001)	1.143(0.089)	2.802(<0.001)
hsa-miR-144	0.519(0.032)	0.255(<0.001)	0.492(0.018)
hsa-let-7d*	1.956(0.001)	0.917(0.103)	0.469(<0.001)
hsa-miR-148a	0.427(<0.001)	0.907(0.1)	2.127(<0.001)
hsa-miR-34a	0.599(0.003)	2.953(<0.001)	4.928(<0.001)
hsa-miR-886-3p	0.629(0.015)	1.316(0.025)	2.091(<0.001)
hsa-miR-181c	1.53(0.024)	0.735(0.009)	0.481(<0.001)
hsa-miR-490-3p	0.595(0.05)	3.572(<0.001)	6.001(<0.001)
hsa-miR-150	1.494(0.262)	0.565(0.052)	0.378(0.007)
hsa-miR-34b*	0.824(0.266)	3.407(<0.001)	4.136(<0.001)
hsa-miR-181c*	1.785(0.004)	0.792(0.039)	0.444(<0.001)
hsa-miR-144*	0.747(0.177)	0.339(<0.001)	0.453(0.001)

**The P values shown in the table are derived from SAM FDRs. The fold changes are calculated in log2 scale and revert back to original values (default SAM settings).

3.5 Supervised classification

Methodology

From the Agilent microarray profiling analysis, a subset of miRNAs was selected and used to delineate between ET and normal cohorts. A further exhaustive searching algorithm was adopted to select a combination of features with optimal discriminant power from the subset of microRNA candidates. This process started with a random selection of 2~3 miRNA biomarkers being tested on their prediction power using 3 popular classifiers: 1) Support Vector Machine (SVM) with Gaussian kernel, 2) Random Forest (RF) and 3) Linear Discriminant Analysis (LDA). The initial screening was conducted with 3 fold cross-validation for higher computational speed. The parameters of each classifier was tuned using a wrapper method [121] and the algorithms were applied to both Agilent microarray and qPCR data. After the initial selection, a set of 'seed miRNAs' with prediction accuracy higher than 90% was retained for a precise classification analysis using leave-one-out cross-validation scheme [32, 122]. The total number of miRNA biomarkers allowed in the supervised classification analysis was confined to a range of 2~4 to prevent over-fitting problem while the prediction rate criterion is set at a reasonably high level to ensure a classification accuracy. Based on this algorithm, a small list of important features was selected from discriminating among 57 instances (30 controls and 27 ET samples) on both microarray and qPCR data. After two stages of evaluation (raw & precise), an optimal feature subset is selected according to the overall classification accuracy for both data sets on all 3 classifiers.

Results of Classification

The initial exhaustive searching algorithm identified 19 combinations of 3 discriminant miRNA biomarkers which generate prediction accuracy higher than 90% in both microarray and

qPCR data sets. With a further precise analysis using leave-one-out cross validation scheme, we eventually selected 3 microRNAs: hsa-miR-10a, hsa-miR-148a and hsa-miR-490-5p, as they resulted in the highest prediction accuracy on both platforms (**Table 3-3**). With these 3 biomarkers, 27 ET samples and 30 Normal controls were well segregated by 3 different classifiers – SVM, Random-Forest and LDA on microarray data, achieving an average classification accuracy of 86% ~ 94.7% (**Table 3-3, left**). The accuracy of discriminant analysis was subsequently verified on qPCR data between ET (n = 10) and Normal (n =10) cohorts. In this validation analysis, the diagnosis information of 19 samples (out of 20) was correctly predicted, resulting in a 90% ~ 95% accuracy rate (**Table 3-3, right**). Among the 3 classifiers, the linear discriminant algorithm (LDA) can result in the best prediction rate on both platforms. (Microarray 94.7% and qPCR 95%), therefore we prefer this classifier for future prediction analysis. In both exploratory and validation stage, the objectivity of discriminant analysis was established using a leave-one-out cross-validation protocol in which each instance is classified by profiles derived from all subjects excluding itself. For classifiers that require parameter tuning, the optimal settings were initially determined from microarray data and directly passed to the same classifier used in qPCR confirmatory analysis, thereby minimizing the inherent classification bias of observational collinearity between training and testing sets. In general scale, the specificity of the each discriminant analysis is better than its sensitivity while the overall prediction accuracy is quite high which reflects the prediction potential of the 3 miRNA features.

Figure 3-4: Boxplots and scatter plots comparing qRT-PCR and microarray microRNA data. (A) Boxplots showing the inter-group relations and intra-group variance to each of the 20 miRNA profiles from microarray data. Each group contains 10 samples (ET [n = 10]; NO [n = 10]) randomly selected from the initial 80 patients pool. Color coded boxes represent the within-group interquartile range, whereas the attached vertical dash lines truncated by upper and lower limits indicates the 95% confidence level; Group median is denoted by the horizontal bar in the middle of boxes. (B) Boxplots showing the miRNA expression profiles from qRT-PCR data. The notation and group arrangement are the same as in (A). (C) Scatter plots showing concordance between the two platforms on each of the 20 miRNA features. The x-axis and y-axis represent the microarray and qRT-PCR data respectively in unified scale (standardized to mean = 0; std = 1). Spearman correlation is given on top of each plot and groups are differentiated by different colors. Figure is generated from R software. Both microarray and qRT-PCR processing completed and provided courtesy of Dr. W. Bahou.

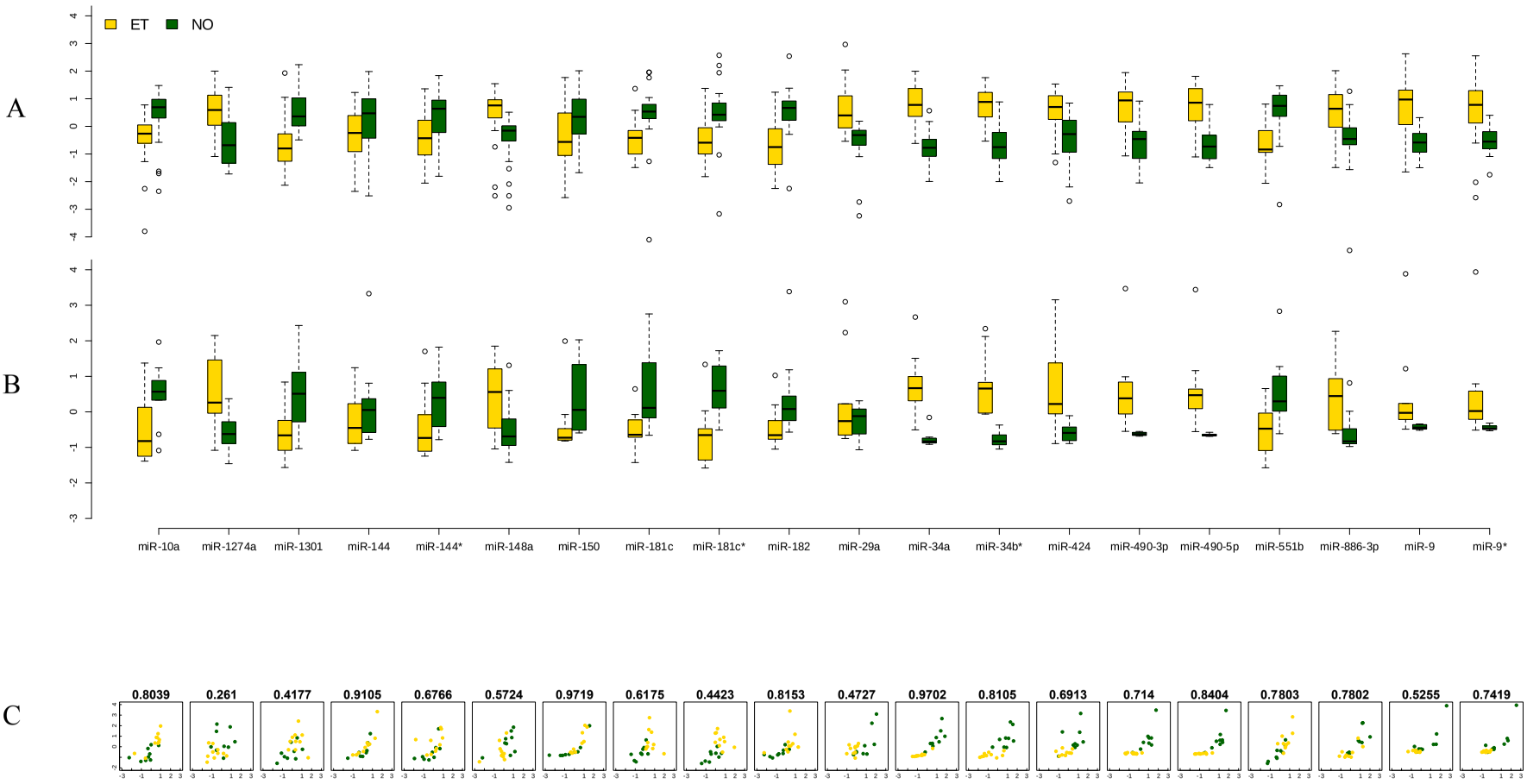
A
B
C
ET
NO
miR-10a
miR-1274a
miR-1301
miR-144
miR-144*
miR-148a
miR-150
miR-181c
miR-181c*
miR-182
miR-29a
miR-34a
miR-34b*
miR-424
miR-490-3p
miR-490-5p
miR-551b
miR-886-3p
miR-9
miR-9*
0.8039
0.261
0.4177
0.9105
0.6766
0.5724
0.9719
0.6175
0.4423
0.8153
0.4727
0.9702
0.8105
0.6913
0.714
0.8404
0.7803
0.7802
0.5255
0.7419

Table 3-3: Class prediction results using 3 biomarkers*

	Class Prediction							
	Agilent microarray				qPCR data			
Phenotypic class	**SVM**	**RF**	**LDA**	**Total**	**SVM**	**RF**	**LDA**	**Total**
ET	24 (88.9%)	24 (88.9%)	25 (92.5%)	27	9 (90%)	9 (90%)	9 (90%)	10
NO	30 (100%)	25 (93.3%)	29 (96.7%)	30	9 (90%)	10 (100%)	10 (100%)	10
Average	54 (94.7%)	49 (86.0%)	54 (94.7%)	57	18 (90%)	19 (95%)	19 (95%)	20

*Classification accuracy was presented in the form of number of corrected predicted samples and the accuracy rate (in parenthesis). SVM, RF and LDA represent Support Vector Machine, Random Forest and Linear Discriminant Analysis. ET denotes essential thrombocythaemia cohort and NO refer to normal controls.

3.6 Discussion and Conclusion

In recent years, scientists have increasingly realized the complexity of biological regulatory network in pathological cellular environment. From the system biology point of view, genomics, epigenomics (including interferomics), transcriptomics and translatomics form the complete and primary regulatory pathway inside the cell [123]. This includes gene expression at DNA level, mRNA production in mediated transcription process, the post-transcriptional regulation of microRNAs and the final translational production of peptides that subsequently fold into functional proteins. For diseases having a strong genetic background, the traditional concentration in molecular genomics is far from being able to provide a comprehensive depiction of the pathogenic process and the disease generation mechanism. Epigenetic factors such as chromatin structure, DNA methylation and RNA interference have such a huge influence in cellular regulatory network that many diseases actually emerge from the deregulation of these factors [124, 125]. In the particular case of human thrombocytosis, we have identified genetic features differentially expressed between disease and control group as well as between two pathological subtypes. A further examination of microRNA interference on cell regulatory network in patient group provides additional insights into the cellular control mechanism under

disease environment. A previous study has set the foundation of our exploratory approach in studying miRNA profiles in platelets model [66]. In this specific study conducted by Landry et al, an abundant and diverse array of miRNAs was identified in human platelets, which are believed to be inherited from their parental megakaryocytes. These miRNAs remain active inside platelets and mediate RNA silencing by coordinating with Dicer and Ago2 complexes [66, 80]. In our analysis of miRNA profiles in human platelet, a total of 392 miRNAs (out of 955) were detected as being nontrivially expressed across three thrombocytosis phenotypes (**Figure 3-5**). The detection criterion was set as more than 70% non-absent flag values detected in particular group based on Agilent present-absence call standard [115]. Among these platelet miRNAs, 232 of them were found in normal controls, bringing the number of recognized platelet miRNAs to a higher level compared to previously reported 219 known miRNAs in Landry et.al. Moreover, they showed 170 published platelet miRNAs among which we found 135 were in common with our results. Together we have identified 257 miRNAs expressed in platelets which were never reported before in normal human platelets. One interesting observation is that both ET and RT cohorts have more expressed microRNAs than the control group. In particular, the RT patient group has 309 microRNAs with their expression level above detection line (**Figure 3-5**).

Before and after the identification of microRNA pathways in human platelets, several studies have been published showing the identification of aberrant miRNA profiles in a subset of chronic MPDs including PV, PMF and ET [61-63, 126]. These studies were primarily focused on hematopoietic cell lines other than platelets, including mononuclear cells, reticulocytes, granulocytes and neutrophils. Some of the prominent observations include the down-regulation of let-7a and up-regulation of miR-182 in PV granulocytes as well as the up-regulation of platelet miR-26b and down-regulation of reticulocyte miR-30b, miR-30c and miR-150 in PV blood. Additionally in PMF granulocytes, the expression of miR-31, miR-150, and miR-95 were

significantly lower, while those of miR-190 significantly greater, than control and PV or ET samples; on the other hand, the level of miR-34a, miR-342, miR-326, miR-105, miR-149, and miR-147 were similarly decreased in patients with PMF, PV, or ET compared to controls [126]. In another study, the PV and ET neutrophils were found to contain significant lower levels of miR-133a and miR-1 [63]. Compared with the aforementioned studies, our own experiments indicated an interesting yet somewhat different set of microRNA profiles among ET patients in the platelets lineage.

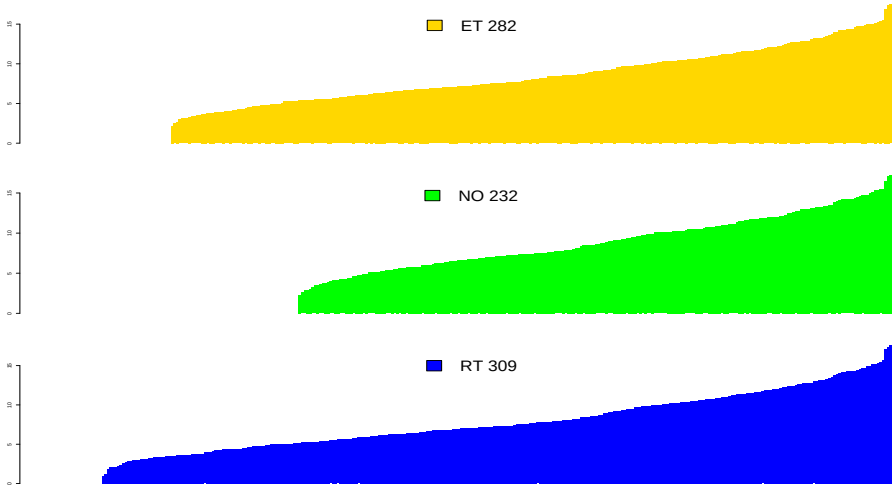

Figure 3-5. Barplots of platelet miRNA profiling comparison among 3 groups. Human platelet miRNAs with less than 30% absent calls in the specified cohort are shown in the order of increasing fluorescence intensities. This figure reflects the number of expressed miRNA in human platelets under different pathological conditions. It is interesting to see that the RT cohort has the most expressed miRNAs. This finding also expanded the previous reported 170 expressed miRNAs found in human platelets [66]. Figure is generated from R software. Experimental data is provided courtesy of Dr. W. Bahou.

Among the 21 microRNA features we identified between ET and control, the hsa-miR-182 was previously found down-regulated in PV granulocytes, its profile in ET platelets is similar according to our observation with a fold change of 0.347 (ET / control). For another microRNA --- the miR-34a, its expression level in granulocytes was lower in ET than controls in

previous study but we have found the platelet miR-34a having a greater expression level of in ET than normal group. Furthermore, we have noticed the significant down-regulation of miR-150 from ET to Normal, which was not observed in previous study of ET platelets on the same microRNA [62]. The fidelity of our findings in the 21 microRNAs was well confirmed by the quantitative PCR experiment, in which 19 out of the 20 miRNAs had consistent pattern across two platforms with an average correlation coefficient up to 0.69. This validation has brought us sufficient confidence for our further pursuit into the miRNA targets. One of the most striking observations in our study is the strong presence of the miRNA complementary strands differentially expressed in a cooperative manner. This is a relatively novel discovery in the microRNA regulation process as the classical perspective assumes that only the mature strand enters the regulatory pathway.

In recent years, the common notion of the preferred miRNA* stand selection in the maturation process was gradually modified with the confirmation of a handful of miRNA/miRNA* regulatorily active in a wide range of vertebrate cells [127]. There have been studies on the collective regulation of miR-155/155* in plasmacytoid dendritic cells [128] and the co-up-regulation of miR-9/miR-9* in primary brain tumor [129]. Among the list of miRNA signatures we identified, hsa-miR-9 and its star-form partner miR-9* as well as hsa-miR-490-3p and miR-490-5p series are showing significant up-regulation in ET cohorts with an average folds of higher than 3. Particularly, the significance and biological novelty of the miR-490 duplex have caught our attention and we decided to explore further into the miR-490 mediated platelet regulatory network in our subsequent analysis. The supervised classification results have suggested the significance of the 3 microRNAs: hsa-miR-10a, hsa-miR-148a and hsa-miR-490-5p for their strong predictive power in ET vs NO delineation. The miR-10a was previously found mediating HOXA1 and plays a key role during human megakaryocytopoiesis [58]. Also, the

down-regulation of miR-10a was reported in chronic myeloid leukemia disease (CML) [130, 131]. The hsa-miR-148a was also identified in significant lower expression level in hematopoietic stem cells and CML [132, 133]. The hsa-miR-490 (-3p and -5p) is a novel biomarker and was never reported before for its relations in hematopoietic cell system. Taking the high fold changes and miR-490 series and the predictive power on one of its members, we decided to go further and examine the potential mRNA targets for both miR-490-3p and 490-5p. This includes the searching for mRNA targets for mir-490 series and the identification of mir-490 regulatory pathways in both normal and thrombocytotic platelets.

Chapter 4:

Correlational Analysis of Platelet microRNA to mRNA Targets

MicroRNAs (miRNA) are a family of short endogenous RNA molecules that mediates regular mRNAs at post-transcriptional level. The miRNA regulates cellular translatomics level by selectively binding to the 3'-noncoding region of specific messenger RNAs through the formation of Watson-Crick pairs [80, 83]. The miRNA-mRNA interaction generally results in a suppressing effect of mRNA activity --- the translational output of mRNA is either blocked or subdued [54, 83]. In both plant and animal cellular systems, there exist distinct miRNA profiles in different tissues (e.g. heart cells and hepatocytes have different miRNA signatures) [134]. These tissue-specific miRNAs are involved in a wide variety of regulatory pathways for cells of specific type such as multipotent stem cell differentiation, cell proliferation cycle and apoptosis control [135, 136]. The deregulation of some key miRNAs were already found to be highly associated with the pathogenic process of certain diseases including lung cancer, lymphoma and leukemia [137-139]. As discussed in the previous chapter, the microRNA were found to exist

extensively in the hematopoietic cellular systems and a subset of them were considered highly relevant to the development of myeloproliferative disorders such as PV, PMF and ET [100, 140, 141]. In our study of platelet interferomics, we have identified a subset of microRNAs that are differentially expressed between ET and Normal using SAM analysis plus the fold change threshold of 2. This stage of analysis was discussed specifically in chapter 3 where we arrived at a list of 21 miRNA features. From the characterization of between-group differences of these miRNAs and the subsequent supervised classification analysis, the compelling significance and relative novelty of hsa-miR-490 family have caught our attention. We therefore decided to explore the mRNA targets for miR-490 series in an effort to unveil the entire regulatory pathway of interest. In this chapter, we hypothesize that there exist one or many functional links between the significant miRNA markers we identified and the genetic profiles between ET and normal phenotypes. By searching and exploring these potential miRNA-mRNA interactions, we will further probe into the underlying biological regulatory pathways relevant to thrombocytosis. For this study, all the experimental data (microarray & qPCR) were provided courtesy of Dr. W. Bahou and his colleagues.

4.1 MicroRNA biogenesis

The miRNAs are a special group of regulatory RNA sequences that are transcribed by RNA polymerase II (Pol II) inside eukaryotic cell nucleus. The biogenesis of miRNAs follows a somewhat different path than other peptides-coding RNAs (**Figure 4-1**). The progenitor of miRNA --- a long (60-100 nucleotide) piece of primary ribonucleic acid sequence called pri-miRNA, is produced initially from transcription process; it then folds into stem-loop structures and subsequently cleaved the RNase-III enzyme: Drosha-DGCR8 complex to form pre-miRNA

structures. Drosha is predominantly localized in the nucleus and contains two tandem RNase-III domains, a dsRNA binding domain and an amino-terminal segment of unknown function [142]. This enzyme cleaves pri-miRNA into ~70 bp pre-miRNAs which consist of a partial stem-loop structure. The pre-miRNAs are subsequently exported to the cytoplasm with the assistance of exportin-5-Ran GTP and are further cleaved into small miRNAs under the catalyzation of Dicer protein. Dicer contains a putative helicase domain, a DUF283 domain, a PAZ (Piwi–Argonaute–Zwille) domain, two tandem RNase-III domains and a dsRNA-binding domain [134]. The Dicer cleavage of pre-miRNA generates an imperfect dsRNA duplex including mature miRNA and its star form complimentary strand -- miRNA*. At this step, some co-functional proteins such as AGO2, AGO1, Dicer and TRBP are assembled with the small miRNA molecules to form RNA-induced silencing complex (RISC). The miRNA loaded RISC (miRISC) binds to target mRNAs and suppress its translation activity by obstructing the functional pathways of translational machinery 38 [80, 134, 143]. The subdued RNA may be released later and re-enter the translational pathway, or being eliminated through the mRNA decay pathways. Among the many questions unanswered on miRNA's biological role and functions, its interaction with target mRNAs in metazoan system is by far one of the most investigated topics because of its complexity and lack of definitive algorithms.

4.2 miRNA Targets Prediction

MicroRNAs regulate virtually every aspect of biological pathways, including developmental timing, stem cell differentiation, proliferation, antiviral defense and metabolism [134]. The miRNA involved cellular regulations were realized through the recognition and interactions between the initial miRNA and its messenger RNA targets. It is estimated that a single miRNA on average regulates hundreds of potential mRNA targets, thereby controlling

multiple downstream cellular cascades [144]. The initial effort of identifying microRNA targets through experimental validation has severe limitations as the number of potential miRNA targets greatly outnumbered the experimental capacities in regular biological routines. Since 2003, computational approaches have been increasingly employed to narrow down the list of putative targets of miRNA regulation and have contributed significantly to the advancement of miRNA mechanism understanding [145].

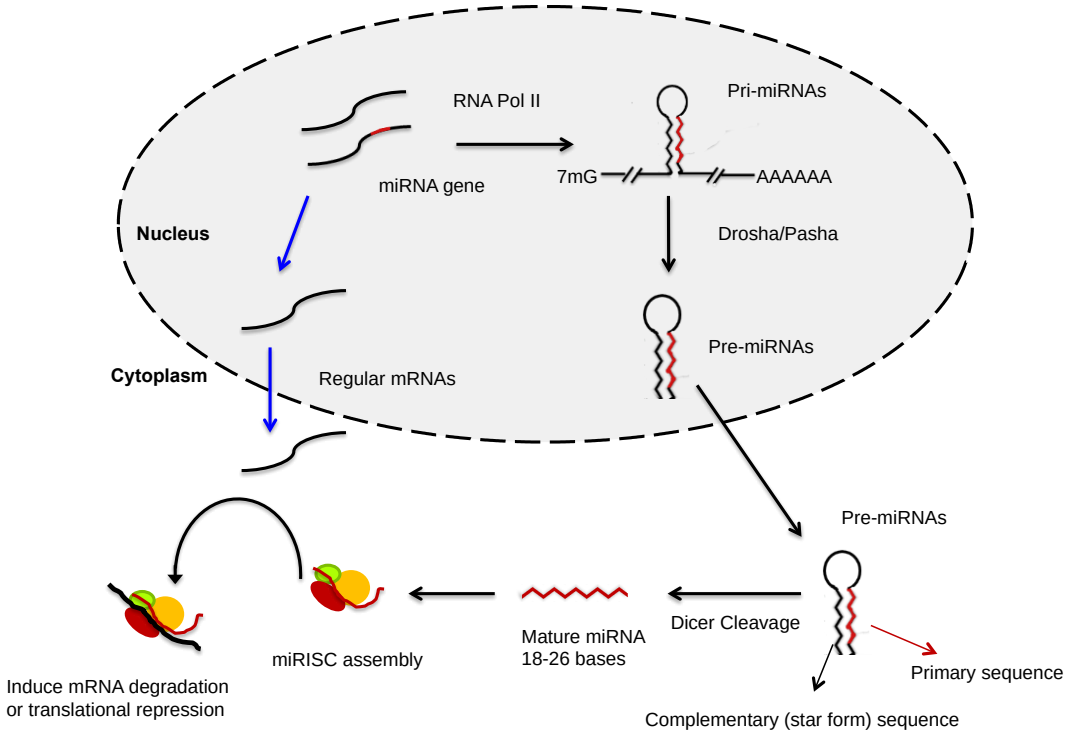

Figure 4-1: The microRNA biosynthesis and functions. The microRNA is transcribed from the special sequence of DNA genome and become the structure of Pri-miRNA inside nucleus. It is then processed by Drosha/Pasha complex to form pre-miRNAs. After being exported to cytoplasm, pre-miRNA is further cleaved by Dicer protein to form mature miRNAs. The final miRNA-mediated regulation occurs between mature miRNA strand and target mRNAs with the assistance of AGO2, Dicer, TRBP protein complex (miRISC assembly).

In a common scheme, the miRNA recognizes its mRNA target by binding to complementary sequences in the three prime untranslated regions (3' UTRs). A variety of factors have to be considered in an computational target prediction algorithm, including 1) primary base

paring pattern; 2) thermodynamics stability of miRNA-mRNA hybrid; 3) comparative sequence similarity through evolution (conservation level); and 4) the presence of multiple target sites. The computational methods in searching for miRNA targets have been fairly successful in plants as the miRNA and target mRNA are often nearly perfectly complementary, therefore the mRNA sequence recognition pattern is predominantly consistent and straightforward [146]. However, only in rare cases can extensive complementarity with consequent cleavage of targeted mRNA be observed in animal miRNA regulations. The computational prediction of miRNA targets in metazoan cells is therefore highly complex and variant. The interaction pattern of miRNA-mRNA pairs contains only short complementary sequence stretches, interrupted by wobbles, gaps and mismatches. To date, definitive rules for functional miRNA–targets recognition that capture all known functional targets have not been clearly described in metazoans [147, 148]. Nevertheless, there exist a handful of fundamental principles in the miRNA target recognition process according to the general consensus in this field. These rules are primarily derived on an empirical base such as those miRNA-mRNA regulatory pairs verified experimentally. These principles used by the mammalian target prediction programs can be generalized into several categories: 1) Base paring at seed region [120, 121]; 2) Thermodynamic stability of miRNA-mRNA hybrid [106]; 3) Comparative sequence analysis on evolutionary conservation [115] and 4) Multiple target sites examination. These rules were quantified and given varying weights in different target prediction algorithms.

To date, most of the popular target prediction programs constantly integrates the seeds matching level, thermodynamic calculations as well as conserved sites in their algorithms (**Table 4-1**) and derive the final prediction score using various statistical models (e.g. Bayesian). In a general scheme, the typical miRNA interaction with specific mRNA targets occurs through two mechanisms: 1) miRNA forms perfect or near perfect Waston-Crick (WC) pairs with target

mRNA in the 3'UTRs. This is the case for most plant miRNAs. 2) miRNA forms incomplete or imperfect WC pairs (mismatch on multiple sites allowed) with target mRNA in the 3'UTRs, as is often the case in animal cells [149].

Since 2003, there has been multiple miRNA target prediction algorithms published in metazoan system. These computational algorithms can be categorized into two generations: 1) the first generation of methods is predominantly based on primary sequence matching with/without structure conservation [150, 151]. Some of the early methods were originated from observations in a handful of microRNAs in *caenorhabditis elegans*, *drosophila melanogaster* and *vertebrates* [89, 152, 153]. The preliminary prediction rules were derived on a very limited set of experimentally identified potential target sites all located in the 3' untranslated regions (UTRs) of mRNAs. Through various in vitro experiments, including the result that the multiplicity of sites in a 3' UTR seemed to exponentially boost the efficacy of target repression [154]. Most methods provided not only scores on complementarity of the miRNA to target sites but they also model the thermodynamic free energy changes for the formation of miRNA/mRNA duplex. Later, it was noticed that miRNA 'seed regions' (include multiple sites) might play a critical role in the miRNA target recognition process [83, 151]. The 'seed region' generally refers to certain consecutive sites located in the 5' end of the miRNA (e.g. 6- to 8- bp position) [93] at which the formation of Watson-Crick base parings between miRNA and target mRNA would greatly enhance the target recognition affinity, and therefore serves as a key indicator for most of the computational prediction programs [155, 156]. It was found that base paring of these 'seed' nuclei can initiate a rapid zip up of the miRNA/mRNA duplex to overcome thermal diffusion, resulting in a stabilized thermodynamic intermediate state for further annealing of the miRNA to the target site [157].

Besides the base paring sites on miRNA sequence, another key methodological advance is to use comparative genomics to filter out hairpins that are not evolutionarily conserved in related species. Potential targets sites that are under selective pressure shall remain constant across species and are therefore considered biological functional. A valid computational algorithm is expected to give higher weights to such conserved target sites [148, 158]. However, such a filtering step may limit the potential of an algorithm to identify novel miRNAs because there might not be known close homologies due to the insufficient coverage of existing databases. A recent study showed that the number of non-conserved miRNAs missed by the comparative genomics strategy was nontrivial [159]. The first generation of miRNA target prediction methods have severe drawbacks as they generally fail to account for factors other than primary genomic sequences; They also have insufficient power to discern false positive sites on 3' UTRs for their high variability between different miRNAs. In general, only a small portion of these computationally predicted miRNA targets were experimentally confirmed and there exists very few overlapped predictions between different methods of this generation [93, 147, 160-162]. Starting from 2005, a new generation of miRNA targets prediction methods has emerged and was experimentally proved to provide more comprehensive and accurate prediction results.

Table 4-1: Popular metazoan miRNA targets Prediction Algorithms

miRNA	Type of Method*	Resource
DIANA-microT	Thermodynamics & Conservation	http://diana.cslab.ece.ntua.gr/microT/
ELMMo	Bayesian Method & Conservation	http://www.mirz.unibas.ch/ElMMo2/
miRanda	Complementary & Thermodynamics	http://www.mirbase.org/
TargetScan	Complementary & Conservation	http://www.targetscan.org/
PicTar	Thermodynamics & Complementary	http://pictar.mdc-berlin.de/
RNAHybrid	Thermodynamics & Complementary	http://bibiserv.techfak.unibielefeld.de/rnahybrid
TarBase	Experimentally Validation	http://diana.cslab.ece.ntua.gr/tarbase

MiTarget	Support Vector Machine	http://cbit.snu.ac.kr/~miTarget

* The type of method involved in the different miRNA target prediction tools basically covered all factors in the current understanding of miRNA-mRNA interaction process. These include thermodynamic stability for the molecular interrelationship, the conservation of sequences across species (evolutionary significance) and the complementary matching of nucleotide sequences. These factors are weighted in different ways and fit into different statistical models (e.g. Bayesian, SVM). Varied in scales and ranges, all these predictions have to be experimentally verified in real cell lines before we can claim a true miRNA target.

This generation of methods combined genome-wide sequencing examination methods and systematic target-site mutation experiments [148]. Through a series of studies, it was found that base paring at some of the 3'- supplementary sites [163, 164] have compensatory effects to the mismatches in the canonical seed region and therefore accounts for a number of microRNA target recognition cases [81, 151]. Sophisticated statistical methods are increasingly employed in the target searching algorithms: one branch of the methods is based on machine learning approaches such as support vector machine (SVM), neural networks, hidden Markov model (HMM) and Naïve Bayes (NB). These methods incorporate multiple distinct sequences and structural 'features' to train experimentally confirmed miRNAs and predict testing miRNAs based on a set of derived formula [93, 147]. Examples of this approach include the initial study conducted by Sewer et al [165], MiRFinder tool [166], RNAmicro program [167] and rna22 algorithm. The rna22 method in particular was developed on recognition of shared miRNA sequence patterns by sampling through all known mature miRNAs to identify statistically over-enriched structure and tentatively fit each known miRNAs to predicted target followed by a ranking of resultant ΔG [168, 169]. The machine-learning method relies heavily on a sufficient number of well characterized miRNAs as training group as well as the complete genome annotation profiles. Therefore, this method is sometimes subjected to severe limitations due to lack of prior information [170, 171].

Other algorithms in this generation may account for target site accessibility rather than only considering free energy for the formation of miRNA-mRNA duplex [172, 173]. This type of algorithms usually confers extra considerations to the thermodynamic cost of the spatial structure changes during mRNA unfolding process. One example is the PITA program [168, 169] in which free energy cost in both miRNA-target pairing process and miRNA-mRNA duplex formation are both taken into consideration with seed match scores. Among the different miRNA target prediction algorithms, the 'seed matching' based methods are still among the most widely used and recognized programs in this field. In several studies where different target prediction algorithms are compared, seed based methods such as TargetScan [93] and PicTar [174] are found to have superior predictive power in terms of specificity and sensitivity [145, 175, 176]. Still, a significant proportion (25%~40%) of mRNAs lack a perfect seed match to their paired miRNAs based on recent discoveries [177-179], therefore a simple filtering of potential genes based on these "seed match" databases might eliminate some of the real targets.

4.3 Mechanisms of microRNA regulation

Although the key step of microRNA mediated regulatory interaction with mRNA is the target recognition process, the downstream translational control initiated by this miRNA-mRNA duplex may follow different pathways. The current understanding of microRNA mediated gene regulation inside metazoan cellular system have suggested the parallel existence of 3 major mechanisms (or regulatory schemes). In the first scheme, the miRNA interferes with translational initiation step by blocking the formation of necessary ribosome assembly [180]. This mechanism was observed from several recent studies: In the experiment conducted by T. Chendrimada et al , *Drosophila* miRNA was found to repress the assembly of 48S translational complexes through

the recruitment of ribosome anti-association factor EIF6 [181] ; In another study, miRISC machinery was found to repress the formation of translational initiation complex through the recruitment of m7 Gppp cap structure [182]. Several further studies have also confirmed the importance of cap structure for its inhibitory involvement with translation initiation machinery [183, 184]. Such a cap structure was also identified in the AGO-2 protein (critical component of RNA-induced silencing complex) [185]. Moreover, The subcellular foci such as PBs were found to involve in the miRNA-mRNA regulatory process at the initiation stage although the details of this process it yet to be determined [186]. Additionally, the miRNA can interfere the translational initiation process through an inhibitory mechanism that prevents Poly-A binding protein (PABP) from concatenating to target mRNA [184]. This event induces shortened poly-A tail of mRNA but does not lead to a further degradation of the targeted mRNA molecule. In the repression of translational initiation mechanism, the regulation of miRNAs also appears to be associated with a special group of cytoplasmic structure -- processing body, such as P-bodies and GW-bodies. These structures temporarily store mRNAs and play a role in 5' to 3'destruction of these RNA molecules [187]. It was experimentally found that miRNA-binding would cause the targeted mRNAs to be trans-located into P-bodies [188] and this happens as a primary event of scheme I for human cells.

In the second scheme, miRNA was found to inhibit mRNA translational activity at post-initiation stage. Evidence of this mechanism comes from the finding of *lin-4* in *C.elegans* where the miRNA blocks the LIN-4 protein synthesis after the initiation of translational process [189]. Further investigation confirmed the presence of active translating polysomes in miRNA-inhibited mRNAs indicating that the repression occurs during the translation process [190]. There are several hypothesized mechanisms in explaining the miRNA-mediated inhibition at post-translation step, including premature ribosome disassociation, nascent polypeptide degradation

and ribosome ‘drop-off’ hypothesis. In these explanations, the mRNA translational activity was either repressed by ribosomal run-off when initiation was inhibited by interfering factors (e.g. hippuristanol [191]) or the nascent polypeptides are shortened and degraded under the catalytic presence of miRNA as indicated in the experiment by Nottrott et al [190]. To date, there’s no consensus in explaining the specific mechanism of post-initiation translational repressing process.

In the two schemes stated above, the microRNA regulates mRNA targets through a ‘translational repression only’ mechanism, where the mRNA expression levels are generally not affected by the microRNA profiles. However, in the third mechanism where miRNA can induce decay or degradation on target mRNAs, the level of miRNA is inversely related to that of the mRNA profiles. This regulatory pathway was first reported by Wu et al in their observation on miR-125b and let-7 mediated accelerated deadenylation experiments [192]. The degradation mechanism induced by miRNA regulation was found to exist in a variety type of species and conditions [180, 193]. The core components of miRISC-ago protein complex have been shown in RNA granules that serve as potential sites for mRNA degradation. Studies have indicated that mRNA degradation mediated by miRNAs is influenced by the enzymes involved in general mRNA decay, including deadenylases, exonuclease and decapping enzymes [82]. This catalytic environment is usually provided in P bodies, therefore the miRNA induced mRNA degradation involves both AGO1 protein and the presence of P-bodies. Additionally, various isoforms of Argonaut proteins carry the capability of interfering with translation and of degrading mRNA through their endonuclease functionalities [194]. However, not all miRNA targets are degraded in this process: Some targets may stay in a translationally silent state, from which they might eventually be released and reenter the regulatory pathway [83, 180]. The cytoplasmic RNA granules such as Stress Granules (SGs) and P-bodies are two important factors in the

posttranscriptional regulation of genes. These structure may temporarily sequestrate untranslated mRNA and direct these stored targets into either degradation or de-repression and release pathways [195]. The P-bodies and SGs were found to have dynamic associations as these two types of granules share a number of common components, such as cap-binding protein eIF4E and translational repressor rck/p54. It was proposed that the two cell granules may not overlap in their functional process with mRNAs (i.e. both degradation or storage). Yet their respective role in miRNA induced mRNA storage and decay process is still under debate [196].

To summarize the miRNA mediated regulatory pathways, there exists 3 regulation schemes in terms of the miRNA induced repression on target mRNAs. Generally the first two mechanisms can be categorized into the same pattern in which miRNA expression level does not notably affect the profiles or mRNA while in the third scheme miRNA level is inversely related to target mRNA abundance (**Figure 4-2**). In all three pathways, the final protein output level is always negatively correlated with miRNA profiles.

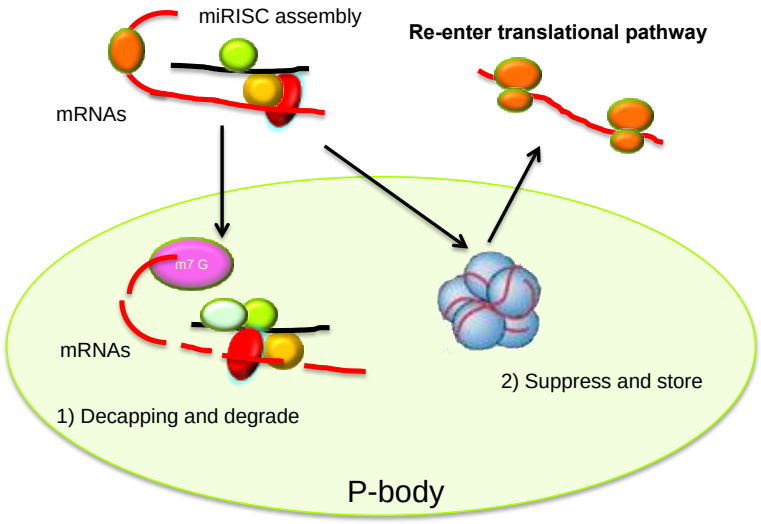

Figure 4-2: The mechanisms of miRNA mediated regulation process.
In the miRNA induced mRNA degradation mechanism, the miRISC machinery binds to target mRNAs and suppress its translational output. It then induces the decapping and degradation of the target. 2) In the second mechanism, the miRISC assembly silence the expression of target mRNAs but does not induce degradation. Instead, it mediates the mRNA to be temporarily stored in P-body. The sequestrated mRNA may or may not re-enter the translational pathway in the end depending on the cellular environment and other regulatory factors.

4.4 Identification of a subset of important miRNA markers.

As discussed in Chapter 3 that we have identified 21 microRNAs being differentially expressed between ET patients and normal controls (**Table 3-2**). Besides, the expression profiles for 20 out of the 21 miRNAs were validated in an independent qPCR experiment with a strong cross-platform correlation (Pearson Correlation coefficient = 0.69, P <0.0001). The hsa-let-7d* was not included in the qPCR validation step because it is a precursor rather than a mature miRNA. Its cDNA generation mechanism is different compared to the others. For that reason we decided to leave this one out for the current stage of analysis. In real biological system, there are hundreds of potential mRNA targets for each of the expressed miRNAs. It is generally not a viable approach to examine and validate all the mRNA targets for the entire 20 miRNAs set identified from microarray experiment.

We tried to narrow down our focus to a small subset of 20 miRNA candidates using several criterions, including significance, novelty and robustness. After examining the entire significant miRNA list, we decided to further pursue the mRNA targets for hsa-miR-490 series (miR-490-3p and miR-490-5p) due to their strong significance and relative novelty. The miR-490 series demonstrated a high fold change (**Figure 3-3:** 490-5p: ET/NO = 9.587; 490-3p: ET/NO = 6.001) in the comparison between ET and NO and they were never reported before in similar cases. In addition, the cross-platform correlation for the miR-490 series is very high, too (**Figure 3-3,** Pearson correlation coefficient: 490-3p = 0.714; 490-5p = 0.840) suggesting a competent robustness in their measured expression level. More importantly, the miR-490 series contain the two complementary sequences -3p and -5p of the pre-miRNA , indicating a combinatorial functional involvement of both individual strand. This observation is somehow contradictory to the traditional view of miRNA regulatory pathway that only the mature strand cleaved from pre-miRNA has regulatory role in cellular system while the complementary

sequence (usually denoted in star form) is degraded and removed from subsequent formation of miRISC silencing complex. However, the current understanding of miRNA biogenesis and regulatory process has been increasingly altered as several recent studies reported the potential sustained activity of miRNA* species in both *Drosophila* and mammalian systems [197, 198]. One explanations proposed the hypothesis that whether primary or star form strand is retained and form miRISC machinery depends on the stability of the 5' UTR of the pre-miRNA. When both 3' UTR and 5' UTR of a particular pre-miRNA structure have similar thermodynamic stability in pre-miRNA duplex, the star strand may have similar if not equal probability to remain active and enter regulatory pathway [199, 200]. Cross-species conservation is commonly considered an important factor in predicting the functionality of a specific miRNA. From this perspective, studies in *Drosophila* showed that a considerable fraction of miRNA genes are highly conserved in both miRNA and miRNA* sequence and seed region matches of both primary and star from strands revealed selective conservation in 3' UTRs [201]. In a more comprehensive study, the Yang et al conducted extensive bioinformatics analysis and demonstrated that a substantial proportion of miRNA complementary strand are stringently conserved over vertebrate evolution, particularly in seed regions [127]. A further test on 22 miRNAs revealed that in majority of the cases, both primary and complementary strands have abundant expression and can effectively repress perfect-match mRNA targets.

The increasing findings of both miRNA strands involved in the mRNA silencing machinery have unambiguously pointed to an alternative mechanism that well-conserved miRNA* (or -5p) strands may potentially contribute to the cellular regulation network. In our own study on miR-490 series，we examined genetic sequences of hsa-miR-490 in stretched full length (-3p & -5p) form using UCSC genome browser (http://genome.ucsc.edu/). From the alignments graph, we found stringent conservation of both 490-3p and 490-5p strands across

evolutionary species (**Figure 4-3**). This observation has conferred strong supports to our microarray/qPCR results that both miR-490-3p and -5p were differentially expressed between ET group and healthy controls. Nevertheless, the specific mechanism whether the two strands coordinate with each other or act independently needs to be further investigated.

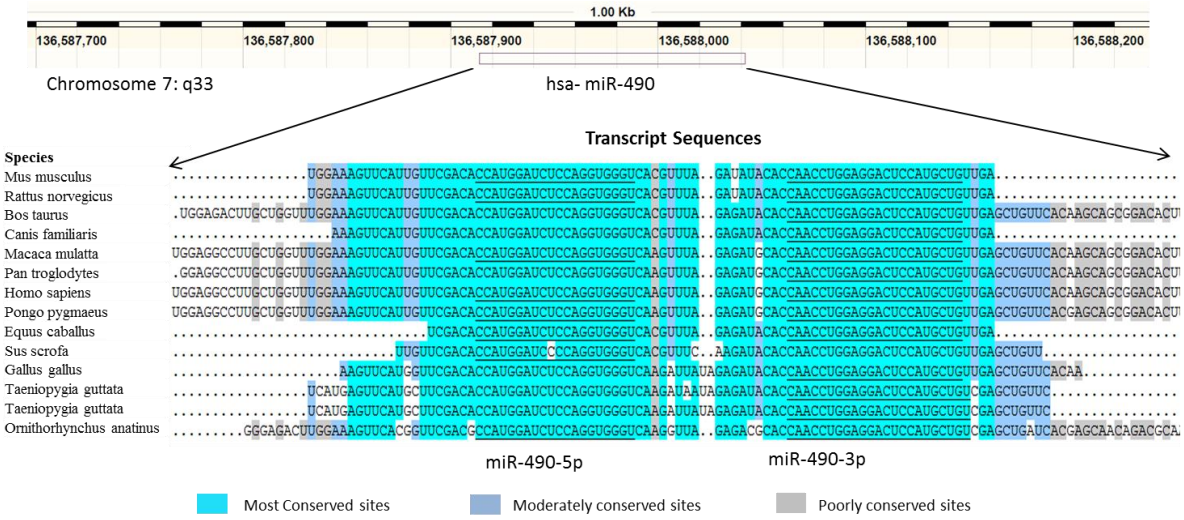

Figure 4−3: Evolutionary profiles of well-conserved hsa-miR-490 gene sequence. The gene encoding hsa-miR-490 series is located on human chromosome 7 (Coordinates GRCh37, 7: 136587914 -136588041). The multi-species alignment of hsa-miR-490 transcript shows that both the primary miR-3p and the complementary miR-5p strands are well-conserved across multiple species, suggesting the importance of this mir-490 series in regulating cytological functions. Figure is generated using USCS genome browser tool (http://genome.ucsc.edu/).

4.5 Searching mRNA targets for hsa-mir-490 series

To identify and validate potential targets for miR-490-3p and miR-490-5p, A searching algorithm comprising two filtering layers was applied in the analysis (**Figure 4-4**) – the first filter is constructed on knowledge-driven database in which four microRNA target prediction algorithms --DIANA-microT 3.0 [202], ELMMo [90], TargetScan 5.1 [203] and mirBase [204] were employed to select putative targets of the 2 miRNA signatures. We chose to use these four algorithms based on a robust methods comparison study in which 5 miRNA target prediction

algorithms were recommended, including the 4 we used in our analysis [175]. The fifth one PicTar [205] was not applied in our analysis because it does not have predictions for the relative novel miR-490 series. Among the four algorithms, DIANA-microT is based on an algorithm combining both computational prediction and experimental validation. The relevant parameters were calculated respectively for each miRNA and each target depending on binding stability and conservation scores. This method relies on a moving window of 38 nucleotide that progressively checking through the 3' UTR of potential targets. The binding energy is calculated at each step and is compared with randomized 3' UTR sequences of the same nucleotide content. 5' seed matches, including G:U wobble pairs are also taken into account in this method. The DIANA-microT algorithm has a high level of precision (66%) while being relatively short in sensitivity. ELMMo is an algorithm based on Bayesian method that derives prediction score from the evolutionary conservation level of miRNA binding sites.

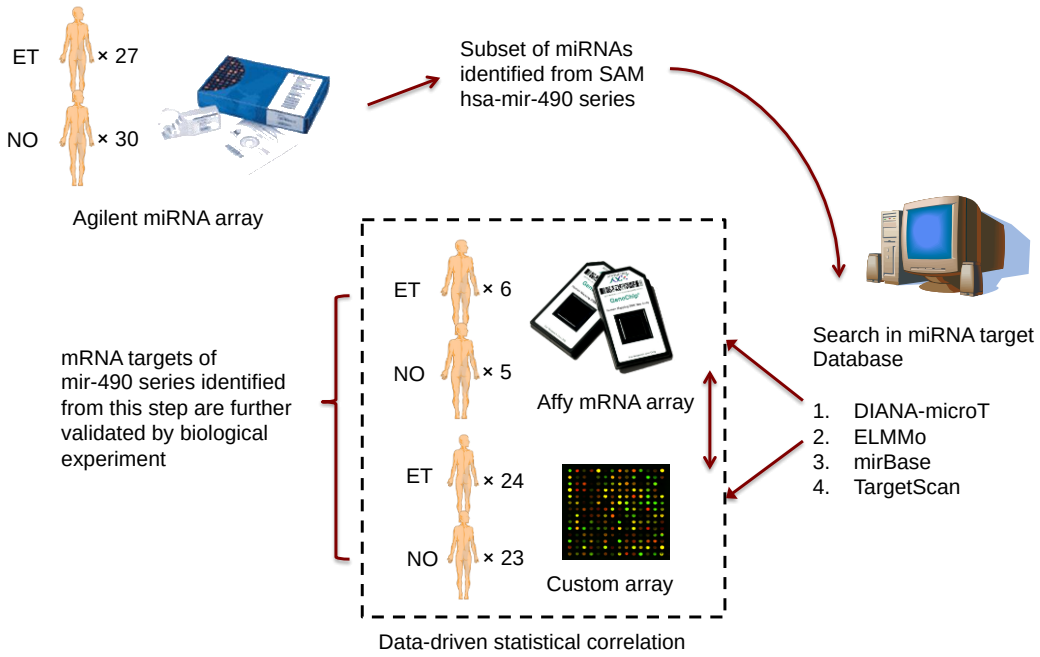

Figure 4-4: A schematic view of mRNA target searching algorithm. Mir-490-3p and mir-490-5p were initially identified from SAM analysis. Their putative mRNA targets were then identified through the pooled results of 4 popular miRNA databases. The preliminary target lists were further verified through correlation to existing genomic

data (two independent platforms as shown in plot) as the data-driven statistical confirmatory analysis. A final, much concentrated list of mRNA targets for the mir-490 series were generated and subjected to further biological experiments. All microarray data is completed and provided courtesy of Dr. W. Bahou.

For each miRNA prediction, an explicit model is constructed to infer the phylogenetic distribution of functional target sites, which is independent of the miRNA itself. This method has achieved a high level of precision and a relatively reasonable sensitivity rate in the methods comparison study [175]. The TargetScan algorithm was developed by Lewis et al in 2003 to identify putative vertebrate miRNA targets. This program incorporates both thermodynamic stability of miRNA-mRNA duplex and cross-species conservation level in deriving prediction scores. The seed matches on the 5' UTR of the miRNA were given considerable weight and a score is assigned to each UTR to reflect its quality. The estimated false-positive rate of this method was reported to vary between 22% and 31% [206] and the precision is comparable to other good methods (~50%), too. miRBase database is constructed from miRanda algorithm which relies on a two-step filter approach that controls on both whole sequence length match and thermodynamic stability of the miRNA-mRNA complex. In this algorithm, strict complementary at 5' seed region is demanded and conservation of target sites at exactly the same location is required, too. The precision of this method is relatively lower than the other three algorithms but its sensitivity is equal or higher in simulated comparisons.

In our analysis, we combined the putative targets of miR-490 series from the pooled results of the 4 programs as the preliminary filter step. This procedure was automated and implemented in a direct R scripting scheme with external database link using Bioconductor package RmiR (http://www.bioconductor.org/packages/2.5/bioc/html/RmiR.html). The resultant mRNA targets from this step were kept with its original sources and subject to the second filter using a data driven method. At this step a statistical model was constructed to examine the expression patterns between miRNA and mRNA experiments. Specifically, the genomic profiles

from a previous study [207] comparing ET and Normal patients was used to correlate between miRNA and mRNA. The mRNA genomic data set contains 6 ET samples and 5 normal controls arrayed on Affymetrix HU133A platform (GEO database accession NO.GPL1716). The 6 ET subjects (THR100, THR101, THR102, THR103, THR105 and THR112) arrayed on this platform were also used in the experiment of custom array experiment (genomic profiles discussed in details in C-2). There are a total of 22283 probe sets hybridized on this platform which included almost all known human genes. The five normal controls from this data set were randomly selected and were never used in a later experiment. Based on the previous discussion of miRNA-mediated regulatory mechanisms, there are two potential patterns between microRNA and its target mRNA in ET vs Normal comparison (**Figure 4-5**).

In the first mechanism, the miRNA silences the target mRNA not only by suppressing its translational activities, it also induces the decay and degradation of the mRNA itself. In this scenario, the miRNA expression level is expected to be inversely related to the target messenger RNA level (**Figure 4-5, A**). For example, if in ET group miR-490-3p has an abundance level twice as much as that in normal controls, a potential mRNA target X of miR-490-3p shall exhibit a reverse pattern in which X expression level in ET patients is significantly less than that in healthy controls and vice versa. In the second mechanism, the miRNA only repress the translational output of target mRNAs but does not notably affect their abundance level. In these cases, we do not expect an observable statistical pattern between the two factors in ET vs NO (**Figure 4-5, B**). However, miRNA-mRNA interactions following this scheme can be traced by the final protein level as the ultimate peptides productions were suppressed by microRNA interference in a definite manner.

In stage one of the analysis, an inverse correlation between miRNA level and target mRNA abundance was built to reflect the miRNA-mediated target degradation mechanism. In

specificity, a two-way ANOVA model was established in formula 4-1, evaluating on group mean level and treating RNA type (miRNA vs mRNA) and Group effects (ET vs control) as two factors

$$RNA_{ik} = \mu + g_i + \psi_k + g_i\psi_k + \varepsilon_{ik} \quad (4\text{-}1)$$

RNA_{ik} is the general RNA expression value, μ denotes the grand mean; g_i refers to group effect; ψ_k is RNA type and $g_i\psi_k$ represents the interaction effect between group (ET vs control) and RNA types. The microRNA and its putative target are considered a valid regulatory pair if the interaction term $g_i\psi_k$ is significant ($P < 0.05$) and the inter-group medians (ET vs NO) are negatively related (Fold change between groups are in opposite direction). All the original p values were corrected by Benjamini-Hochberg multiple testing correction procedure. In this way, we were able to narrow down from a long list of putative mRNA targets to a handful of most important candidates (**Table 4-2**) for final biological validation.

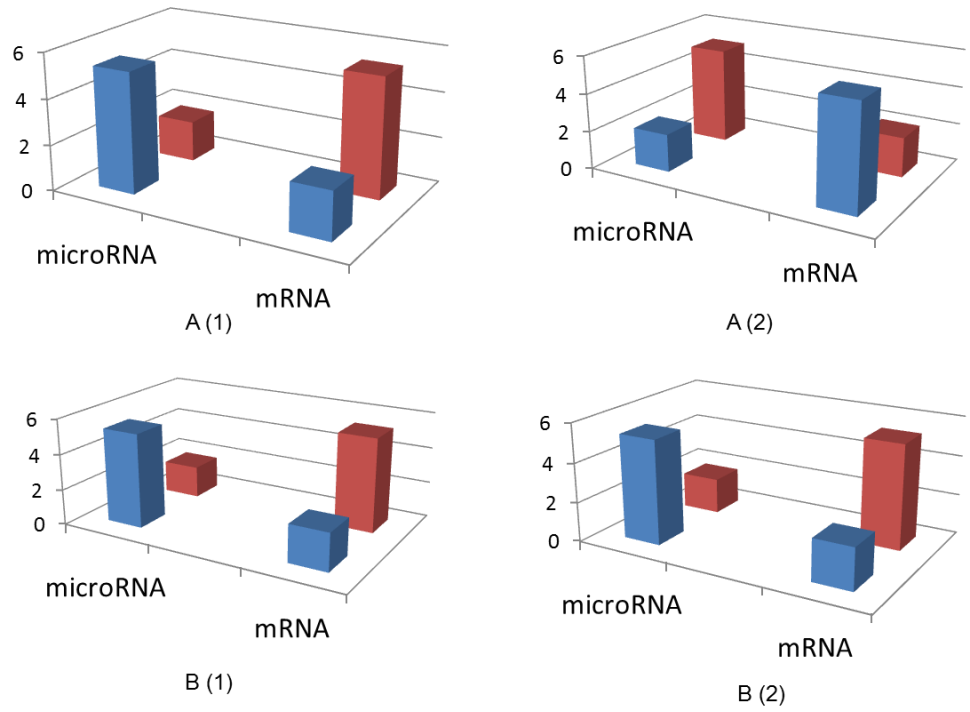

Figure 4-5: The two expression patterns between miRNA and its target in ET vs NO comparison. The first two patterns A(1) and A(2) represent the inverse correlation between miRNA and mRNA targets comparing between ET

and NO. The pattern can be in either way but the group means need to follow the opposite directions. The two patterns in second row (B 1 and 2) denote the alternative mechanism in which miRNA only induce translational suppression without further degradation or decay on the mRNA targets.

In addition to this pattern filter step, we also utilized another mRNA genomic data set to validate putative targets for mir-490 series. This data set was based on the platelet-specific genomic profiles we used in our study of thrombocytosis class prediction model as discussed in chapter 2 [32]. An advantage of using this data set is that the majority of the patient samples in genomic study were also characterized in the miRNA experiment (A total of 57 samples have both miRNA and mRNA values). Therefore, we modified our regular two-way ANOVA model and treated the group factor (ET or NO) as repeated measure in our analysis as demonstrated in formula 4-2.

$$RNA_{ikm} = m + g_i + y_{k(i)} + g_i y_{k(i)} + e_{m(ik)} \quad (4\text{-}2)$$

RNA_{ikm} is the general RNA expression value, μ denotes the grand mean; g_i represents group effect; $\psi_{k(i)}$ is RNA type (repeated within group factor) and $g_i\psi_{k(i)}$ represents the interaction effect between group (ET vs control) and RNA types. MicroRNA and its putative target are considered a valid regulatory pair if the interaction term $g_i\psi_{k(i)}$ is significant ($P < 0.05$) with an additional criterion dictating that between-group fold changes are in opposite direction (p values were corrected by Benjamini-Hochberg multiple testing correction procedure). However, this genomic data platform based on our custom microarray was initially hybridized with only 506 platelet restricted genes. In this regard, some of the relevant biomarkers not incorporated in our platelet exclusive setting may not be detected through the two-way repeated measure ANOVA model. Moreover, the candidate gene list is further compressed due to the presence of missing values in a fraction of the spots (we used 70% presence-absence call standard to filter the data set) Nevertheless, we were still able to identify a list of mRNA candidates inversely correlated

between the two RNA types (**Table 4-3**) and some of the targets are overlapped with the findings from Affymetrix GPL1716 platform. We examined the biological background for each potential target and decided to pursue biological validation on a subset of these biomarkers.

Table 4-2 Putative targets for miR-490-3p and miR-490-5p (two-way ANOVA model)

DataBase	miRNA Name	Gene Targets
DIANAmicroT	hsa-miR-490-3p	HDAC2*, CYCS
	hsa-miR-490-5p	CXCL12
TargetScan	hsa-miR-490-3p	RBM12, SHANK2, PPME1**, YWHAE, PTPN12
	hsa-miR-490-5p	NTRK2
EMMo	hsa-miR-490-3p	HDAC2*
miRBASE	hsa-miR-490-3p	RPL35, SF3B3, RPS15, RPLP2, FUS, ATP6V1F, UBE2L6, PRKCD, MYLK, KIAA0391, FBLN1, RHOG, ASF1A, IPO13, ACOT8, DCT, CALB2, SULT2B1, ITGA10, MASP2, MUC1, GORASP2, ARHGDIA, CDKN1C, GAPVD1, SPG21, TGIF2, GMPPA, DAPP1, METTL4, FBXL12, OR1D2, MOSC2, PPME1**, SUPT4H1, RGS10, MYL4, RABGAP1L, RHOBTB3, DYNLRB1, MPM1
	hsa-miR-490-5p	RPL21, WARS, SF3B3, DYNLL1, FUS, PSMD1, PLOD3, FBLN1, PSMB9, FGL1, NFRKB, MASP2, MUSK, SETD3, GAPVD1, CCDC92, PARP6, CCRN4L, LEPRE1, OR1D2, ARMC6, HOXB5, TAGLN2, FOLR1, PPIA, TROVE2, LDHB, RABGAP1L, MLH3

*HDAC was identified as common targets of miR-490-3p in EMMLo and DIANA-micro T databases.
** PPME1 was identified as common targets of miR-490-3p in EMMLo, TargetScan and miRBase.

Among the 78 mRNA targets (**Table 4-2**) we identified from correlating miRNA with GPL1716 data set, there are only two biomarkers appeared in at least two target prediction databases, suggesting a lack of consistency among existing miRNA target prediction algorithms. These two biomarkers are histone deacetylase 2 (HDAC2) and protein phosphatase methylesterase-1 (PPME1). We decided to further investigate their interactions with miR-490 series by validating their expression profiles in cell lines transfected with an over-expression of

miR-490 series. The other 8 putative targets we checked on include NPM1, RABGAP1, PDZK1IP1, DYNLL1, HBE1, TUBB2C, MLH3 and CCDC92. These 10 miRNA biomarkers were selected based on a careful examination of their biological background and supportive evidence from statistical pattern analysis (overlapped in both ANOVA results). The biological validation experiments are underway now and we expect to identify a subset of genuine mRNA targets for mir-490 series.

Table 4-3 Putative Targets for mir-490 series (2-way ANOVA model using repeated measure, custom array)

miRNA	mRNA targets	ANOVA P value	Overlap with Table 4-2
hsa-miR-490-3p	NPM1	3.57E-05	Y
hsa-miR-490-3p	FUS*	6.14E-05	Y
hsa-miR-490-3p	SNRPG	1.23E-05	N
hsa-miR-490-3p	RABGAP1L*	1.24E-05	Y
hsa-miR-490-3p	PDZK1IP1	5.23E-06	N
hsa-miR-490-5p	DYNLL1	2.24E-07	Y
hsa-miR-490-5p	FUS*	1.50E-06	Y
hsa-miR-490-5p	HBE1	3.26E-08	N
hsa-miR-490-5p	TUBB2C	1.91E-08	N
hsa-miR-490-5p	RABGAP1L*	2.14E-07	Y
hsa-miR-490-5p	MLH3	3.24E-07	Y
hsa-miR-490-5p	CCDC92	2.35E-07	Y

*FUS and RABGAP1L are mutual targets for both mir-490 strands

4.6 The GLM model for Alternative Patterns

Aside from the two-way ANOVA w/o repeated measures that are based on group mean (ET & Normal) structure of miRNA and mRNA profiles, we also tried to explore alternative mRNA-miRNA interaction pattern using the model of general linear model (GLM) where not

only the group mean, but also the subject level correlations are taken into account (**Figure 4-6**). We assumed that there exists a potential mechanism that some of the miRNA maybe mediating mRNA targets in one experimental group (ET or NO) but experience a certain level of deregulation in the other cohort. Alternatively, some more extreme patterns might exist so that miRNA induce down-regulation of the target in one condition while causing up-regulation of the same mRNA target in the other condition. The two-way ANOVA model has severe limitations as it only reflects the relative position between group means. To investigate all these potential within-group patterns, we employed a general linear model to test the hypothesis as shown in formula 4-3

$$y = b_0 + b_1 x + b_2 g + b_3 g x + e \tag{4-3}$$

y in the formula denotes the mRNA expression level and x represents the miRNA level. g is the grouping factor (ET vs NO, 1 vs 0) and ε stands for error. β_0 is the default intercept of the formula; β_1 reflects the general association between miRNA and mRNA profiles; β_2 indicates a nontrivial group effect and the final β_3 is the interaction term between group effect and miRNA level.

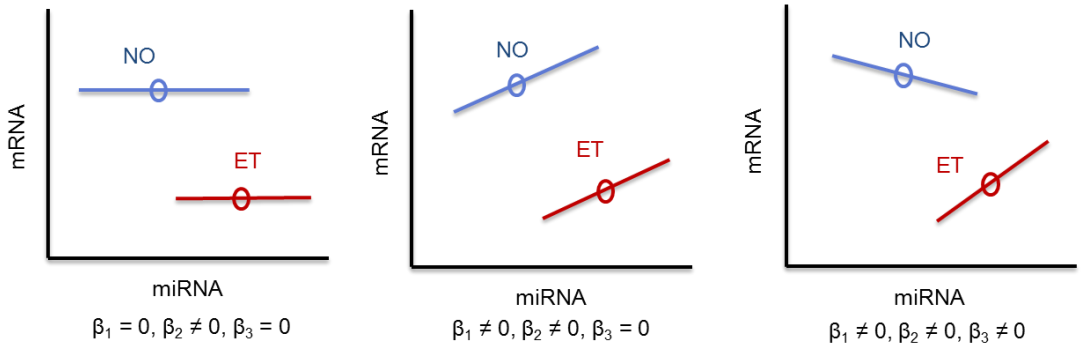

Figure 4-6: An illustration of intra-group correlation explained by GLM model. The GLM model can explain the correlation between miRNA and mRNA expression profiles with/without the presence of group effect. This model investigates the intra-group correlation and reflects both general and group-specific expression patterns. The figure provides examples on three cases with the captions closed related to **formula 4-3**: (A) there exists only group effect. (B) Both group effect and general associations (positive) are significant. (C) The expression pattern when all three effects (1.general association between miRNA and mRNA; 2. group effect and 3. Interaction term) are significant. The biological background for each of these patterns needs to be investigated.

The GLM model can be illustrated either in terms of the original model with the interaction term (**Figure 4-7 A**) or in terms of a covariate structural equation model (cSEM) where in addition to have a direct effect on the mRNA levels, "Group" also serves as a covariate modulating the influence from the miRNA to the mRNA (**Figure 4-7 B**). The latter is perhaps a more visually intuitive way to understand the interaction term in the GLM model. We are interested in all three βs as each represents a unique association between the miRNA and its targets. The significance of each coefficients was controlled by its p value (P <= 0.05) with Benjamini-Hochberg correction and the entire algorithm was implemented in R with supportive packages. This method was only used in correlating miRNA profiles with genomic data in our custom array experiment since only these two platforms share the same patient subjects.

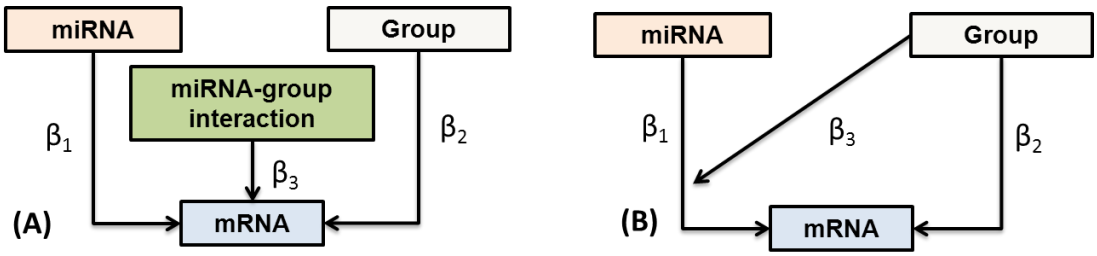

Figure 4-7: Path diagrams for the GLM model for illustration of the original model (formula 4-3) and a covariate structural equation model: (A) The thrombocytotic miRNA-mRNA correlation scheme in terms of the original model with the interaction term, and (B) The thrombocytotic miRNA-mRNA correlation scheme in terms of a covariate structural equation model (cSEM) where the covariate "Group" is modeled to have a potential effect on the path between miRNA to mRNA.

However, the same limitation of using the custom array is reflected in this analysis: the initial candidate mRNAs are all pre-determined platelet-specific transcripts, therefore we may potentially loose a fraction of real mRNA targets. Indeed, the final mRNA results from GLM model seem to corroborate our concern as we only uncovered 4 miRNA-mRNA pairs (**Table 4-4**): the first pair is the miR-490-5p with GSS with a significant β_1 at P = 0.005; The other 3 pairs: miR-490-5p & AGPAT1 (β_3, P = 0.07); miR-490-5p & TAGLN2 (β_3, P = 0.077); and miR-490-

3p & PDLIM7 ($\beta_{2,}$ P = 0.083) are all marginal findings as they all have p values 0.05 < P < 0.1. We have documented these findings and leave it for future validation.

Table 4-4: The four mir-490 series and gene targets pairs identified from GLM model

miRNA-mRNA pairs identified from GLM Model	Coefficient			P value		
	miRNA	Group	Interaction	miRNA	Group	Interaction
hsa-miR-490-5p vs GSS	0.502	0.256	-0.294	0.005**	0.862	0.285
hsa-miR-490-3p vs PDLIM7	0.04	-0.892	0.11	0.626	0.083*	0.346
hsa-miR-490-5p vs AGPAT1	-0.177	-2.116	0.369	0.156	0.056*	0.07
hsa-miR-490-5p vs TAGLN2	-0.106	-0.858	0.249	0.22	0.256	0.077*

** mir-490 and mRNA pairs identified at significance level of 95%.
* mir-490 and mRNA pairs identified at significance level of 90% (marginal).

A genuine valid analysis correlating between miRNA and mRNA through this model may require a complete human genome array to begin with – a condition we do not have at current stage. The sequent biological validation will test on both target gene profiles with changing miRNA level as well as the final protein output. A cell line closely resemble the physiological feature of platelets shall be carefully picked at this step of validation.

4.7 Results and Discussion

The miR-490 series are relative novel miRNAs that were never reported before in the study of hematopoietic system. The comparative genomics study has shown a high level of conservation across 13 species in both of its -3p and -5p strands, suggesting a potential significance of the miR-490 series for their role in cellular regulation process. In the initial putative mRNA targets identification step, we relied on 4 popular computational prediction algorithms (DIANA-microT 3.0, ELMMo, TargetScan and mirBase) to identify 1860

preliminary targets for 490-3p and 490-5p with roughly equal hits for each strand. The criterion we used in selecting gene candidates in each database was based on author recommended cutoff. Specifically, a score threshold of 7.3 was used in DIANA-microT database; a confidence score of 0.6 was applied in ELMMo program; In TargetScan algorithm, only conserved targets were kept in analysis and the same criterion was used in mirBase algorithms. One interesting observation is that a fraction of mRNAs was found to be mutual targets for both -3p and -5p strands and this phenomenon is quite notable in mirBase predictions. A possible explanation is that the two stretches of miR-490 series share different but functional seed regions that point to the same 3' UTR of mRNA sequences such as those in FUS and RABGAP1L genes. The subsequent data-driven filter relies on two previously published thrombocytosis genomic data set from our group. The Affymetrix GPL1716 platform contains a much larger genetic candidates pool than the custom array platform but has fewer samples and is not directly related to the patients in miRNA experiment.

The ANOVA algorithm with repeated measures can provide more statistical power than the regular ANOVA model in the analysis. However, one limitation in our study is that quite a few genes (30% of all the probes on custom array) do not have a consistent cross-group pattern between the two microarray data sets. Therefore, we do not expect a high overlap rate of the mRNA target results between the two platforms. In practice, 6 out of 10 genes were found to appear in both lists (**Table 4-2 & Table 4-3**) and we will conduct biological validation on all of the 6 targets.

Comparing to the ANOVA model, the GLM method the GLM method is a more sophisticated model focusing not only the group mean patterns but also on the within group subject level correlations. At current stage, we only identified 4 mRNA targets using custom array genomic data to correlate with miRNA levels. The limitation may come from two sources:

1) our initial use of knowledge base filter where only putative miRNA targets confirmed by prediction programs are considered in statistical validation; and 2) the fact that custom array platform only contains limited number of platelet restricted genes. In the future, we will try the GLM model on whole genome profiling without the restriction of pre-defined putative targets pool. From our analysis, miR-490-5p and glutathione synthetase (GSS) have a general positive association (beta1 = 0.502, P = 0.005) regardless of the group assignment. The other 3 pairs (-5p with AGPAT1, -5p with TAGLN2 and -3p with PDLIM7) all contain marginal significance in either group effect or interaction terms. A further examination may be focused on the biological correlation between miR-490-5p and GSS but it is equally possible that there exists an unrelated coincidence in their expression profile changes. Correlation analysis of miRNA-mRNA in thrombocytosis at this stage only accounts for the potential targets following mechanism 1 (miRNA induced mRNA degradation), a further and more robust approach has to rely on the protein pattern which is more likely be modulated by miRNA mediator even without the decay of intermediate mRNAs (Mechanism 2). However, it is still difficult to delineate between direct targets and indirect targets, because some of the changes at transcriptomics level can be incurred by the altering profiles of other genes (in the same pathway). Therefore in the future biological validation step, we plan to use both positive controls (introduce an empty particle into cell) and negative controls (no procedure is applied at all) to compare with the experimental group (introduce an particle containing miR-490 series) in our experiment. In this way, a direct negative response of mRNA/protein is expected when miR-490 sequences are transfected into the cell. This suggests the direct causal relationship between the two factors while excluding potential confusions brought in by indirect targets.

Chapter 5:

Correlation of MicroRNA to Proteomics Data

Although the miRNA induced target mRNA degradation is a widespread phenomenon in miRNA-mediated regulatory process, alternative mechanism in which targeted mRNAs are simply suppressed by miRISC machinery without being led to a decaying pathway is presumably even more prevalent [54, 180]. In our previous analysis that miRNA was correlated with two thrombocytotic genomic experiments, the putative targets may not reflect the genuine biological regulations of the mir-490 series we studied. A substantial amount of 'real' mRNA targets maybe lost as they follow the alternative miRNA-mediated regulatory pathway. A more robust research scheme is planned as we attempt to correlate miRNA profiles to the final protein output level. This way our statistical model will not be affected by the complication of biological alternatives. In this chapter, we hypothesize that the existence of miRNA-mRNA interactions can be ultimately reflected in the final protein level and by uncovering such linkage and interrelating all three regulatory layers (miRNA, mRNA and protein), a complete miRNA-mediated regulatory pathway can be constructed under a unified scheme of thrombocytotic system

biology. However, at the current stage, we are still limited by the availability of proteomics data in ET patient groups, considering that ET is a rare disease and recruiting patients for new experiment is expensive and time-consuming. In the preliminary analysis, we performed Mass Spectrometry (MS) experiment have collected peptide spectral counts data from one ET patient and 3 normal controls to represent the protein output level. In Chapter 5, we report our analysis and results in correlating miRNAs (mir-490-series in particular) to protein profiles depending on existing knowledge base as well as mathematical validation algorithms (fold change and group means). Moreover, the entire research scheme in hematopoietic cell system under thrombocytosis settings is reviewed and future research directions are discussed, too. The microarray data on both mRNA and miRNA involved in this study was provided courtesy of Dr. W. Bahou's lab while the mass spectrometry experiment was conducted courtesy of Dr. Dwight Martin.

5.1 MiRNA-mediated target degradation and translational repression

In Chapter 4, we studied the thrombocytotic miRNA-mRNA correlation based on the miRNA-mediated target degradation mechanism. Under this scheme, the mRNA targets of mir-490 series are expected to be inversely related to the changes of miRNA due to the lack of on-going transcriptional activity in platelet model. From biological perspective, this regulatory pathway requires key components inside the cytoplasm environment or P-bodies, including a member of the Argonaute protein family, the P-body component GW182, the CAF1-CCR4-NOT deadenylase complex, the decapping enzyme DCP2, and several decapping activators including DCP1, Ge-1, EDC3, and RCK/p54. Besides, several enzymatic foci were also found in P-bodies which are important sites in the miRISC further processing pipeline. While the deadenylation is

considered a principle effect in miRNA-mediated mRNA decay process [208], the alternative mechanism such as eIF4E for binding to the 5'cap structure is also supported by experimental evidences [185]. On the other hand, previous studies have proved that translational repression by miRNA is not a consequence of the decreased mRNA levels [182, 186, 191]. Rather, the down-regulation of mRNA occurs after the translational suppression step. Furthermore, for targets that are degraded, translational output is still active even when degradation is suppressed by depleting components of the deadenylase complex [209]. It has been widely observed that the primary effect of miRNAs is to inhibit translation and mRNA degradation is a consequence of this primary effect [208]. Following the initial step of translational inhibition, the miRISC bound mRNA is assumed to be translocated into P-bodies where it can be further degraded or temporarily stored. The stored mRNA may later be sequestrated again into degradation track or released back into cytoplasmic translational pathway. Although the details of the mechanism is still under debate, it was widely accepted that whether or not miRNAs trigger target mRNA degradation may depend on the presence of specific proteins associated with a given target and/or on the miRNA-binding site conditions and its RNA context [210]. In vitro whole genome study has found that miRNA-mediated mRNA degradation only accounts for a fraction of total cases in miRNA-mRNA interactions [211]. A considerable amount of target mRNAs are simply repressed by miRNAs in their translational activities and are temporarily 'freezed' in P-bodies. Therefore, the research at final proteomics level is an essential supplement to our previous focus on miRNA-mRNA correlation.

In terms of the relationship between protein and miRNAs, the mRNA is an important intermediate factor that links the transcriptional variation to translational output. Based on our previous discussion, there exists two statistical patterns through the miRNA-mRNA-protein regulatory process: 1) mRNA and protein expressions are consistent while both factors are

significantly suppressed by miRNA output (miRNA-mediated target degradation); and 2) protein expression profiles are negatively correlated to miRNA level while both factors are irrelevant to intermediate mRNA profiles (miRNA-mediated translational repression only). In practice, we planned our analysis following the two theoretical schemes and have generated some potentially important regulatory targets for mir-490 series.

5.2 Quantification of protein using MASS Spectrometry

Sample Preparation

Human platelets were obtained from donors and centrifuged into a pellet. Each of the pellets contains approximately 0.5 mg protein and were re-suspended in water containing protease inhibitors. The proteins in the suspension were subsequently precipitated and dissolved according to established protocol [212]. Protein concentrations were determined using the Peterson modification of the Lowry [213]. The protein solution was then subjected to trypsin digestion and later brought to 5% formic acid.

Mass Spectrometry

The samples were analyzed for protein content using a modification of the multi-dimensional protein identification technology (MUDPIT) method [214]. Specifically, samples were pressure-bomb loaded through the proximal end of a MUDPIT column constructed of 250 μm ID fused silica tubing. The MUDPIT column was connected with a microtee to a fritless electrospray interface feed column for automated micro-capillary liquid chromatography-tandem mass spectrometry. The feed column to the mass spectrometer consisted of a fused-silica capillary (100 μm ID) packed with 10 cm of 5 μm C18 material using a pressure bomb. The dual column construct was placed in line with an HPLC unit flowing at about 300nl/min. The

HPLC separation was provided by a 12 step, three components gradient as described elsewhere [215]. The application of a 1.8 kV distal voltage electro-sprayed the eluted peptides directly into a Thermo Fisher Scientific LTQ ion trap mass spectrometer. Full mass (MS) spectra were recorded on the peptides over a 400-2000 m/z range, followed by five tandem mass (MS/MS) events sequentially generated in a data-dependent manner on the first, second, third, fourth and fifth most intense ions selected from the full MS spectrum (at 35% collision energy). Mass spectrometer scanning functions and HPLC solvent gradients were controlled by the Xcalibur data system (ThermoFinnigan®, NJ). The output MS/MS spectra were extracted from the RAW file with Readw.exe (http://sourceforge.net/projects/sashimi/). The resulting mzXML data file was searched with Inspection against a human IPI database with optional modifications: +16 on Methionine, +57 on Cysteine, +80 on Threonine, Serine and Tyrosine. Only peptides with at least a P-value of 0.01 were analyzed further. Common contaminants (e.g. keratins) were removed from the database. Proteins identified by at least 2 distinct peptides within a sample were considered valid.

5.3 Searching protein targets along mir-490-mediated regulatory pathway.

Methodology

Spectral count data from MS experiments are first normalized using a quantile method [216] to compensate for varied total spectral count in each MUDPIT run. There are totally 3 normal controls and 1 ET samples quantified on their protein level in this experiment (ET sample size is limited due to the difficulty in recruiting new ET patients). In the initial stage of analysis, we checked on the consistency of within group measurements (for Normal controls only indeed) using a criterion that any spectral value beyond 2 standard deviations of group mean is

considered an outlier. Given the fact that we only have one sample in ET group, we constructed the analysis primarily based on fold change. The mRNA genomic profile from Affymetrix GPL1716 platform (6 ET vs 5 Normal) was used in the analysis to link protein level to miRNA profiles.

In scheme one, assuming that protein level and mRNA expressions follow a consistent pattern but are opposite to miRNA changes between ET and NO (**Figure 5-1 A**). The mRNA-protein relationship was first established from knowledge base using R Bioconductor annotation package hgu133a.db (the entire process was automated and implemented in R environment), their correlation was evaluated based on a SAM analysis for mRNA levels and fold change for protein entries. The final consistency score is derived by the multiplication product of cross-group fold changes and is ranked from high to low for each mRNA-protein pair. Subsequently, we employed the same target prediction programs (DIANA-microT, EMMLo, TargetScan and mirBase) to identify putative mRNA targets for mir-490-3p and mir-490-5p. Since the mir-490 series are all up-regulated in ET compared to Normal (-3p: ET/NO = 6.001 & -5p: ET/NO = 9.587), we only search results in the down-regulated list (from ET to NO) of mRNA-protein pairs. Specifically, the individual target results of mir-490 series from the four prediction algorithms were pooled into one putative gene list. On the other end, only mRNA-protein pairs with lower group mean in ET than normal were kept to correlate with miRNAs. The putative targets of mir-490 series were then matched to the filtered mRNA-protein pairs and only corresponding entries were included in the final report.

Alternatively, a more straightforward pathway was established in which protein level is directly linked to mir-490 series following mechanism 2 (miRNA-mediated translational suppression) (**Figure 5-1 B**). In this approach, we assume that mRNA expression level is not related to either miRNA or protein output therefore we only use mRNA as an intermediate

connector to construct regulatory pathway between protein and miRNAs. Specifically, the miRNA and protein pairs were first identified using R Bioconductor annotation package hgu133a.db without any consistency check. The pair list was filtered by protein fold changes that only down-regulated entries (from normal to ET) are kept in the candidate list. Subsequently, pooled results from the same 4 target prediction programs were used to link mir-490 series to mRNA-protein pairs. The final miRNA-mRNA-protein regulatory pathways were generated in this way and included in the final report. A general view of the entire experimental pipeline is illustrated in **Figure 5-1**.

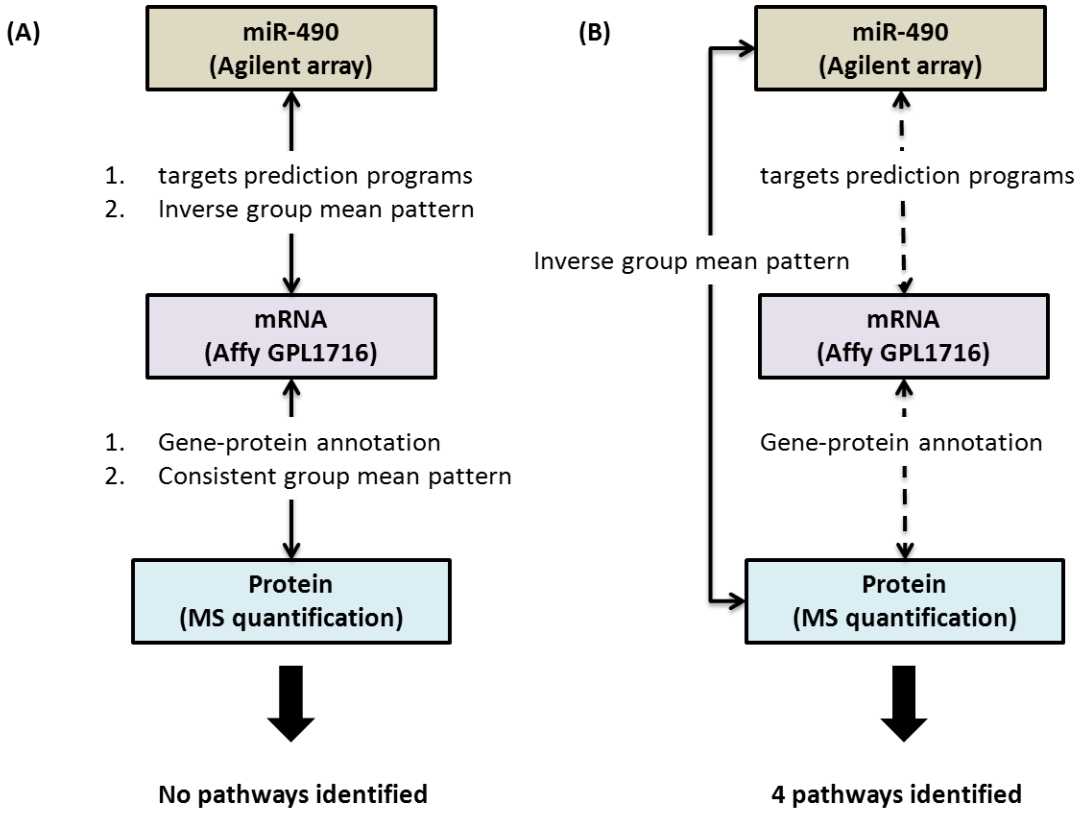

Figure 5-1. An illustration of the two mechanisms for miRNA mediated mRNA-protein regulatory pathway. (A) In the first mechanism, miRNA (**mir-490 in our analysis**) induces target mRNA degradation after suppressing its translational activity. Therefore consistent group mean pattern (ET vs NO) are expected between mRNA and protein output while both mRNA and protein profiles are inversely correlated with miRNA level. (B) In the alternative mechanism, miRNA induces translational suppression only on the target mRNA. The mRNA expression level is thus unrelated to miRNA or final protein output. In this pathway, protein expression level is expected to be inversely related to miRNA profile alone. The results for mir-490 series are indicated in the figure, too (bottom text)

Results

Based on scheme 1, we couldn't identify any regulatory targets from miRNA to mRNA and proteins as very few mRNA-protein pairs are consistently down-regulated (ET/NO) while correlating to the up-regulated mir-490 series based on existing knowledge base. Using the alternative scheme where protein entries and mir-490 series were directly linked through intermediate mRNAs, we were able to identify at least 4 miRNA-mRNA-protein regulatory pathways (**Table 5-1, Figure 5-1 A&B**). Among the 4 targets we found, the RAS homolog gene (RHOG) and (Stromal interaction molecule 1) STIM1 are considered potentially important gene targets in the thrombocytotic pathology. The RHOG is a small monomeric GTP-binding protein involved in a variety of intracellular signaling pathways, including cell motility, gene transcription, endocytosis and particularly platelet activation process [217]. STIM1 is a human protein localized to the endoplasmic reticulum and plasma membrane. It was found having a function as calcium sensor in the endoplasmic reticulum [218]. In human platelet model, the STIM1 was recognized as the key regulatory factor in thrombin-dependent pro-coagulant activity and thrombus formation [219].

The preliminary between group fold changes of the two targets have suggested a notable difference in protein level (FC: ET/NO < 0.5) as compared to the significantly up-regulated mir-490 series. One interesting observation is that mRNA expression level in all of the four target pathways are roughly unchanged between ET and Normal controls, suggesting the miRNA-mediated translational repression without further targets decaying is the primary mechanism in the regulation process. One limitation in our study is that we only have one ET sample in the experiment. Therefore we plan to perform additional validation study to confirm the spectral quantity of the identified protein entries.

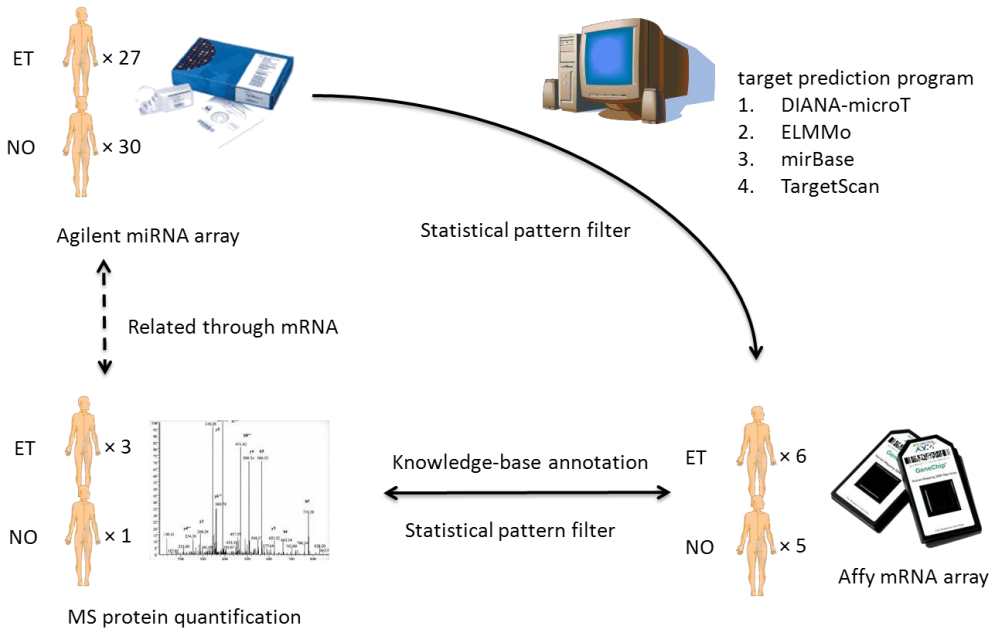

Figure 5-2: A schematic view of the miRNA-mRNA-proteomics correlational analysis. In the proteomics study, the protein output level was quantified using Mass Spectrometry technique on the whole genomic base. The miRNA and mRNA expression profiles were correlated to the protein levels through both mechanisms (I and II). The correlations between any two of the three factors were constructed based on both knowledge base and statistical pattern filtering (consistent or inverse relation). In knowledge-based filter, the mRNA and protein pairs were identified through hgu133a microarray annotation file (R, Bioconductor) while mRNA target prediction programs were used to link miRNA and mRNA (and further to protein entries). For the statistical filter process, group means (ET or Normal) are the major indicator for expression levels. Mass spectrometry experiment was conducted courtesy of Dr. Dwight Martin associated with Dr. W. Bahou's lab.

Table 5-1: Identified mRNA and protein targets for mir-490 series.

miRNA*	Gene Symbol	Protein	Gene Fold Change**	Protein Fold Change**	Gene Names
hsa-miR-490-5p	PSMD1	IPI00299608	0.896	0.661	Proteasome 26S subunit, non-ATPase,1
hsa-miR-490-3p	RHOG	IPI00017342	0.984	0.344	RAS homolog gene
hsa-miR-490-5p	STIM1	IPI00299063	1.156	0.251	Stromal interaction molecule 1
hsa-miR-490-5p	GSTM3	IPI00246975	1.084	0.610	Glutathione Transferase, GST

*The fold change of the miR-490 series can be found in table 3-2 column 4.

**The fold changes are all calculated through dividing ET group mean by that of normal controls. It is notable that the gene fold changes are quite moderate while the miRNA and protein expression levels are inversely related.

5.4 Additional thoughts on TRAP stimulated experiment

Besides the protein spectral quantification in 3 normal controls and 1 ET sample, we also performed an additional test by introducing thrombin-receptor-activating peptides (TRAP) stimulation to the G protein-coupled receptor pathway. The serine proteases thrombin and trypsin are both powerful platelet agonists that remove the tail of thrombin receptor and open the new C-terminal to stimulate the receptor. The synthetic peptides TRAP can mimic the functionality of thrombin and serve as a perfect agonist to activate the G-protein Coupled Receptor Signaling pathway (GCRS) which subsequently leads to calcium release and platelet aggregation (**Figure 5-3**).

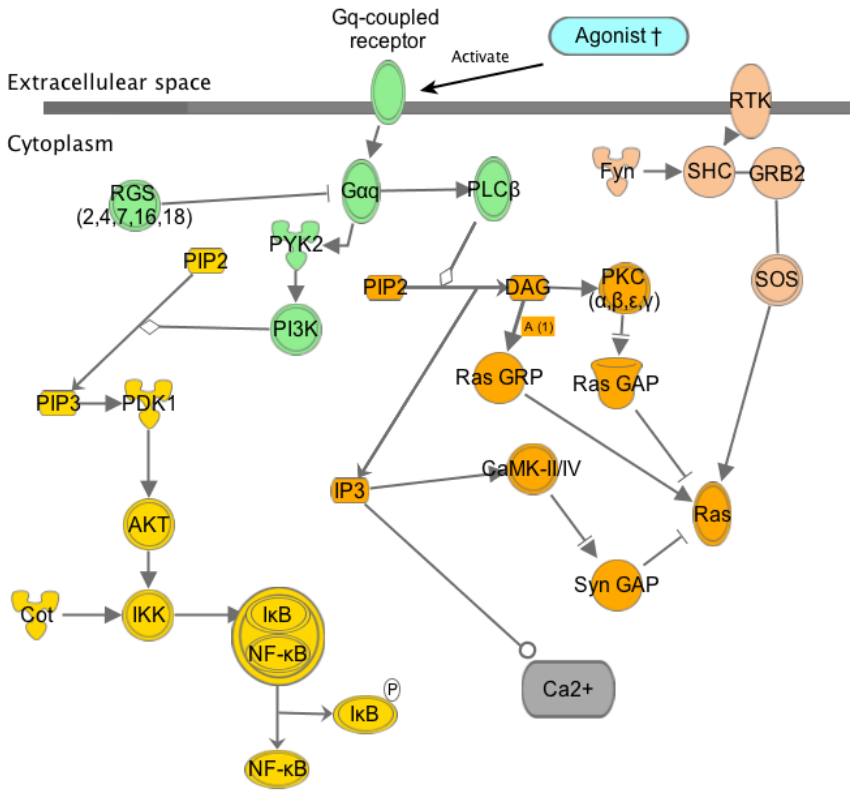

Figure 5-3 The activation of platelet G-protein coupled receptor pathway under the TRAP stimulation. In this experiment, thrombin-receptor-activating peptides is used as the agonist to activate the G-protein coupled receptor on platelet membrane. In this particular experiment, the target is the protease-activated receptor 1 (PAR1) marked in

blue color in graph. The overall effect of this TRAP stimulation process is the release of calcium ions (in grey color) from platelet cytoplasm and the subsequent initiation of platelet aggregation cascade. This pathway figure is generated from Ingenuity® software (Redwood, CA).

This experiment is similar to the MS procedure under regular conditions except that the protein level of platelets was measured the second time 5 min after the TRAP stimulation procedure. This way we can observe the translational changes of platelets in both regular and TRAP stimulated environment (**Figure 5-4**). This study was initially planned as the a separate research scheme, however, we noticed an interesting phenomenon in the 4 protein targets we identified using mir-490 series: in the ET samples of almost all 4 proteins, the TRAP stimulated protein output is significantly higher than that in regular conditions. Also, for protein IPI00299608, the TRAP stimulated translational output is highly unstable. Generally, it is interesting to find a highly different protein output level with/without TRAP stimulation for the given proteins. These preliminary observations have prompted us to think of the interactive regulatory effect of mir-490 series and TRAP stimulation. However, we still need to perform further confirmatory studies to validate the protein patterns showed in this preliminary study. All these results and assumptions will lend incredible assistance for us to look at the entire pathway in a system biological scheme.

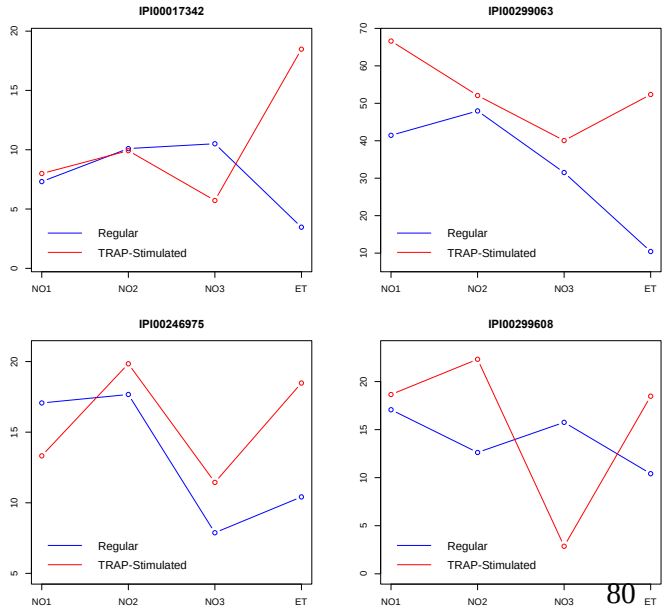

Figure 5-4. The expression levels of 4 protein targets under regular and TRAP stimulated condition (3 ET vs 1 NO). The first 3 data points indicate the expression level of each protein in normal controls and the last value represents protein expression level in ET. The red line denotes the TRAP-stimulated condition while the blue line signifies the regular expression level. On the general scale, the 3 normal samples have shown a consistent expression profile except for IPI00299608. However, the ET sample has demonstrated a notable difference between regular and TRAP-stimulated

5.5 Beyond miR-490 series: Other potential miRNA pathways

Throughout the study of the miRNA-mediated regulatory pathways in thrombocytosis, we have mostly focused on the two specific biomarker: hsa-mir-490-3p and hsa-mir-490-5p due to their significant cross-group differences and biological novelty. However, we are also interested in other potentially important miRNAs that might play a role in the pathological development of thrombocytosis. In the discussion of chapter 3, we have shown the biological relevance of some miRNAs in the hematopoietic system unveiled in previous studies. At this stage, miRNAs other than mir-490 series can also be studied on their targets given both genomic and proteomics data. On one hand, this part serves as a preliminary analysis for future expansion into searching of more valid miRNA targets; on the other hand, the correlational study from all miRNAs to protein targets can also be considered as a good validation step in the study of thrombocytosis. The basic research scheme of this section is mostly similar to our searches of protein target for mir-490 series. The only difference is that we no longer limit ourselves to 2 miRNAs (490-3p and 490-5p); instead we utilized the entire miRNA expression profiles to correlate with mRNA and proteins. This algorithm begins with a significance filter on all miRNA profiles from the Agilent miRNA microarray platform. A flag absence-call standard was enforced with more than 70% non-absent instances in any of the 2 groups (ET and Normal) be kept for subsequent procedure. A following SAM analysis without any fold change criterion was performed to select differentially expressed miRNAs and assigned into two groups: (in ET) the up-regulated and down-regulated cohorts. Simultaneously, the fold change of group means were documented for consistency score derivation. A similar procedure was repeated for mRNA profiles from Affymetrix GPL1716 platform, resulting in two lists of genomic features, one

being higher in ET than normal and the other is in opposite direction. Likewise, the fold change of group means are recorded, too. The final thread starts with proteomics data. Since we have very limited sample size from this platform (Normal, n = 3; ET, n =1), a simple fold change criterion is applied to divide all the spectral count entries into two groups 1) ET/NO: FC > 1 and 2) ET/NO: FC < 1. As a common practice, a filter (70% within-group presence) was applied before this step to remove protein candidates with too many missing values. The above 3 tracks converged under 2 correlational schemes: (1) mRNA and protein form consistent pairs (high positive correlation between the two profiles) and are inversely correlated to miRNA level (mechanism 1) as well as (2) protein and miRNA level are in reverse directions. In either of the correlational schemes, the regulatory pathway between mRNA and protein and from miRNA to mRNA has to be built on knowledge bases (gene annotation + target prediction algorithms see section 5.3). In this way, we have identified 2 lists of putative miRNA-mediated pathways following mechanism 1 (**Table 5-2**) and 2 (**Table 5-3**) respectively. These identified pathways are ranked by fold change consistency score (multiplicative product of fold changes) in descending order. Based on a minimum score criterion (score >= 40 in mechanism I and score >= 20 in mechanism II), top miRNA-mRNA-protein pathways were shown in this chapter as preliminary results for future exploration. One interesting observation from the results of mechanism 1 is that the absolute majority of the identified pathways have their miRNA profiles lower in ET group than normal with protein levels in opposite direction. This pattern is contradictory to the expression profiles of miR-490 series and that may partly explains why it is very difficult to identify miR-490 targets following mechanism 1. Based on the results of mechanism 2 (**Table 5-3**) where mRNA expression levels are not considered in the direct miRNA-protein linkage, the protein targets Alpha-endosulfine (ENSA) and epidermal surface antigen 2 (FLOT-2) recurrently appear in the top 20 pathway list. These two proteins appear to

be a common target to multiple miRNAs that are differentially expressed between ET and Normal. So far their relevance to the pathological development of ET is unclear and therefore the biological roles of these two targets need to be further explored in the setting of thrombocytosis. In summary, the exploration of mRNA and protein targets based on entire miRNA profiles can serve as a preliminary but important step towards the search of more miRNA-mediated regulatory pathways in thrombocytotic platelets. So far, we are still limited by the sample sizes of proteomics data and the lack of robustness in available mRNA targets prediction algorithms. Our next step will be set towards collecting more data at proteomics level and incorporate other computational targets prediction programs in our analysis.

Table 5-2 Identified miRNA-mRNA-protein pathways under mechanism 1 (pathways with score >= 40)

miRNA	Genes	Protein	Gene Fold-Change*	Protein Fold-Change*	miRNA Fold-Change*	Score**
hsa-miR-144*	SEP15	IPI00030877	7.44	6.61	0.43	113.21
hsa-miR-330-3p	SEP15	IPI00030877	7.44	6.61	0.5	98.87
hsa-miR-101*	SEP15	IPI00030877	7.44	6.61	0.58	85.09
hsa-miR-182	DAP	IPI00018117	9.86	2.75	0.32	84.05
hsa-miR-33a	SEP15	IPI00030877	7.44	6.61	0.65	76.02
hsa-miR-425*	SEP15	IPI00030877	7.44	6.61	0.7	70.71
hsa-miR-182	PROS1	IPI00294004	17.65	1.16	0.32	63.25
hsa-miR-182	ARHGAP6	IPI00011219	4.93	4.13	0.32	63.03
hsa-let-7d*	SH3GLB1	IPI00006558	3.09	10.42	0.51	62.84
hsa-miR-144	SH3GLB1	IPI00006558	3.09	10.42	0.52	62.28
hsa-miR-551b	ARHGAP6	IPI00011219	4.93	4.13	0.36	57.26
hsa-miR-181c	PSAP	IPI00012503	4.68	5.5	0.49	52.81
hsa-miR-144	DENND4C	IPI00100453	2.6	10.42	0.52	52.36
hsa-miR-330-3p	PSAP	IPI00012503	4.68	5.5	0.5	51.78
hsa-miR-1	DAP	IPI00018117	9.86	2.75	0.56	48.19
hsa-miR-181d	PSAP	IPI00012503	4.68	5.5	0.54	47.86
hsa-miR-181a	PSAP	IPI00012503	4.68	5.5	0.55	46.92
hsa-miR-454	DENND4C	IPI00100453	2.6	10.42	0.6	45.35
hsa-miR-330-3p	ATP6V1B2	IPI00007812	7.65	2.88	0.5	44.32
hsa-miR-423-5p	DAP	IPI00018117	9.86	2.75	0.62	44.06
hsa-miR-454	PSAP	IPI00012503	4.68	5.5	0.6	43.16
hsa-miR-183	DAP	IPI00018117	9.86	2.75	0.66	40.83
hsa-miR-183	DENND4C	IPI00100453	2.6	10.42	0.66	40.75

hsa-miR-423-5p	PSCD2	IPI00015228	2.37	10.42	0.62	40.11
hsa-miR-381	RAB2A	IPI00031169	0.18	0.81	1.83	12.53***

*All the gene fold changes are calculated through dividing ET group mean by normal controls.
** The concordance score is derived by multiplying all fold changes of mRNA and protein as well as the reciprocal of fold changes from miRNA.
*** This is the only pathway identified in which miRNA is down-regualted from Normal to ET. Although its score is far below 40, it is still listed as reference

Table 5-3 Identified miRNA-mRNA-protein pathways under mechanism 2 (pathways with score >= 20)

miRNA	Genes	Protein	Protein Fold-Change*	miRNA Fold-Change*	Score**
hsa-miR-34a	FLOT2	IPI00789008	0.07	4.90	71.65
hsa-miR-150	ENSA	IPI00220797	13.89	0.35	39.46
hsa-miR-424	HTRA2	IPI00001663	0.08	2.70	35.44
hsa-miR-148a	FLOT2	IPI00789008	0.07	2.06	30.16
hsa-miR-10a	ENSA	IPI00220797	13.89	0.47	29.78
hsa-miR-150	ARRB2	IPI00180375	10.42	0.35	29.60
hsa-miR-503	FLOT2	IPI00789008	0.07	1.96	28.67
hsa-miR-181c	ENSA	IPI00220797	13.89	0.49	28.46
hsa-miR-330-3p	ENSA	IPI00220797	13.89	0.50	27.90
hsa-miR-181d	ENSA	IPI00220797	13.89	0.54	25.79
hsa-miR-143*	ENSA	IPI00220797	13.89	0.54	25.77
hsa-miR-181a	ENSA	IPI00220797	13.89	0.55	25.28
hsa-miR-23a*	ENSA	IPI00220797	13.89	0.55	25.10
hsa-miR-30b*	ENSA	IPI00220797	13.89	0.56	24.73
hsa-miR-10b	ENSA	IPI00220797	13.89	0.60	23.09
hsa-miR-625	FLOT2	IPI00789008	0.07	1.55	22.67
hsa-miR-449a	FLOT2	IPI00789008	0.07	1.54	22.54
hsa-miR-143	ENSA	IPI00220797	13.89	0.63	22.17
hsa-miR-30c-1*	ENSA	IPI00220797	13.89	0.63	21.88
hsa-miR-30d*	ENSA	IPI00220797	13.89	0.64	21.54
hsa-miR-487b	HTRA2	IPI00001663	0.08	1.64	21.49
hsa-miR-33a	ENSA	IPI00220797	13.89	0.65	21.45
hsa-miR-181c	ARRB2	IPI00180375	10.42	0.49	21.34
hsa-miR-181b	ENSA	IPI00220797	13.89	0.66	20.92
hsa-miR-342-3p	ENSA	IPI00220797	13.89	0.68	20.43
hsa-miR-625	HTRA2	IPI00001663	0.08	1.55	20.35
hsa-miR-766	ENSA	IPI00220797	13.89	0.68	20.32
hsa-miR-769-5p	ENSA	IPI00220797	13.89	0.69	20.19
hsa-miR-423-3p	ENSA	IPI00220797	13.89	0.69	20.09

*All the gene fold changes are calculated through dividing ET group mean by normal controls.

** The concordance score is derived by multiplying fold changes of protein and the reciprocal of fold changes from miRNA expression.

5.6 Summary of Our Study on Thrombocytosis in A Systems Biology Framework

Throughout the study of thrombocytosis in human platelet model, we have worked out a research plan to view this problem from different perspectives and sought to construct a complete thread to track the biological changes at every developmental stage of a blood cell. In our previous study of thrombocytotic class prediction model, we developed a linear discriminant algorithm sampling on the genomic profiles of a cohort study to delineate among normal controls and patients with ET or RT, as well as between ET and RT subtypes only. In this particular study, we initially formulated the goal of our research as a biomarker searching process in response to the current clinic challenge of lacking in differential gene features between ET and RT. The study was based on the characterization of platelet-specific genomes on a two-channel custom array with quantitative PCR experiment as a validation platform. Using step-wise feature selection method, we identified 11 genetic biomarkers that differentiate among the 3 groups or between the two thrombocytosis subtypes. Another set of 4 gene features was also confirmed for their role of classification between ET JAK2 617F positive and JAK2 617 negative. This layer of analysis in thrombocytotic genome has served as a critical link in the entire system biology scheme that it associates clinical phenotypic features to the underlying genetic cause. Based on available evidences (discussed in C-3), we advanced our exploration of thrombocytosis into the next step where the microRNA modulation was studied and correlated to the already established genomic profile. At this stage of analysis, we sought to study the important miRNA features between the ET group and normal controls and this led us to a list of 21 miRNA biomarkers. A further quantitative PCR validation experiment has well confirmed the robustness of our finding.

Down in the same track, a supervised classification analysis using an ensemble of 3 classifiers was applied to narrow down our focus to a subset of miRNAs of optimal prediction power. After carefully reviewing the biological relevance, novelty and statistical significance, the mir-490 series were singled out and brought into the subsequent mRNA target searching process. This miRNA-mRNA correlational analysis is the most important link of the entire system biology scheme as it connects the layer of miRNA regulation to mRNA targets and further with the final protein output.

Equally critical as its significance, the current miRNA target prediction algorithms have severe limitations, including low sensitivity, lack of robustness, short in consistency as well as the prevalence of false discoveries in some of the programs [175, 220]. We sought to compensate for the uncertainty of computationally identified targets by applying additional mathematical filters. In particular, the profiles of mir-490s between ET and Normal were correlated to existing genomics data based on two data sources: 1) custom microarray and 2) Affymetrix GPL 1716 set. However, in real biological system, the relationship between miRNAs and mRNA targets are complicated by alternative competing mechanisms (I and II), therefore we can only model the expression pattern of miRNA-mRNA pairs following one assumption: miRNA-mediated target degradation. In the case of miRNA-mediated translational repression (mechanism II), the targets of miRNA can only be observed by monitoring the protein level. In either of the cases, a final biological validation is indispensable and this procedure is still undergoing right now. In addition, another on-going experiment: the thrombocytotic proteomics quantification has generated some preliminary results on 3 normal controls and 1 ET sample, which enabled us to, at least speculatively, probe into the miRNA protein targets and putative regulatory pathways for both mechanisms. So far the protein data source is still limited by the small sample size and lack of robustness. We are expected to expand this part of analysis and include more ET patients to

correlate with mir-490 series. Moreover, with the preliminary analysis done in correlating entire miRNA profiles with mRNA and protein output, we are also trying to project out attention into biomarkers other than mir-490 series. In the end, the new interfering factor: TRAP stimulation is associated with the primary activation pathway in platelets and studying its effect in both normal and thrombocytotic samples should help unveil the underlying mechanism of platelet activation process in ET context. All the interactions of genetic and epigenetic factors are to be studied in close association with each other. Together, these elements form the whole system biology research network for our study of thrombocytosis and platelet functionalities (**Figure 1-2**). There are other branches of study we conducted before on hemocytoblast cells or other interfering factor, but those are not included in this dissertation due to the direction of my own research focus.

5.7 Future Work

At the current stage, we have studied the genomic profiles, microRNA expression and protein output levels on both thrombocytotic platelets (ET and RT) as well as those in normal controls. Under the system biology scheme, we planned to explore the entire cellular regulatory pathway originated from miRNA to final peptide production under both normal condition and the pathological environment of thrombocytosis. The current analysis has been mainly focused on the model of platelet system alone the $CD34^+$ hematopoietic stem cell differentiation lineage. In the next step, we are going to expand our research concentration of miRNA mediated mRNA-protein output pathway into the upper stream megakaryocyte (MK) system and correlate the expression profiles of miRNA, mRNA and protein to their counterparts in platelets. In fact, we have already studied the genomic profiles in MK cell under normal conditions using the platelet-

specific custom microarray platform (unpublished manuscript). This section of study is not included in this thesis but it will be an important link in the entire system biology scheme of CD 34^+ hematopoietic stem cell lineage. Once we studied the complete regulatory pathway in MK cell system in both thrombocytotic patients and normal controls, a further expansion of the study will be conducted to investigate the cellular regulation response during TRAP stimulation. At current stage, we have finished a preliminary MS experiment to quantify protein output in TRAP stimulated platelets, a more comprehensive investigation is already planned to measure the miRNA and mRNA expression level in the same system. Eventually, our goal is to correlate the expression profiles alone the miRNA→mRNA→protein pathway in each of the cellular systems through the hematopoietic stem cell differentiation process. This way we hope to unveil the complete biological regulatory pathways underlying the development of thrombocytosis as well as other physiological conditions.

Reference

1. Hoffman, *Hematology:Basic Principles and Practice*. 4 ed. Chapter 17. Vol. 71. 2005: Churchill Livingstone.
2. Briere, J.B., *Essential thrombocythemia.* Orphanet journal of rare diseases, 2007. **2**(3).
3. Zhan, H. and J.L. Spivak, *The diagnosis and management of polycythemia vera, essential thrombocythemia, and primary myelofibrosis in the JAK2 V617F era.* Clinical advances in hematology & oncology : H&O, 2009. **7**(5): p. 334-42.
4. Kutti, J. and B. Ridell, *Epidemiology of the myeloproliferative disorders: essential thrombocythaemia, polycythaemia vera and idiopathic myelofibrosis.* Pathol Biol (Paris), 2001. **49**(2): p. 164-6.
5. Tefferi, A. and S. Murphy, *Current opinion in essential thrombocythemia: pathogenesis, diagnosis, and management.* Blood Rev, 2001. **15**(3): p. 121-31.
6. Vannucchi, A.M., *Insights into the pathogenesis and management of thrombosis in polycythemia vera and essential thrombocythemia.* Intern Emerg Med, 2010. **5**(3): p. 177-84.
7. Johansson, P., *Epidemiology of the myeloproliferative disorders polycythemia vera and essential thrombocythemia.* Semin Thromb Hemost, 2006. **32**(3): p. 171-3.
8. Beer, P.A. and A.R. Green, *Pathogenesis and management of essential thrombocythemia.* Hematology Am Soc Hematol Educ Program, 2009: p. 621-8.
9. Barbui, T. and G. Finazzi, *Treatment indications and choice of a platelet-lowering agent in essential thrombocythemia.* Current hematology reports, 2003. **2**(3): p. 248-56.
10. Chistolini, A., et al., *Essential thrombocythemia: a retrospective study on the clinical course of 100 patients.* Haematologica, 1990. **75**(6): p. 537-40.
11. Colombi, M., et al., *Thrombotic and hemorrhagic complications in essential thrombocythemia. A retrospective study of 103 patients.* Cancer, 1991. **67**(11): p. 2926-30.
12. Michiels, J.J., et al., *Current diagnostic criteria for the chronic myeloproliferative disorders (MPD) essential thrombocythemia (ET), polycythemia vera (PV) and chronic idiopathic myelofibrosis (CIMF).* Pathologie-biologie, 2007. **55**(2): p. 92-104.
13. Harrison, C.N. and S.J. Machin. *What is thrombocytosis?* Thrombocytosis and essential thrombocythaemia Netdoctor 2006 [cited 2011; Available from: http://www.netdoctor.co.uk/diseases/facts/thrombocytosis.htm.
14. Hellmann, A., *Myeloproliferative syndromes: diagnosis and therapeutic options.* Polskie Archiwum Medycyny Wewnetrznej, 2008. **118**(12): p. 756-60.
15. Khan, P.N., et al., *Postsplenectomy reactive thrombocytosis.* Proceedings, 2009. **22**(1): p. 9-12.
16. Saha, S.A., *Thrombocytosis in patients with polycystic ovary syndrome: reactive or culpable?* Southern medical journal, 2011. **104**(4): p. 252-3.
17. Epstein, E. and A. Goedel, *Hammorhagische thrombocythamie bei vasculerer schrumpfmilz.* Virch Arch (Pathol Anat). 1934: p. 292:233.
18. Levine, R.L., et al., *Role of JAK2 in the pathogenesis and therapy of myeloproliferative disorders.* Nature reviews. Cancer, 2007. **7**(9): p. 673-83.

19. Levine, R.L., et al., *Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis.* Cancer Cell, 2005. **7**(4): p. 387-97.
20. Oh, S.T. and J. Gotlib, *JAK2 V617F and beyond: role of genetics and aberrant signaling in the pathogenesis of myeloproliferative neoplasms.* Expert review of hematology, 2010. **3**(3): p. 323-37.
21. Saemann, M.D., et al., *Janus kinase-3 (JAK3) inhibition: a novel immunosuppressive option for allogeneic transplantation.* Transplant international : official journal of the European Society for Organ Transplantation, 2004. **17**(9): p. 481-9.
22. Nashan, B., *Review of T-cell activation: impact of Janus kinase 3 inhibition.* Transplantation, 2003. **75**(11): p. 1783-5.
23. Yeh, T.C. and S. Pellegrini, *The Janus kinase family of protein tyrosine kinases and their role in signaling.* Cellular and molecular life sciences : CMLS, 1999. **55**(12): p. 1523-34.
24. Ihle, J.N., *The Janus protein tyrosine kinase family and its role in cytokine signaling.* Advances in immunology, 1995. **60**: p. 1-35.
25. Bellucci, S. and J.J. Michiels, *The role of JAK2 V617F mutation, spontaneous erythropoiesis and megakaryocytopoiesis, hypersensitive platelets, activated leukocytes, and endothelial cells in the etiology of thrombotic manifestations in polycythemia vera and essential thrombocythemia.* Seminars in thrombosis and hemostasis, 2006. **32**(4 Pt 2): p. 381-98.
26. Yamada, M., et al., *Thrombopoietin supports in vitro erythroid differentiation via its specific receptor c-Mpl in a human leukemia cell line.* Cell growth & differentiation : the molecular biology journal of the American Association for Cancer Research, 1998. **9**(6): p. 487-96.
27. Briere, J. and F. Guilmin, *Management of patients with essential thrombocythemia: current concepts and perspectives.* Pathologie-biologie, 2001. **49**(2): p. 178-83.
28. Finazzi, G. and T. Barbui, *Evidence and expertise in the management of polycythemia vera and essential thrombocythemia.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2008. **22**(8): p. 1494-502.
29. Spivak, J.L. and R.T. Silver, *The revised World Health Organization diagnostic criteria for polycythemia vera, essential thrombocytosis, and primary myelofibrosis: an alternative proposal.* Blood, 2008. **112**(2): p. 231-9.
30. Michiels, J.J., et al., *WHO bone marrow features and European clinical, molecular, and pathological (ECMP) criteria for the diagnosis of myeloproliferative disorders.* Leukemia research, 2007. **31**(8): p. 1031-8.
31. Tefferi, A., *Essential thrombocythemia, polycythemia vera, and myelofibrosis: current management and the prospect of targeted therapy.* American journal of hematology, 2008. **83**(6): p. 491-7.
32. Gnatenko, D.V., et al., *Class prediction models of thrombocytosis using genetic biomarkers.* Blood, 2010. **115**(1): p. 7-14.
33. Tefferi, A., M.N. Silverstein, and H.C. Hoagland, *Primary thrombocythemia.* Seminars in oncology, 1995. **22**(4): p. 334-40.
34. Klinger, M.H. and W. Jelkmann, *Role of blood platelets in infection and inflammation.* Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research, 2002. **22**(9): p. 913-22.
35. Franchini, M., et al., *Iron and thrombosis.* Annals of hematology, 2008. **87**(3): p. 167-73.

36. Keung, Y.K. and J. Owen, *Iron deficiency and thrombosis: literature review.* Clinical and applied thrombosis/hemostasis : official journal of the International Academy of Clinical and Applied Thrombosis/Hemostasis, 2004. **10**(4): p. 387-91.
37. Greist, A., *The role of blood component removal in essential and reactive thrombocytosis.* Therapeutic apheresis : official journal of the International Society for Apheresis and the Japanese Society for Apheresis, 2002. **6**(1): p. 36-44.
38. Michiels, J.J., et al., *WHO bone marrow features and European clinical, molecular, and pathological (ECMP) criteria for the diagnosis of myeloproliferative disorders.* Leuk Res, 2007. **31**(8): p. 1031-8.
39. Rothlisberger, B., et al., *JAK2 exon 12 mutation in JAK2V617F-negative polycythemia vera.* Leukemia & lymphoma, 2008. **49**(3): p. 586-8.
40. Inami, M., et al., *Analysis of the exon 12 and 14 mutations of the JAK2 gene in Philadelphia chromosome-positive leukemia.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2008. **22**(1): p. 216.
41. Scott, L.M., et al., *JAK2 exon 12 mutations in polycythemia vera and idiopathic erythrocytosis.* The New England journal of medicine, 2007. **356**(5): p. 459-68.
42. Percy, M.J. and M.F. McMullin, *The V617F JAK2 mutation and the myeloproliferative disorders.* Hematol Oncol, 2005. **23**(3-4): p. 91-3.
43. Martinez-Aviles, L., et al., *JAK2 exon 12 mutations in polycythemia vera or idiopathic erythrocytosis.* Haematologica, 2007. **92**(12): p. 1717-8.
44. Saint-Martin, C., et al., *Analysis of the ten-eleven translocation 2 (TET2) gene in familial myeloproliferative neoplasms.* Blood, 2009. **114**(8): p. 1628-32.
45. Tefferi, A., K.H. Lim, and R. Levine, *Mutation in TET2 in myeloid cancers.* The New England journal of medicine, 2009. **361**(11): p. 1117; author reply 1117-8.
46. Tefferi, A., *Novel mutations and their functional and clinical relevance in myeloproliferative neoplasms: JAK2, MPL, TET2, ASXL1, CBL, IDH and IKZF1.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2010. **24**(6): p. 1128-38.
47. Naramura M, N.S., et al., *Mutant Cbl proteins as oncogenic drivers in myeloproliferative disorders.* Oncotarget, 2011. **2**(3): p. 245-250.
48. Skoda, R.C., *Thrombocytosis.* Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program, 2009: p. 159-67.
49. Levine, R.L., *Mechanisms of mutations in myeloproliferative neoplasms.* Best practice & research. Clinical haematology, 2009. **22**(4): p. 489-94.
50. Passamonti, F., et al., *Myeloproliferative neoplasms: From JAK2 mutations discovery to JAK2 inhibitor therapies.* Oncotarget, 2011.
51. Melis, S., et al., *JAK2 V617F mutation and PRV-1 overexpression: relevance in the diagnosis of polycythaemia vera and essential thrombocythaemia.* Acta clinica Belgica, 2009. **64**(5): p. 429-33.
52. Tutaeva, V., et al., *Application of PRV-1 mRNA expression level and JAK2V617F mutation for the differentiating between polycytemia vera and secondary erythrocytosis and assessment of treatment by interferon or hydroxyurea.* Hematology, 2007. **12**(6): p. 473-9.
53. Teofili, L., et al., *Markers of myeloproliferative diseases in childhood polycythemia vera and essential thrombocythemia.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2007. **25**(9): p. 1048-53.

54. Bushati, N. and S.M. Cohen, *microRNA functions.* Annu Rev Cell Dev Biol, 2007. **23**: p. 175-205.
55. Merkerova, M., M. Belickova, and H. Bruchova, *Differential expression of microRNAs in hematopoietic cell lineages.* Eur J Haematol, 2008. **81**(4): p. 304-10.
56. Yendamuri, S. and G.A. Calin, *The role of microRNA in human leukemia: a review.* Leukemia, 2009. **23**(7): p. 1257-63.
57. Chen, C.Z., et al., *MicroRNAs modulate hematopoietic lineage differentiation.* Science, 2004. **303**(5654): p. 83-6.
58. Garzon, R., et al., *MicroRNA fingerprints during human megakaryocytopoiesis.* Proc Natl Acad Sci U S A, 2006. **103**(13): p. 5078-83.
59. Zhou, B., et al., *miR-150, a microRNA expressed in mature B and T cells, blocks early B cell development when expressed prematurely.* Proc Natl Acad Sci U S A, 2007. **104**(17): p. 7080-5.
60. Lu, J., et al., *MicroRNA-mediatedcontrol of cell fate in megakaryocyte-erythrocyte progenitors.* Dev. Cell, 2008. **14**: p. 843-853.
61. Bruchova, H., M. Merkerova, and J.T. Prchal, *Aberrant expression of microRNA in polycythemia vera.* Haematologica, 2008. **93**(7): p. 1009-16.
62. Bruchova, H., et al., *Regulated expression of microRNAs in normal and polycythemia vera erythropoiesis.* Exp Hematol, 2007. **35**(11): p. 1657-67.
63. Slezak, S., et al., *Gene and microRNA analysis of neutrophils from patients with polycythemia vera and essential thrombocytosis: down-regulation of micro RNA-1 and -133a.* J Transl Med, 2009. **7**: p. 39.
64. Gnatenko, D.V., et al., *Transcript profiling of human platelets using microarray and serial analysis of gene expression.* Blood, 2003. **101**(6): p. 2285-93.
65. McRedmond, J.P., et al., *Integration of Proteomics and Genomics in Platelets: A profile of platelet proteins and platelet-specific genes.* Mol Cell Proteomics, 2004. **3**(2): p. 133-144.
66. Landry, P., et al., *Existence of a microRNA pathway in anucleate platelets.* Nat Struct Mol Biol, 2009. **16**(9): p. 961-6.
67. Gnatenko, D.V., et al., *Identification of Platelet microRNA Signatures from Patients with Thrombocytosis*, 2011, Stony Brook University: Stony Brook.
68. Ofosu, F.A. and K.A. Nyarko, *Human platelet thrombin receptors. Roles in platelet activation.* Hematol Oncol Clin North Am, 2000. **14**(5): p. 1185-98, x.
69. Patel, S.R., J.H. Hartwig, and J.E. Italiano, Jr., *The biogenesis of platelets from megakaryocyte proplatelets.* The Journal of clinical investigation, 2005. **115**(12): p. 3348-54.
70. Bang, S.M., et al., *Diagnostic usefulness of the Janus kinase 2 mutation in non BCR/ABL myeloproliferative disorders.* The Korean journal of internal medicine, 2006. **21**(4): p. 219-24.
71. el-Kassar, N., et al., *Clonality analysis of hematopoiesis in essential thrombocythemia: advantages of studying T lymphocytes and platelets.* Blood, 1997. **89**(1): p. 128-34.
72. Ammatuna, E., et al., *Diagnostic refinement of chronic myeloproliferative disorders and thrombocytoses of unknown origin by multiple RT-PCR and capillary electrophoresis of BCR-ABL rearrangements and JAK2 (V617F) mutation.* Annals of hematology, 2007. **86**(5): p. 355-61.
73. Dmitri, G., et al., *Class prediction models of thrombocytosis using genetic biomarkers.* blood, 2010. **115**: p. 7-14.

74. Puigdecanet, E., et al., *Gene expression profiling distinguishes JAK2V617F-negative from JAK2V617F-positive patients in essential thrombocythemia.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2008. **22**(7): p. 1368-76.
75. Vainchenker, W., et al., *New mutations and pathogenesis of myeloproliferative neoplasms.* Blood, 2011.
76. Xing, E.P., M.I. Jordan, and R.M. Karp, *Feature Selection for High-Dimensional Genomic Microarray Data,* in *Machine Learning -International Workshop*2001. p. 601-608.
77. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.
78. Krol, J., I. Loedige, and W. Filipowicz, *The widespread regulation of microRNA biogenesis, function and decay.* Nat Rev Genet, 2010. **11**(9): p. 597-610.
79. Lee, E.J., et al., *Systematic evaluation of microRNA processing patterns in tissues, cell lines, and tumors.* RNA, 2008. **14**(1): p. 35-42.
80. Conrad, R., M. Barrier, and L.P. Ford, *Role of miRNA and miRNA processing factors in development and disease.* Birth defects research. Part C, Embryo today : reviews, 2006. **78**(2): p. 107-17.
81. Wang, Y., et al., *Mechanism of microRNA-target interaction: molecular dynamics simulations and thermodynamics analysis.* PLoS Comput Biol, 2010. **6**(7): p. e1000866.
82. Behm-Ansmant, I., J. Rehwinkel, and E. Izaurralde, *MicroRNAs silence gene expression by repressing protein expression and/or by promoting mRNA decay.* Cold Spring Harbor symposia on quantitative biology, 2006. **71**: p. 523-30.
83. Jackson, R.J. and N. Standart, *How do microRNAs regulate gene expression?* Science's STKE : signal transduction knowledge environment, 2007. **2007**(367): p. re1.
84. Sun, Y.H., et al., *Computational prediction of plant miRNA targets.* Methods in molecular biology, 2011. **744**: p. 175-86.
85. Kuhn, D.E., et al., *Experimental validation of miRNA targets.* Methods, 2008. **44**(1): p. 47-54.
86. Zhang, Y., *miRU: an automated plant miRNA target prediction server.* Nucleic acids research, 2005. **33**(Web Server issue): p. W701-4.
87. Dai, X., Z. Zhuang, and P.X. Zhao, *Computational analysis of miRNA targets in plants: current status and challenges.* Briefings in Bioinformatics, 2011. **12**(2): p. 115-21.
88. Joung, J.G. and Z. Fei, *Computational identification of condition-specific miRNA targets based on gene expression profiles and sequence information.* BMC Bioinformatics, 2009. **10 Suppl 1**: p. S34.
89. Kiriakidou, M., et al., *A combined computational-experimental approach predicts human microRNA targets.* Genes & development, 2004. **18**(10): p. 1165-78.
90. Gaidatzis, D., et al., *Inference of miRNA targets using evolutionary conservation and pathway analysis.* BMC Bioinformatics, 2007. **8**: p. 69.
91. John, B., et al., *Human MicroRNA targets.* PLoS biology, 2004. **2**(11): p. e363.
92. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.
93. Lewis, B.P., et al., *Prediction of mammalian microRNA targets.* Cell, 2003. **115**(7): p. 787-98.

94. Lall, S., et al., *A genome-wide map of conserved microRNA targets in C. elegans.* Current biology : CB, 2006. **16**(5): p. 460-71.
95. Long, D., C.Y. Chan, and Y. Ding, *Analysis of microRNA-target interactions by a target structure based hybridization model.* Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2008: p. 64-74.
96. Hsu, S.D., et al., *miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes.* Nucleic acids research, 2008. **36**(Database issue): p. D165-9.
97. Lindow, M. and A. Krogh, *Computational evidence for hundreds of non-conserved plant microRNAs.* BMC Genomics, 2005. **6**: p. 119.
98. Monticelli, S., et al., *MicroRNA profiling of the murine hematopoietic system.* Genome biology, 2005. **6**(8): p. R71.
99. Merkerova, M., M. Belickova, and H. Bruchova, *Differential expression of microRNAs in hematopoietic cell lineages.* European journal of haematology, 2008. **81**(4): p. 304-10.
100. Ramkissoon, S.H., et al., *Hematopoietic-specific microRNA expression in human cells.* Leukemia research, 2006. **30**(5): p. 643-7.
101. Shivdasani, R.A., *MicroRNAs: regulators of gene expression and cell differentiation.* Blood, 2006. **108**(12): p. 3646-53.
102. Chen, C.Z. and H.F. Lodish, *MicroRNAs as regulators of mammalian hematopoiesis.* Seminars in immunology, 2005. **17**(2): p. 155-65.
103. Lu, J., et al., *MicroRNA-mediated control of cell fate in megakaryocyte-erythrocyte progenitors.* Dev Cell, 2008. **14**(6): p. 843-53.
104. Felli, N., et al., *MicroRNAs 221 and 222 inhibit normal erythropoiesis and erythroleukemic cell growth via kit receptor down-modulation.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(50): p. 18081-6.
105. Havelange, V. and R. Garzon, *MicroRNAs: emerging key regulators of hematopoiesis.* American journal of hematology, 2010. **85**(12): p. 935-42.
106. Calin, G.A., et al., *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 of the United States of America, 2002. **99**(24): p. 15524-9.
107. Girardot, M., et al., *miR-28 is a thrombopoietin receptor targeting microRNA detected in a fraction of myeloproliferative neoplasm patient platelets.* Blood, 2010. **116**(3): p. 437-45.
108. Garzon, R. and C.M. Croce, *MicroRNAs in normal and malignant hematopoiesis.* Current opinion in hematology, 2008. **15**(4): p. 352-8.
109. Ota, A., et al., *Identification and characterization of a novel gene, C13orf25, as a target for 13q31-q32 amplification in malignant lymphoma.* Cancer research, 2004. **64**(9): p. 3087-95.
110. Bahou, W.F., *Genetic disorders of platelets*, in *Genomics and Clinical Medicine*, D. Kumar, Editor 2006, Oxford Univesity Press: Oxford.
111. Rio, D.C., et al., *Purification of RNA using TRIzol (TRI reagent).* Cold Spring Harb Protoc, 2010. **2010**(6): p. pdb prot5439.
112. Gaikwad, A., et al., *In vitro expansion of erythroid progenitors from polycythemia vera patients leads to decrease in JAK2 V617F allele.* Exp Hematol, 2007. **35**(4): p. 587-95.
113. Caramuta, S., et al., *MicroRNA expression profiles associated with mutational status and survival in malignant melanoma.* J Invest Dermatol, 2010. **130**(8): p. 2062-70.
114. Agilent, T., *miRNA Microarray System with miRNA Complete Labeling and Hyb Kit 2.3*, in *Characterization of Total RNA*2011, Agilent Literature in press.

115. Lopez-Romero, P., et al., *Processing of Agilent microRNA array data.* BMC Res Notes, 2010. **3**: p. 18.
116. Livak, K.J. and T.D. Schmittgen, *Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.* Methods, 2001. **25**(4): p. 402-8.
117. Delenstarr, G., S. Vacha, and M. Harnett, *Evaluating the Reproducibility of Microarray Technology*, 2011, Agilent, Technologies.
118. Troyanskaya, O., et al., *Missing value estimation methods for DNA microarrays.* Bioinformatics, 2001. **17**(6): p. 520-5.
119. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.
120. Pradervand, S., et al., *Impact of normalization on miRNA microarray expression profiling.* RNA, 2009. **15**(3): p. 493-501.
121. Blanco, R., et al., *Gene selection for cancer classification using wrapper approaches.* International Journal of Pattern Recognition and Artificial Intelligence, 2004. **18**(8): p. 1373-1390.
122. Cawley, G.C. and N.L. Talbot, *Fast exact leave-one-out cross-validation of sparse least-squares support vector machines.* Neural Netw, 2004. **17**(10): p. 1467-75.
123. Werner, E., *The music of life - Biology beyond the genome.* Science, 2007. **317**(5839): p. 753-754.
124. Ballestar, E., *An introduction to epigenetics.* Advances in experimental medicine and biology, 2011. **711**: p. 1-11.
125. Herceg, Z. and T. Ushijima, *Introduction: epigenetics and cancer.* Advances in genetics, 2010. **70**: p. 1-23.
126. Guglielmelli, P., et al., *MicroRNA expression profile in granulocytes from primary myelofibrosis patients.* Experimental hematology, 2007. **35**(11): p. 1708-18.
127. Yang, J.S., et al., *Widespread regulatory activity of vertebrate microRNA* species.* RNA, 2011. **17**(2): p. 312-26.
128. Zhou, H., et al., *miR-155 and its star-form partner miR-155* cooperatively regulate type I interferon production by human plasmacytoid dendritic cells.* Blood, 2010. **116**(26): p. 5885-94.
129. Nass, D., et al., *MiR-92b and miR-9/9* are specifically expressed in brain primary tumors and can be used to differentiate primary from metastatic brain tumors.* Brain Pathol, 2009. **19**(3): p. 375-83.
130. Hussein, K., et al., *Expression of myelopoiesis-associated microRNA in bone marrow cells of atypical chronic myeloid leukaemia and chronic myelomonocytic leukaemia.* Annals of hematology, 2011. **90**(3): p. 307-13.
131. Agirre, X., et al., *Down-regulation of hsa-miR-10a in chronic myeloid leukemia CD34+ cells increases USF2-mediated cell growth.* Molecular cancer research : MCR, 2008. **6**(12): p. 1830-40.
132. Suresh, S., et al., *MicroRNAs 130a/b are regulated by BCR-ABL and downregulate expression of CCN3 in CML.* Journal of cell communication and signaling, 2011.
133. Merkerova, M., et al., *MicroRNA expression profiles in umbilical cord blood cell lineages.* Stem cells and development, 2010. **19**(1): p. 17-26.
134. He, L. and G.J. Hannon, *MicroRNAs: small RNAs with a big role in gene regulation.* Nature reviews. Genetics, 2004. **5**(7): p. 522-31.

135. Inui, M., G. Martello, and S. Piccolo, *MicroRNA control of signal transduction.* Nature reviews. Molecular cell biology, 2010. **11**(4): p. 252-63.
136. Shcherbata, H.R., et al., *The MicroRNA pathway plays a regulatory role in stem cell division.* Cell Cycle, 2006. **5**(2): p. 172-5.
137. Lawrie, C.H., *MicroRNA expression in lymphoma.* Expert opinion on biological therapy, 2007. **7**(9): p. 1363-74.
138. Lawrie, C.H., *microRNA expression in erythropoiesis and erythroid disorders.* British journal of haematology, 2010. **150**(2): p. 144-51.
139. Marcucci, G., et al., *MicroRNA expression profiling in acute myeloid and chronic lymphocytic leukaemias.* Best practice & research. Clinical haematology, 2009. **22**(2): p. 239-48.
140. Kluiver, J., et al., *The role of microRNAs in normal hematopoiesis and hematopoietic malignancies.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2006. **20**(11): p. 1931-6.
141. Pelosi, E., C. Labbaye, and U. Testa, *MicroRNAs in normal and malignant myelopoiesis.* Leukemia research, 2009. **33**(12): p. 1584-93.
142. Lee, Y., et al., *The nuclear RNase III Drosha initiates microRNA processing.* Nature, 2003. **425**(6956): p. 415-9.
143. Chen, K. and N. Rajewsky, *The evolution of gene regulation by transcription factors and microRNAs.* Nature reviews. Genetics, 2007. **8**(2): p. 93-103.
144. Brodersen, P. and O. Voinnet, *Revisiting the principles of microRNA target recognition and mode of action.* Nature reviews. Molecular cell biology, 2009. **10**(2): p. 141-8.
145. Rajewsky, N., *microRNA target predictions in animals.* Nature genetics, 2006. **38 Suppl**: p. S8-13.
146. Dugas, D.V. and B. Bartel, *MicroRNA regulation of gene expression in plants.* Current opinion in plant biology, 2004. **7**(5): p. 512-20.
147. Maziere, P. and A.J. Enright, *Prediction of microRNA targets.* Drug discovery today, 2007. **12**(11-12): p. 452-8.
148. Brennecke, J., et al., *Principles of microRNA-target recognition.* PLoS biology, 2005. **3**(3): p. e85.
149. Long, D. *ANALYSIS OF MICRORNA-TARGET INTERACTIONS BY A TARGET STRUCTURE BASED HYBRIDIZATION MODEL*. in *Pacific Symposium on Biocomputing*. 2008.
150. Li, L., et al., *Computational approaches for microRNA studies: a review.* Mammalian genome : official journal of the International Mammalian Genome Society, 2010. **21**(1-2): p. 1-12.
151. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.
152. Enright, A.J., et al., *MicroRNA targets in Drosophila.* Genome biology, 2003. **5**(1): p. R1.
153. Stark, A., et al., *Identification of Drosophila MicroRNA targets.* PLoS biology, 2003. **1**(3): p. E60.
154. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.
155. Laurent, L.C., et al., *Comprehensive microRNA profiling reveals a unique human embryonic stem cell signature dominated by a single seed sequence.* Stem Cells, 2008. **26**(6): p. 1506-16.

156. Grimson, A., et al., *MicroRNA targeting specificity in mammals: determinants beyond seed pairing.* Molecular cell, 2007. **27**(1): p. 91-105.
157. Rajewsky, N. and N.D. Socci, *Computational identification of microRNA targets.* Developmental biology, 2004. **267**(2): p. 529-35.
158. Lin, S., et al., *Computational identification and characterization of primate-specific microRNAs in human genome.* Computational biology and chemistry, 2010. **34**(4): p. 232-41.
159. Bentwich, I., et al., *Identification of hundreds of conserved and nonconserved human microRNAs.* Nature genetics, 2005. **37**(7): p. 766-70.
160. John, B., C. Sander, and D.S. Marks, *Prediction of human microRNA targets.* Methods in molecular biology, 2006. **342**: p. 101-13.
161. Martin, G., et al., *Prediction and validation of microRNA targets in animal genomes.* Journal of biosciences, 2007. **32**(6): p. 1049-52.
162. Rehmsmeier, M., *Prediction of microRNA targets.* Methods in molecular biology, 2006. **342**: p. 87-99.
163. Wang, X., et al., *Identification of phylogenetically conserved microRNA cis-regulatory elements across 12 Drosophila species.* Bioinformatics, 2008. **24**(2): p. 165-71.
164. Thomas, L.F., T. Saito, and P. Saetrom, *Inferring causative variants in microRNA target sites.* Nucleic acids research, 2011.
165. Sewer, A., et al., *Identification of clustered microRNAs using an ab initio prediction method.* BMC Bioinformatics, 2005. **6**: p. 267.
166. Huang, T.H., et al., *MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans.* BMC Bioinformatics, 2007. **8**: p. 341.
167. Hertel, J. and P.F. Stadler, *Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data.* Bioinformatics, 2006. **22**(14): p. e197-202.
168. Witkos, T.M., E. Koscianska, and W.J. Krzyzosiak, *Practical Aspects of microRNA Target Prediction.* Current molecular medicine, 2011. **11**(2): p. 93-109.
169. Sahoo, S. and A.A. Albrecht, *Ranking of microRNA target prediction scores by Pareto front analysis.* Computational biology and chemistry, 2010. **34**(5-6): p. 284-92.
170. Yousef, M., et al., *Learning from positive examples when the negative class is undetermined--microRNA gene identification.* Algorithms for molecular biology : AMB, 2008. **3**: p. 2.
171. Wang, X. and I.M. El Naqa, *Prediction of both conserved and nonconserved microRNA targets in animals.* Bioinformatics, 2008. **24**(3): p. 325-32.
172. Kertesz, M., et al., *The role of site accessibility in microRNA target recognition.* Nature genetics, 2007. **39**(10): p. 1278-84.
173. Obernosterer, G., H. Tafer, and J. Martinez, *Target site effects in the RNA interference and microRNA pathways.* Biochemical Society transactions, 2008. **36**(Pt 6): p. 1216-9.
174. Grun, D., et al., *microRNA target predictions across seven Drosophila species and comparison to mammalian targets.* PLoS computational biology, 2005. **1**(1): p. e13.
175. Alexiou, P., et al., *Lost in translation: an assessment and perspective for computational microRNA target identification.* Bioinformatics, 2009. **25**(23): p. 3049-55.
176. Baek, D., et al., *The impact of microRNAs on protein output.* Nature, 2008. **455**(7209): p. 64-71.
177. Zisoulis, D.G., et al., *Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans.* Nature structural & molecular biology, 2010. **17**(2): p. 173-9.

178. Lal, A., et al., *miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements.* Molecular cell, 2009. **35**(5): p. 610-25.
179. Chi, S.W., et al., *Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps.* Nature, 2009. **460**(7254): p. 479-86.
180. Zhao, S. and M.F. Liu, *Mechanisms of microRNA-mediated gene regulation.* Science in China. Series C, Life sciences / Chinese Academy of Sciences, 2009. **52**(12): p. 1111-6.
181. Chendrimada, T.P., et al., *MicroRNA silencing through RISC recruitment of eIF6.* Nature, 2007. **447**(7146): p. 823-8.
182. Mathonnet, G., et al., *MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F.* Science, 2007. **317**(5845): p. 1764-7.
183. Humphreys, D.T., et al., *MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(47): p. 16961-6.
184. Wakiyama, M., et al., *Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system.* Genes & development, 2007. **21**(15): p. 1857-62.
185. Kiriakidou, M., et al., *An mRNA m7G cap binding-like motif within human Ago2 represses translation.* Cell, 2007. **129**(6): p. 1141-51.
186. Bhattacharyya, S.N., et al., *Relief of microRNA-mediated translational repression in human cells subjected to stress.* Cell, 2006. **125**(6): p. 1111-24.
187. Ferraiuolo, M.A., et al., *A role for the eIF4E-binding protein 4E-T in P-body formation and mRNA decay.* The Journal of cell biology, 2005. **170**(6): p. 913-24.
188. Pillai, R.S., et al., *Inhibition of translational initiation by Let-7 MicroRNA in human cells.* Science, 2005. **309**(5740): p. 1573-6.
189. Olsen, P.H. and V. Ambros, *The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation.* Developmental biology, 1999. **216**(2): p. 671-80.
190. Nottrott, S., M.J. Simard, and J.D. Richter, *Human let-7a miRNA blocks protein production on actively translating polyribosomes.* Nature structural & molecular biology, 2006. **13**(12): p. 1108-14.
191. Petersen, C.P., et al., *Short RNAs repress translation after initiation in mammalian cells.* Molecular cell, 2006. **21**(4): p. 533-42.
192. Wu, L., J. Fan, and J.G. Belasco, *MicroRNAs direct rapid deadenylation of mRNA.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(11): p. 4034-9.
193. Nilsen, T.W., *Mechanisms of microRNA-mediated gene regulation in animal cells.* Trends in genetics : TIG, 2007. **23**(5): p. 243-9.
194. Filipowicz, W., S.N. Bhattacharyya, and N. Sonenberg, *Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?* Nature reviews. Genetics, 2008. **9**(2): p. 102-14.
195. Anderson, P. and N. Kedersha, *RNA granules.* The Journal of cell biology, 2006. **172**(6): p. 803-8.
196. Thomas, M.G., et al., *RNA granules: the good, the bad and the ugly.* Cellular signalling, 2011. **23**(2): p. 324-34.
197. Nykanen, A., B. Haley, and P.D. Zamore, *ATP requirements and small interfering RNA structure in the RNA interference pathway.* Cell, 2001. **107**(3): p. 309-21.

198. Elbashir, S.M., et al., *Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate.* The EMBO journal, 2001. **20**(23): p. 6877-88.
199. Guo, L. and Z. Lu, *The fate of miRNA* strand through evolutionary analysis: implication for degradation as merely carrier strand or potential regulatory molecule?* PLoS One, 2010. **5**(6): p. e11387.
200. Stark, A., et al., *Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes.* Genome research, 2007. **17**(12): p. 1865-79.
201. Okamura, K., et al., *The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila.* Cell, 2007. **130**(1): p. 89-100.
202. Maragkakis, M., et al., *Accurate microRNA target prediction correlates with protein repression levels.* BMC Bioinformatics, 2009. **10**: p. 295.
203. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs.* Genome Res, 2009. **19**(1): p. 92-105.
204. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res, 2006. **34**(Database issue): p. D140-4.
205. Krek, A., et al., *Combinatorial microRNA target predictions.* Nature genetics, 2005. **37**(5): p. 495-500.
206. Min, H. and S. Yoon, *Got target? Computational methods for microRNA target prediction and their extension.* Exp Mol Med, 2010. **42**(4): p. 233-44.
207. Gnatenko, D.V., et al., *Platelets express steroidogenic 17beta-hydroxysteroid dehydrogenases. Distinct profiles predict the essential thrombocythemic phenotype.* Thromb Haemost, 2005. **94**(2): p. 412-21.
208. Eulalio, A., et al., *Deadenylation is a widespread effect of miRNA regulation.* RNA, 2009. **15**(1): p. 21-32.
209. Behm-Ansmant, I., et al., *mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes.* Genes & development, 2006. **20**(14): p. 1885-98.
210. Eulalio, A., et al., *Target-specific requirements for enhancers of decapping in miRNA-mediated gene silencing.* Genes & development, 2007. **21**(20): p. 2558-70.
211. Lim, L.P., et al., *Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.* Nature, 2005. **433**(7027): p. 769-73.
212. Wessel, D. and U.I. Flugge, *A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids.* Analytical biochemistry, 1984. **138**(1): p. 141-3.
213. Peterson, G.L., *A simplification of the protein assay method of Lowry et al. which is more generally applicable.* Analytical biochemistry, 1977. **83**(2): p. 346-56.
214. Washburn, M.P., D. Wolters, and J.R. Yates, 3rd, *Large-scale analysis of the yeast proteome by multidimensional protein identification technology.* Nature biotechnology, 2001. **19**(3): p. 242-7.
215. Gatlin, C.L., et al., *Protein identification at the low femtomole level from silver-stained gels using a new fritless electrospray interface for liquid chromatography-microspray and nanospray mass spectrometry.* Analytical biochemistry, 1998. **263**(1): p. 93-101.
216. Fei, S.S., et al., *Protein database and quantitative analysis considerations when integrating genetics and proteomics to compare mouse strains.* Journal of proteome research, 2011. **10**(7): p. 2905-12.
217. Monypenny, J., et al., *Cdc42 and Rac family GTPases regulate mode and speed but not direction of primary fibroblast migration during platelet-derived growth factor-dependent chemotaxis.* Molecular and cellular biology, 2009. **29**(10): p. 2730-47.

218. Roos, J., et al., *STIM1, an essential and conserved component of store-operated Ca2+ channel function.* The Journal of cell biology, 2005. **169**(3): p. 435-45.
219. Gilio, K., et al., *Roles of platelet STIM1 and Orai1 in glycoprotein VI- and thrombin-dependent procoagulant activity and thrombus formation.* The Journal of biological chemistry, 2010. **285**(31): p. 23629-38.
220. Mendes, N.D., A.T. Freitas, and M.F. Sagot, *Current tools for the identification of miRNA genes and their targets.* Nucleic acids research, 2009. **37**(8): p. 2419-33.