STRUCTURAL ANALYSIS OF GENE CO-EXPRESSION IN CHRONIC MYELOGENOUS LEUKEMIA

FENGFENG WANG

Ph.D

The Hong Kong Polytechnic University

2015

The Hong Kong Polytechnic University

Department of Health Technology and Informatics

Structural Analysis of Gene Co-expression in Chronic Myelogenous Leukemia

Fengfeng Wang

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

August 2014

CRETIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

WANG Fengfeng (Name of student)

Abstract

Background: Chronic myelogenous leukemia (CML) is a clonal myeloproliferative disorder characterized by an increased proliferation of granulocytes in bone marrow. The characteristics of CML at the cellular level include increased proliferation, increased resistance to apoptosis and alterations in adhesion properties. Co-expression analysis has been used to study functionally related genes, since the co-expressed genes are more likely to participate in the similar biological processes and signal pathways. Moreover, researchers found that genes with similar mRNA expression profiles tend to be regulated via the same mechanism (s), e.g. the same regulator. We plan to explore the differences between the normal and the CML groups in the co-expression patterns of those genes involved in a functional gene set, regulated by the same regulators, and covering from the whole genome, in order to further explore the altered biological pathways and novel mechanisms in CML. Nucleophosmin 1 (NPM1) is important in ribosomal synthesis and malignancies. The NPM1-associated gene set was chosen as the candidate gene set for the co-expression analysis. We wonder if NPM1-associated genes can affect the ribosomal synthesis and translation process in CML. E2F1–3 and MYC are important transcription factors (TFs) reciprocally regulated in the transcription process to form positive feedback loops. Target genes regulated by E2F1–3 or MYC are related to cell proliferation and apoptosis. MicroRNAs (miRNAs) are post-transcriptional regulators regulating target gene expression. Mature miRNAs from the miR-17-92 cluster are overexpressed in chronic-phase CML patients compared with normal individuals. The overexpression can promote cell cycle progression and proliferation,

and inhibit apoptosis. We wonder what the co-expression patterns of the target genes directly regulated by E2F1–3 and MYC, or by miRNAs in the normal and the CML groups are.

Result and conclusion: We presented a distribution-based approach for gene pair classification by identifying a disease-specific cutoff point that classified the co-expressed gene pairs into strong and weak classes. Our developed method effectively identified the differences in the co-expression patterns from the overall structure: a) whole genome co-expression analysis: p-value < 0.05 for the maximum deviation D = 0.041; b) NPM1-associated gene set co-expression analysis: p-value = 1.71×10^{-22} < 0.05 for the maximum deviation D = 0.109; c) E2F and MYC target genes co-expression analysis: p-value = 2.00×10^{-34} < 0.05 for the maximum deviation D = 0.0577; d) miR-17-92 cluster target genes co-expression analysis: p-value = 2.62×10^{-58} < 0.05 for the maximum deviation D = 0.0567. The distribution-based classification divided the co-expressed gene pairs into specific and common groups, forming the co-expression structures. Functional annotation showed that ribosomal protein (RP) genes were more likely to be co-expressed in the CML group compared to the normal group. In addition, genes involved in the ribosomal synthesis and translation process tended to be co-expressed in the CML group. While, genes related to cell adhesion and angiogenesis properties, as well as metabolism processes were more likely to be co-expressed in the normal group. The co-expression pattern in the normal group represents the healthy pathological balance. Our findings may be helpful in exploring the underlying mechanisms of CML, and provide useful information in cancer treatment.

List of Publications

International journal papers

Published or accepted journal papers:

1. **Fengfeng Wang**, Lawrence W.C. Chan, William C.S. Cho, Petrus Tang, Jun Yu, Chi-Ren Shyu, Nancy B.Y. Tsui, S.C. Cesar Wong, Parco M. Siu, S.P. Yip and Benjamin Y.M. Yung, "Novel Approach for Co-expression Analysis of E2F1–3 and MYC Target Genes in Chronic Myelogenous Leukemia", BioMed Research International, vol. 2014, Article ID 439840, 2014.

2. **Fengfeng Wang**, S.C. Cesar Wong, Lawrence W.C. Chan, William C.S. Cho, S.P. Yip and Benjamin Y.M. Yung, "Multiple Regression Analysis of mRNA-miRNA Associations in Colorectal Cancer Pathway", BioMed Research International, vol. 2014, Article ID 676724, 2014.

3. **Fengfeng Wang**, Lawrence W.C. Chan, Helen K.W. Law, William C.S. Cho, Petrus Tang, Jun Yu, Chi-Ren Shyu, S.C. Cesar Wong, S.P. Yip, Benjamin Y.M. Yung, "Exploring MicroRNA-Mediated Alteration of EGFR Signaling Pathway in Non-Small Cell Lung Cancer Using an mRNA:miRNA Regression Model Supported by Target Prediction Databases", Genomics. vol. 104, pp. 504–511, 2014.

4. **Fengfeng Wang**, William C.S. Cho, Lawrence W.C. Chan, S.C. Cesar Wong, Nancy B.Y. Tsui, Parco M. Siu, S.P. Yip and Benjamin Y.M. Yung, "Gene Network Exploration of Crosstalk between Apoptosis and Autophagy in

Chronic Myelogenous Leukemia", paper accepted, to be published in BioMed Research International.

5. Lawrence Wing-Chi Chan, Connie Hiu-Ching Ngo, **Fengfeng Wang**, Moss Y. Zhao, Mengying Zhao, Helen Ka-Wai Law, Sze Chuen Cesar Wong, and Benjamin Yat-Ming Yung, "Disease-Specific Target Gene Expression Profiling of Molecular Imaging Probes: Database Development and Clinical Validation", Molecular Imaging, vol. 13, pp. 1–12, 2014.
6. Lawrence WC Chan, **Feng F Wang** and William CS Cho, "Genomic Sequence Analysis of EGFR Regulation by MicroRNAs in Lung Cancer", Current Topics in Medicinal Chemistry, vol. 12, pp. 920–926, 2012.

Submitted journal paper:

7. **Fengfeng Wang**, Lawrence W.C. Chan, Nancy B.Y. Tsui, S.C. Cesar Wong, Parco M. Siu, S.P. Yip and Benjamin Y.M. Yung, "Co-expression Pattern Analysis of NPM1-associated Genes in Chronic Myelogenous Leukemia", submitted to BioMed Research International.

International conference papers

8. **Fengfeng Wang**, Lawrence Chan, Helen KW Law, William CS Cho, Cesar Wong, SP Yip, Benjamin YM Yung, "In-Silico Analysis of EGFR-Associated MicroRNA Signature in Cancer", IEEE Interactional Conference on Bioinformatics and Biomedicine (BIBM), pp.7–12, Dec. 2013, Shanghai, China.

9. **Feng-Feng Wang**, Lawrence WC Chan, SP Yip, Benjamin YM Yung, "MicroRNA-Mediated Alteration of TET2 Interaction Network in Myeloproliferative Neoplasms", IEEE Interactional Conference on e-Health Networking, Applications and Services (Healthcom), pp.241–245, Jun. 2011, Missouri, USA.

Acknowledgements

I would like to express my deepest gratitude to my chief supervisor Dr. Lawrence WC Chan, for his excellent guidance, patience, immense knowledge and providing me this precious chance to pursue my Ph.D. degree. His endless support helped me in all the time of research and the writing of thesis. His hard-working style and immense enthusiasm on research have impressed me very much. This has inspired me to work hard on my study and thesis.

I would also like to thank my co-supervisors Prof. Shea Ping Yip and Prof. Benjamin YM Yung for guiding my research and giving me previous suggestions on my study for the past several years.

I would like to thank Dr. Parco M Siu, Dr. SC Cesar Wong and Dr. Helen KW Law for their guidance in our Joint Group Lab meetings.

I would like to thank my advisors Dr. William CS Cho, Dr. Petrus Tang, Prof. Jun Yu and Prof. Chi-Ren Shyu for their help and suggestions on my study.

I would also like to thank my team members and all other students and staffs in our department for their help and the sharing of ideas and experience.

I would like to express my gratitude to the Department of Health Technology and Informatics for providing me a comfortable studying environment. I would also like

to thank the Hong Kong Polytechnic University for the financial support to my research project.

Last but not the least, I would like to thank my family for their great love and constant encouragement at all times. Their understanding and patience helped me complete this research study.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABL	Abelson
ACVR2A	Activin A receptor, type IIA
AML	Acute myeloid leukemia
AP	Accelerated phase
BCR	Breakpoint cluster region
BP	Blastic phase
C	Cutoff point
CDFs	Cumulative distribution functions
CML	Chronic myelogenous leukemia
CP	Chronic phase
CRMs	Cis-regulatory modules
CS	Strongly co-expressed gene pairs in the CML group
CTBP2	C-terminal binding protein 2
CW	Weakly co-expressed gene pairs in the CML group
D	Maximum deviation
DAVID	Database for Annotation, Visualization and Integrated Discovery

EASE score	Expression Analysis Systematic Explorer score
EEF2	Eukaryotic translation elongation factor 2
EFNA5	Ephrin-A5
EFNB2	Ephrin-B2
EGFR	Epidermal growth factor receptor
EIF3F	Eukaryotic translation initiation factor 3, subunit F
EPHA4	EPH receptor A4
EZH2	Enhancer of zeste homolog 2 protein
FBS	Fetal bovine serum
FDR	False discovery rate
GEO	Gene expression omnibus
GO	Gene ontology
HP1	Heterochromatin protein 1
HTATIP2	HIV-1 Tat interactive protein 2
INHBA	Inhibin, beta A
ITGA2	Integrin, alpha 2
KS	Kolmogorov-Smirnov
MI	Molecular interaction
MiRNAs	MicroRNAs
MSigDB	Molecular Signature Database

NCBI	National Center for Biotechnology Information
NPM1	Nucleophosmin 1
NPM2	Nucleoplasmin 2
NPM3	Nucleoplasmin 3
NRP2	Neuropilin 2
NS	Strongly co-expressed gene pairs in the normal group
NW	Weakly co-expressed gene pairs in the normal group
Ph	Philadelphia
PPP2R3A	Protein phosphatase 2, regulatory subunit B", alpha
Pre-miRNAs	Precursor miRNAs
PReMod	Prediction of transcriptional regulatory modules
PREP	Prolyl endopeptidase
Pri-miRNAs	Primary miRNAs
PRKACB	Protein kinase, cAMP-dependent, catalytic, beta
r	Pearson correlation coefficient
rDNA	Ribosomal DNA
RISC	RNA-induced silencing complex
RP	Ribosomal protein

RPL5	Ribosomal protein L5
RPS28	Ribosomal protein S28
rRNA	Ribosomal RNA
RT-PCR	Reverse transcription polymerase chain reaction
SMAD2	Mothers against decapentaplegic homolog 2
SSTR2	Somatostatin receptor 2
TCF4	Transcription T cell factor 4
TFBSs	Transcription factor binding sites
TFs	Transcription factors
TGFBR2	Transforming growth factor, beta receptor II
TGs	Target genes
tRNAs	Transfer RNAs
UTR	Untranslated region

Chapter 1 Literature Review

1.1 Chronic myelogenous leukemia

Chronic myelogenous leukemia (CML) is a clonal myeloproliferative disorder that is characterized by an increased proliferation of granulocytes in the bone marrow. The annual incidence rate of CML is about 1-2 per 100,000, which accounts for 20% of all leukemias affecting adults with a median age of 45 to 55 years (Faderl *et al*., 1999; Frazer *et al*., 2007). Some characteristics of CML at the cellular level were found: increased proliferation, increased resistance to apoptosis and alterations in adhesion properties of leukemic progenitors (Salesse and Verfaillie, 2002). In total, there are three phases for CML: chronic phase (CP), accelerated phase (AP) and blastic phase (BP) (Kalidas *et al*., 2001). The majority of CML patients are diagnosed in the first chronic phase. With the symptoms becoming worse and the immature blasts increasing, the CP patients will progress to the accelerated phase. While, patients with more than 20% blasts in the bone marrow or peripheral blood belong to the final and fatal blastic phase (Vardiman *et al*., 2002).

1.1.1 Philadelphia chromosome

The hallmark of CML is Philadelphia (Ph) chromosome, which results from a balanced reciprocal translocation event t(9;22)(q34;q11), originating from a single hematopoietic stem cell (Nowell and Hungerford, 1960) (Figure 1.1). In 1960, an abnormal G group chromosome with a deletion on the q arm was identified in CML

(Nowell and Hungerford, 1960). Later, Rowley showed that a translocation from chromosome 22 to another chromosome, often chromosome 9, was found (Rowley, 1973). The BCR-ABL oncogene is generated by the translocation that combines the Abelson oncogene (ABL) at 9q34 with the breakpoint cluster region (BCR) at 22q11.2, forming a fusion BCR-ABL oncoprotein (Melo and Barnes, 2007). Such fusion can prohibit the BCR-ABL oncoprotein to shuttle between the nucleus and the cytoplasm, as well as increase the tyrosine kinase activity of ABL and the autophosphorylation of BCR-ABL oncoprotein, creating more binding sites on the SH2 regulatory domains of the interacting proteins (Melo and Barnes, 2007). Hence, the dysregulated physiological and molecular properties of ABL protein result in the leukemogensis of CML.

1.1.2 Current research on CML using microarray analysis

Recently, more studies have been performed on the analysis of microarray gene expression profiles in CML. Microarray analysis is powerful for extracting useful information for the diagnosis, prognosis and therapy of acute and chronic leukemias (Haferlach *et al.*, 2005; Wadlow and Ramaswamy, 2005). At present, most of microarray analyses focus on differentially expressed genes, for example, the study exploring the relationship between pathways and differentially expressed genes using the data from untreated CML patients in the chronic phase (Diaz-Blanco *et al.*, 2007). In another study, researchers applied the differential gene expression microarray analysis to compare CML patients with normal and variant Ph-chromosome (Albano *et al.*, 2013). Nevertheless, few studies of gene co-expression analysis have been investigated.

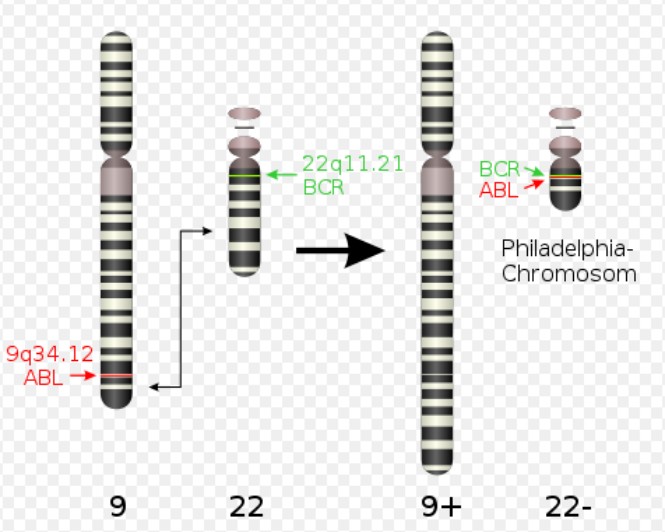


Figure 1.1: Schematic representation of Ph-chromosome formation (http://en.wikipedia.org/wiki/Philadelphia_chromosome#mediaviewer/File:Philadelphia_Chromosom.svg).

1.2 Gene co-expression analysis

Genes with similar expression levels tend to be co-expressed through the microarray data analysis. On one hand, co-expression analysis can be used to study functionally related genes, since the co-expressed genes are more likely to participate in similar biological processes and signal pathways (Eisen *et al.*, 1998; Spellman *et al.*, 1998). On the other hand, genes with similar mRNA expression profiles are likely to be regulated via the same mechanism(s), such as the same regulatory mechanism (Allocco *et al.*, 2004; Altman and Raychaudhuri, 2001; Schulze and Downward, 2001). One limitation of differential expression analysis compared to co-expression analysis is that the former only reflects the upregulation or downregulation of existing components in the well-known pathways, which cannot identify the functionally associated linkages among genes during signal transduction. However, co-expression analysis is very powerful for grouping genes and further analyzing the mechanisms of diseases. Besides, gene co-expression is different in different states (e.g. the normal and the disease states) and cell types (Torkamani *et al.*, 2010), leading to different co-expression patterns. Hence, the different co-expression pattern can be regarded as the signature of a disease.

1.2.1 Functionally related genes are co-expressed

Genes participating in the same molecular mechanism are more likely to exhibit similar expression profiles. Co-expressed genes tend to participate in the similar regulatory and signaling circuits, forming complexes, pathways and network modules (Fuller *et al.*, 2007; Ideker *et al.*, 2002). Co-expression analysis of gene-gene interactions has identified genes involved in controlling the human T helper cell

differentiation process (Elo *et al*., 2007). Genetic modules from co-expressed genes were discovered to be functionally related and highly conserved among different species (Stuart *et al*., 2003). Furthermore, strongly co-expressed genes were observed to have higher gene ontology similarity and protein-protein interaction when compared to the randomized gene pairs (Elo *et al*., 2007).

1.2.2 Genes shared the same regulators are co-expressed

It was found that genes with similar mRNA expression profiles are more likely to be regulated by the same mechanisms (Altman and Raychaudhuri, 2001; Schulze and Downward, 2001). Several studies have demonstrated that the co-expressed genes in a cluster often share common upstream sequence motifs by analyzing their mRNA expression profiles (Brazma *et al*., 1998; Tavazoie *et al*., 1999; Wolfsberg *et al*., 1999). Other researchers observed that the expression profiles of genes regulated by the same transcription factor are strongly correlated (Ideker *et al*., 2001). Allocco *et al*. discovered that genes tend to have a common transcription factor binding to their promoter regions if they have strongly correlated mRNA expression profiles (Allocco *et al*., 2004). Yu *et al*. revealed that genes targeted by the same transcription factor tend to be co-expressed, and the degree of co-expression is increased if genes share more than one transcription factor (Yu *et al*., 2003).

1.2.3 Aim of the study

In this study, we plan to explore the differences between the normal and the CML groups in the co-expression patterns of those genes involved in a functional gene set, regulated by the same regulators (transcription factors and microRNAs), and

covering from the whole genome. The different co-expression pattern indicated the alteration of molecular interactions in CML when compared to the normal state. By analyzing the biological meaning of strongly co-expressed gene pairs, we can further elucidate the underlying mechanisms of CML.

1.3 Nucleophosmin 1

Nucleophosmin 1 (NPM1), also called nucleolar phosphoprotein B23, is a member of the NucleoPhosMin/NucleoPlasMin family of nuclear chaperones. This protein family can be divided to four classes according to protein sequence similarities: nucleophosmin (NPM1), nucleoplasmin 2 (NPM2), nucleoplasmin 3 (NPM3) and NPM-like invertebrate proteins (Federici and Falini, 2013; Frehlick *et al*., 2006). NPM1 is well studied in the whole family. In 1989, a human NPM1 cDNA was cloned, and its encoded protein has 294 amino acids (Chan *et al*., 1989). Later, a shorter isoform with a different C-terminus was identified in 2002, encoding a 259-amino-acid protein (Dalenc *et al*., 2002). The expression of this gene is frequently altered in solid tumors, and its mutation and translocation are also found in hematological malignancies (Grisendi *et al*., 2006). The protein product encoded by this gene is a phosphoprotein that travels between the nucleus and cytoplasm. It is a versatile protein that plays multiple roles in ribosomal RNA (rRNA) processing, ribosome assembly, transport of ribosomal subunits, centrosome duplication, regulation of p53, as well as cell growth and proliferation (Naoe *et al*., 2006; Pelletier *et al*., 2007; Yao *et al*., 2010). In addition, the NPM1 gene often participates in chromosomal translocation, mutation and deletion in hematological malignancies (Naoe *et al*., 2006).

1.3.1 NPM1 in ribosomal synthesis

It has been reported that NPM1 is overexpressed in proliferating cells and tumor cells. A possible reason is the increased ribosomal synthesis (Naoe *et al*., 2006). NPM1 protein was reported to have a molecular chaperone-like activity (Szebeni and Olson,

1999). It can inhibit the aggregation of some proteins, and the inhibition degree is proportional to NPM1 concentration. NPM1 protein can also protect some enzymes from aggregation in the thermal denaturation process, and preserve the enzyme activity. Moreover, NPM1 preferentially forms a complex with denatured proteins to further expose hydrophobic regions (Szebeni and Olson, 1999). Yu *et al*. have discovered the direct interaction between NPM1 and ribosomal protein L5 (RPL5), which is the first reported physical link between NPMs and ribosomal subunits (Yu *et al*., 2006b). Knockdown of NPM1 can impair the processing of 28S rRNA, a component of ribosome 60S subunit, from the 32S rRNA precursor (Itahana *et al*., 2003). NPM1 was discovered to mediate the nuclear export of ribosomal large and small subunits, and colocalize with the ribosomal subunit proteins in the nucleolus, nucleus and cytoplasm (Maggi *et al*., 2008). Besides, NPM1 has a direct interaction with various ribosomal proteins, including RPS9 and RPL23 (Lindström, 2011).

1.3.2 NPM1 in AML and CML

NPM1 can both promote cell growth and repress tumor cells. Its overexpression increases cell division and growth, possibly owing to the effects on ribosomal DNA (rDNA) transcription, ribosomal subunit export and S-phase DNA replication (Lindström, 2011). Falini *et al*. indicated that cytoplasmic NPM1 is regarded as the hallmark of patients with acute myeloid leukemia (AML) who have a normal karyotype, NPM gene mutations, and responsiveness to induction chemotherapy, which accounts for about one third of primary AML in adults (Falini *et al*., 2005). Federici and Falini reviewed that NPM1 is characterized as the most frequently mutated gene in patients with AML (Federici and Falini, 2013). The mutations of NPM1 are usually heterozygous and mutually exclusive in AML patients with

recurrent genetic abnormalities (Federici and Falini, 2013). The cytoplasmic mutated NPM1 was identified for the first time in a blast-crisis CML patient, indicating that NPM1 gene mutation may function in the blastic transformation of CML (Piccaluga *et al*., 2009). Interestingly, in another study researchers did not detect any NPM1 mutations in the analyzed blast-crisis and chronic-phase CML patients (Watkins *et al.*, 2013).

1.3.3 NPM1-associated genes

Based on the gene list curated by Brentani *et al.*, NPM1 is one of 380 cancer-associated genes internally and from a published cancer gene database (Brentani *et al.*, 2003). In Subramanian *et al.*'s study, the neighborhoods around these cancer-associated genes were selected according to four large gene expression datasets. These datasets were collected from various cancer projects mainly on primary tumors, including prostate, breast, lung, lymphoma, leukemia and so on (Subramanian *et al.*, 2005). Pearson correlation coefficient (r) between every gene in these four datasets and one of the cancer-associated genes (e.g. NMP1) was calculated. The calculation was performed independently in each dataset. A gene can be selected as the neighborhood if $r \geq 0.85$ in at least one out of four datasets. The cancer-associated genes with no less than 25 selected neighborhoods are stored in the database (Subramanian *et al.*, 2005). The NPM1 associated genes (neighborhoods) were selected based on the above criteria, describing a set of genes highly associated with NPM1. The NPM1-associated gene set (GCM_NPM1), in total 116 genes including NPM1, is stored as one of the neighborhood sets in the *Molecular Signature Database* (*MSigDB*) (Subramanian *et al*., 2005).

1.3.4 Aim of the study

NPM1 plays an important role in ribosomal synthesis and malignancies. We wonder if NPM1-associated genes can affect the ribosomal synthesis and translation process in CML. Co-expression analysis has been used to study functionally related genes, since the co-expressed genes are more likely to participate in the similar biological processes and signal pathways (Eisen *et al*., 1998; Spellman *et al*., 1998). Therefore, we plan to explore the differences in the co-expression patterns of those NPM1-associated genes between the normal and the CML states, in order to further investigate the altered ribosome activities in CML.

1.4 Transcription factor

Gene expression largely depends on the effect of trans-acting factors binding to the cis-acting elements located on the regulatory regions of genes. Among all these trans-acting factors and the cis-acting elements, transcription factors (TFs) and their binding sites (TFBSs) play an important role in gene expression regulation (Cui *et al*., 2007). TFs are proteins that bind to specific sites on DNA sequences, which control the transcription of genetic information from DNA to mRNA (Figure 1.2). Cells make response to the outside signals by combining them with the internal state during signaling transduction process (Farkas *et al*., 2006). Importantly, transcriptional regulators localize in a key position in the process of signal transduction (Farkas *et al*., 2006). Interactions between TFs and their target genes make adjustments to the transcriptional activities of DNA and further regulate the global gene expression of a living cell (Chen *et al*., 2011).

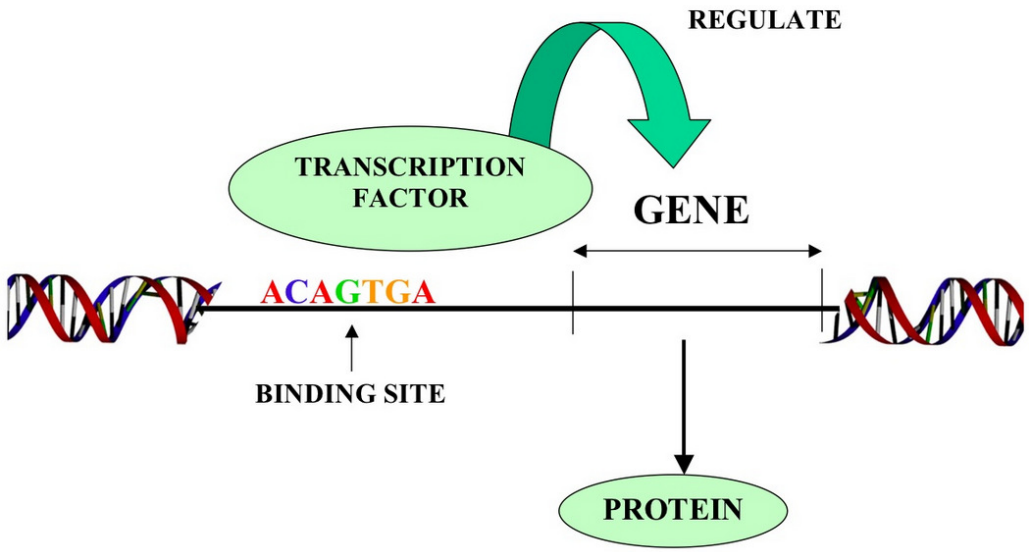


Figure 1.2: Transcription process from a gene to a protein starting from the binding of transcription factors to the specific sites of DNA sequence (http://galleryhip.com/transcription-factor.html).

1.4.1 Candidate transcription factors

1.4.1.1 E2F family of transcription factors

1.4.1.1.1 Category of E2Fs

The E2F family of transcription factors is important to control cellular proliferation by regulating the transcription of various genes related to DNA replication, DNA repair, mitosis and cell cycle progression (Timmers *et al*., 2007). Based on structure-function studies and amino acid sequence analysis, members of the E2F family are classified into two main subclasses: activators E2F1–3 and repressors E2F4–8 (Chong *et al*., 2009; Timmers *et al*., 2007). The transcription activators E2F1–3 play a vital role in cell cycle progression, particularly in the G1/S transition process (Wu *et al*., 2001). Recently, for a more detailed classification, the eight members can be divided into three subclasses based on their transcriptional activity (Figures 1.3 and 1.4): i) activators E2F1–E2F3a, they can activate the target gene expression during the late G1/S phase of cell cycle progression; ii) repressors E2F3b–E2F5, they repress the target gene expression by cooperating with E2F inhibitory pocket proteins and the repressive histone deacetylases in the early G1 phase; and iii) inhibitors E2F6–E2F8, they have the similar functions with repressors, and the only difference is that they do not interact with pocket proteins (Hazar-Rethinam *et al*., 2011; Zhan *et al*., 2014). It is noted that this kind of classification is broadly relied on the *in vitro* study, no *in vivo* support, which may be not enough to demonstrate the important roles of E2Fs (Zhan *et al*., 2014).

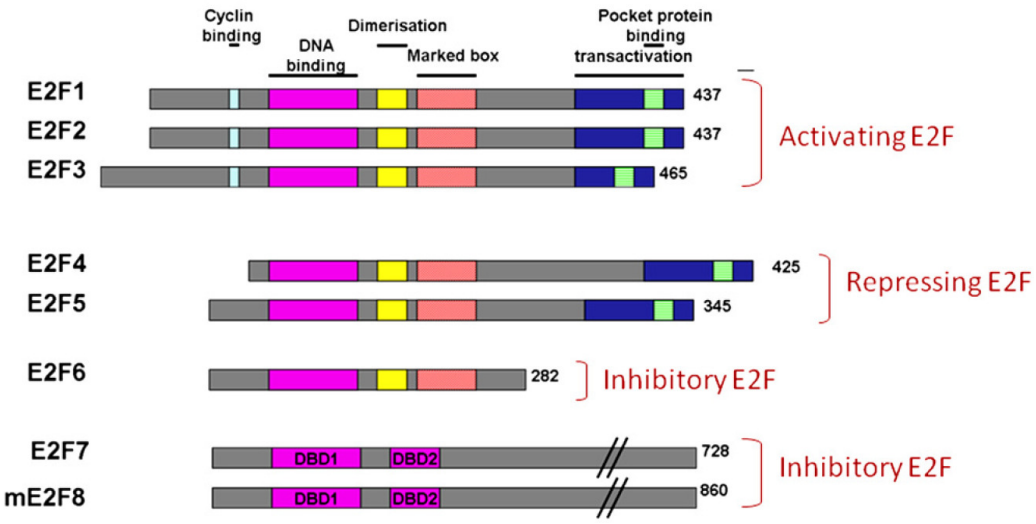


Figure 1.3: Classification and structure of E2Fs. The same colors indicate the homologues regions. The number on the right is for the amino acid counting (Hazar-Rethinam *et al*., 2011).

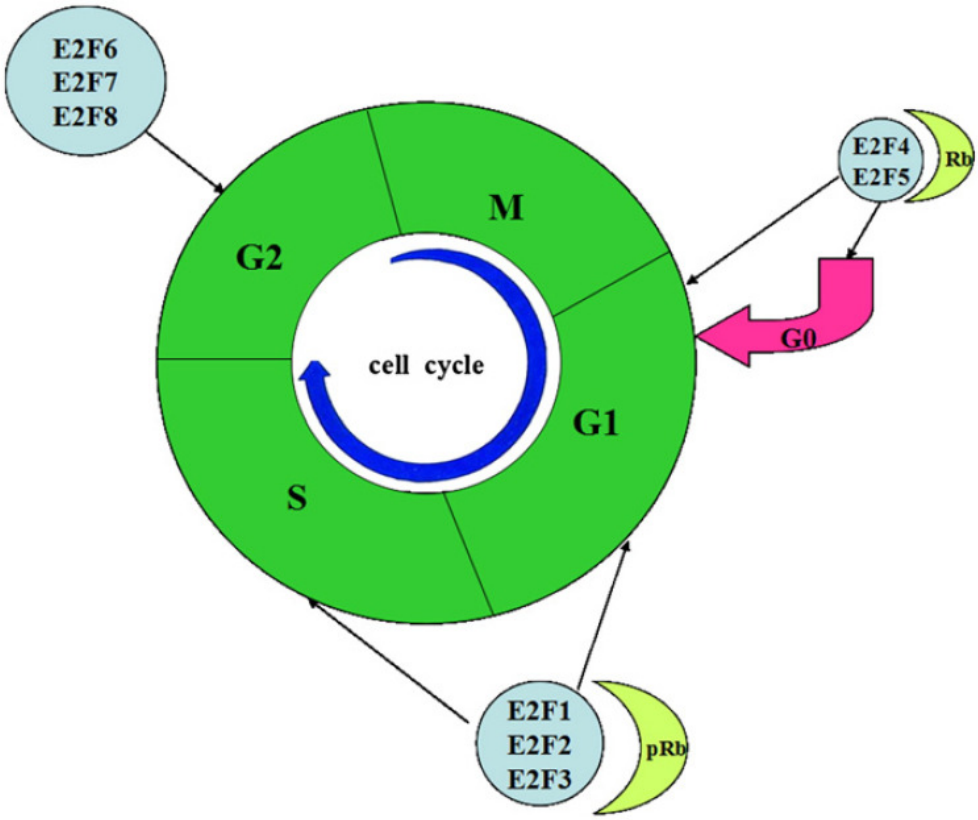


Figure 1.4: E2Fs function in different phases of cell cycle progression. The repressors E2F4–E2F5 regulate gene expression in the G0 and early G1 phases. In the late G1 phase, the activators E2F1–E2F3 function to replace the repressors. As cells enter into the G2 phase, the activators are decreased, and the inhibitors E2F6–E2F8 appear to perform the function (Zhan *et al*., 2014).

1.4.1.1.2 E2Fs in cancer

Since E2Fs are the downstream molecules of cell cycle signaling pathways, they play key roles in regulating a variety of cellular functions, such as cell proliferation, apoptosis, differentiation, senescence, DNA-damage response and DNA repair (Zhan *et al*., 2014). E2F1 is regarded as a possible oncogene or tumor suppressor gene through *in vitro* studies. The expression level of E2F1 encoded protein is significantly overexpressed in chronic viral liver disease and hepatocellular carcinoma (HCC) compared to adjacent cirrhotic liver parenchyma (Palaiologou *et al*., 2012). The E2F1 and E2F2 genetic variants have been proved to play an important role in head and neck carcinogenesis (Lu *et al*., 2012). E2F3a was found to be overexpressed in liver cancer tissues, and it can induce apoptosis in HepG2 cells (Li *et al*., 2010b). E2F4 transcript level is overexpressed in the neoplastic cells compared to the normal tissues, and its upregulation may positively regulate target gene expression (Schwemmle and Pfeifer, 2000). E2F7 expression is upregulated in primary blasts AML patients, and it can significantly enhance cell cycle progression and inhibit monocytic differentiation in AML cells (Salvatori *et al*., 2012). E2F7 and E2F8 have been identified as the novel and critical regulators of angiogenesis (Bakker *et al*., 2013).

1.4.1.2 Transcription factor MYC

Another important gene, proto-oncogene c-myc, encodes a transcription factor (MYC) that can induce both cell proliferation and apoptosis (Pelengaris *et al*., 2000). The MYC transcription factor can both activate and repress the transcription of many target genes. High-throughput techniques have demonstrated that MYC-activated genes are involved in some important cellular processes, such as growth, protein

synthesis and mitochondrial function. The majority of MYC-repressed genes participate in the interaction and communication between cells and their external environment, while several genes have anti-proliferative or anti-metastatic properties (O'Connell *et al.*, 2003; Pelengaris and Khan, 2003). Most importantly, E2F1, E2F2, E2F3 and MYC are reciprocally regulated to form positive feedback loops among them in their transcription processes (Figure 1.5) (Aguda *et al.*, 2008; Coller *et al.*, 2007). They can activate one another's transcription to well control their expression levels.

1.4.2 Aim of the study

Usually, one gene can be simultaneously regulated by several TFs, and each TF can also simultaneously regulate several genes. Genes targeted by the same TF tend to be co-expressed, and the degree of co-expression is increased if genes share more than one TF (Yu *et al.*, 2003). Since target genes regulated by E2F1–3 or MYC are related to cell proliferation and apoptosis, we wonder what the co-expression patterns of the target genes regulated directly and concurrently by E2F1–3 and MYC in the normal and the CML groups are. Therefore, we plan to explore the difference in the co-expression patterns of those candidate target genes between the normal and the CML groups.

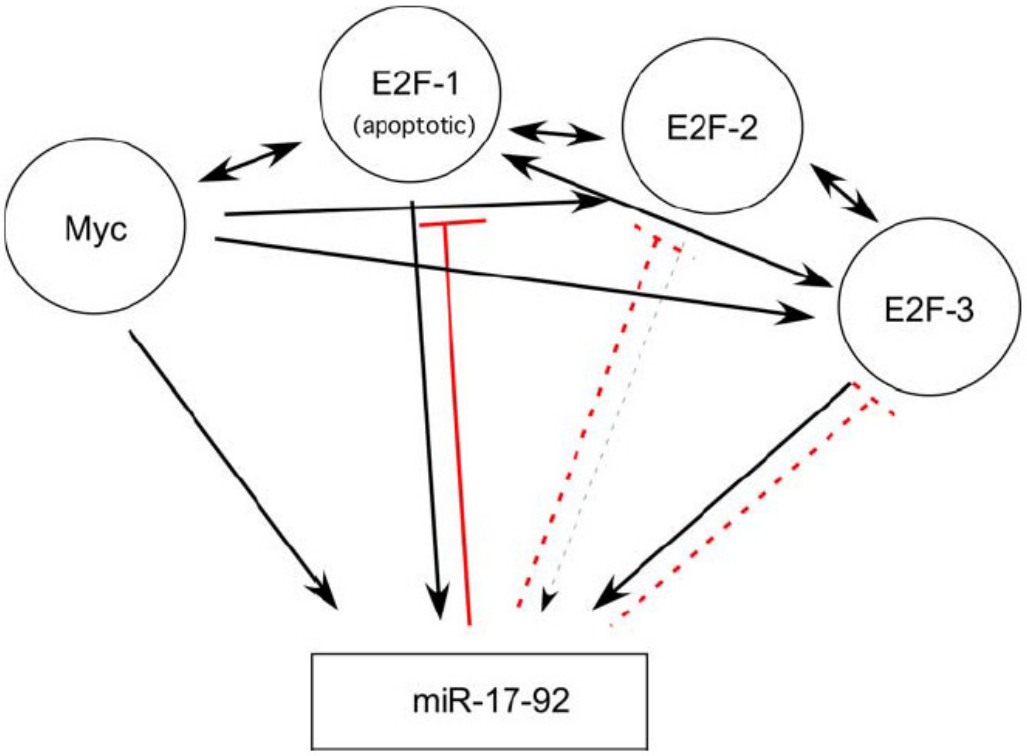


Figure 1.5: The interactions among transcription factors (E2F1, E2F2, E2F3 and MYC) and microRNAs (miR-17-92 cluster). Lines with black arrows represent a transcriptional induction. Bidirectional arrows refer to mutual transcriptional induction. Red lines represent post-transcriptional inhibition by microRNAs (Coller *et al*., 2007).

1.5 MicroRNA

MicroRNAs (miRNAs) are regarded as a new class of gene regulatory factors regulating the expression of human genes during the post-transcriptional process in recent years. MiRNAs are short, endogenous and noncoding RNA molecules, with ~22 nucleotides long, which can regulate gene expression by binding to the complementary sequences in the 3' untranslated region (3'UTR) of mRNAs (Kumar *et al*., 2007). MiRNAs have two important functions when regulating target genes: repressing the translation of mRNAs to inhibit protein expression, and directly degrading mRNAs (Bagga *et al*., 2005; Seggerson *et al*., 2002; Wu and Belasco, 2005).

1.5.1 Biogenesis of miRNAs

There are three forms of miRNAs: primary miRNAs (pri-miRNAs), precursor miRNAs (pre-miRNAs) and mature miRNAs (Figure 1.6). The conversion from pri-miRNAs to pre-miRNAs is catalyzed by RNase III enzyme (Drosha) and the double-stranded-RNA-binding protein (Pasha/DGCR8), leading to an approximately 70-nucleotide long pre-miRNAs with imperfect stem-loop structures (Denli *et al*., 2004; Han *et al*., 2004). Under the help of exportin 5 and Ran-GTP complex, the pre-miRNAs are exported into the cytoplasm (Yi *et al*., 2003). The pre-miRNAs are then processed in the cytoplasm to the mature miRNA duplexes by the help of the RNAse III enzyme Dicer (Bernstein *et al*., 2001). Usually, the mature miRNA duplexes are double-stranded RNA molecules. Only one strand of the duplex can be incorporated into the RNA-induced silencing complex (RISC), which will lead to the cleavage of

target mRNAs or repression of protein expression (Carmell and Hannon, 2004; Meister and Tuschl, 2004).

1.5.2 MiRNAs in solid tumors

It has been reported that miRNAs can be served as rheostats to adjust protein output (Baek *et al*., 2008). Some miRNAs were observed to play an important role in human tumors by affecting target oncogenes or tumor suppressor genes (Medina and Slack, 2008). The enhancer of zeste homolog 2 protein (EZH2), a component of polycomb repressive complex 2 (PRC2), results in epigenetic silencing by tri-methylating histone H3 lysine 27 (H3K27) at promoter regions of target genes (Yu *et al*., 2007). EZH2 was discovered to be overexpressed in aggressive solid tumors. Varambally *et al.* demonstrated that miR-101 can repress EZH2 protein expression and further affect EZH2 function by binding to 3'UTR of EZH2 (Varambally *et al*., 2008). Moreover, the genomic loss of miR-101 results in the overexpression of EZH2 in cancer (Varambally *et al*., 2008). Epidermal growth factor receptor (EGFR) signaling pathway is important in the maintenance and growth of epithelial tissues (Scagliotti *et al*., 2004). EGFR is often overexpressed or mutated in non-small cell lung cancer (NSCLC) patients (Irmer *et al*., 2007). Chou *et al.* have identified that deregulation of EGFR in signaling pathway results in the increase of miR-7, which in turn downregulates ERF (Est2 transcriptional repression factor) in lung cancer cells (Chou *et al*., 2010). Other studies revealed that miRNAs participate in a number of pathological and biological processes, including cell proliferation, cell differentiation and apoptosis (Ambros, 2004; Bartel, 2004; Hammond, 2006). The aberrant expression of miRNAs has been found in various cancers, and they can be regarded as oncogenes or tumor suppressor genes (Cho, 2007). As a result, miRNAs are

emerging as an important area of study in varied signaling pathways involved in cancer research.

1.5.3 MiRNAs in hematological malignancies

MiRNAs were also reported to be involved in multiple steps of myeloid differentiation, for example, the differentiation of common progenitor from the early stage to the terminal stage (Bhagavathi and Czader, 2010). The miRNAs, miR-17, miR-24, miR-146, miR-155, miR-128 and miR-181, are possibly responsible for sustaining the early stem-progenitor cell phenotype, which regulate the transition of multipotent progenitor cells to common myeloid and lymphoid progenitors (Bhagavathi and Czader, 2010). Therefore, miRNA signature may be served as a candidate biomarker to differentiate different types of myeloid and lymphoid malignancies, as well as a useful prognostic indicator for hematological diseases (Calin *et al.*, 2004; Cammarata *et al.*, 2010; Garzon *et al.*, 2008; Gordon *et al.*, 2013; Marcucci *et al.*, 2011).

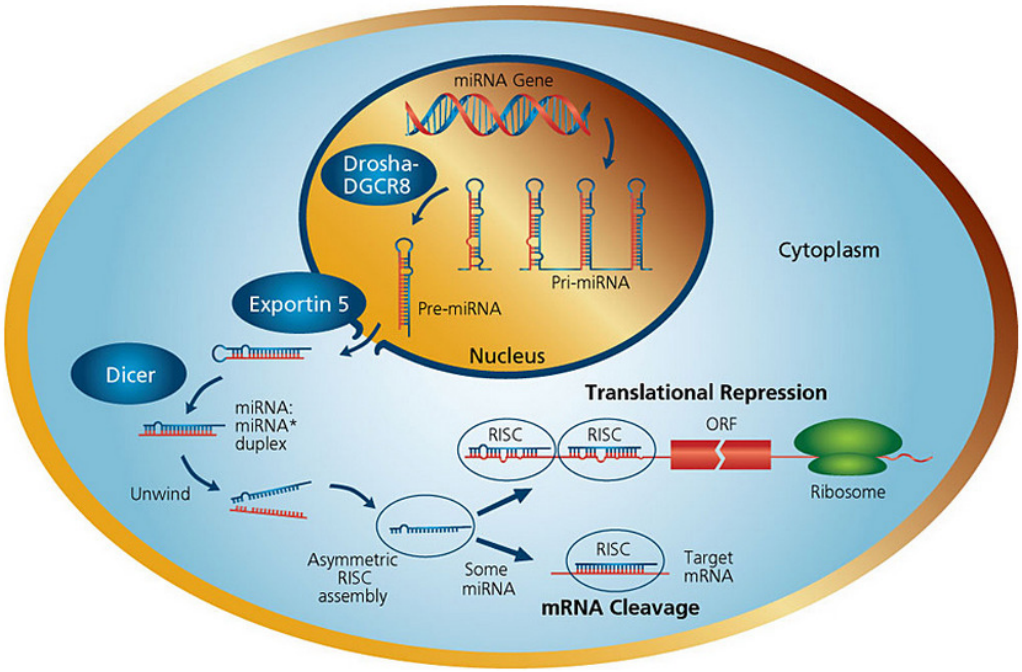


Figure 1.6: The biogenesis of miRNA and its function process (http://www.sigmaaldrich.com/life-science/functional-genomics-and-rnai/mirna/learning-center/mirna-introduction.html).

1.5.4 MiR-17-92 cluster

In our study, we plan to explore the role of a set of miRNAs transcribed from the same primary miRNA cluster in CML. The miR-17-92 cluster located in chromosome 13 consists of 7 mature miRNAs (miR-17-5p, miR-17-3p, miR-18a, miR-19a, miR-20a, miR-19b and miR-92-1) (Aguda *et al.*, 2008; Coller *et al*., 2007). These mature miRNAs have similar expression patterns in hematopoietic cell lines (Coller *et al*., 2007; Yu *et al.*, 2006a). Expression of these miRNAs can promote cell proliferation, inhibit apoptosis, and induce tumor angiogenesis in cancer cells (Mendell, 2008). Most importantly, the miR-17-92 cluster is overexpressed in chronic-phase CML patients compared with normal individuals, and its overexpression can promote cell cycle progression and proliferation, and inhibit apoptosis (Mendell, 2008; Venturini *et al*., 2007). Researchers also found that the expression of miR-17-92 cluster is BCR-ABL-dependent using miRNA-specific quantitative real-time reverse transcriptase-polymerase chain reaction (miR-qRT-PCR) in CML cell lines (Venturini *et al*., 2007). In other words, the BCR-ABL tyrosine kinase activity can affect this miRNA cluster.

1.5.5 Feedback loops among miR-17-92, E2Fs and MYC

Moreover, the feedback loops among miR-17-92, E2F1, E2F2, E2F3 and MYC have been well studied (Figure 1.5). E2F1, E2F2, E2F3 and MYC are reciprocally regulated by each other in the transcription process to form the positive feedback loops. Most importantly, E2F1, E2F2 and E2F3 are experimentally validated targets of some members of miR-17-92 cluster (Aguda *et al*., 2008; Coller *et al*., 2007). In return, E2F1, E2F2, E2F3 and MYC can induce the expression of miR-17-92 cluster transcriptionally, forming the negative feedback loops (Aguda *et al*., 2008; Coller *et*

al., 2007). The expression levels of these molecules are well controlled by each other through the feedback loops.

1.5.6 Aim of the study

As the validated data about miRNA targets are not complete at the moment, a variety of computational prediction databases have been used to analyze the profile of miRNA targets (Chen *et al*., 2011). The computational approaches demonstrate that nearly one third of human genes are potentially targeted by miRNAs. Usually, one gene can be regulated by a number of miRNAs. Accordingly, one miRNA could regulate over 200 genes on average (Cui *et al*., 2007). Researchers found that genes with similar mRNA expression profiles are likely to be regulated via the same mechanism(s), e.g. the same regulators (Allocco *et al*., 2004; Altman and Raychaudhuri, 2001; Schulze and Downward, 2001). In our study, we hypothesize that target genes regulated by the same miRNA should be co-expressed. We plan to explore the differences in the co-expression patterns of those target genes directly regulated by miR-17-92 cluster between the normal and the CML groups.

1.6 Research questions and project objectives

We have presented a detailed method to identify a disease-specific cutoff point for co-expression levels that classified the co-expressed gene pairs into strong and weak co-expression classes so that the class was best coherent with the disease state. Previous studies on the co-expression analysis calculate a p-value of correlation coefficient for each gene pair individually to identify significantly co-expressed gene pairs, which cannot reflect the overall difference in two different groups. In our study, we calculated all the correlation coefficients in each group (the normal group or the CML group) to form two different distributions, which can identify the difference between two groups from the overall structure. The different co-expression pattern indicated the alteration of molecular interactions in CML when compared to the normal state. By analyzing the biological meaning of strongly co-expressed gene pairs, we can further elucidate the underlying mechanisms of CML.

This project included four parts. First, the whole genome co-expression analysis was performed on a CML microarray dataset. Since functionally related genes tend to be co-expressed, we then explored the NPM1-associated gene set co-expression in the normal and the CML groups. In addition, genes shared the same regulators are more likely to be co-expressed. The third and fourth studies were to investigate E2F1–3 and MYC target genes co-expression, and miR-17-92 cluster target genes co-expression in the normal and the CML groups.

1.6.1 Whole genome co-expression analysis

Research questions:

- Is there difference in the co-expression patterns from the whole genome between the normal and CML groups?
- How to identify the genome-wide co-expression structures in CML?

Objectives:

- To investigate the whole genome co-expression differences between the normal and the CML groups;
- To verify and elucidate the system-level co-expression structure in the whole genome of CML.

1.6.2 NPM1-associated gene set co-expression analysis

Research questions:

- Is there difference in the co-expression patterns of NPM1-associated genes between the normal and the CML groups?
- Are the ribosome synthesis and translation process dysregulated in the CML state?

Objectives:

- To explore the differences in the co-expression patterns of NPM1-associated genes between the normal and the CML groups;
- To investigate the biological alterations related to the ribosome synthesis and translation process in the CML state.

1.6.3 E2F1–3 and MYC target genes co-expression analysis

Research questions:

- Is there difference in the co-expression patterns of those candidate target genes regulated directly and concurrently by E2F1–3 and MYC between the normal and the CML groups?
- Does the co-expression difference lead to biological alterations related to the cellular characteristics of CML?

Objectives:

- To explore the differences in the co-expression patterns of those candidate target genes between the normal and the CML groups;
- To investigate the biological alterations due to the transcriptional regulation of E2F1–3 and MYC in the CML state.

1.6.4 MiR-17-92 cluster target genes co-expression analysis

Research questions:

- Is there difference in the co-expression patterns of those candidate target genes directly regulated by miR-17-92 cluster between the normal and the CML groups?
- Does the co-expression difference lead to biological alterations in CML?

Objectives:

- To explore the differences in the co-expression patterns of those candidate target genes between the normal and the CML groups;
- To investigate the biological alterations due to the post-transcriptional regulation of miR-17-92 cluster in the CML state.

1.7 Chapter summary

The rest of the thesis is organized as follows. In Chapter 2, the whole genome co-expression analysis is performed. The system-level co-expression structure in the whole genome of CML is identified. In Chapter 3, we explore the differences in the co-expression patterns of NPM1-associated genes between the normal and the CML groups. The dysregulated ribosome synthesis and translation process are further investigated in the CML state. In Chapter 4, we explore the differences in the co-expression patterns of those candidate target genes regulated directly and concurrently by E2F1–3 and MYC between the normal and the CML groups. The biological alterations due to the transcriptional regulation of E2F1–3 and MYC are further investigated in the CML state. In chapter 5, the co-expression analysis for those candidate target genes directly regulated by miR-17-92 cluster between the normal and the CML groups is performed. We further investigate the biological alterations due to the post-transcriptional regulation of miR-17-92 cluster in the CML state. Finally, the overall discussion and conclusion are made, and the possible future directions of this study are given in Chapter 6.

Chapter 2 Whole Genome Co-expression Analysis

2.1 Methods

2.1.1 Microarray expression data

Microarray technology has been applied to monitor the expression levels of thousands of genes in cells simultaneously (Dudoit *et al.*, 2002). Gene expression analysis across different conditions, such as the normal and the disease states, provides much more information to elucidate the deep mechanisms of diseases. In this study, we analyzed the microarray dataset GSE5550, which is publicly available on the *Gene Expression Omnibus* (*GEO*) repository database (Diaz-Blanco *et al.*, 2007). To ensure the normality of the data, the expression intensity measured by each probe is log transformed and normalized by variance stabilizing transformations (VSN) method across the samples (Diaz-Blanco *et al.*, 2007). The data included in this dataset were obtained from gene expression measurements of 8,537 unique mRNAs. In that study, CD34+ hematopoietic stem and progenitor cells were collected from bone marrows of untreated CML patients in the first chronic phase and health controls (Diaz-Blanco *et al.*, 2007). All these CML patients are Ph positive with BCR-ABL/G6PDH ratios larger than 4% (the laboratory-specific median baseline (Press *et al.*, 2006)). The subjects recruited for this dataset were Caucasians from Germany. There are two groups in this dataset: i) the CML group: nine patients; and ii) the control group: eight normal individuals. In this dataset, sometimes a gene is interrogated by more than one probe. We took the average of all

the probes for the same mRNA to handle this situation (Breslin *et al.*, 2005; Kapp *et al.*, 2006).

2.1.2 Co-expression measure

Gene co-expression can be quantified by a similarity measure evaluating how similar their expression patterns are across the biological samples. Correlations between gene expression profiles have been used to identify interactions coherent to microarray samples. Pearson correlation coefficient (r) was chosen as the similarity measure to calculate the correlation coefficients for this study. Pearson correlation coefficient is represented by the direction cosine between two vectors normalized by the subtraction of their own means. Its value accounts for the angle between two feature vectors instead of the vector lengths. Furthermore, Pearson correlation coefficient demonstrates the biological relationship of two genes numerically but does not emphasize the magnitude of their expression profiles (Eisen *et al.*, 1998; Horvath and Dong, 2008). Generally, similarity measure is regarded as a kernel function between two feature vectors.

In this study, each feature vector contained the expression profiles of a gene across all the samples in the normal group or the CML group respectively. The absolute values of correlation coefficients (|r| values) were considered, since the co-expression measure output a scalar in the range from 0 to 1 where a high output indicated a strong biological relationship in either positive or negative direction, and a low output represented a weak biological relationship. The co-expression level was denoted by $C_d(i, j)$ if the expression profiles of two genes were extracted from the

disease (CML) group, and $C_n(i, j)$ for the normal group, as shown in Formulas 1 and 2.

$$C_d(i,j) = |cor(x_{di}, x_{dj})| \quad (1)$$

$$C_n(i,j) = |cor(x_{ni}, x_{nj})| \quad (2)$$

where $C_d(i,j)$ and $C_n(i,j)$ are defined as the absolute values of correlation coefficients between the expression profiles of gene i and gene j in the CML group and the normal group, respectively (Horvath and Dong, 2008); x_{di} and x_{dj} represent the expression profiles of the i^{th} and j^{th} genes in the CML group; x_{ni} and x_{nj} are the expression profiles of the i^{th} and j^{th} genes in the normal group; $cor(x_{di},x_{dj})$ stands for the Pearson correlation coefficient between the i^{th} and j^{th} genes in the CML group; $cor(x_{ni},x_{nj})$ represents the Pearson correlation coefficient between the i^{th} and j^{th} genes in the normal group.

2.1.3 Identification of the disease-specific cutoff point

Two sets of correlation coefficients in the normal and the CML groups were obtained, considering two-gene combinations among all the 8,537 mRNAs. These two sets of data formed two different cumulative distributions. In the next step, we performed two-sample Kolmogorov-Smirnov (KS) test to exam if these two sets of correlation coefficients significantly differed in terms of the overall distributions between two different conditions. The significance for KS test is represented by the p-value for the maximum deviation between two cumulative distributions of C_d and C_n (Formulas 3-5). At the maximum deviation a threshold was identified to classify the co-expressed gene pairs into strong and weak co-expression classes, called the disease-specific

cutoff point (C), so that the class was significantly associated with the disease (CML) group. Different conditions (e.g. different diseases and different datasets) had different cutoff points, depending on the cumulative distributions from the collected data. The cutoff point represented a co-expression level, at which F_d and F_n were extremely deviated.

$$D = \max_{C} | F_d (C) - F_n (C) | \tag{3}$$

$$F_d (C) = Prob (C_d \geq C) \tag{4}$$

$$F_n (C) = Prob (C_n \geq C) \tag{5}$$

where F_d and F_n represent the cumulative distribution functions (CDFs) of C_d and C_n, respectively; D is defined as the maximum deviation; C represents the disease-specific cutoff point.

2.1.4 Distribution-based classification of co-expressed gene pairs

After the disease-specific cutoff point was identified, the gene pairs were classified into four co-expression classes according to the distributions: i) strongly co-expressed gene pairs in the normal group (NS) with $|r| \geq C$ in the normal group; ii) strongly co-expressed gene pairs in the CML group (CS) with $|r| \geq C$ in the CML group; iii) weakly co-expressed gene pairs in the normal group (NW) with $|r| < C$ in the normal group; and iv) weakly co-expressed gene pairs in the CML group (CW) with $|r| < C$ in the CML group (Table 2.1). Chi-square test was applied to determine if the proportions of strongly and weakly co-expressed gene pairs significantly differed between the normal and the CML groups.

From the above classification, some pairs may be strongly co-expressed in both the normal and the CML groups, or weakly co-expressed in both these two groups. For better illustration of the groups' characteristics, the co-expressed gene pairs can be further grouped into specific and common pairs (Figure 2.1). The normal-specific strongly co-expressed pairs were the gene pairs strongly co-expressed only in the normal group, which were regarded as the potential molecular interactions maintaining physiological balance in healthy individuals and the impairment of these connections may lead to diseases. Obviously, these pairs were the CML-specific weakly co-expressed pairs, which were weakly co-expressed only in the CML group. The common strongly co-expressed pairs were those gene pairs strongly co-expressed in both these two groups. The CML-specific strongly co-expressed pairs were the gene pairs strongly co-expressed only in the CML group, which represented the characteristics of the disease and may be the pathogenic alternatives when the corresponding normal-specific pairs were not co-expressed in response to stress. It is the same situation that these pairs were regarded as the normal-specific weakly co-expressed pairs. The common weakly co-expressed pairs were those gene pairs weakly co-expressed in both the normal and the CML groups.

Table 2.1: The classification of co-expressed gene pairs

Group	Strongly co-expressed gene pairs	Weakly co-expressed gene pairs
Normal	NS	NW
CML	CS	CW

The total number of gene pairs = NS+NW = CS+CW

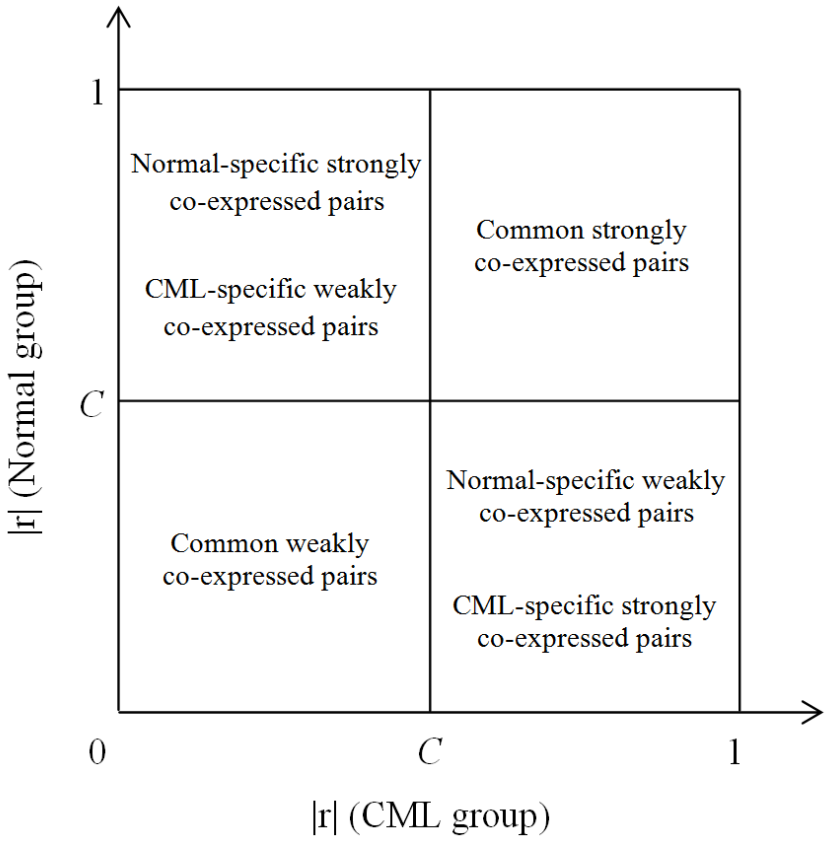


Figure 2.1: Co-expression regions partitioned by the disease-specific cutoff point (*C*). The x-axis is the |r| from 0 to 1 in the CML group. The y-axis is the |r| from 0 to 1 in the normal group.

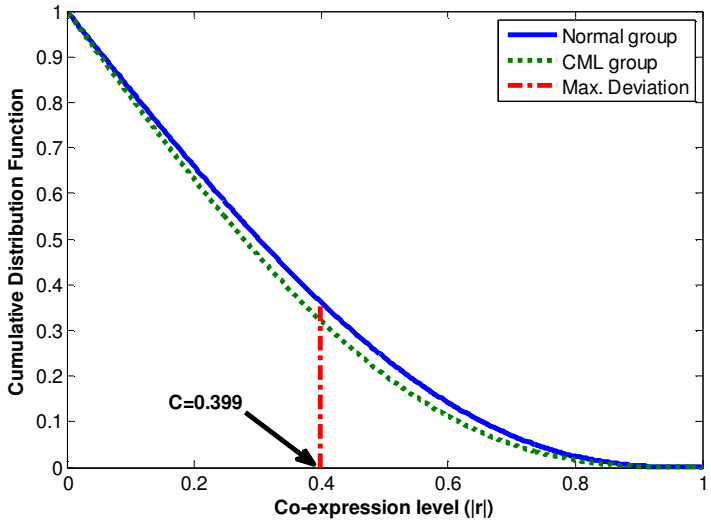
2.2 Results

2.2.1 Identification of the co-expression difference and disease-specific cutoff point from the whole genome

All genes from the CML microarray dataset were considered to form the genome-wide expression matrices for the normal and the CML groups. We calculated the correlation coefficients of all the possible gene pair combinations in these two groups. In each group, there was a set of correlation coefficients of 36,435,916 gene pairs. The cumulative distributions of these two sets of data were plotted (Figure 2.2). Two-sample KS test was performed to identify the difference from the overall structure. The results showed that the two cumulative distributions were significantly different between the normal and the CML groups (p-value < 0.05 for the maximum deviation $D = 0.041$).

The disease-specific cutoff point, $C = 0.399$, was identified at the maximum deviation (Figure 2.2). Figure 2.2 (b) illustrates that the deviation was small at the two extremes, and the peak ($D = 0.041$) was found at the disease-specific cutoff point. Two co-expression patterns were so distinct that the normal group had more strongly co-expressed (level above ~0.399) gene pairs than that in the CML group. The cutoff point classified gene pairs into four co-expression classes (Table 2.2). From the results, we could also observe that the number of strongly co-expressed gene pairs in the normal group (13,270,647) was larger than that in the CML group (11,788,939). Chi-square test indicated that the proportions of strongly and weakly co-expressed gene pairs significantly differed between the normal and the CML groups (p-value < 0.05 for the statistic $\chi^2 = 133,528$).

a.



b.

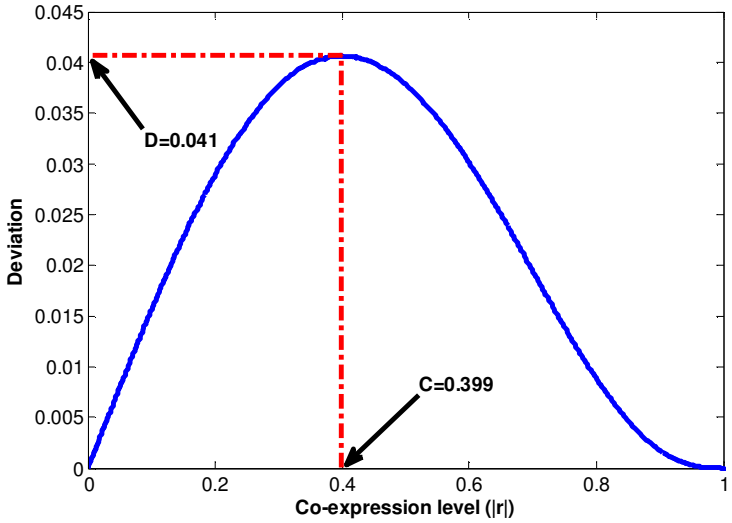


Figure 2.2: Plots of distributions for genome-wide co-expression analysis. (a) Cumulative distribution functions of co-expression levels in the normal and the CML groups. (b) Deviation distribution against different co-expression cutoff points.

Table 2.2: Cross-tabulation of gene pair counts in the genome-wide analysis

Group	# of strongly co-expressed gene pairs	# of weakly co-expressed gene pairs
Normal	13,270,647	23,165,269
CML	11,788,939	24,646,977

2.2.2 Genome-wide co-expression galaxy and structures

To account for the differential structure, the co-expression galaxy was constructed on the genomic scale and partitioned into four regions (Figures 2.3 and 2.4): i) normal-specific strongly co-expressed pairs (CML-specific weakly co-expressed pairs): the percentage was 24.445%; ii) common strongly co-expressed pairs: the percentage was 11.977%; iii) CML-specific strongly co-expressed pairs (normal-specific weakly co-expressed pairs): the percentage was 20.378%; and iv) common weakly co-expressed pairs: the percentage was 43.200%. This kind of distribution-based classification of co-expressed gene pairs well demonstrated the co-expression structures. Figures 2.3 and 2.4 illustrate that most of gene pairs belonged to the common weakly co-expressed pairs. The gene pairs specifically co-expressed in a particular group had important biological meanings, which were totally different in the normal and the CML groups. From the results, we can see that there were more normal-specific strongly co-expressed pairs than CML-specific strongly co-expressed pairs.

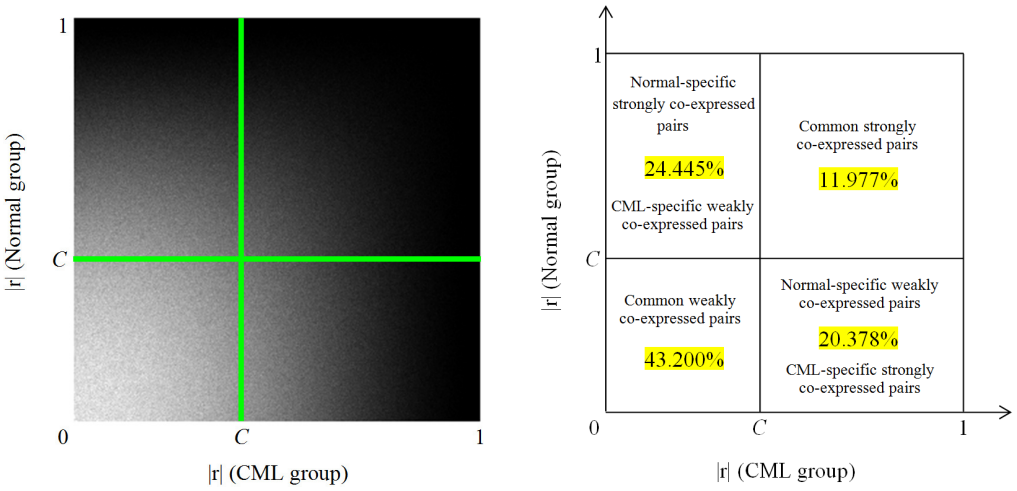


Figure 2.3: Co-expression galaxy (left) and four regions partitioned by the disease-specific cutoff point, C = 0.399 (right). Each correlation coefficient (|r|) is represented by one white dot in the galaxy. More dots mean that there are more |r| values located in that region.

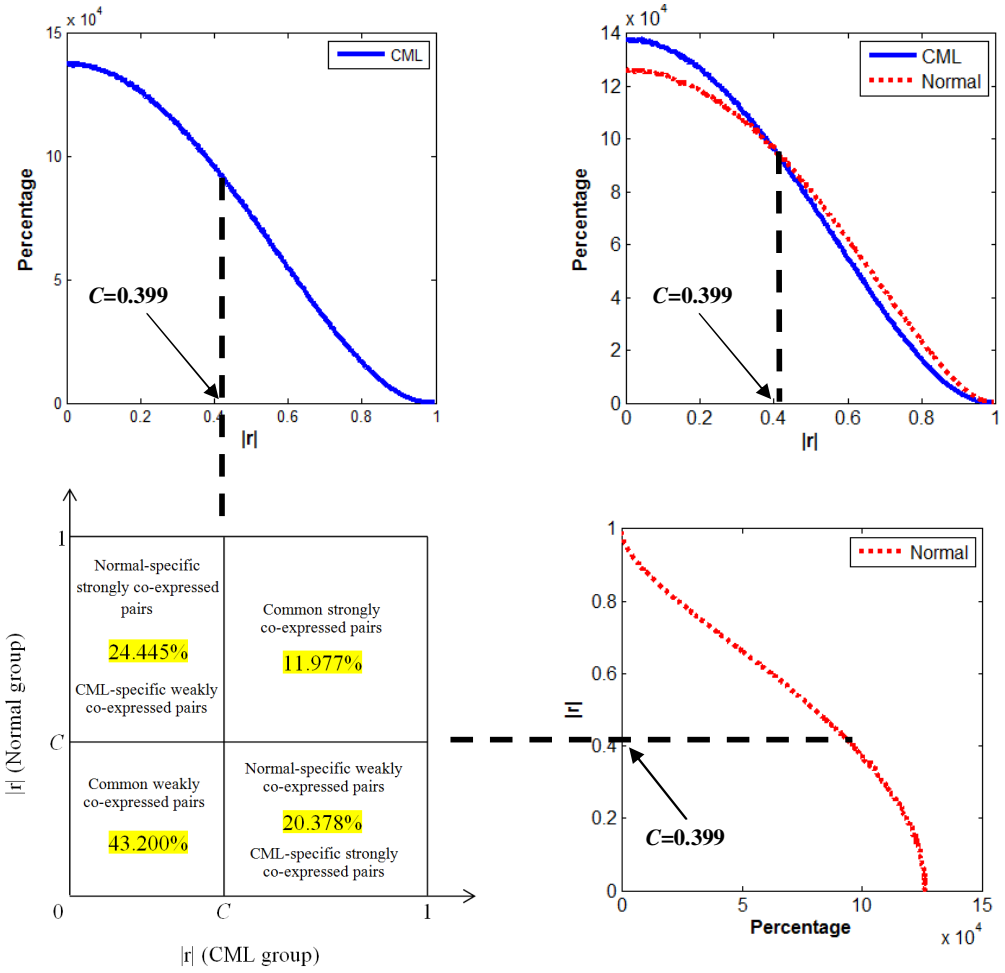


Figure 2.4: Relationship between co-expression structures and partitioned four regions in the co-expression galaxy. The red dot line represents the distribution of percentage for correlation coefficients (|r| values) in the normal group. The blue solid line stands for the distribution of percentage for correlation coefficients (|r| values) in the CML group.

2.3 Discussion and conclusion

In this chapter, we identified the overall differences in the co-expression patterns between the normal and the CML groups from the whole genome. Correlation coefficients for all the possible gene pairs were considered to form two different cumulative distributions. The co-expression pattern differences were reflected from the overall structure, not only considering pair by pair independently. Two-sample KS test was performed to identify the difference (Figure 2.2). Firstly, the maximum deviation (D = 0.041) between two cumulative distributions indicated the difference between the normal and the CML groups structurally. Then, a disease-specific cutoff point (C = 0.399) was identified at the maximum deviation to classify the co-expressed gene pairs so that the class was best coherent with the disease phenotype.

The distribution-based classification divided co-expressed gene pairs into four regions based on their locations in the co-expression galaxy, forming the co-expression structures (Figures 2.3 and 2.4): i) the normal-specific strongly co-expressed pairs (the CML-specific weakly co-expressed pairs); ii) the common strongly co-expressed pairs; iii) the CML-specific strongly co-expressed pairs (the normal-specific weakly co-expressed pairs); and iv) the common weakly co-expressed pairs. This kind of classification considered all the gene pairs to locate them to different locations based on their different co-expression levels (|r| values) and different groups (the normal group or the CML group).

The specifically co-expressed gene pairs had special biological meanings. The normal-specific strongly co-expressed pairs indicated the potential molecular interactions maintaining physiological balance in healthy individuals, which was

regarded as the inter-gene linkages. The CML-specific strongly co-expressed pairs represented the characteristics of the disease. Our results showed that there were more normal-specific strongly co-expressed pairs than CML-specific strongly co-expressed pairs (Figures 2.3 and 2.4). In other words, genes were more likely to be co-expressed in the normal group when compared to the CML group, resulting in the different co-expression patterns. The specifically co-expressed gene pairs and the different co-expression pattern may be associated with the CML disease.

Chapter 3 NPM1-Associated Gene Set Co-expression Analysis

3.1 Method

3.1.1 NPM1-associated co-expression networks

Besides the structural genome-wide co-expression analysis, another important research question is if the normal and the CML states exhibit different co-expression patterns over a set of functional genes related to a particular physiological function or biological process. The NPM1-associated gene set (GCM_NPM1), in total 116 genes including NPM1, was chosen as the candidate gene set for the co-expression analysis. There were 93 out of 116 NPM1-associated genes found in the CML microarray dataset GSE5550 (Appendix A1). We extracted the expression profiles of these 93 genes from the expression matrix for the co-expression analysis. The reduced expression matrix was in dimension of 93x17, where each row represented the relative expression levels of a gene across all the samples (8 normal and 9 CML samples).

Using the same approach as the genome-wide co-expression analysis, the correlation coefficients for all the possible gene pairs of 93 genes were calculated (Sections 2.1.2 and 2.1.3). Following the same distribution-based approach for gene pair classification as described above (Section 2.1.4), the gene pairs were also classified into the normal-specific, CML-specific and common co-expressed pairs. The normal-

specific strongly co-expressed pairs represented the gene-gene associations, e.g. protein-protein interactions, which can maintain the corresponding physiological and pathological balance in the normal state. The CML-specific strongly co-expressed pairs indicated the pathologically altered gene-gene associations, or protein interactions.

Gene-gene co-expression networks of the normal-specific strongly co-expressed pairs and CML-specific strongly co-expressed pairs were constructed to visualize and explain the underlying mechanisms of CML related to NPM1. These networks focused on the connections between NPM1 and its strongly co-expressed genes, as well as between NPM1 and ribosomal protein (RP) genes to elucidate the altered associations of NPM1 with ribosome biogenesis and activities in CML. Centered with NPM1, these networks described how many genes co-expressed with NPM1 were and what these genes were. To visualize gene co-expression networks, we used nodes to represent genes and edges to indicate the strong correlation between nodes.

3.1.2 Gene ontology annotation for NPM1-assoicated genes

3.1.2.1 Flow chart

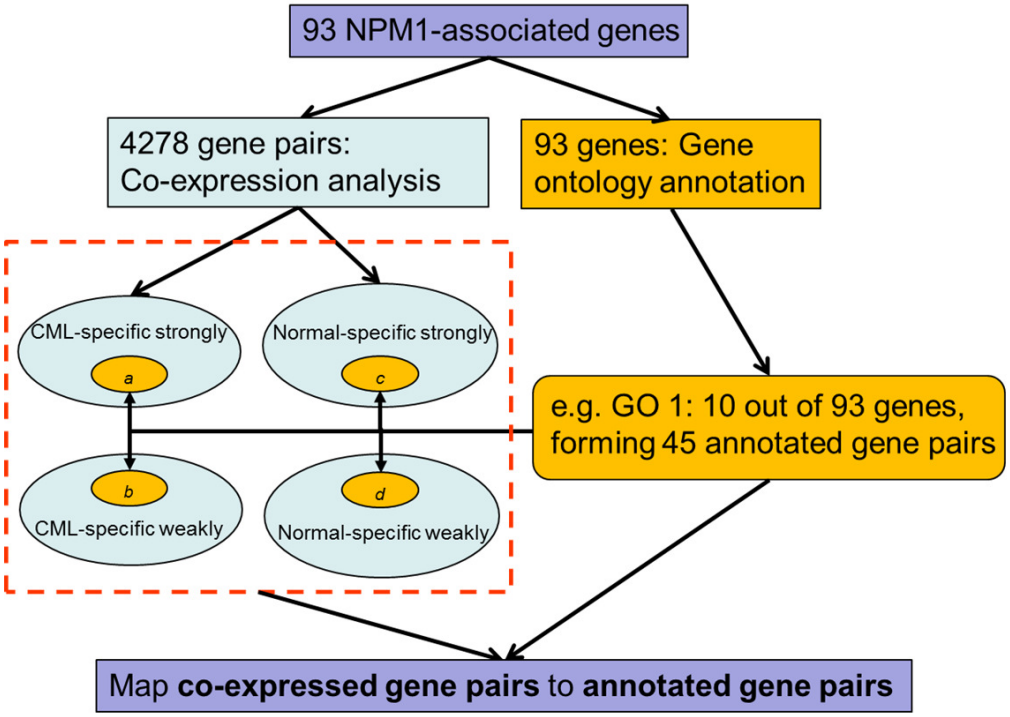


Figure 3.1: Flow chart for the gene ontology (GO) annotation of NPM1-assoicated genes. "*a*" refers to the mapped CML-specific strongly co-expressed pairs. "*b*" represents the mapped CML-specific weakly co-expressed pairs. "*c*" stands for the mapped normal-specific strongly co-expressed pairs. "*d*" represents the mapped normal-specific weakly co-expressed pairs.

3.1.2.2 Gene ontology annotation

3.1.2.2.1 Gene ontology

Gene ontology (GO) provides a systematic language, or ontology to describe gene and gene product attributes across all species (Gene Ontology Consortium, 2008). It can be divided into three categories (Ashburner *et al.*, 2000): i) biological process: a set of molecular events with a defined beginning and end, often including a chemical or physical transformation; ii) molecular function: the elemental activities of a gene product at the molecular level, including specific binding to ligands and catalysis; and iii) cellular component: the parts of a cell or the extracellular environment where a gene product is active. In this study, we applied gene ontology to classify the NPM1-associated genes into different groups, to further explore the biological meaning for the co-expressed gene pairs in the CML state.

3.1.2.2.2 Gene ontology annotation using *DAVID* database

We applied the *Database for Annotation, Visualization and Integrated Discovery* (*DAVID*) to annotate these 93 genes. Data analysis for gene lists is a very important task to understand the underlying biological mechanisms. *DAVID* is useful to extract the biological meaning by combining an integrated biological knowledgebase and multiple analytic tools (Huang *et al*., 2009). Functional annotation chart was chosen to identify the significant batch annotation and GO terms that were most pertinent to the input data. When we uploaded the NPM1-associated gene list to *DAVID* for functional annotation, the annotation chart provided the significantly enriched GO terms. The significance of GO term enrichment is calculated based on a modified Fisher exact test, Expression Analysis Systematic Explorer (EASE) score. The EASE score is regarded as a more conservative and robust adjustment than the Fisher exact

probability, which depends on the concept of jackknifing (Hosack *et al*., 2003). Jackknifing refers to a procedure in which the stability of a statistic can be ascertained. The statistic is recalculated many times to obtain a distribution of probabilities by removing a single data point (Hosack *et al*., 2003; Tukey, 1958). The EASE score is calculated by removing one gene from a category (e.g. a GO term) and calculating the Fisher exact probability for the remaining genes in that category (Hosack *et al*., 2003).

A hypothetical example to calculate the EASE score (Huang *et al*., 2009):
There are 40 out of 30,000 whole genome genes involved in p53 signaling pathway. In a list of input genes of interest, 3 out of 300 genes participate in this pathway. We want to exam if 3/300 is more than random chance when compared to the whole genome 40/30000.

A 2x2 contingency table is shown according to this situation:

	Genes of interest	Whole genome
In Pathway	2 (=3-1)	40
Not In Pathway	297	29,960

Traditionally, we directly put "3" into the first box of contingency table, and the Fisher exact test p-value = 0.008 < 0.05. The conclusion is that this gene list is significantly associated with p53 signaling pathway. When applying the modified Fisher exact test, we put "2" (=3-1) into the first box of contingency table, and the EASE score = 0.06 > 0.05. The result from EASE score is not significant for this case. Therefore, EASE score is more conservative and stringent to get the significant result. *DAVID* also provides false discovery rate (FDR) to control the expected proportion of false positives for the multiple hypotheses.

Using *DAVID*, we annotated these 93 NPM1-associated genes to get the annotated genes involved in the significantly enriched GO terms. All these three GO categories (biological process, molecular function and cellular component) were considered in our study. The selection criteria for the significantly enriched GO terms were: i) EASE score < 0.05; and ii) FDR < 0.05.

3.1.2.2.3 Mapping co-expressed gene pairs to annotated gene pairs

The annotated genes in each GO term were paired with all the possible combinations to form the annotated gene pairs. The annotated gene pairs from each GO term were mapped to the identified co-expressed gene pairs: the mapped CML-specific strongly co-expressed (*a*), the mapped CML-specific weakly co-expressed (*b*), the mapped normal-specific strongly co-expressed (*c*) and the mapped normal-specific weakly co-expressed pairs (*d*). Fisher exact test was used to verify if there were more mapped CML-specific strongly co-expressed pairs than mapped normal-specific strongly co-expressed pairs in each GO term. In other words, we want to test if genes were more likely to be co-expressed in the CML group compared to the normal group. As a result, one-sided p-value was chosen. The mapped gene pairs were slotted into a contingency table for Fisher exact test (Table 3.1). The multiple-hypothesis correction was performed for a list of mapped GO terms by following a more stringent criterion, Bonferroni correction. That is, the p-value of Fisher exact test for each GO term was multiplied by the total number of considered GO terms. A GO term was significantly mapped if its corrected p-value was still smaller than the error rate (0.05).

Table 3.1: The classification of mapped gene pairs for Fisher exact test

Group	Mapped strongly specific pairs	Mapped weakly specific pairs
CML	*a*	*b*
Normal	*c*	*d*

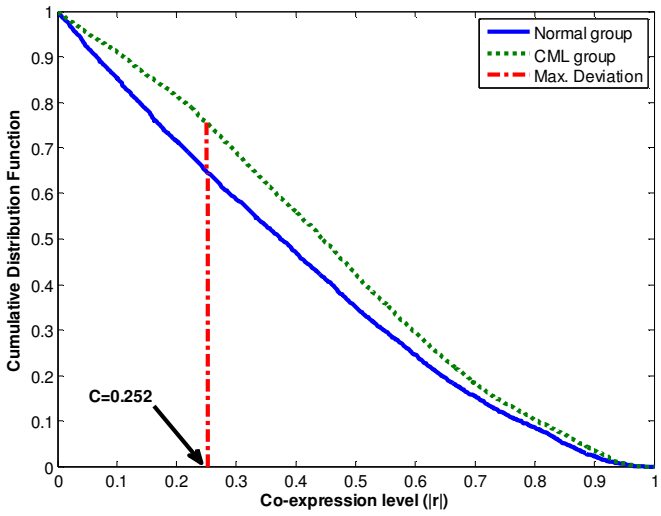
3.2 Results

3.2.1 Identification of structural co-expression difference

The correlation coefficients for all the possible gene pair combinations of these 93 NPM1-associated genes were calculated. In each group, there was a set of correlation coefficients of 4,278 gene pairs. The cumulative distributions of these two sets of data were plotted (Figure 3.2). Two-sample KS test was performed to identify the difference from the overall structure. The results showed that the two distributions in the normal and the CML groups were significantly different (p-value = 1.71×10^{-22} < 0.05 for the maximum deviation $D = 0.109$).

The disease-specific cutoff point, $C = 0.252$, was identified at the maximum deviation (Figure 3.2). Two co-expression patterns were so distinct that the CML group had more strongly co-expressed (level above ~0.252) gene pairs than that in the normal group. The cutoff point classified gene pairs into four co-expression classes (Table 3.2). Chi-square test indicated that the proportions of strongly and weakly co-expressed gene pairs significantly differed between the normal and the CML groups (p-value = $5.20 \times 10^{-28} < 0.05$ for the statistic $\chi^2 = 120$).

a.



b.

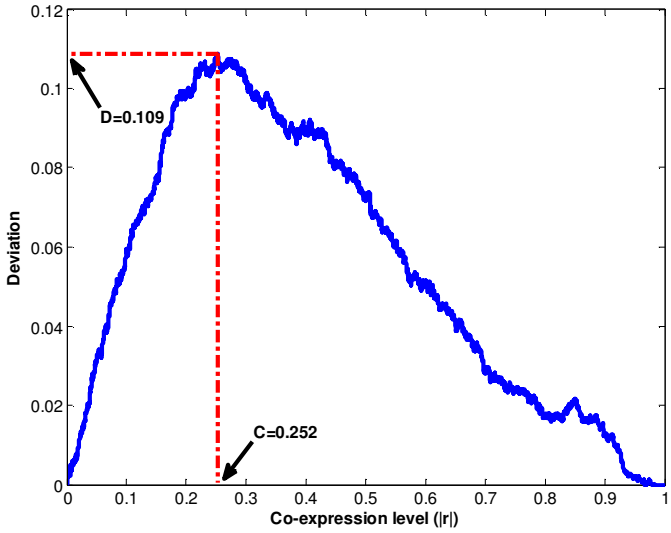


Figure 3.2: Plots of distributions for the 93 NPM1-associated genes co-expression analysis. (a) Cumulative distribution functions of co-expression levels in the normal and the CML groups. (b) Deviation distribution against different co-expression cutoff points.

Table 3.2: Cross-tabulation of gene pair counts in the NPM1-associated genes co-expression analysis

Group	# of strongly co-expressed gene pairs	# of weakly co-expressed gene pairs
Normal	2,763	1,515
CML	3,228	1,050

3.2.2 Co-expression galaxy and structures for NPM1-associated genes

The co-expression galaxy was plotted and partitioned into four regions, according to the same procedures with the genome-wide co-expression analysis (Section 2.2.2): i) normal-specific strongly co-expressed pairs (CML-specific weakly co-expressed pairs): the percentage was 13.628%; ii) common strongly co-expressed pairs: the percentage was 50.958%; iii) CML-specific strongly co-expressed pairs (normal-specific weakly co-expressed pairs): the percentage was 24.497%; and iv) common weakly co-expressed pairs: the percentage was 10.916% (Figures 3.3 and 3.4). From the results, we observed that there were more CML-specific strongly co-expressed pairs than normal-specific strongly co-expressed pairs.

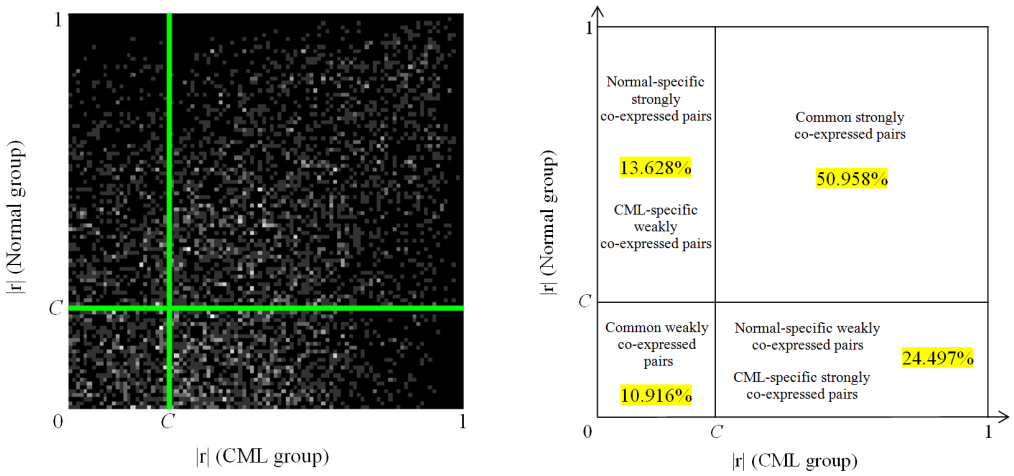


Figure 3.3: Co-expression galaxy (left) and four regions partitioned by the disease-specific cutoff point, C = 0.252 (right). Each correlation coefficient (|r|) is represented by one white dot in the galaxy. More dots mean that there are more correlation coefficients located in that region.

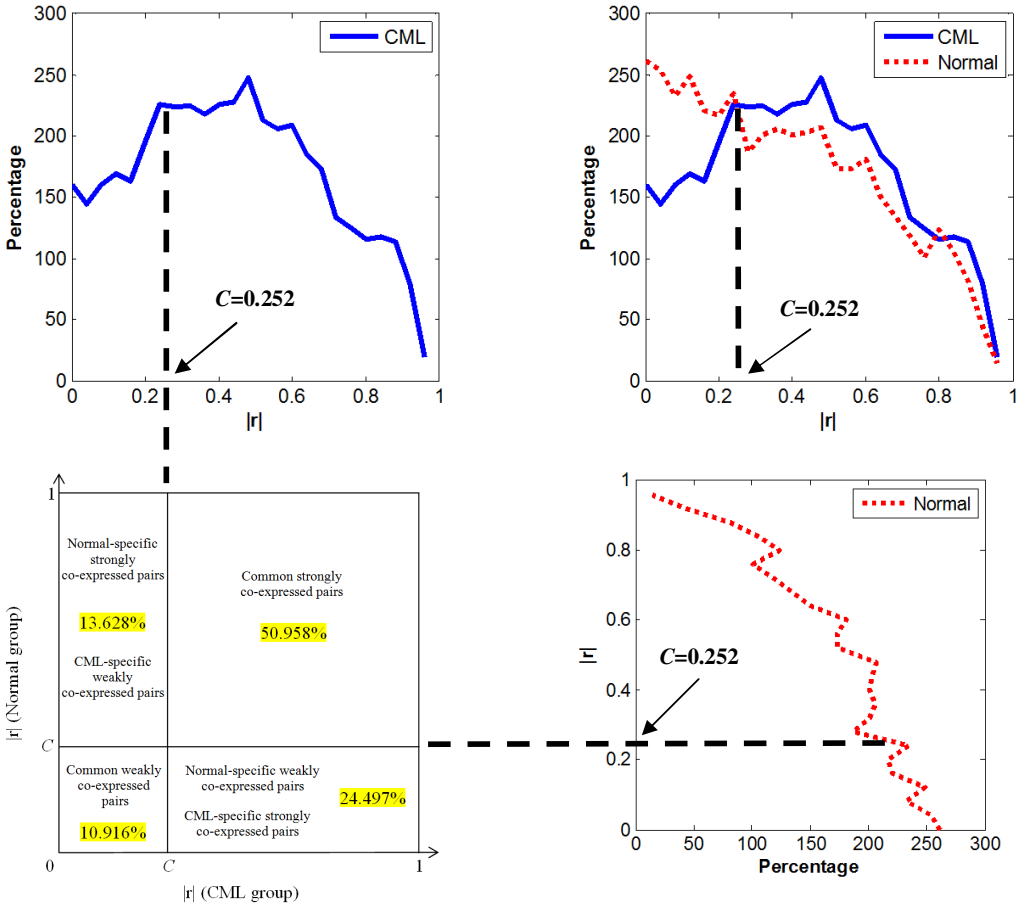


Figure 3.4: Relationship between co-expression structures and partitioned four regions in the co-expression galaxy for the NPM1-associated genes co-expression analysis. The red dot line represents the distribution of percentage for correlation coefficients (|r| values) in the normal group. The blue solid line stands for the distribution of percentage for correlation coefficients (|r| values) in the CML group.

3.2.3 Co-expression networks centered with NPM1

The co-expression networks centered with NPM1 for the normal-specific and CML-specific strongly co-expressed pairs were constructed (Figures 3.5, 3.6 and 3.7). From the results we can see that there were more genes strongly co-expressed with NPM1 in the CML group compared to the normal group (Figures 3.5 and 3.6). In addition, in the CML group, three RP genes were found to be strongly co-expressed with NPM1 (Figure 3.6): RPL31, PRL36A and RPL10A. However, no RP genes were found to be co-expressed with NPM1 in the normal group (Figure 3.5). We further expanded the CML co-expression network from NPM1 with RP genes (Figure 3.7). The results showed that RP genes were more likely to be co-expressed with NPM1 in the CML group when compared to the normal group.

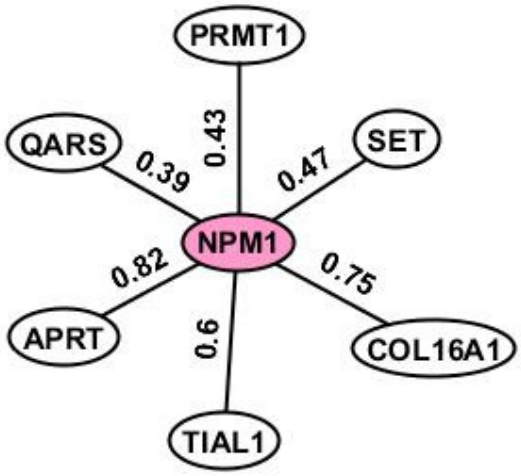


Figure 3.5: Normal-specific co-expression network of NPM1 (using yEd). Values next to the lines indicate the |r| values.

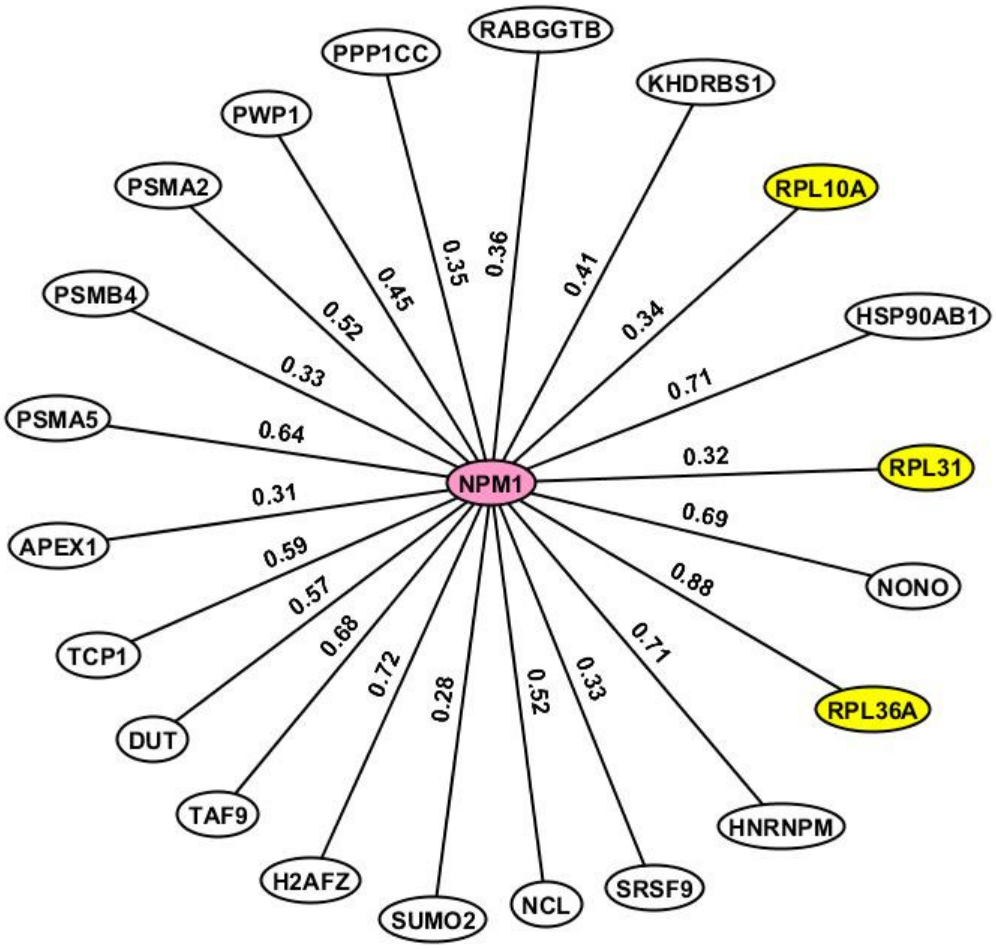


Figure 3.6: CML-specific co-expression network of NPM1 (using yEd). Values next to the lines indicate the |r| values. Genes with yellow colors are RP genes.

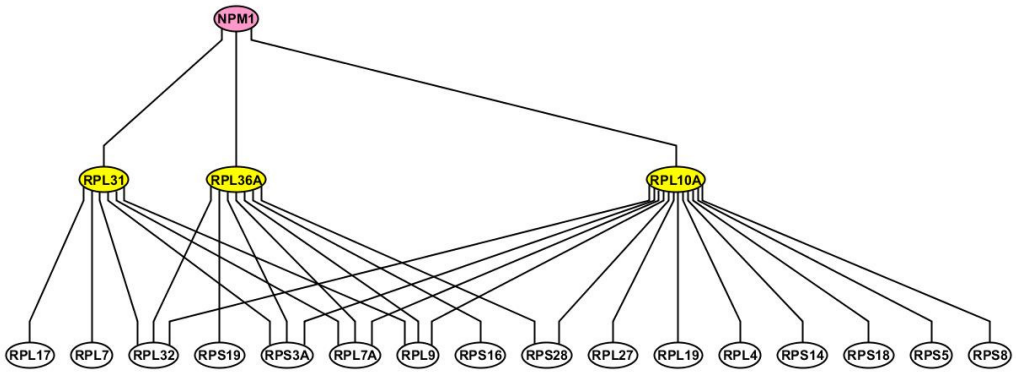


Figure 3.7: CML-specific RP gene co-expression network expanded from NPM1 (using yEd). Genes with yellow colors are those RP genes strongly co-expressed with NPM1.

3.2.4 *David* annotation for enriched gene ontology

3.2.4.1 Biological process

According to the selection criteria (EASE score < 0.05 and FDR < 0.05), eight significantly enriched GO terms for biological processes were identified (Table 3.3). We obtained the annotated genes involved in each biological process and formed the annotated gene pairs. In the next step, the co-expressed gene pairs were mapped to the annotated gene pairs. The results showed that all these eight processes had more mapped CML-specific strongly co-expressed pairs (Table 3.4). In other words, genes were more likely to be co-expressed in the CML group when compared to the normal group. Fisher exact test was used to indicate the significance. The results showed that *Translational elongation*, *Translation*, *Cellular protein metabolic process*, *RNA processing* and *RNA metabolic process* were significantly mapped (p-values < 0.05, and corrected p-values < 0.05).

GO terms *Translational elongation* and *Translation* were related to gene translation process. *Translational elongation* is defined as the successive addition of amino acid residues to a nascent polypeptide chain in the protein biosynthesis process. *Translation* refers to the cellular metabolic process to form a protein by using a mature mRNA molecule to determine the amino acids sequence in a polypeptide chain. We further plotted the co-expression networks for the strongly co-expressed gene pairs in the normal and the CML groups (Figures 3.8 and 3.9). From the co-expression networks, we also observed that there were more connections in the CML group compared to the normal group. Genes identified in the co-expression networks were classified to two major classes: i) RP genes, such as ribosomal protein L6 (RPL6) and ribosomal protein S28 (RPS28); and ii) translation factors, such as

eukaryotic translation elongation factor 2 (EEF2) and eukaryotic translation initiation factor 3, subunit F (EIF3F). The results revealed that nearly all the co-expressed genes were RP genes, which are responsible for encoding the ribosomal small and large subunits.

The basic information for the identified translation factors was obtained from *National Center for Biotechnology Information* (*NCBI*) database. Protein products from EEF2 and EEF1B2 belong to translation elongation factors. EEF2 is a member of the GTP-binding translation elongation factor family, which is very important for protein synthesis. This protein can mediate the process of GTP-dependent translocation of the nascent protein chain from A-site to P-site on the ribosome. The encoded protein of EEF1B2 is a guanine nucleotide exchange factor responsible for the transfer of aminoacylated transfer RNAs (tRNAs) to the ribosome. Eukaryotic translation initiation factor 3, subunit F and initiation factor 4B (EIF3F and EIF4B) are translation initiation factors, which are vital to initiate the translation.

Table 3.3: Biological process_Enriched GO terms for the functional annotation of NPM1-associated genes

#	Enriched GO terms	Genes found in our data	EASE score	FDR
1	Translational elongation	35	6.80×10^{-51}	9.00×10^{-48}
2	Translation	38	3.20×10^{-36}	4.20×10^{-33}
3	Cellular protein metabolic process	53	6.00×10^{-18}	7.90×10^{-15}
4	RNA processing	23	8.60×10^{-12}	1.10×10^{-08}
5	RNA metabolic process	26	1.70×10^{-09}	2.20×10^{-06}
6	mRNA processing	14	2.40×10^{-07}	3.20×10^{-04}
7	RNA splicing	13	4.70×10^{-07}	6.20×10^{-04}
8	mRNA metabolic process	14	1.20×10^{-06}	1.60×10^{-03}

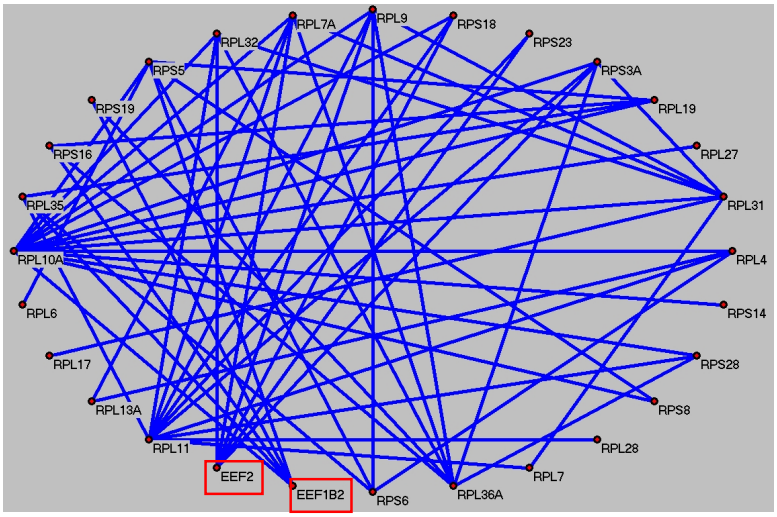
GO: gene ontology. EASE score: a modified Fisher exact test, Expression Analysis Systematic Explorer score. FDR: false discovery rate.

Table 3.4: Mapping co-expressed gene pairs to annotated gene pairs from each biological process

#	GO terms	Fisher exact test					Corrected
		a	*b*	*c*	*d*	p-value	p-value
1	**Translational elongation**	59	5	5	59	**<0.001**	**<0.008**
2	**Translation**	89	10	10	89	**<0.001**	**<0.008**
3	**Cellular protein metabolic process**	299	116	116	299	**<0.001**	**<0.008**
4	**RNA processing**	63	28	28	63	**<0.001**	**<0.008**
5	**RNA metabolic process**	84	39	39	84	**<0.001**	**<0.008**
6	mRNA processing	18	11	11	18	0.057	0.456
7	RNA splicing	16	10	10	16	0.082	0.656
8	mRNA metabolic process	18	11	11	18	0.057	0.456

GO: gene ontology. GO terms highlighted in bold text are significantly mapped. *a*: mapped CML-specific strongly co-expressed pairs. *b*: mapped CML-specific weakly co-expressed pairs. *c*: mapped normal-specific strongly co-expressed pairs. *d*: mapped normal-specific weakly co-expressed pairs.

a.



b.

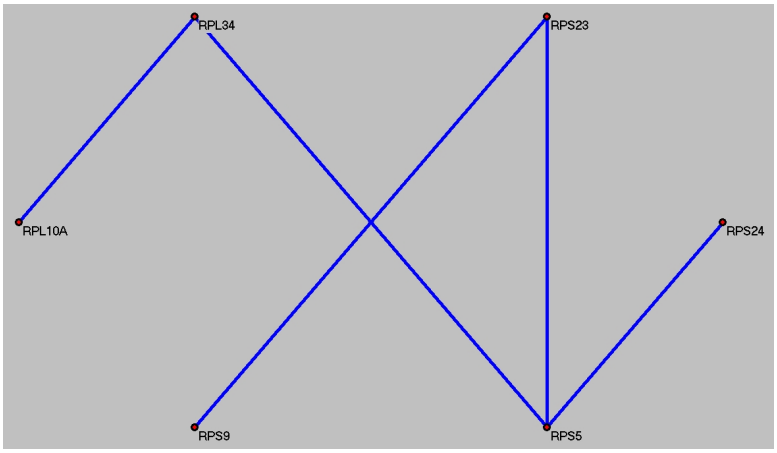
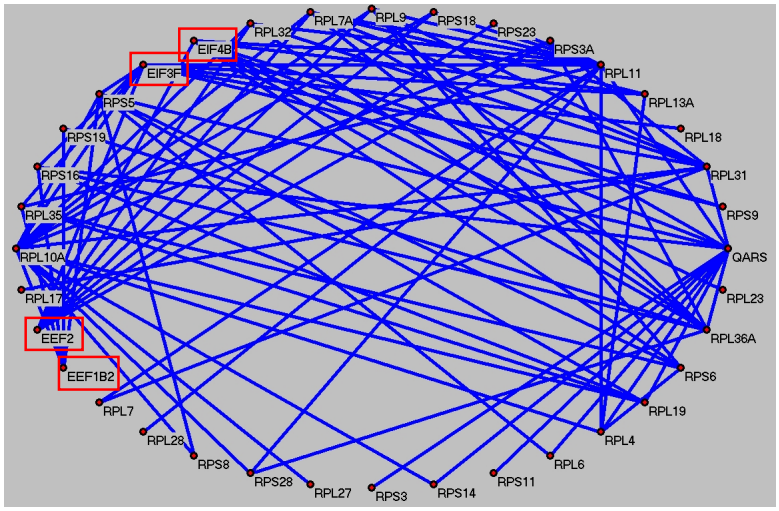


Figure 3.8: Co-expression networks for the mapped strongly co-expressed pairs in the *Translational elongation* biological process (using Pajek). Genes with red rectangles are not RP genes. (a) Mapped CML-specific strongly co-expressed pairs. (b) Mapped normal-specific strongly co-expressed pairs.

a.



b.

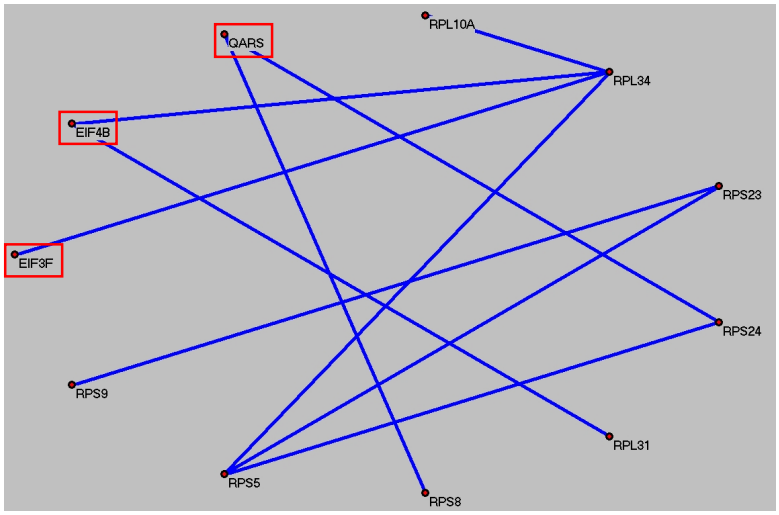


Figure 3.9: Co-expression networks for the mapped strongly co-expressed pairs in the *Translation* biological process (using Pajek). Genes with red rectangles are not RP genes. (a) Mapped CML-specific strongly co-expressed pairs. (b) Mapped normal-specific strongly co-expressed pairs.

3.2.4.2 Cellular component

Based on the same selection criteria (EASE score < 0.05 and FDR < 0.05), 21 significantly enriched GO terms for cellular components were identified (Table 3.5). The annotated genes involved in each GO term were obtained and formed the annotated gene pairs. We also mapped the co-expressed gene pairs to the annotated gene pairs. The results demonstrated that genes were more likely to be co-expressed in the CML group when compared to the normal group among 18 out of 21 GO terms (Table 3.6). Fisher exact test showed that *Ribonucleoprotein complex*, *Ribosome*, *Cytosolic ribosome*, *Ribosomal subunit*, *Cytosol*, *Cytosolic part*, *Intracellular non-membrane-bounded organelle*, *Intracellular organelle part*, *Cytosolic large ribosomal subunit*, *Cytoplasmic part*, *Cytoplasm*, *Intracellular organelle*, *Nuclear part*, *Nuclear lumen*, *Intracellular organelle lumen* and *Nucleolus* were significantly mapped (p-values < 0.05, and corrected p-values < 0.05).

In these significantly mapped GO terms, five of them were related to ribosome: *Ribonucleoprotein complex*, *Ribosome*, *Cytosolic ribosome*, *Ribosomal subunit* and *Cytosolic large ribosomal subunit*. *Ribonucleoprotein complex* refers to a macromolecular complex consisting of both proteins and RNA molecules. *Ribosome* contains large and small subunits, as well as other proteins and RNAs, which is regarded as a machine for protein biosynthesis. *Cytosolic ribosome* describes a ribosome that is located in the cytosol. *Ribosomal subunit* consists of ribosomal large and small subunits. *Cytosolic large ribosomal subunit* refers to the large subunit that is located in the cytosol. There were more connections in the CML group compared to the normal group (Table 3.6). In addition, most of the co-expressed genes belong to RP genes encoding the ribosomal large and small subunits.

The nucleolus is very important for ribosome biogenesis, containing the proteins for ribosome production (Maggi and Weber, 2005; Maggi *et al*., 2008). A number of nucleoli were found to be centered around rDNAs that are transcribed to rRNAs for ribosome (Maggi *et al*., 2008; Warner, 1990). In addition, various proteins responsible for the processing and assembly of ribosomal large and small subunits are also included in the nucleolus (Maggi *et al*., 2008). We found that genes encoding small nuclear ribonucleoproteins were well connected with other genes in the CML group: small nuclear ribonucleoprotein D2 polypeptide 16.5kDa (SNRPD2), D3 polypeptide 18kDa (SNRPD3), polypeptide E (SNRPE) and polypeptide F (SNRPF) (Figures 3.10 and 3.11). It was reported that NPM1 can shuttle from the nucleus to the cytoplasm (Brady *et al*., 2004). NPM1 was also found to direct the nuclear export of ribosome (Maggi *et al*., 2008). When exported to the cytoplasm, the small and large subunits are combined together to form functional subunits (Maggi *et al*., 2008). In our result, NPM1 was found in both *cytoplasm* and *nucleolus* GO terms for cellular components (Figures 3.10 and 3.11). Most importantly, NPM1 was co-expressed with more genes in the CML group than that in the normal group, including the RP genes RPL10A and RPL36A.

Table 3.5: Cellular component_Enriched GO terms for the functional annotation of NPM1-associated genes

#	Enriched GO terms	Genes found in our data	EASE score	FDR
1	Ribonucleoprotein complex	51	6.40×10^{-49}	7.60×10^{-46}
2	Ribosome	35	1.80×10^{-39}	2.10×10^{-36}
3	Cytosolic ribosome	26	7.90×10^{-37}	9.40×10^{-34}
4	Ribosomal subunit	29	1.50×10^{-36}	1.80×10^{-33}
5	Cytosol	53	5.90×10^{-31}	7.10×10^{-28}
6	Cytosolic part	27	6.80×10^{-31}	8.10×10^{-28}
7	Cytosolic small ribosomal subunit	14	9.10×10^{-20}	1.10×10^{-16}
8	Small ribosomal subunit	15	1.40×10^{-18}	1.60×10^{-15}
9	Large ribosomal subunit	15	3.50×10^{-18}	4.10×10^{-15}
10	Intracellular non-membrane-bounded organelle	54	4.00×10^{-18}	4.80×10^{-15}
11	Intracellular organelle part	65	2.10×10^{-16}	2.70×10^{-13}
12	Cytosolic large ribosomal subunit	12	3.10×10^{-16}	4.00×10^{-13}
13	Cytoplasmic part	67	2.20×10^{-14}	2.60×10^{-11}
14	Cytoplasm	79	5.90×10^{-13}	7.00×10^{-10}
15	Intracellular organelle	82	2.50×10^{-09}	3.00×10^{-06}
16	Nuclear part	33	2.00×10^{-08}	2.30×10^{-05}
17	Nuclear lumen	29	2.90×10^{-08}	3.40×10^{-05}
18	Intracellular organelle lumen	32	4.30×10^{-08}	5.20×10^{-05}
19	Spliceosome	10	1.30×10^{-07}	1.50×10^{-04}
20	Nucleolus	19	2.40×10^{-07}	2.90×10^{-04}
21	Small nuclear ribonucleoprotein complex	5	1.40×10^{-05}	1.60×10^{-02}

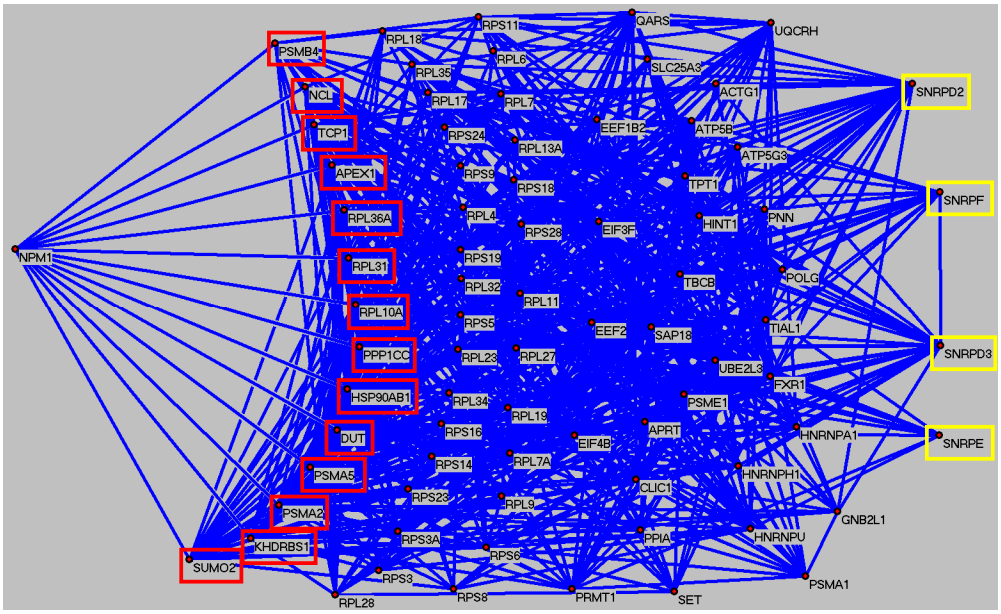
GO: gene ontology. EASE score: a modified Fisher exact test, Expression Analysis Systematic Explorer score. FDR: false discovery rate.

Table 3.6: Mapping co-expressed gene pairs to annotated gene pairs from each GO term for cellular component

GO terms	Fisher exact test					Corrected
	a	*b*	*c*	*d*	p-value	p-value
Ribonucleoprotein complex	271	107	107	271	**<0.001**	**<0.018**
Ribosome	62	9	9	62	**<0.001**	**<0.018**
Cytosolic ribosome	22	4	4	22	**<0.001**	**<0.018**
Ribosomal subunit	26	4	4	26	**<0.001**	**<0.018**
Cytosol	273	127	127	237	**<0.001**	**<0.018**
Cytosolic part	34	8	8	34	**<0.001**	**<0.018**
Large ribosomal subunit	14	10	10	14	0.193	3.474
Intracellular non-membrane-bounded organelle	281	157	157	281	**<0.001**	**<0.018**
Intracellular organelle part	459	265	265	459	**<0.001**	**<0.018**
Cytosolic large ribosomal subunit	10	0	0	10	**<0.001**	**<0.018**
Cytoplasmic part	481	273	273	481	**<0.001**	**<0.018**
Cytoplasm	704	416	416	704	**<0.001**	**<0.018**
Intracellular organelle	819	413	413	819	**<0.001**	**<0.018**
Nuclear part	138	73	73	138	**<0.001**	**<0.018**
Nuclear lumen	103	57	57	103	**<0.001**	**<0.018**
Intracellular organelle lumen	123	80	80	123	**<0.001**	**<0.018**
Spliceosome	11	7	7	11	0.159	2.862
Nucleolus	41	19	19	41	**<0.001**	**<0.018**

GO: gene ontology. GO terms highlighted in bold text are significantly mapped. *a*: mapped CML-specific strongly co-expressed pairs. *b*: mapped CML-specific weakly co-expressed pairs. *c*: mapped normal-specific strongly co-expressed pairs. *d*: mapped normal-specific weakly co-expressed pairs.

a.



b.

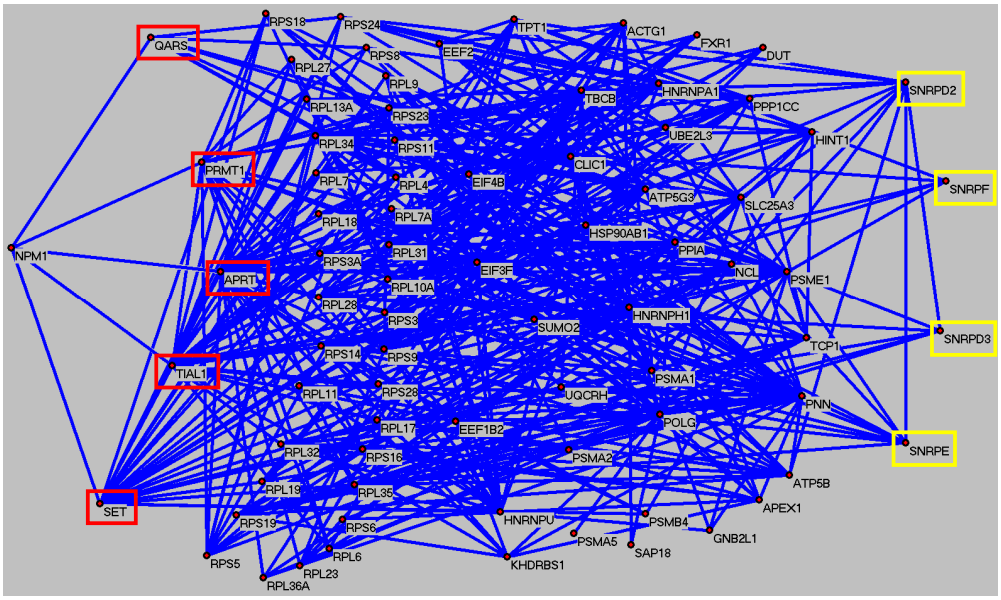
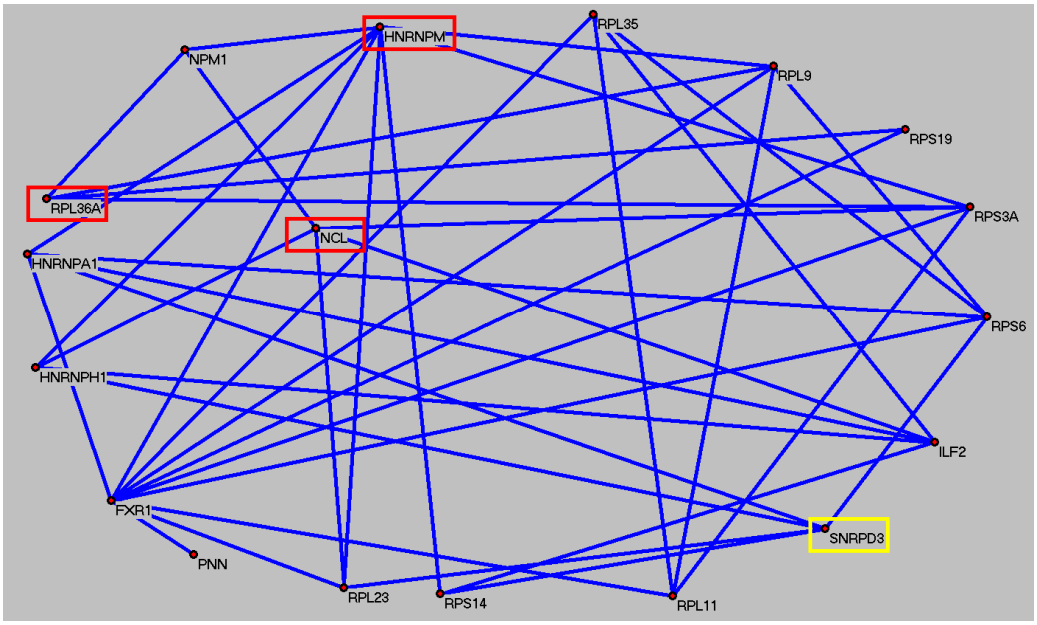


Figure 3.10: Co-expression networks for the mapped strongly co-expressed pairs in the *Cytoplasm* cellular component (using Pajek). Genes with red rectangles are those genes co-expressed with NPM1. Genes with yellow rectangles refer to those genes encoding small nuclear ribonucleoproteins. (a) Mapped CML-specific strongly co-expressed pairs. (b) Mapped normal-specific strongly co-expressed pairs.

a.



b.

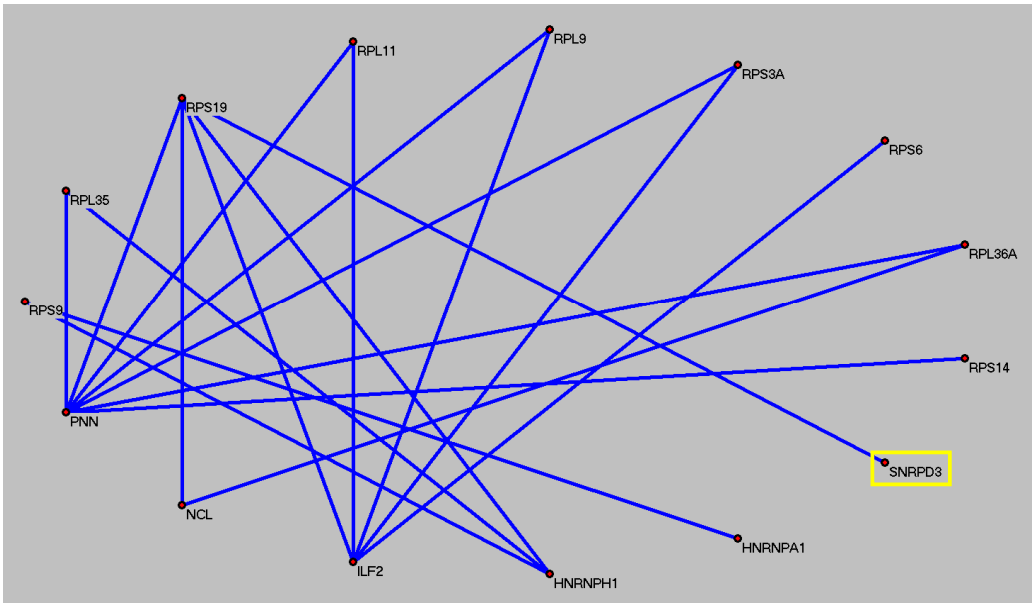


Figure 3.11: Co-expression networks for the mapped strongly co-expressed pairs in the *Nucleolus* cellular component (using Pajek). Genes with red rectangles are those genes co-expressed with NPM1. Genes with yellow rectangles refer to those genes encoding small nuclear ribonucleoproteins. (a) Mapped CML-specific strongly co-expressed pairs. (b) Mapped normal-specific strongly co-expressed pairs.

3.2.4.3 Molecular function

There was no significantly enriched GO term for molecular function identified according to the same selection criteria (EASE score < 0.05 and FDR < 0.05).

3.3 Discussion and conclusion

In this chapter, we have identified the overall differences in the co-expression patterns of those NPM1-associated genes between the normal and the CML groups. Correlation coefficients for all the possible gene pairs among these 93 genes were considered to form two different cumulative distributions. Two-sample KS test was performed to identify the difference (Figure 3.2). Firstly, the maximum deviation (D = 0.109) between two cumulative distributions indicated the difference between the normal and the CML groups structurally. Then, a disease-specific cutoff point (C = 0.252) was discovered at the maximum deviation to classify the co-expressed gene pairs. We further identified the specifically co-expressed gene pairs in the normal and the CML groups (Figures 3.3 and 3.4).

We also discovered that there were more genes co-expressed with NPM1 in the CML group compared to the normal group (Figures 3.5 and 3.6). Most importantly, the RP genes were more likely to be co-expressed in the CML group (Figure 3.7). NPM1 was reported to mediate the nuclear export of ribosomal large and small subunits, and colocalize with the ribosomal subunit proteins in the nucleolus, nucleus and cytoplasm (Maggi *et al*., 2008).

David annotation for enriched biological process gene ontology demonstrated that genes involved in *Translational elongation*, *Translation*, *Cellular protein metabolic process*, *RNA processing* and *RNA metabolic process* were more likely to be co-expressed in the CML group (Table 3.4). From the results, we observed that the first two gene ontology terms were related to translation process. The co-expressed genes participated in these two biological processes covered RP genes (e.g. RPL6 and

RPS28) and translation factors (e.g. EEF2 and EIF3F) (Figures 3.8 and 3.9). The RP genes are responsible for encoding the ribosomal large and small subunits. Ribosome is regarded as a machine for protein biosynthesis. During the translation process, some factors are needed to assist the translation, such as initiation factors and elongation factors. In the significantly mapped GO terms for cellular components, some of them were related to ribosome, nucleolus and cytoplasm (Table 3.6, Figures 3.10 and 3.11).

Altered mRNA translation is involved in the pathogenesis of various human cancers, including CML (Zhang *et al.*, 2008). Ly *et al.* reported that the translational regulators, ribosomal protein S6 and 4E-BP1 (a negative regulator in cap-dependent mRNA translation process), are constitutively phosphorylated in CML cells (Ly *et al.*, 2003). The encoded protein by eukaryotic translation initiation factor 4E (EIF4E) is regarded as both a key translation factor and a promoter for nucleocytoplasmic transport of specific transcripts (Topisirovic *et al.*, 2003). Overexpression of EIF4E has been found in CML patients, suggesting its possible role in neoplastic transformation and the feasibility as a novel therapeutic approach (Hagner *et al.*, 2010; Topisirovic *et al.*, 2003).

The large and small subunits, as well as the rRNAs are generated in the nucleolus. After exported to the cytoplasm, these components are combined together to form functional ribosome to perform the translation function. Therefore, both the biological processes and cellular components are important. We found that genes involved in the translation processes, ribosome, nucleolus and cytoplasm were more likely to be co-expressed in the CML group compared to the normal group. We can

infer that the ribosome biogenesis and translation process may be more active in the CML state.

Chapter 4 E2F1–3 and MYC Target Genes Co-expression Analysis

4.1 Method

4.1.1 Flow chart

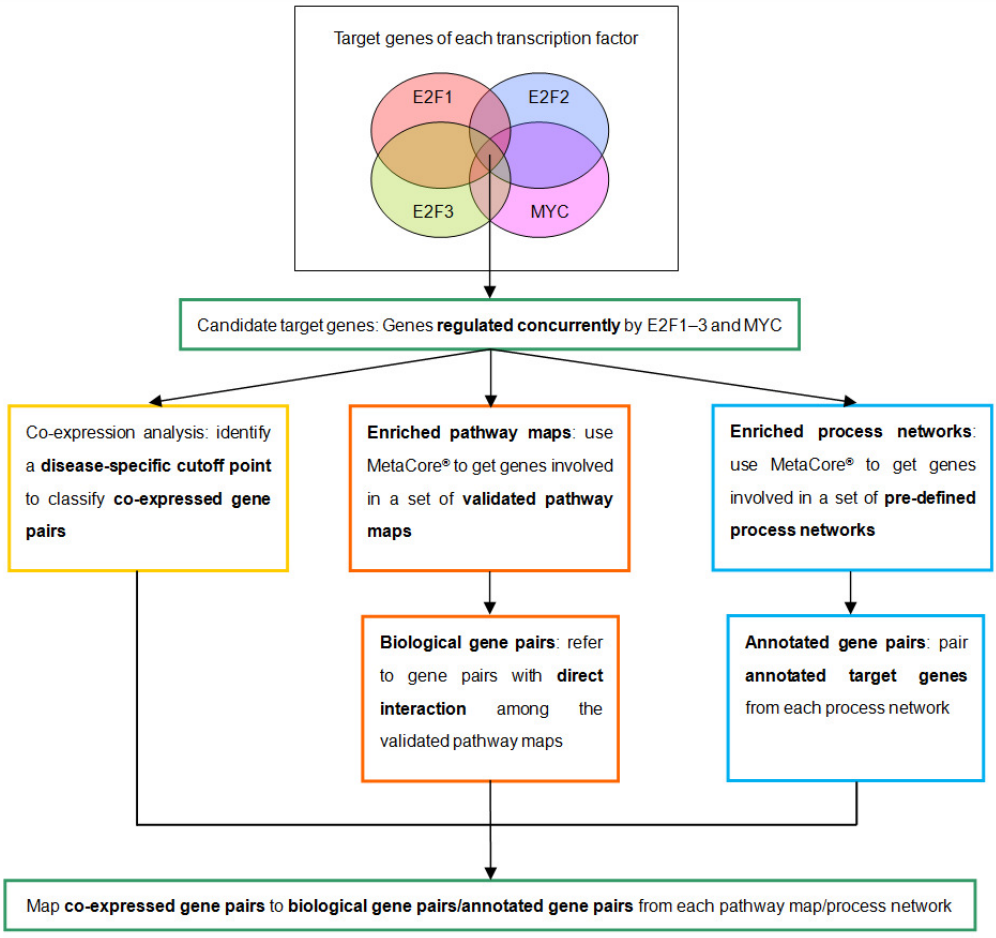


Figure 4.1: Flow chart for the E2F1–3 and MYC target genes co-expression analysis.

4.1.2 Identification of candidate target genes regulated directly and concurrently by E2F1–3 and MYC

The interactions between TFs (E2F1, E2F2, E2F3 and MYC) and their target genes (TGs) were obtained from *prediction of transcriptional regulatory modules* (*PReMod*) database (Ferretti *et al*., 2007). The binding sites of TFs are often clustered together, called cis-regulatory modules (CRMs). *PReMod* database predicts interactions between TFs and their TGs according to the binding affinity and conservation of CRMs. This database contains more than 100,000 computationally predicted transcriptional regulatory modules describing the relationship between TFs and TGs within the human genome (Ferretti *et al*., 2007). These modules describe 229 potential transcription factor families, which are the first genome-wide collection of predicted regulatory modules for the human genome (Blanchette *et al*., 2006). In our study, the set of binding predictions (TF-TG pairs) was called a molecular interaction (MI) set, which was regarded as the reference data. After obtaining the TGs of each TF (E2F1, E2F2, E2F3 and MYC) individually, we identified the common TGs of these four TFs, which were served as the candidate target genes for the following co-expression analysis.

4.1.3 Co-expression analysis for candidate target genes

Using the same approach as the genome-wide analysis, correlation coefficients for all the possible gene pairs of the candidate target genes were calculated (Sections 2.1.2 and 2.1.3). Following the same distribution-based approach for gene pair classification as described above (Section 2.1.4), the gene pairs were also classified into the normal-specific, CML-specific and common pairs.

4.1.4 *MetaCore* functional annotation

MetaCore® from GeneGo Inc. is an integrated software with high-quality to functionally analyze various data, such as microarrays, SAGE, proteomics and other experimental data. This database provides functional context to validate the results for research. The information in this database is curated according to human protein-protein interaction, protein-DNA interaction, transcriptional factors, signaling, metabolism and other aspects. This database is useful for data visualization, mapping and in-silico analysis. We applied *MetaCore*® to obtain the enriched pathway maps and process networks for the functional annotation of candidate target genes.

Enrichment analysis for pathway maps and process networks in *MetaCore*® was performed based on the p-value of hypergeometric intersection between the input data and the target list pre-existed in this database. The lower the p-value is obtained, the higher the relevance of this pathway map/process network to the candidate target genes and the rating of this pathway map/process network are indicated. Only the top 10 significantly enriched pathway maps/process networks are shown according to the sorted p-values in *MetaCore*®.

4.1.4.1 *MetaCore* biological analysis for pathway maps

Firstly, we applied *MetaCore* to map the candidate target genes regulated directly and concurrently by E2F1–3 and MYC to a set of validated pathways. Specifically, when we uploaded the candidate target genes to this database, it mapped these genes to a set of validated signaling and metabolic maps, which are created by Thomson

Reuters scientists on the basis of published peer-reviewed literatures with a high-quality curation process manually. In the next step, we identified gene pairs that had the direct interactions in the top 10 significantly enriched pathways, which were regarded as the biological gene pairs. The identified co-expressed gene pairs were further mapped to the biological gene pairs from each pathway map. In this part, we mainly focused on the strongly co-expressed gene pairs, since these pairs had strong biological relationships.

4.1.4.2 *MetaCore* annotation for process networks

4.1.4.2.1 Functional annotation for candidate target genes

According to the same procedure (Section 4.1.4.1), we uploaded the candidate target genes to this database. This time, it mapped these genes to a set of cellular and molecular process networks, which are defined and annotated by Thomson Reuters scientists. Each process is defined as a pre-set network describing the protein interactions among them. We obtained the annotated target genes in the top 10 significantly enriched process networks. In each process network, the annotated target genes referred to the uploaded candidate target genes that can be found in this process network.

4.1.4.2.2 Mapping co-expressed gene pairs to annotated gene pairs

The annotated target genes in each process network were paired with all the possible combinations to form the annotated gene pairs. The annotated gene pairs from each process network were mapped to the identified co-expressed gene pairs: the normal-specific strongly, normal-specific weakly, CML-specific strongly and CML-specific

weakly co-expressed pairs. Fisher exact test was used to verify if there were more mapped normal-specific strongly co-expressed pairs than mapped CML-specific strongly co-expressed pairs in each process network (Table 4.1). False discovery rates (FDRs) are usually used to control the expected proportion of false positives for the multiple hypotheses. In this study, the FDRs were calculated based on the p-values obtained from Fisher exact test (Storey, 2002). A process network was significantly mapped, if its FDR value was smaller than 0.05 (Fu *et al*., 2012). The FDR values were estimated via the *Matlab* function, *mafdr* (Haskins *et al*., 2011).

Table 4.1: The classification of mapped gene pairs for Fisher exact test

Group	Mapped strongly specific gene pairs	Mapped weakly specific gene pairs
Normal	*a*	*b*
CML	*c*	*d*

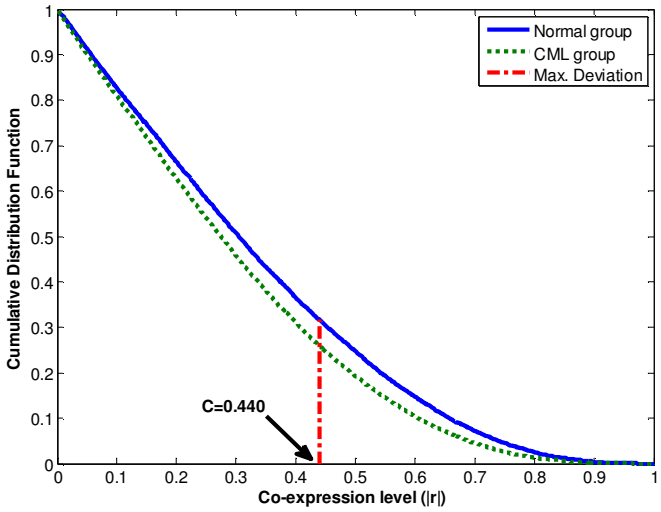
4.2 Results

4.2.1 Identification of structural co-expression difference

Firstly, the TGs of each TF (E2F1, E2F2, E2F3 and MYC) were obtained from MI set in the *PReMod* database individually. In the next step, we identified the common TGs of these four TFs. In order to calculate the correlation coefficients, the TGs should be found in the microarray dataset. In total, we identified 217 common TGs in the microarray dataset GSE5550 (Appendix A2). We further extracted the available expression profiles of these 217 TGs and calculated the correlation coefficients in both the normal and the CML groups. In each group, there was a set of correlation coefficients of 23,436 gene pairs. We then plotted the cumulative distributions for these two sets of data. Two-sample KS test was performed to identify the difference from the overall structure. The results showed that these two distributions between the normal and the CML groups were significantly different (p-value = 2.00×10^{-34} < 0.05 for the maximum deviation $D = 0.0577$) (Figure 4.2).

The disease-specific cutoff point, $C = 0.440$, was identified at the maximum deviation (Figure 4.2). Two co-expression patterns were so distinct that the normal group had more strongly co-expressed (level above ~0.440) gene pairs compared to the CML group (Table 4.2). In other words, these genes were more likely to be co-expressed in the normal group when compared to the CML group. Chi-square test indicated that the proportions of strongly and weakly co-expressed gene pairs significantly differed between the normal and the CML groups (p-value = 2.74×10^{-43} < 0.05 for the statistic $\chi^2 = 190$).

a.



b.

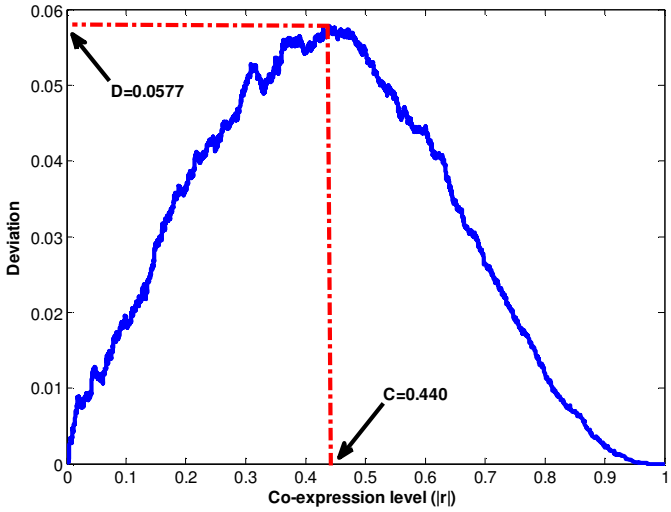


Figure 4.2: Plots of distributions for the co-expression analysis of candidate target genes regulated directly and concurrently by E2F1–3 and MYC. (a) Cumulative distribution functions of co-expression levels in the normal and the CML groups. (b) Deviation distribution against different co-expression cutoff points.

Table 4.2: Cross-tabulation of gene pair counts in the co-expression analysis of candidate target genes regulated directly and concurrently by E2F1–3 and MYC

Group	# of strongly co-expressed gene pairs	# of weakly co-expressed gene pairs
Normal	7,436	16,000
CML	6,083	17,353

4.2.2 Co-expression galaxy and structures for candidate target genes regulated directly and concurrently by E2F1–3 and MYC

The co-expression galaxy was plotted and partitioned into four regions, according to the same procedures with the genome-wide co-expression analysis (Section 2.2.2): i) normal-specific strongly co-expressed pairs (CML-specific weakly co-expressed pairs): the percentage was 23.464%; ii) common strongly co-expressed pairs: the percentage was 8.265%; iii) CML-specific strongly co-expressed pairs (normal-specific weakly co-expressed pairs): the percentage was 17.691%; and iv) common weakly co-expressed pairs: the percentage was 50.580% (Figures 4.3 and 4.4). From the results, we observed that there were more normal-specific strongly co-expressed pairs than CML-specific strongly co-expressed pairs.

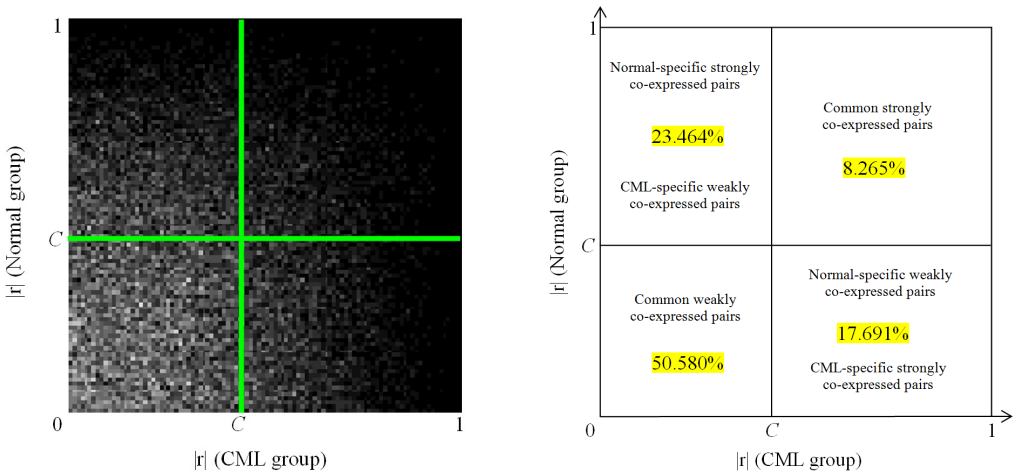


Figure 4.3: Co-expression galaxy (left) and four regions partitioned by the disease-specific cutoff point, C = 0.440 (right). Each correlation coefficient (|r|) is represented by one white dot in the galaxy. More dots mean that there are more correlation coefficients located in that region.

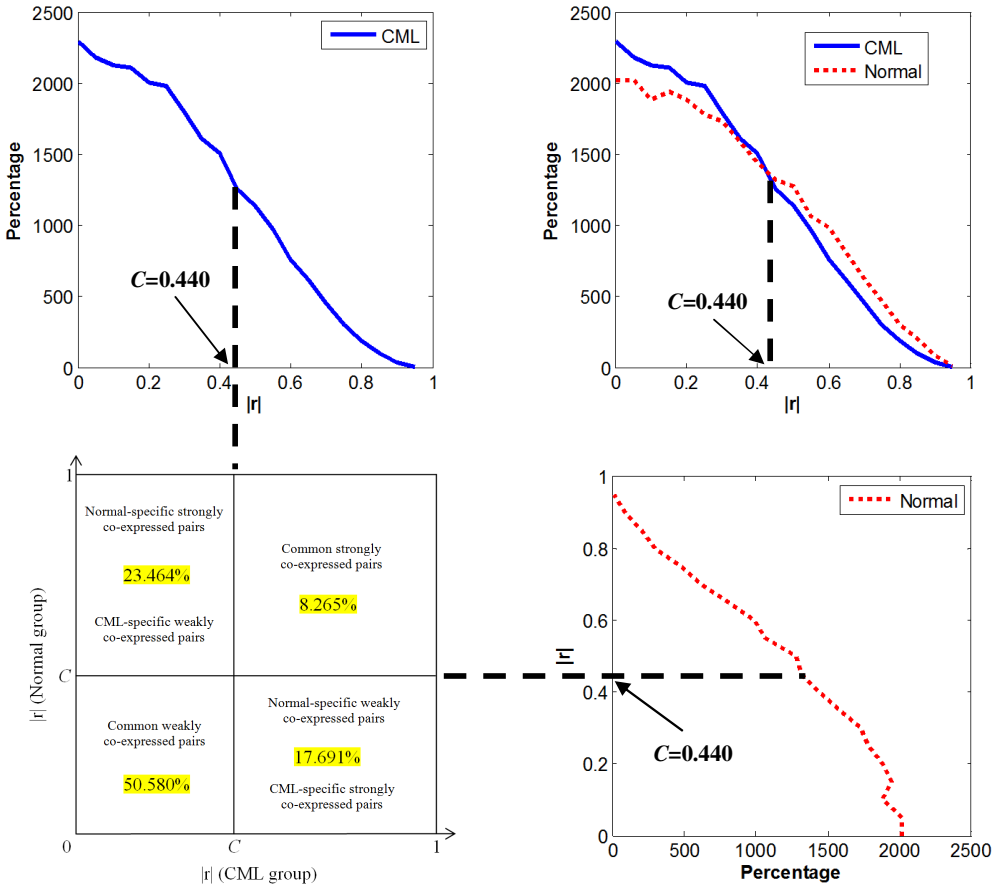


Figure 4.4: Relationship between co-expression structures and partitioned four regions in the co-expression galaxy. The red dot line represents the distribution of percentage for correlation coefficients (|r| values) in the normal group. The blue solid line stands for the distribution of percentage for correlation coefficients (|r| values) in the CML group.

4.2.3 *MetaCore* analysis for functional annotation

4.2.3.1 Enriched pathway maps

We applied *MetaCore* to map the strongly and specifically co-expressed gene pairs to the validated pathways. The top 10 significantly enriched pathway maps are shown in Table 4.3. Here, we mainly focused on the direct interactions among the strongly and specifically co-expressed gene pairs found in the enriched pathways. The important findings are shown in Table 4.4 and Figure 4.5. These strongly and specifically co-expressed pairs represented the characteristics of the normal state and the CML state, respectively.

The first two matched pathways were "*Development_Role of Activin A in cell differentiation and proliferation*" and "*Signal transduction_Activin A signaling regulation*". One normal-specific strongly co-expressed pair was identified as the direct interaction in these two pathways: INHBA (Inhibin, beta A) connected with ACVR2A (Activin A receptor, type IIA) ($|r| = 0.593$) (Figure 4.5a and b). The inhibin beta A subunit forms a homodimer, Activin A. It was reported that Activin A binding to the type II serine kinase receptor results in the recruitment, phosphorylation and subsequent activation of the type I receptor (Attisano *et al*., 1996; Deli *et al*., 2008), finally inhibiting cell growth and proliferation and triggering apoptosis. Figure 4.5a and b show that the binding of Activin A to Activin A receptor, type IIA, activates Activin receptor type-1B (ALK-4). The activated ALK-4 phosphorylates mothers against decapentaplegic homolog 2 (SMAD2) and 3 (SMAD3). SMAD4 then binds to the phosphorylated complex. The whole complex is translocated into the nucleus to regulate the transcription of various genes (Abe *et al*., 2004). It can inhibit the transcription of c-myc (Matsuo *et al*., 2006), and

stimulate transcription of cell cycle inhibitors p15 and p21 (Burdette *et al*., 2005; Panopoulou *et al*., 2005). The Activin A signaling pathway has been found to be deregulated in some kinds of primary bone tumors, including osteosarcoma, chondrosarcomas and osteochondromas (Leto, 2010).

In "*Cell adhesion_Ephrin signaling*", one normal-specific link was found: EFNA5 (Ephrin-A5) connected with EPHA4 (EPH receptor A4) ($|r| = 0.720$) (Figure 4.5c). Ephrin-A receptors are the largest subfamily of receptor tyrosine kinases that regulate cell shape, mobility and attachment (Himanen *et al*., 2007). The interaction between Ephrin-A receptors and ligands plays a vital role in cell-cell communication, which initiates the unique bi-directional signaling cascades to transduce the information in both the receptor- and ligand-expressing cells (Himanen and Nikolov, 2003).

In "*Development_Degradation of beta-catenin in the absence WNT signaling*" and "*Transcription_Role of heterochromatin protein 1 (HP1) family in transcriptional silencing*", a normal-specific connection from CTBP2 (C-terminal binding protein 2) to TCF4 (Transcription T cell factor 4) was found with $|r| = 0.537$ (Figure 4.5d and e). β-catenin is a key factor in the WNT signaling pathway, which plays an important role in the regulation of cell growth and differentiation (Huang *et al*., 2010). Hyperactivation of β-catenin is regarded as a common cause of carcinoma (Huang *et al*., 2010). In the presence of WNT signal, transcription T cell factor (TCF) interacts with β-catenin to activate transcription of various target genes involved in cellular proliferation, differentiation, survival and apoptosis, such as Cyclin D1, Cyclin D2 and c-myc (He *et al*., 1998; Huang *et al*., 2010; Tetsu and McCormick, 1999;

Valenta *et al*., 2003). While, in the absence of the WNT signal, TCF functions as a transcriptional repressor by recruiting its co-repressors histone deacetylase 1 (HDAC1), C-terminal binding protein (CTBP) and Groucho/transducin-like enhancer of Split (TLE) to silence the expression of target genes (Huang *et al.*, 2010; Valenta *et al.*, 2003). In our study, the two mapped pathways were relating to transcriptional gene silencing.

Importantly, a CML-specific connection from PRKACB (Protein kinase, cAMP-dependent, catalytic, beta) to PPP2R3A (Protein phosphatase 2, regulatory subunit B", alpha) was found in "*Signal transduction_PKA signaling*" with $|r| = 0.585$ (Figure 4.5f). The phosphorylation of PP2A regulatory subunit (PPP2R3A) by PKA-cat (PRKACB) is one step of HTR1A signaling pathway, leading to cell survival. Figure 4.5f shows that PKA-cat is the hub connecting other proteins involved in this pathway. In the other enriched pathways, there was no direct interaction found from the strongly and specifically co-expressed gene pairs.

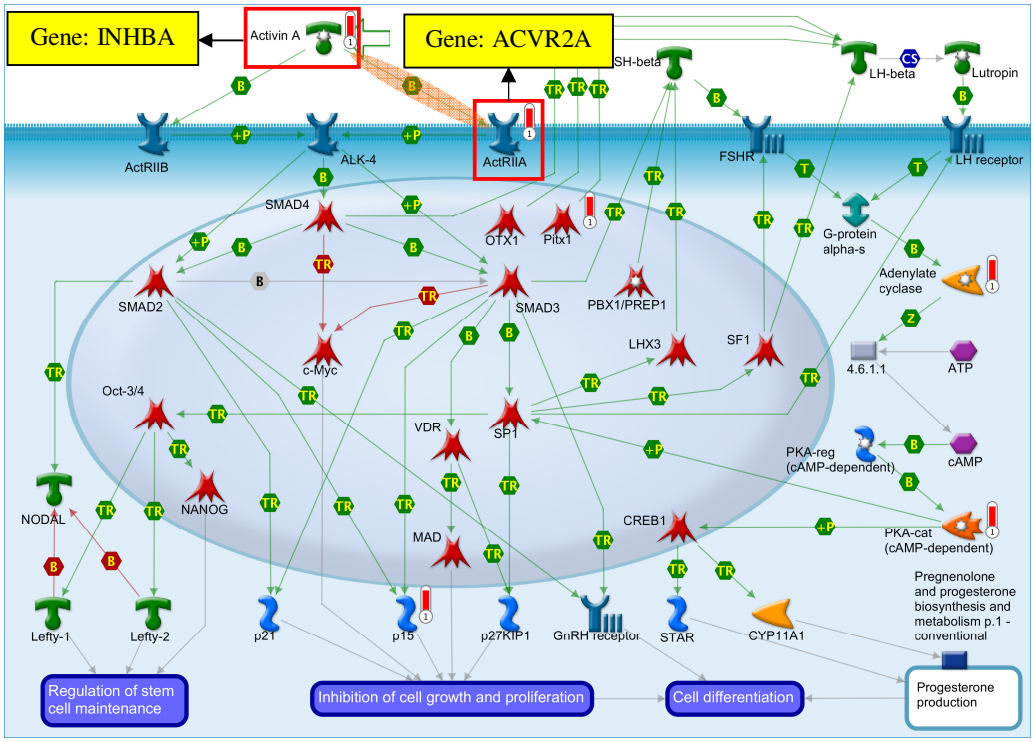
Table 4.3: The top 10 enriched pathway maps from *MetaCore*

#	Enriched pathway maps	Total genes	In our data	p-value
1	Development_Role of Activin A in cell differentiation and proliferation	40	6	6.36×10^{-06}
2	Signal transduction_Activin A signaling regulation	33	4	5.70×10^{-04}
3	Development_Regulation of epithelial-to-mesenchymal transition (EMT)	64	5	9.01×10^{-04}
4	Development_Degradation of beta-catenin in the absence WNT signaling	19	3	1.34×10^{-03}
5	Neurophysiological process_Melatonin signaling	43	4	1.57×10^{-03}
6	Cell adhesion_Ephrin signaling	45	4	1.86×10^{-03}
7	Transcription_Role of heterochromatin protein 1 (HP1) family in transcriptional silencing	22	3	2.07×10^{-03}
8	Development_Melanocyte development and pigmentation	49	4	2.56×10^{-03}
9	Signal transduction_PKA signaling	51	4	2.96×10^{-03}
10	Neurophysiological process_Dopamine D2 receptor transactivation of PDGFR in CNS	26	3	3.38×10^{-03}

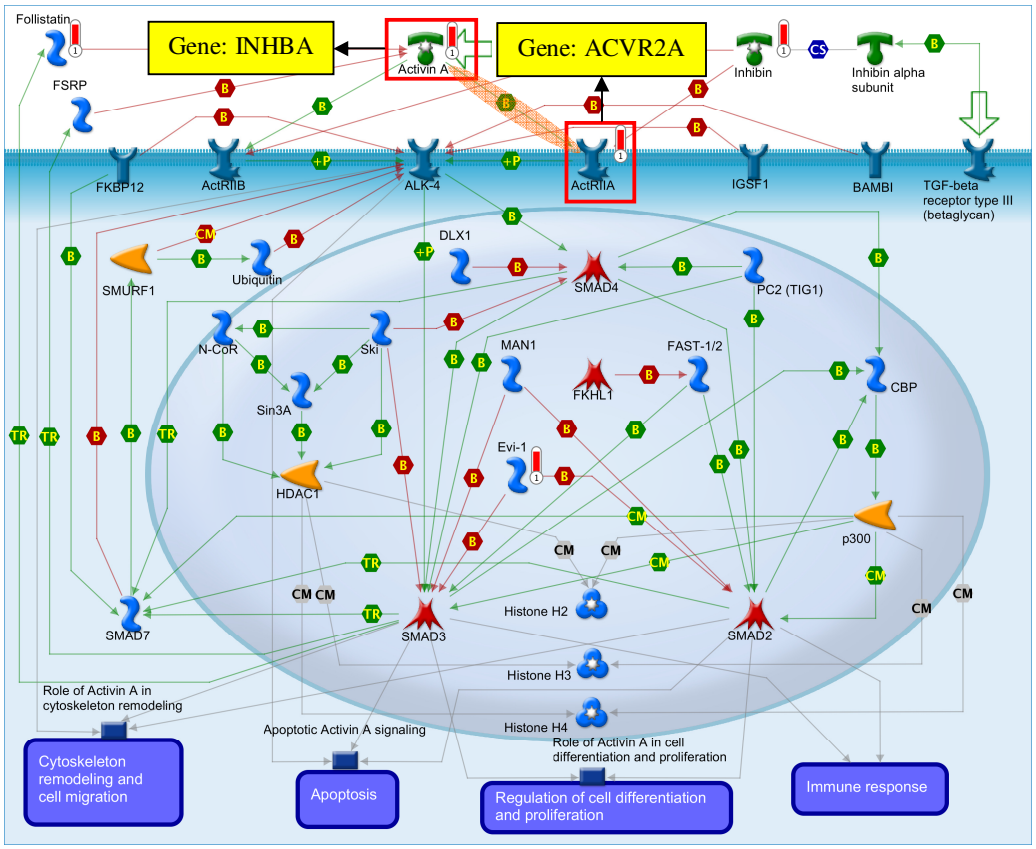
Table 4.4: Strongly and specifically co-expressed gene pairs mapped to the pathways from *MetaCore*

Group	Links found	|r|	Pathway name	Localization and importance
Normal	INHBA to ACVR2A	0.593	i) Development_Role of Activin A in cell differentiation and proliferation; ii) Signal transduction_ Activin A signaling regulation	The first step in these two pathways to inhibit cell growth and proliferation, and trigger apoptosis
	EFNA5 to EPHA4	0.720	Cell adhesion_Ephrin signaling	The key step in cell-cell communication
	CTBP2 to TCF4	0.537	i) Development_Degradation of beta-catenin in the absence WNT signaling; ii) Transcription_Role of heterochromatin protein 1 (HP1) family in transcriptional silencing	CTBP functions as a co-suppressor by binding to TCF to silence the expression of various genes involved in cellular proliferation and apoptosis
CML	PRKACB to PPP2R3A	0.585	Signal transduction_PKA signaling	This connection is one step of HTR1A signaling pathway, leading to cell survival; PKA-cat (Gene: PRKACB) is the hub connecting other proteins involved in this pathway

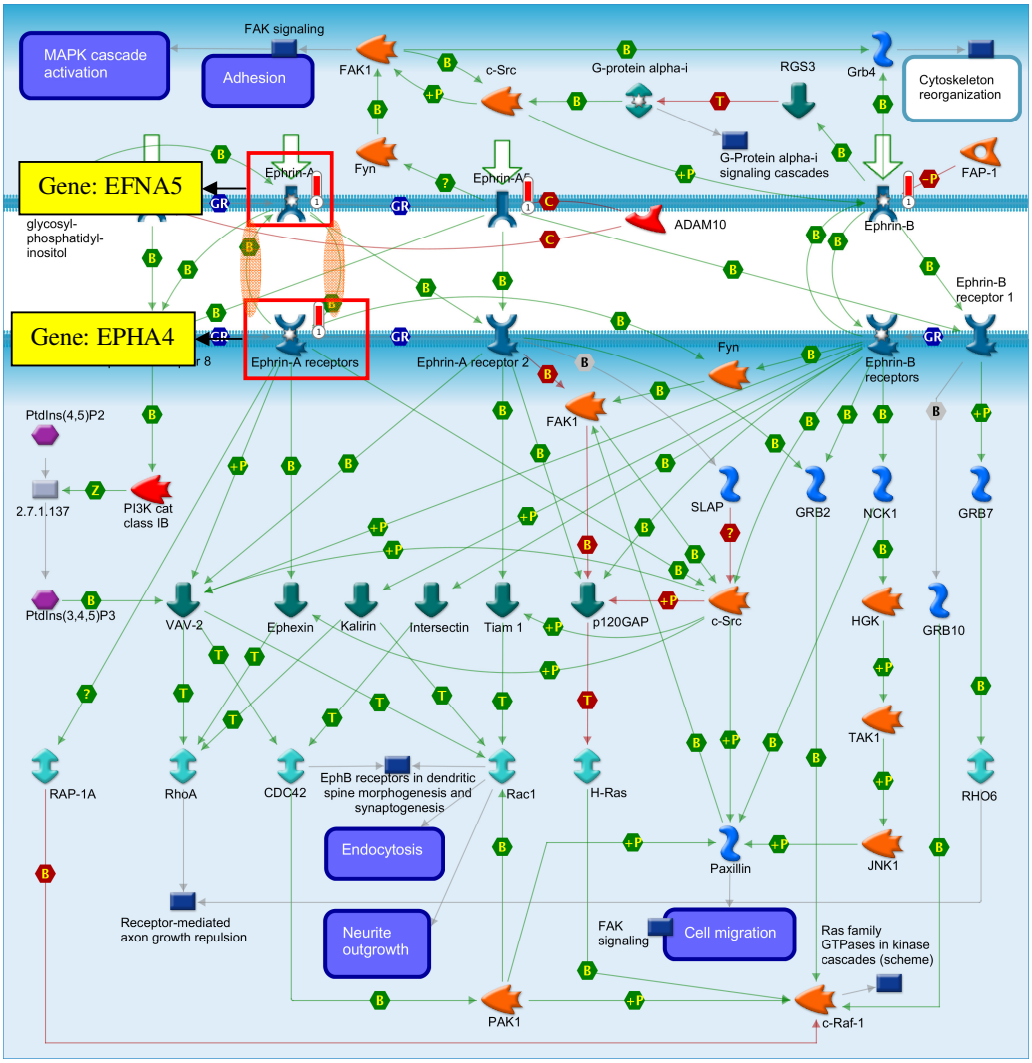
a.


Gene: INHBA
Activin A
Gene: ACVR2A
SH-beta
LH-beta
Lutropin
ActRIIB
ALK-4
ActRIIA
FSHR
LH receptor
SMAD4
OTX1
Pitx1
G-protein alpha-s
Adenylate cyclase
SMAD2
SMAD3
PBX1/PREP1
LHX3
SF1
4.6.1.1
ATP
c-Myc
Oct-3/4
VDR
SP1
PKA-reg (cAMP-dependent)
cAMP
NODAL
NANOG
MAD
CREB1
PKA-cat (cAMP-dependent)
Pregnenolone and progesterone biosynthesis and metabolism p.1 - conventional
Lefty-1
Lefty-2
p21
p15
p27KIP1
GnRH receptor
STAR
CYP11A1
Regulation of stem cell maintenance
Inhibition of cell growth and proliferation
Cell differentiation
Progesterone production

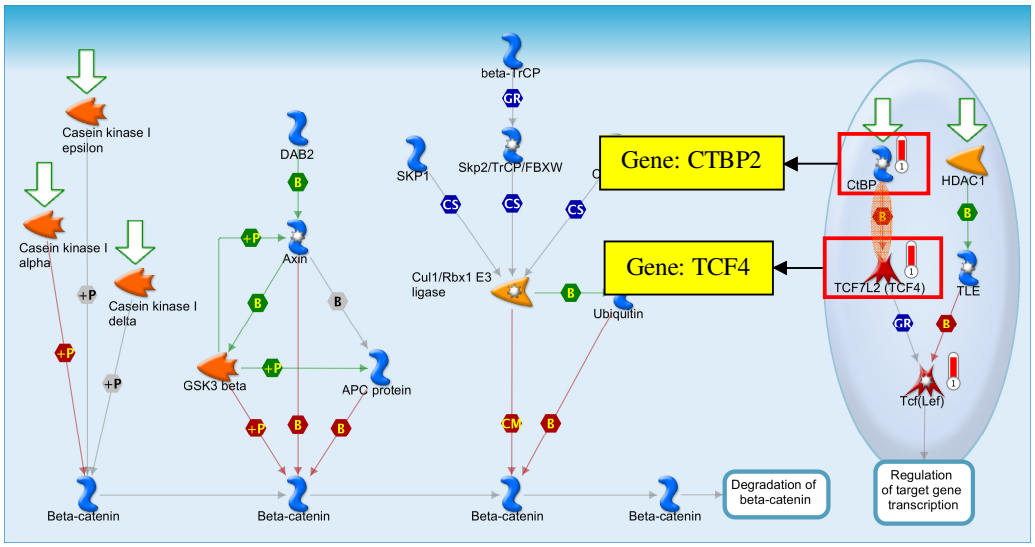
b.


Follistatin
Gene: INHBA
Activin A
Gene: ACVR2A
Inhibin
Inhibin alpha subunit
FSRP
FKBP12
ActRIIB
ALK-4
ActRIIA
IGSF1
BAMBI
TGF-beta receptor type III (betaglycan)
SMURF1
Ubiquitin
DLX1
SMAD4
PC2 (TIG1)
N-CoR
Ski
Sin3A
MAN1
FKHL1
FAST-1/2
CBP
HDAC1
Evi-1
p300
SMAD7
SMAD3
SMAD2
Histone H2
Histone H3
Histone H4
Role of Activin A in cytoskeleton remodeling
Apoptotic Activin A signaling
Role of Activin A in cell differentiation and proliferation
Cytoskeleton remodeling and cell migration
Apoptosis
Regulation of cell differentiation and proliferation
Immune response

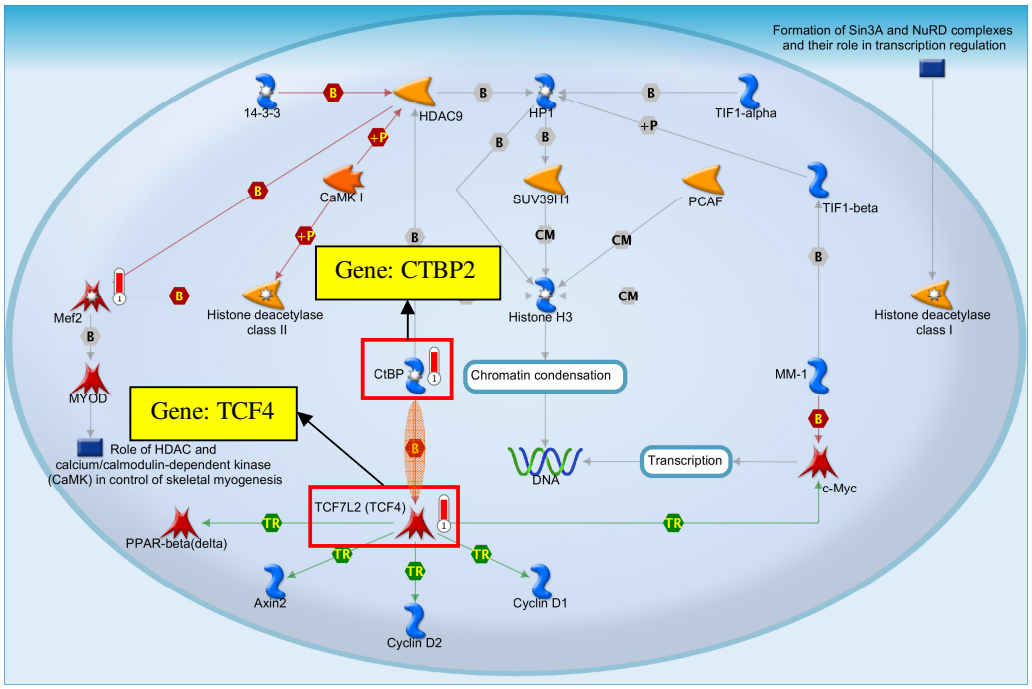
c.


MAPK cascade activation
FAK signaling
Adhesion
FAK1
c-Src
G-protein alpha-i
RGS3
Grb4
Cytoskeleton reorganization
Gene: EFNA5
Ephrin-A
Fyn
Ephrin-A5
G-Protein alpha-i signaling cascades
FAP-1
glycosyl-phosphatidyl-inositol
ADAM10
Ephrin-B
Gene: EPHA4
Ephrin-B receptor 1
Ephrin-A receptors
Ephrin-A receptor 2
Fyn
Ephrin-B receptors
FAK1
PtdIns(4,5)P2
2.7.1.137
PI3K cat class IB
SLAP
GRB2
NCK1
GRB7
PtdIns(3,4,5)P3
VAV-2
Ephexin
Kalirin
Intersectin
Tiam 1
p120GAP
c-Src
HGK
GRB10
TAK1
RAP-1A
RhoA
CDC42
EphB receptors in dendritic spine morphogenesis and synaptogenesis
Rac1
H-Ras
RHO6
Endocytosis
Paxillin
JNK1
Receptor-mediated axon growth repulsion
Neurite outgrowth
FAK signaling
Cell migration
Ras family GTPases in kinase cascades (scheme)
PAK1
c-Raf-1

d.


Gene: CTBP2
Gene: TCF4
Casein kinase I epsilon
Casein kinase I alpha
Casein kinase I delta
DAB2
Axin
GSK3 beta
APC protein
beta-TrCP
SKP1
Skp2/TrCP/FBXW
Cul1/Rbx1 E3 ligase
Ubiquitin
CtBP
HDAC1
TCF7L2 (TCF4)
TLE
Tcf(Lef)
Beta-catenin
Beta-catenin
Beta-catenin
Beta-catenin
Degradation of beta-catenin
Regulation of target gene transcription

e.


Formation of Sin3A and NuRD complexes and their role in transcription regulation
Gene: CTBP2
Gene: TCF4
14-3-3
HDAC9
HP1
TIF1-alpha
CaMK I
SUV39H1
PCAF
TIF1-beta
Mef2
Histone deacetylase class II
Histone H3
Histone deacetylase class I
CtBP
Chromatin condensation
MM-1
MYOD
Role of HDAC and calcium/calmodulin-dependent kinase (CaMK) in control of skeletal myogenesis
DNA
Transcription
c-Myc
TCF7L2 (TCF4)
PPAR-beta(delta)
Axin2
Cyclin D1
Cyclin D2

f.

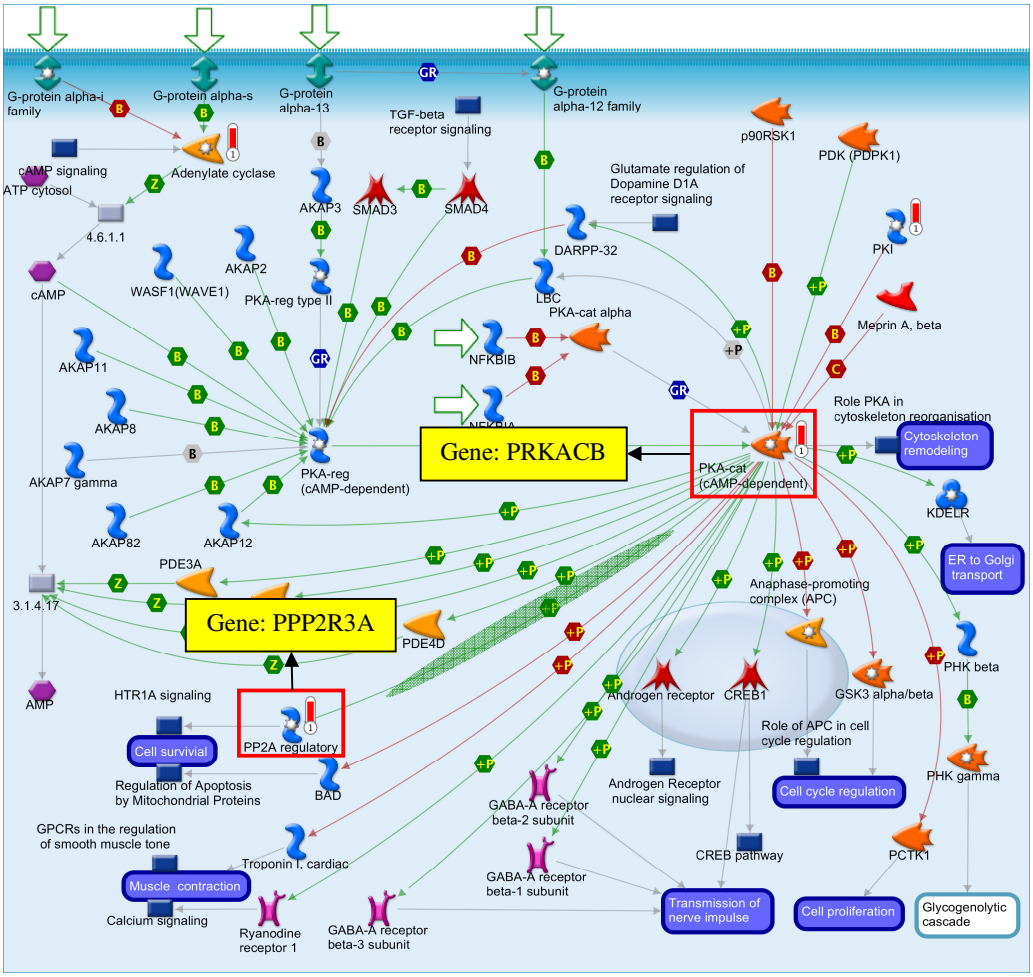


Figure 4.5: Mapped pathways from *MetaCore*. The red bar ( ) means that this gene from 217 target genes is found in the pathway. B indicates that the interaction is binding. P indicates that the interaction is phosphorylation. The red rectangles mean that these genes have direct interactions and are found in the strongly co-expressed gene pairs. (a) *Development_Role of Activin A in cell differentiation and proliferation*. (b) *Signal transduction_Activin A signaling regulation*. (c) *Cell adhesion_Ephrin signaling*. (d) *Development_Degradation of beta-catenin in the absence WNT signaling*. (e) *Transcription_Role of heterochromatin protein 1 (HP1) family in transcriptional silencing*. (f) *Signal transduction_PKA signaling*.

4.2.3.2 Enriched process networks

The top 10 significantly enriched process networks for the functional annotation of 217 candidate target genes are shown in Table 4.5. All the p-values for hypergeometric intersection test were smaller than 0.05. We got the annotated target genes involved in each process network. Next, the annotated target genes were paired to form the annotated gene pairs. The annotated gene pairs were further mapped to the identified co-expressed gene pairs. Table 4.6 shows that 8 out of 10 process networks had more mapped normal-specific strongly co-expressed pairs than mapped CML-specific strongly co-expressed pairs. Fisher exact test showed that "*Cell adhesion_Attractive and repulsive receptors*" and "*Development_Regulation of angiogenesis*" process networks were significantly mapped (p-values = 0.001 and 0.012, < 0.05, and FDR values were 0.004 and 0.026, < 0.05). The conclusion was that the candidate target genes related to these two process networks were more likely to be co-expressed in the normal group compared to the CML group.

In order to explore what genes accounted for this co-expression difference between the normal and the CML groups, we further plotted the co-expression networks for the mapped strongly co-expressed pairs among these two significantly mapped process networks (Figures 4.6 and 4.7). Figure 4.7a and b show the normal-specific strongly co-expressed pairs in these two process networks, both of which had ephrin-B2 (EFNB2), ephrin-A5 (EFNA5) and EPH receptor A4 (EPHA4). We obtained the basic information for these genes/proteins from *NCBI* database. EFNB2 and EFNA5 are two members of the ephrin gene family. The protein product from EPHA4 is an ephrin receptor. The ephrins (EPH) and EPH-related receptors belong to the largest subfamily of receptor protein-tyrosine kinases, which are very important to mediate

developmental events. Figure 4.7a and b demonstrate that the connection from EFNA5 to EPHA4 was identified as a strongly co-expressed gene pair for these two process networks in the normal group. Moreover, proteins encoded by neuropilin 2 (NRP2), transforming growth factor, beta receptor II (TGFBR2) and somatostatin receptor 2 (SSTR2) also belong to receptors, which play an important role in signal transduction process. In addition, the encoded protein from integrin, alpha 2 (ITGA2) is involved in the leukocyte intercellular adhesion process. Three enzymes were also found in these two co-expression networks: i) the protein product from protein kinase, cAMP-dependent, catalytic, beta (PRKACB) is a protein kinase; ii) the protein encoded by prolyl endopeptidase (PREP) is a protease, whose function is to cleave peptide bonds on the C-terminal domain of prolyl residues within peptides that have up to nearly 30 amino acids long; and iii) the protein product from HIV-1 Tat interactive protein 2 (HTATIP2) is an oxidoreductase required for tumor suppression.

There was no CML-specific strongly co-expressed pair found in the "*Cell adhesion_Attractive and repulsive receptors*" process network. Figure 4.7c shows the CML-specific strongly co-expressed pairs identified in the "*Development_Regulation of angiogenesis*" process network. Comparing with the CML group, there were more connections in the normal group. That is to say, the involved genes were more likely to be co-expressed in the normal group. We can infer that these genes/proteins were well connected with each other to transduce signals and maintain physiological balance in healthy individuals. However, in the CML group most of the connections among these molecules were impaired in these two process networks.

Table 4.5: The top 10 enriched process networks from *MetaCore*

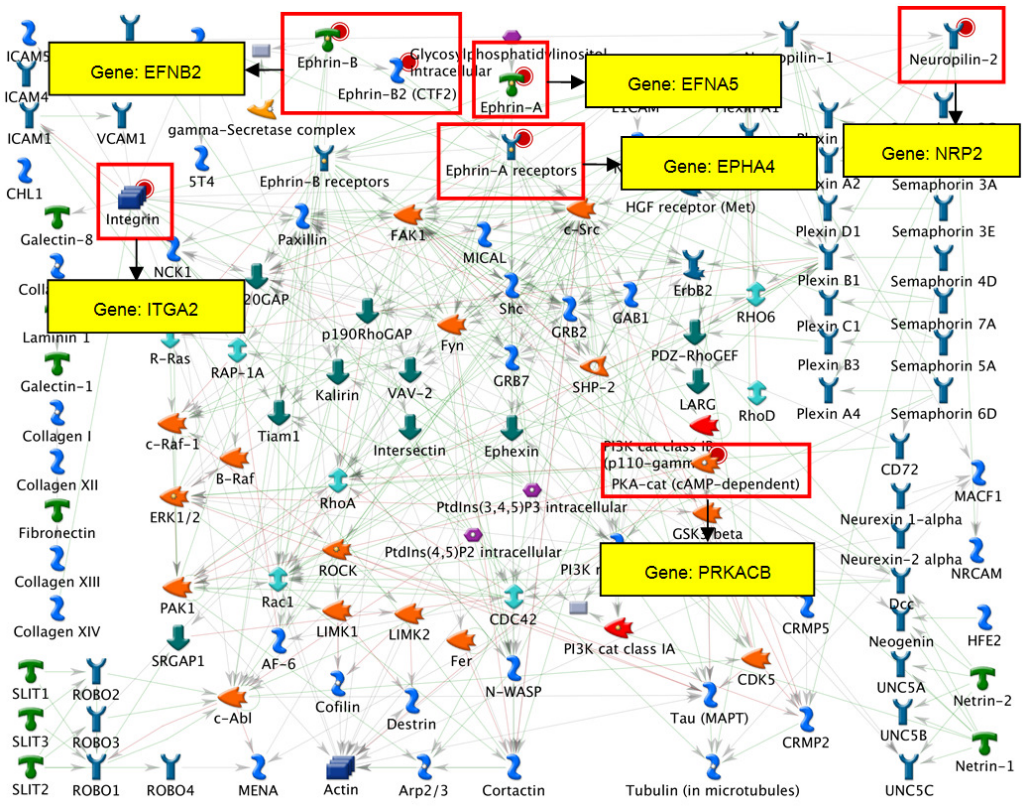
#	Enriched process networks	Total genes	In our data	p-value
1	Cell adhesion_Synaptic contact	184	13	1.47×10^{-4}
2	Development_Neurogenesis in general	192	12	8.06×10^{-4}
3	Development_Hedgehog signaling	254	14	1.04×10^{-3}
4	Signal transduction_WNT signaling	177	11	1.39×10^{-3}
5	Signal Transduction_TGF-beta, GDF and Activin signaling	154	10	1.65×10^{-3}
6	Cell adhesion_Attractive and repulsive receptors	175	10	4.20×10^{-3}
7	Reproduction_FSH-beta signaling pathway	160	9	7.21×10^{-3}
8	Development_Regulation of angiogenesis	223	11	8.17×10^{-3}
9	Cardiac development_BMP_TGF_beta_signaling	117	7	1.25×10^{-2}
10	Neurophysiological process_Melatonin signaling	43	4	1.30×10^{-2}

Table 4.6: Mapping co-expressed gene pairs to annotated gene pairs from each process network

Process networks	Fisher exact test					FDR
	a	*b*	*c*	*d*	p-value	
Development_Neurogenesis in general	14	11	11	14	0.286	0.251
Development_Hedgehog signaling	22	15	15	22	0.081	0.118
Signal transduction_WNT signaling	10	7	7	10	0.247	0.270
Signal Transduction_TGF-beta, GDF and Activin signaling	6	5	5	6	0.500	0.365
Cell adhesion_Attractive and repulsive receptors	6	0	0	6	**0.001**	**0.004**
Development_Regulation of angiogenesis	8	2	2	8	**0.012**	**0.026**
Cardiac development_BMP_TGF_beta_signaling	2	1	1	2	0.500	0.313
Neurophysiological process_Melatonin signaling	3	2	2	3	0.500	0.274

Process networks highlighted in bold text are significantly mapped. FDR: false discovery fate. *a*: mapped normal-specific strongly co-expressed pairs. *b*: mapped normal-specific weakly co-expressed pairs. *c*: mapped CML-specific strongly co-expressed pairs. *d*: mapped CML-specific weakly co-expressed pairs.

a.


Gene: EFNB2
Gene: EFNA5
Gene: EPHA4
Gene: NRP2
Gene: ITGA2
Gene: PRKACB
Ephrin-B
Ephrin-B2 (CTF2)
Ephrin-A
Ephrin-A receptors
Ephrin-B receptors
gamma-Secretase complex
Integrin
Paxillin
FAK1
c-Src
HGF receptor (Met)
Neuropilin-2
Semaphorin 3A
Semaphorin 3E
Semaphorin 4D
Semaphorin 7A
Semaphorin 5A
Semaphorin 6D
Plexin D1
Plexin B1
Plexin C1
Plexin B3
Plexin A4
ICAM4
ICAM1
VCAM1
CHL1
Galectin-8
Laminin 1
Galectin-1
Collagen I
Collagen XII
Fibronectin
Collagen XIII
Collagen XIV
5T4
NCK1
MICAL
Shc
GRB2
GAB1
ErbB2
RHO6
PDZ-RhoGEF
LARG
RhoD
SHP-2
GRB7
Fyn
p190RhoGAP
20GAP
R-Ras
RAP-1A
Kalirin
VAV-2
Tiam1
Intersectin
Ephexin
c-Raf-1
B-Raf
ERK1/2
RhoA
ROCK
PAK1
Rac1
SRGAP1
LIMK1
LIMK2
AF-6
CDC42
Fer
N-WASP
Cofilin
Destrin
c-Abl
PtdIns(3,4,5)P3 intracellular
PtdIns(4,5)P2 intracellular
PKA-cat (cAMP-dependent)
GSK3beta
PI3K cat class IA
CRMP5
CDK5
Tau (MAPT)
CRMP2
Tubulin (in microtubules)
CD72
MACF1
Neurexin 1-alpha
Neurexin-2 alpha
NRCAM
Dcc
Neogenin
HFE2
UNC5A
Netrin-2
UNC5B
Netrin-1
UNC5C
SLIT1
SLIT3
SLIT2
ROBO2
ROBO3
ROBO1
ROBO4
MENA
Actin
Arp2/3
Cortactin

b.

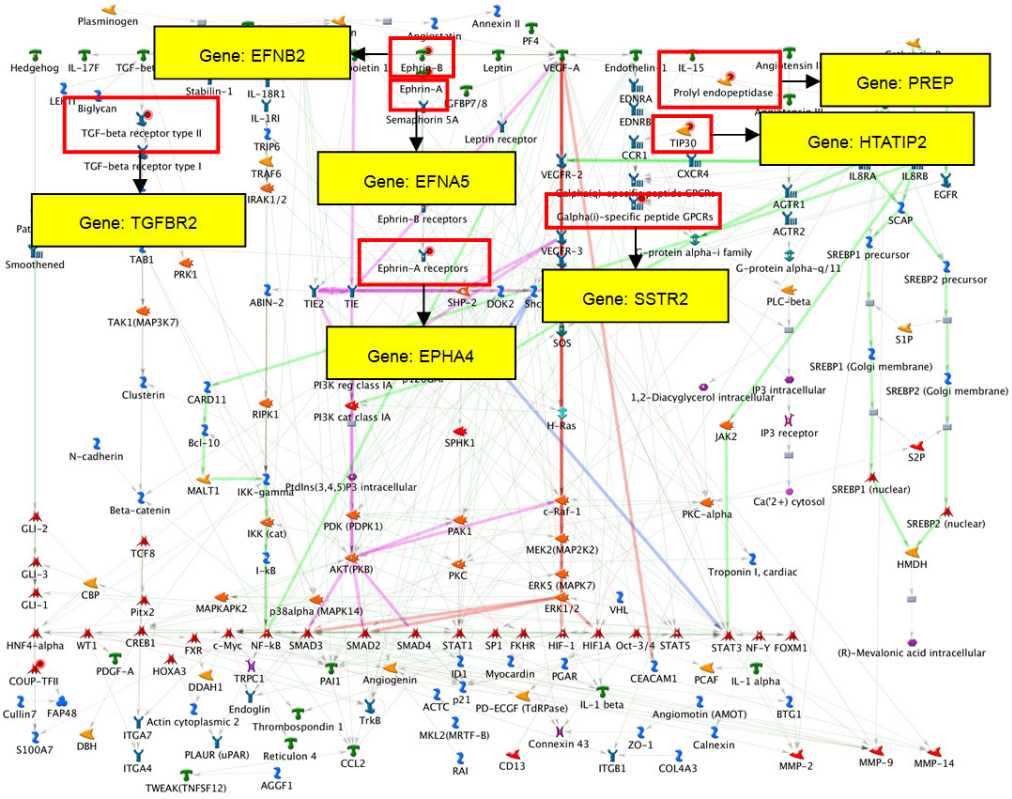


Figure 4.6: Functional annotation for candidate target genes in *MetaCore* process networks. The red rectangles mean that these genes are found in the mapped normal-specific strongly co-expressed pairs. (a) "*Cell adhesion_Attractive and repulsive receptors*" process network. (b) "*Development_Regulation of angiogenesis*" process network.

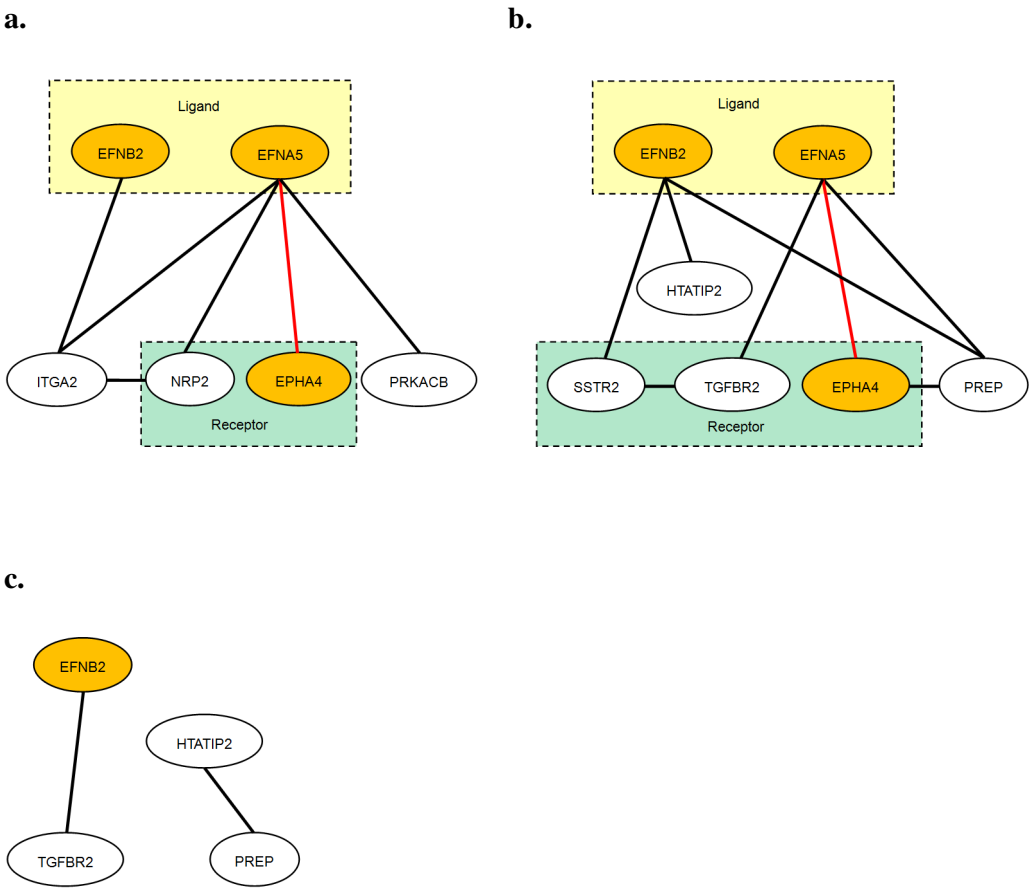


Figure 4.7: Co-expression networks for the mapped strongly and specifically co-expressed pairs. The yellow ellipses are those genes found in both process networks. (a) Mapped normal-specific strongly co-expressed pairs in "*Cell adhesion_Attractive and repulsive receptors*" process network. (b) Mapped normal-specific strongly co-expressed pairs in "*Development_Regulation of angiogenesis*" process network. (c) Mapped CML-specific strongly co-expressed pairs in "*Development_Regulation of angiogenesis*" process network.

4.3 Discussion and conclusion

In this chapter, we have identified the statistical differences in the co-expression patterns of those candidate target genes regulated directly and concurrently by E2F1–3 and MYC between the normal and the CML groups from the overall structure. Two-sample KS test was performed in this study. Firstly, the maximum deviation (D = 0.0577) between two cumulative distributions was identified to indicate the difference between the normal and the CML groups structurally (Figure 4.2). Then, a disease-specific cutoff point (C = 0.440) for correlation coefficients was identified at the maximum deviation to classify the co-expressed gene pairs (Figure 4.2 and Table 4.2). There were more strongly co-expressed gene pairs in the normal group than that in the CML group (Figure 4.2 and Table 4.2). The co-expression galaxy and structures were analyzed to identify the specifically co-expressed gene pairs in the normal and the CML groups (Figure 4.3), in order to further investigate the alterations of biological properties.

MetaCore functional annotation for enriched pathway maps showed that these two links (INHBA to ACVR2A and CTBP2 to TCF4) in the related pathways can be found only in the normal group (Figure 4.5a, b, d and e), the dysregulation of these two links in the CML group is a possible reason leading to cell proliferation and inhibition of apoptosis. Moreover, the connection from EFNA5 to EPHA4 is important in cell-cell communication and adhesion, which also can be found in the normal group (Figure 4.5c), not in the CML group. This is a possible reason leading to alterations in adhesion properties of leukemic progenitors in CML. Most importantly, one link from PRKACB to PPP2R3A can be found only in the CML group. This link is one step of HTR1A signaling pathway, leading to cell survival

(Figure 4.5f). Although only one co-expressed gene pair was found in each signaling pathway, all these identified gene pairs located in the important steps in the whole pathways (Table 4.4): i) Activin A (INHBA) binding to the type II serine kinase receptor (ACVR2A) is the first step in the "*Development_Role of Activin A in cell differentiation and proliferation*" and "*Signal transduction_Activin A signaling regulation*"; ii) Ephrin-A (EFNA5) binding to Ephrin-A receptor (EPHA4) is the key step in cell-cell communication and adhesion; iii) CTBP functions as a co-suppressor by binding to TCF to silence the expression of various genes involved in cellular proliferation and apoptosis; iv) the phosphorylation of PP2A regulatory subunit (PPP2R3A) by PKA-cat (PRKACB) is one step of HTR1A signaling pathway, leading to cell survival; Also PKA-cat is the hub connecting other proteins involved in the "*Signal transduction_PKA signaling pathway*".

MetaCore analysis for enriched process networks revealed that genes involved in "*Cell adhesion_Attractive and repulsive receptors*" and "*Development_Regulation of angiogenesis*" process networks were more likely to be co-expressed in the normal group than that in the CML group (Table 4.6, Figures 4.6 and 4.7). The alteration in adhesion properties of leukemic progenitors is one of CML characteristics at the cellular level (Salesse and Verfaillie, 2002). BCR-ABL oncogene is generated by the translocation between chromosome 9 and 22, forming a fusion protein. The tyrosine kinase activity of BCR-ABL protein is increased due to the translocation. It can phosphorylate many molecules related to cell adhesion, leading to the dysregulation of adhesion property (Salesse and Verfaillie, 2002) (Figure 4.8).

Angiogenesis is the process forming new blood from the preexisting vasculature, which includes the degradation of extracellular matrix proteins, as well as the activation, proliferation and migration of endothelial cells (Aguayo *et al.*, 2000; Cines *et al*., 1998). Moreover, solid tumor growth, dissemination and metastasis are related to angiogenesis (Aguayo *et al*., 2000). In leukemia, the hematopoietic cells are supported from the normal vascular bed in bone marrow (Aguayo *et al.*, 2000). It has been reported that increased vascularity is found in acute myeloid leukemia (AML) patients (Hussong *et al*., 2000). Importantly, in CML the number of blood vessels and vascular area were found to be increased when compared to control bone marrows (Aguayo *et al*., 2000) (Figure 4.9).

Among these two process networks, the connection from EFNA5 to EPHA4 was identified as a strongly co-expressed gene pair in the normal group (|r| values were 0.720 and 0.013 in the normal group and the CML group, respectively) (Figure 4.7). Ephrin-A receptors are the largest subfamily of receptor tyrosine kinases that regulate cell shape, mobility and attachment (Himanen *et al*., 2007). The interaction between Ephrin-A receptors and the ligands plays a vital role in cell-cell communication, which initiates the unique bi-directional signaling cascades to transduce the information in both receptor- and ligand-expressing cells (Himanen and Nikolov, 2003). There may be some relationships between adhesion and angiogenesis. Moreover, these two process networks were found to be well controlled in the normal group compared to the CML group. The adhesion and angiogenesis properties may be dysregulated in the CML state.

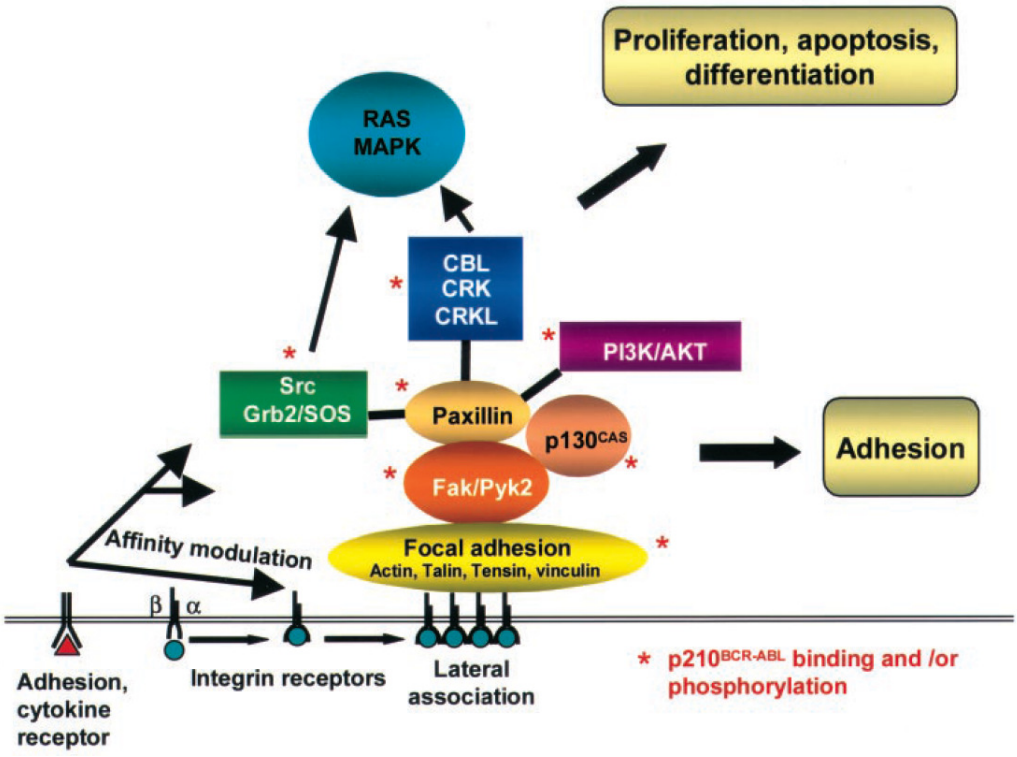


Figure 4.8: Dysregulated adhesion signaling pathway in CML. The binding of integrins to receptors results in the assembling of focal adhesion that recruits various signaling molecules. In CML, the fusion oncoprotein $p210^{BCR-ABL}$ can phosphorylate and/or bind to many intracellular signaling molecules, such as CRKL, paxillin and PI3-K to affect cell adhesion (Salesse and Verfaillie, 2002).

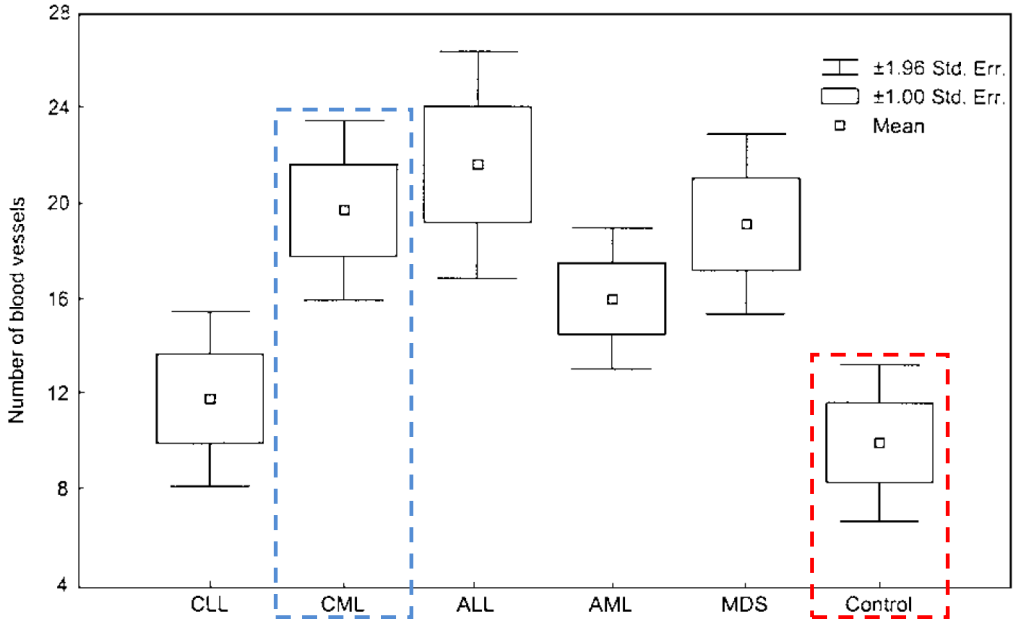


Figure 4.9: Number of blood vessels compared to control. The box plots indicate the significant differences among various diseases (P = .0005, Kruskal-Wallis test) (Aguayo *et al.*, 2000).

Chapter 5 MiR-17-92 Cluster Target Genes Co-expression Analysis

5.1 Method

5.1.1 Identification of candidate target genes directly regulated by miR-17-92 cluster

The systematic search for genes directly targeted by miR-17-92 cluster (miR-17-5p, miR-17-3p, miR-18a, miR-19a, miR-20a, miR-19b and miR-92-1) was performed on five miRNA prediction databases (*DIANA-microT*, *MicroCosm-Targets*, *miRWalk*, *TargetScan* and *miRDB*) (Table 5.1). Some prediction databases (e.g. *DIANA-microT* and *TargetScan*) predict the miRNA targets based on three basic criteria: i) complementarity when miRNA binds to mRNA in seed regions; ii) free energy to fold the miRNA-mRNA duplex; and iii) conservation among different species (Chan *et al*., 2012; Li *et al*., 2010a; Wang *et al*., 2014). In the first step, we obtained the target genes regulated by each mature miRNA from the miR-17-92 cluster using these five prediction databases. In order to increase the prediction accuracy, we selected miRNA targets predicted by at least four out of five databases. In the next step, we combined the genes from all the seven mature miRNAs as the candidate target genes directly regulated by miR-17-92 cluster for the following co-expression analysis.

Table 5.1: MiRNA prediction databases for identifying the candidate target genes of miR-17-92 cluster

#	Database	Species	Website
1	DIANA-microT	Any	http://diana.cslab.ece.ntua.gr/microT/
2	MicroCosm-Targets	Any	http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/
3	miRWalk	Human, mouse, rat	http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk/predictedmirnagene.html
4	TargetScan	Any	http://www.targetscan.org/
5	miRDB	Human, mouse, rat, dog, chicken	http://mirdb.org/miRDB/

5.1.2 Co-expression analysis for candidate target genes

Using the same approach as the genome-wide analysis, correlation coefficients for all the possible gene pairs of the candidate target genes directly regulated by miR-17-92 cluster were calculated (Sections 2.1.2 and 2.1.3). Following the same distribution-based approach for gene pair classification, the gene pairs were also classified into the normal-specific, CML-specific and common pairs (Section 2.1.4).

5.1.3 Gene ontology annotation for miR-17-92 target genes

5.1.3.1 Flow chart

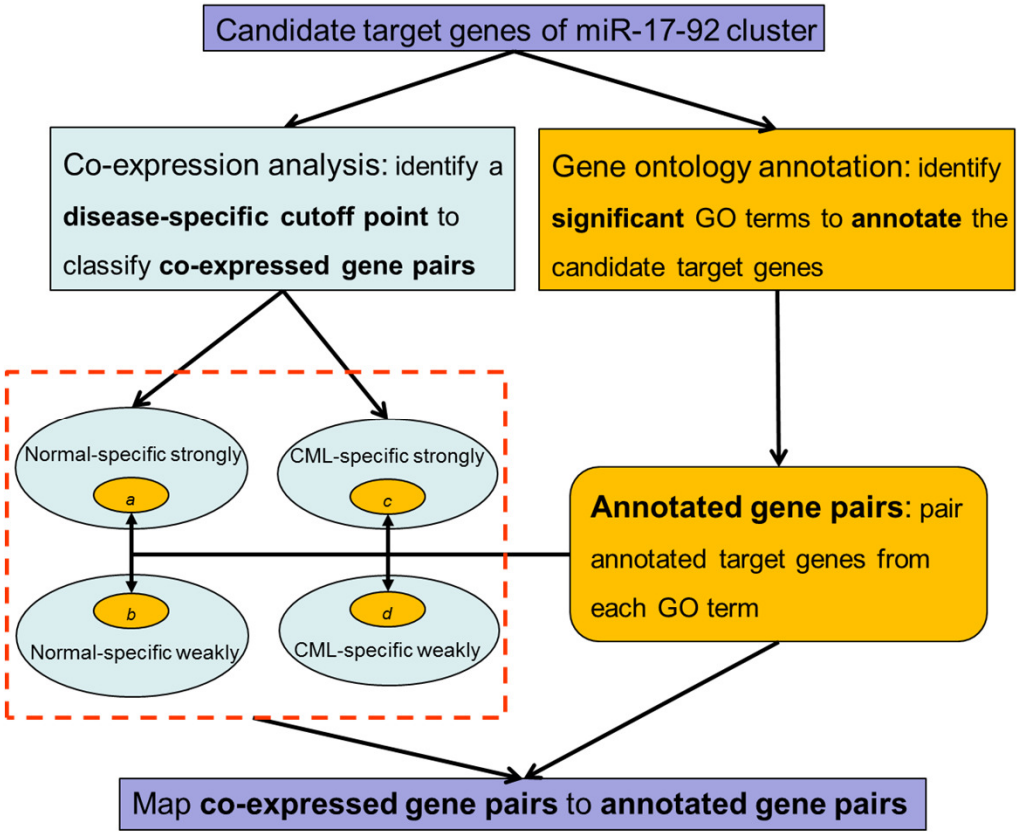


Figure 5.1: Flow chart for the gene ontology (GO) annotation of candidate target genes of miR-17-92. "*a*" refers to the mapped normal-specific strongly co-expressed pairs. "*b*" represents the mapped normal-specific weakly co-expressed pairs. "*c*" stands for the mapped CML-specific strongly co expressed pairs. "*d*" represents the mapped CML-specific weakly co-expressed pairs.

5.1.3.2 Gene ontology annotation

In this study, we also applied the *DAVID* database to annotate the candidate target genes directly regulated by miR-17-92 cluster (Section 3.1.2.2.2). Functional annotation chart was chosen to demonstrate the significant batch annotation and GO terms that were most pertinent to the input data. The significance of GO term enrichment is calculated based on a modified Fisher exact test (EASE score). Using *DAVID*, we annotated the candidate target genes involved in the significantly enriched GO terms. All these three GO categories (molecular function, biological process and cellular component) were considered. The selection criteria for the significantly enriched GO terms were: i) EASE score < 0.05; and ii) FDR < 0.05, for multiple-hypothesis correction. Candidate target genes identified in each significantly enriched GO term were called the annotated target genes.

5.1.3.3 Mapping co-expressed gene pairs to annotated gene pairs

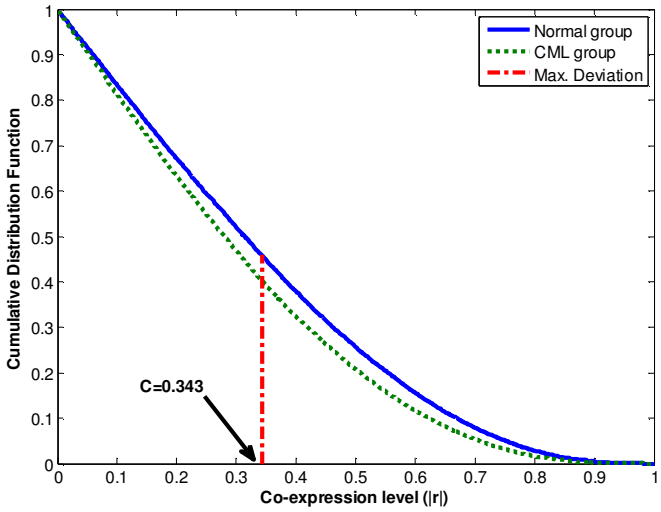
The annotated target genes in each GO term were to form the annotated gene pairs. The annotated gene pairs from each GO term were mapped to the identified co-expressed gene pairs: mapped normal-specific strongly (*a*), mapped normal-specific weakly (*b*), mapped CML-specific strongly (*c*) and mapped CML-specific weakly co-expressed pairs (*d*). Fisher exact test was used to identify if there were more mapped normal-specific strongly co-expressed pairs than mapped CML-specific strongly co-expressed pairs in each GO term. The multiple-hypothesis correction was performed by following Bonferroni correction (Section 3.1.2.2.3).

5.2 Results

5.2.1 Identification of structural co-expression difference

The candidate target genes directly regulated by miR-17-92 cluster were collected from the five prediction databases. In order to calculate the correlation coefficients, the target genes should be found in the microarray dataset. In total, we identified 288 candidate target genes in the microarray dataset GSE5550 (Appendix A3). We further extracted the available expression profiles of these 288 genes and calculated the correlation coefficients. In each group, there was a set of correlation coefficients of 41,328 gene pairs. We then plotted the cumulative distributions for these two sets of data. Two-sample KS test was performed to identify the difference from the overall structure. The results showed that these two distributions between the normal and the CML groups were significantly different (p-value = $2.62 \times 10^{-58} < 0.05$ for the maximum deviation $D = 0.0567$) (Figure 5.2). The disease-specific cutoff point, $C = 0.343$, was identified at the maximum deviation (Figure 5.2). The cutoff point classified gene pairs into four co-expression classes based on the co-expression level (Table 5.2). Two co-expression patterns were so distinct that the normal group had more strongly co-expressed (level above ~0.343) gene pairs compared to the CML group. Chi-square test indicated that the proportions of strongly and weakly co-expressed gene pairs significantly differed between the normal and the CML groups (p-value = $6.12 \times 10^{-61} < 0.05$ for the statistic $\chi^2 = 271$).

a.



b.

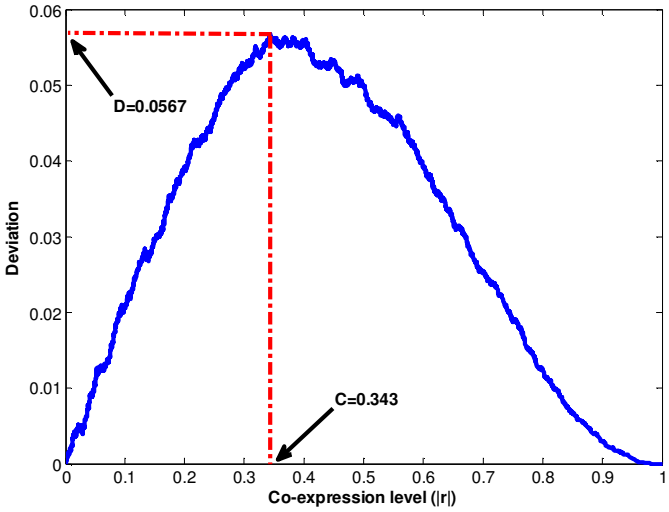


Figure 5.2: Plots of distributions for the co-expression analysis of candidate target genes directly regulated by miR-17-92 cluster. (a) Cumulative distribution functions of co-expression levels in the normal and the CML groups. (b) Deviation distribution against different co-expression cutoff points.

Table 5.2: Cross-tabulation of gene pair counts in the co-expression analysis of candidate target genes directly regulated by miR-17-92 cluster

Group	# of strongly co-expressed gene pairs	# of weakly co-expressed gene pairs
Normal	18,999	22,329
CML	16,654	24,674

5.2.2 Co-expression galaxy and structures for the candidate target genes directly regulated by miR-17-92 cluster

The co-expression galaxy was plotted and partitioned into four regions, according to the same procedures with the genome-wide co-expression analysis (Section 2.2.2): i) normal-specific strongly co-expressed pairs (CML-specific weakly co-expressed pairs): the percentage was 27.277%; ii) common strongly co-expressed pairs: the percentage was 18.694%; iii) CML-specific strongly co-expressed pairs (normal-specific weakly co-expressed pairs): the percentage was 21.603%; and iv) common weakly co-expressed pairs: the percentage was 32.426% (Figures 5.3 and 5.4). From the results, we observed that there were more normal-specific strongly co-expressed pairs than CML-specific strongly co-expressed pairs.

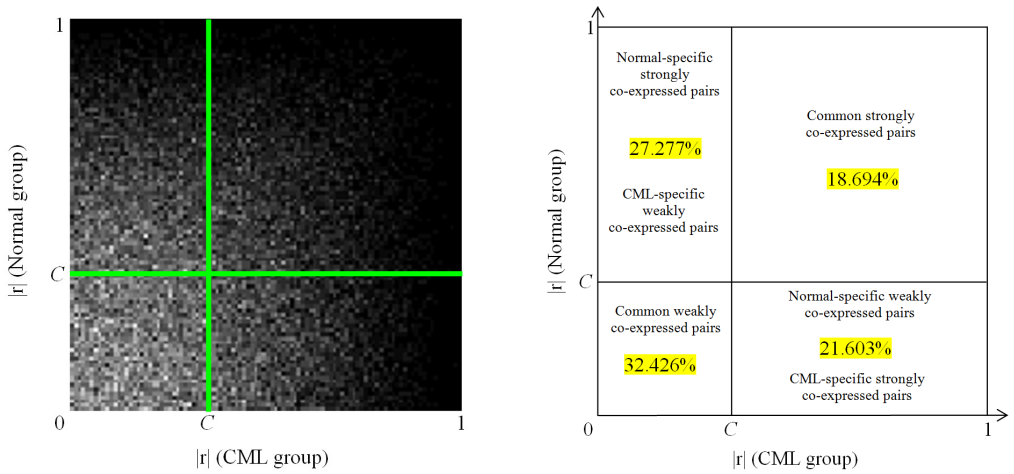


Figure 5.3: Co-expression galaxy (left) and four regions partitioned by the disease-specific cutoff point, C = 0.343 (right). Each correlation coefficient (|r|) is represented by one white dot in the galaxy. More dots mean that there are more correlation coefficients located in that region.

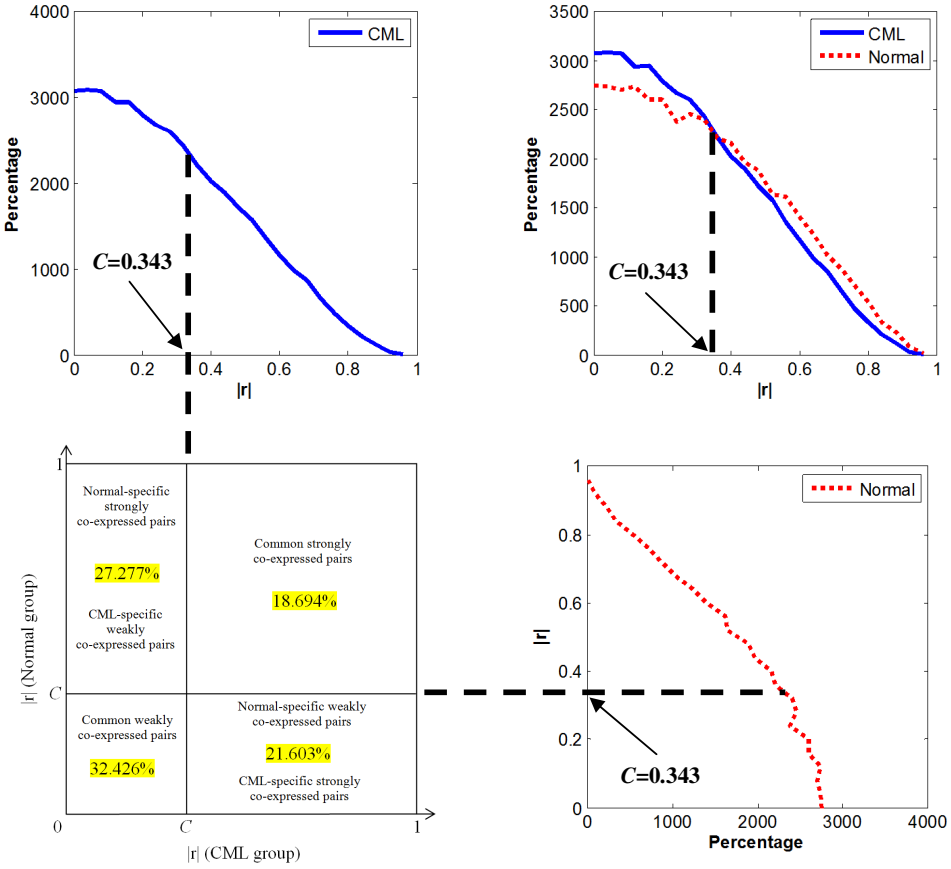


Figure 5.4: Relationship between co-expression structures and partitioned four regions in the co-expression galaxy. The red dot line represents the distribution of percentage for correlation coefficients (|r| values) in the normal group. The blue solid line stands for the distribution of percentage for correlation coefficients (|r| values) in the CML group.

5.2.3 *David* annotation for enriched gene ontology

5.2.3.1 Biological process

According to the selection criteria (EASE score < 0.05 and FDR < 0.05), 11 significantly enriched GO terms for biological processes were identified (Table 5.3). We obtained the annotated target genes involved in each biological process and formed the annotated gene pairs. In the next step, the co-expressed gene pairs were mapped to the annotated gene pairs. The results showed that all these 11 processes had more mapped normal-specific strongly co-expressed pairs than mapped CML-specific strongly co-expressed pairs (Table 5.4). Fisher exact test demonstrated that 8 of 11 biological processes were significantly mapped (p-values < 0.05, corrected p-values < 0.05): *Positive regulation of nitrogen compound metabolic process*, *Positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process*, *Positive regulation of biosynthetic process*, *Positive regulation of cellular biosynthetic process*, *Positive regulation of macromolecule biosynthetic process*, *Positive regulation of transcription, DNA-dependent*, *Positive regulation of RNA metabolic process* and *Positive regulation of cellular metabolic process*. From the results, we observed that nearly all the processes were related to metabolism, such as nitrogen compound metabolic process, cellular biosynthetic process and RNA metabolic process. Only one process was an exception that was related to transcription: *Positive regulation of transcription, DNA-dependent*. Moreover, all these significant biological processes perform positive regulation function. Our results discovered that genes involved in these processes were more likely to be co-expressed in the normal group when compared to the CML group.

Table 5.3: Biological process_Enriched GO terms for the functional annotation of candidate target genes directly regulated by miR-17-92 cluster

#	Enriched GO terms	Genes found in our data	EASE score	FDR
1	Positive regulation of nitrogen compound metabolic process	37	$4.50x10^{-08}$	$7.30x10^{-05}$
2	Positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	36	$6.50x10^{-08}$	$1.10x10^{-04}$
3	Positive regulation of biosynthetic process	38	$1.00x10^{-07}$	$1.60x10^{-04}$
4	Positive regulation of cellular biosynthetic process	37	$2.10x10^{-07}$	$3.40x10^{-04}$
5	Positive regulation of transcription	32	$6.20x10^{-07}$	$1.00x10^{-03}$
6	Positive regulation of macromolecule biosynthetic process	35	$6.20x10^{-07}$	$1.00x10^{-03}$
7	Positive regulation of gene expression	32	$1.20x10^{-06}$	$1.90x10^{-03}$
8	Positive regulation of transcription, DNA-dependent	28	$2.00x10^{-06}$	$3.30x10^{-03}$
9	Positive regulation of RNA metabolic process	28	$2.30x10^{-06}$	$3.80x10^{-03}$
10	Positive regulation of cellular metabolic process	40	$4.50x10^{-06}$	$7.30x10^{-03}$
11	Positive regulation of macromolecule metabolic process	38	$1.50x10^{-05}$	$2.40x10^{-02}$

GO: gene ontology. EASE score: a modified Fisher exact test, Expression Analysis Systematic Explorer score. FDR: false discovery rate.

Table 5.4: Mapping co-expressed gene pairs to annotated gene pairs from each biological process

GO terms	Fisher exact test					Corrected
	a	*b*	*c*	*d*	p-value	p-value
Positive regulation of nitrogen compound metabolic process	186	148	148	186	**0.002**	**0.022**
Positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	177	141	141	177	**0.003**	**0.033**
Positive regulation of biosynthetic process	196	156	156	196	**0.002**	**0.022**
Positive regulation of cellular biosynthetic process	192	146	146	192	**<0.001**	**<0.011**
Positive regulation of transcription	137	111	111	137	0.012	0.132
Positive regulation of macromolecule biosynthetic process	169	129	129	169	**0.001**	**0.011**
Positive regulation of gene expression	137	111	111	137	0.012	0.132
Positive regulation of transcription, DNA-dependent	110	79	79	110	**0.001**	**0.011**
Positive regulation of RNA metabolic process	110	79	79	110	**0.001**	**0.011**
Positive regulation of cellular metabolic process	216	170	170	216	**0.001**	**0.011**
Positive regulation of macromolecule metabolic process	188	153	153	188	0.005	0.055

GO: gene ontology. GO terms highlighted in bold text are significantly mapped. *a*: mapped normal-specific strongly co-expressed pairs. *b*: mapped normal-specific weakly co-expressed pairs. *c*: mapped CML-specific strongly co-expressed pairs. *d*: mapped CML-specific weakly co-expressed pairs.

5.2.3.2 Cellular component

There was no significantly enriched GO term for cellular component identified according to the same selection criteria (EASE score < 0.05 and FDR < 0.05).

5.2.3.3 Molecular function

There was no significantly enriched GO term for molecular function identified according to the same selection criteria (EASE score < 0.05 and FDR < 0.05).

5.3 Discussion and conclusion

In this chapter, we have identified the overall differences in the co-expression patterns of those candidate target genes directly regulated by miR-17-92 cluster between the normal and the CML groups. Two-sample KS test was performed to identify the difference. Firstly, the maximum deviation (D = 0.0567) between two cumulative distributions indicated the difference between the normal and the CML groups structurally (Figure 5.2). Then, a disease-specific cutoff point (C = 0.343) was found at the maximum deviation to classify the co-expressed gene pairs. We further identified the specifically co-expressed gene pairs in the normal and the CML groups to investigate the alterations of biological processes (Figures 5.3 and 5.4).

In order to explore which biological process had more strongly co-expressed gene pairs in the normal group, we applied *David* database to annotate the candidate target genes. The GO annotation for enriched biological process showed that genes related to metabolism, such as nitrogen compound metabolic process, cellular biosynthetic process and RNA metabolic process, were more likely to be co-expressed in the normal group compared to the CML group (Table 5.4). Moreover, all these significantly mapped biological processes perform positive regulation function (Table 5.4). However, the positive function was dysregulated in the CML group, since there were less strongly co-expressed gene pairs in the CML group than that in the normal group.

Dysregulated mRNA metabolism is regarded as a feature for many human cancers, including CML (Perrotti and Neviani, 2007). BCR-ABL oncoprotein was reported to affect the basal mRNA translation machinery by regulating the function of

translation factors eukaryotic translation initiation factor 4E and its binding protein (Perrotti and Neviani, 2007). Other researchers found that the metabolic patterns of untreated CML patients are different from healthy controls, indicating the metabolic dysregulation in CML patients (A *et al.*, 2010). Compared to health individuals, CML patients had lower levels of tricarboxylic acid cycle and lipid metabolism, as well as higher levels of urea cycle metabolites and amino acid turnover (A *et al.*, 2010; Denkert *et al.*, 2008; Pelicano *et al.*, 2006). Purines (urate) were also found to be increased in CML patients, indicating a higher capacity for DNA synthesis (A *et al.*, 2010).

Chapter 6 Overall Discussion and Conclusion

6.1 Co-expression galaxy and structures

In this study, we presented a novel approach for gene co-expression analysis based on the co-expression structure. We identified the overall differences in the co-expression patterns between the normal and the CML groups. The co-expression pattern differences were reflected from the overall structure, not only considering pair by pair independently. Correlation coefficients for all the possible gene pairs were considered to form two different cumulative distributions. Two-sample KS test was performed to identify the difference of these two distributions. Firstly, the maximum deviation between two cumulative distributions indicated the difference structurally. Then, a disease-specific cutoff point was identified at the maximum deviation to classify the co-expressed gene pairs so that the class was best coherent with the CML state.

After identifying the disease-specific cutoff point, the co-expressed gene pairs were partitioned into four regions based on their locations in the co-expression galaxy, forming the co-expression structures: i) the normal-specific strongly co-expressed pairs (the CML-specific weakly co-expressed pairs); ii) the common strongly co-expressed pairs; iii) the CML-specific strongly co-expressed pairs (the normal-specific weakly co-expressed pairs); and iv) the common weakly co-expressed pairs (Figure 2.1). This kind of classification considered all the gene pairs to locate them to different locations based on their different co-expression levels and different

groups (the normal group or the CML group), called the distribution-based classification.

The divided gene pairs had special biological meanings. The normal-specific strongly co-expressed pairs indicated the potential molecular interactions maintaining physiological balance in healthy individuals, which was regarded as the inter-gene linkages. The CML-specific strongly co-expressed pairs represented the characteristics of the disease and may be the pathogenic alternatives when the corresponding normal-specific pairs were not co-expressed in response to stress. By analyzing the biological meaning of these gene pairs, we further explored the possible reasons leading to the CML disease.

Compared with the differential expression analysis, co-expression analysis can identify the functionally associated linkages among genes during signal transduction. In addition, analyses of differential expression do not take account of the level of correlations that may exist between gene expression patterns (Torkamani *et al.*, 2010). As a result, the co-expression analysis is more useful for analyzing the underlying mechanisms of diseases.

Previous studies on the co-expression analysis identify significantly co-expressed gene pairs by calculating a p-value of correlation coefficient for each gene pair individually, which cannot reflect the overall difference in two different groups. Several algorithms have been proposed to analyze the co-expressed genes. One of them is the two-stage screening procedure, which was applied to select statistically and biologically significant gene pairs in Zhu *et al.*'s study (Zhu *et al.*, 2005). In

Gupta *et al.*'s study, they proposed a method to determine the correlation threshold using the clustering coefficient. R^2 metric was applied as a measure of similarity between two different genes (Gupta *et al.*, 2006). These two studies cannot reflect the overall difference in two different groups. While, our approach calculated all the correlation coefficients in each group (the normal group or the CML group) to form two different cumulative distributions, which can identify the difference of two different groups from the overall structure.

In summary, we have presented a detailed method to identify a disease-specific cutoff point for co-expression levels that classified the co-expressed gene pairs. This distribution-based classification considered all the gene pairs to partition them to different locations based on their different co-expression levels and different groups. We applied this method to explore the difference between the normal and the CML groups in the co-expression patterns of those genes involved in a functional gene set, regulated by the same regulators (TFs and miRNAs), and covering from the whole genome. Our method effectively identified the statistical differences from the overall structure. The different co-expression pattern can reflect the biological alterations in CML.

6.2 Gene set co-expression analysis and functional annotation

After identifying the statistical differences in the co-expression patterns of the NPM1-associated genes, candidate target genes regulated by E2F1–3 and MYC, and candidate target genes regulated by miR-17-92 cluster, we further performed the functional annotation to explore the biological differences. Our findings provided important information to understand the underlying mechanisms of the CML state.

6.2.1 NPM1-associated gene set

It has been reported that NPM1 is overexpressed in proliferating cells and tumor cells. A possible reason is that NPM1 is related to the increased ribosomal synthesis (Naoe *et al.*, 2006). NPM1 can both promote cell growth and repress tumor cells. Its overexpression increases cell division and growth, possibly owing to the effects on rDNA transcription, ribosome subunit export and S-phase DNA replication (Lindström, 2011). The NPM1-associated gene set, 93 out of 116 genes found in the CML microarray dataset, was chosen in this study. We wonder if the co-expression analysis can demonstrate the dysregulated ribosomal synthesis and translation process in the CML state that are different from the normal state.

The co-expression networks revealed that the RP genes were more likely to be co-expressed in the CML group compared to the normal group (Figures 3.5, 3.6 and 3.7). The GO annotation showed that: i) genes involved in *Translational elongation* and *Translation* biological processes tended to be co-expressed in the CML group, including RP genes (e.g. RPL6 and RPS28) and translation factors (e.g. EEF2 and EIF3F) (Table 3.4, Figures 3.8 and 3.9); ii) genes related to ribosome, nucleolus and cytoplasm tended to be co-expressed in the CML group (Table 3.6, Figures 3.10 and 3.11). The co-expression pattern in the normal group was regarded as the inter-gene linkages which represented the healthy pathological balance. Compared to the normal group, these genes were more likely to be co-expressed in the CML group. As a result, the co-expression pattern for these genes related to ribosome synthesis and translation process may be dysregulated in the CML disease.

We also found some literature supports for these findings. Altered mRNA translation is involved in the pathogenesis of CML (Zhang *et al*., 2008). Ly *et al*. reported that the translational regulators, ribosomal protein S6 and 4E-BP1 (a negative regulator in cap-dependent mRNA translation process), are constitutively phosphorylated in CML cells (Ly *et al*., 2003). The encoded protein by eukaryotic translation initiation factor 4E (EIF4E) is regarded as both a key translation factor and a promoter for nucleocytoplasmic transport of specific transcripts (Topisirovic *et al*., 2003). Overexpression of EIF4E has been found in CML patients, suggesting its possible role in neoplastic transformation and the feasibility as a novel therapeutic approach (Hagner *et al*., 2010; Topisirovic *et al*., 2003).

6.2.2 E2F1–3 and MYC Target Genes

E2F1–3 and MYC are important transcription factors that can regulate both cell proliferation and apoptosis. Most importantly, E2F1–3 and MYC are reciprocally regulated to form the positive feedback loops in the transcription process to well control their expression levels (Figure 1.5) (Aguda *et al*., 2008; Coller *et al*., 2007). Since target genes transcriptionally regulated by E2F1–3 or MYC are related to cell proliferation and apoptosis, we wonder what the co-expression patterns of the target genes regulated directly and concurrently by E2F1–3 and MYC in the normal and the CML groups are.

MetaCore functional annotation for enriched pathway maps revealed that some potentially dysregulated signal pathway links were identified. These three links, INHBA connected with ACVR2A, EFNA5 connected with EPHA4 and CTBP2 connected with TCF4, can be found only in the normal group, not in the CML group

(Figure 4.5a, b, c, d and e). While, one CML-specific link was found that was the link from PRKACB to PPP2R3A (Figure 4.5f). These identified links located in the important steps in the whole pathways (Table 4.4): i) Activin A (INHBA) binding to the type II serine kinase receptor (ACVR2A) is the first step in the "*Development_Role of Activin A in cell differentiation and proliferation*" and "*Signal transduction_Activin A signaling regulation*"; ii) Ephrin-A (EFNA5) binding to Ephrin-A receptor (EPHA4) is the key step in cell-cell communication and adhesion; iii) CTBP functions as a co-suppressor by binding to TCF to silence the expression of various genes involved in cellular proliferation and apoptosis; iv) The phosphorylation of PP2A regulatory subunit (PPP2R3A) by PKA-cat (PRKACB) is one step of HTR1A signaling pathway, leading to cell survival; Also PKA-cat is the hub connecting other proteins involved in the "*Signal transduction_PKA signaling pathway*". These links may be dysregulated in the CML disease.

MetaCore analysis for enriched process networks showed that genes involved in "*Cell adhesion_Attractive and repulsive receptors*" and "*Development_Regulation of angiogenesis*" process networks were more likely to be co-expressed in the normal group than that in the CML group (Table 4.6, Figures 4.6 and 4.7). The alteration in adhesion properties of leukemic progenitors is one of CML characteristics at the cellular level (Salesse and Verfaillie, 2002). In CML, the BCR-ABL fusion protein can phosphorylate many molecules related to cell adhesion, leading to the dysregulation of adhesion property (Salesse and Verfaillie, 2002) (Figure 4.8). In addition, Bhatia *et al*. hypothesized that decreased integrin-mediated adhesion of CML progenitors to stroma can lead to continuous cell proliferation (Bhatia *et al*., 1996). They treated the cells with Interferon-α (IFN-α). The results showed that the

treatment restored the CML progenitor adhesion to stroma, and also the regulation of CML progenitor proliferation (Bhatia *et al*., 1996). In leukemia, the hematopoietic cells are supported from the normal vascular bed in bone marrow (Aguayo *et al*., 2000). Importantly, in CML the number of blood vessels and vascular area were found to be increased when compared to control bone marrows (Aguayo *et al*., 2000) (Figure 4.9). Dysregulation of adhesion and angiogenesis properties may be associated with CML.

6.2.3 MiR-17-92 Cluster Target Genes

MiRNAs are regarded as a new class of gene regulatory factors regulating the expression of human genes during the post-transcriptional process in recent years. The mature miRNAs from miR-17-92 cluster have similar expression patterns in hematopoietic cell lines (Coller *et al*., 2007; Yu *et al*., 2006a). Most importantly, the miR-17-92 cluster is overexpressed in chronic-phase CML patients compared with normal individuals, and its overexpression can promote cell cycle progression and proliferation, and inhibit apoptosis (Mendell, 2008; Venturini *et al*., 2007). Researchers also found that the BCR-ABL tyrosine kinase activity can affect this miRNA cluster (Venturini *et al*., 2007). Genes with similar mRNA expression profiles are likely to be regulated via the same mechanism(s), e.g. the same regulator (Allocco *et al*., 2004; Altman and Raychaudhuri, 2001; Schulze and Downward, 2001). In our study, we hypothesize that target genes regulated by the same miRNA should be co-expressed. We explored the differences in the co-expression patterns of those target genes directly regulated by miR-17-92 cluster between the normal and the CML groups.

David gene ontology annotation demonstrated that genes related to metabolism, such as nitrogen compound metabolic process, cellular biosynthetic process and RNA metabolic process, were more likely to be co-expressed in the normal group compared to the CML group (Table 5.4). Moreover, all the significantly mapped biological processes perform the positive regulation function (Table 5.4). However, the positive function was dysregulated in the CML group, since there were less strongly co-expressed gene pairs in the CML group. Dysregulated mRNA metabolism is regarded as a feature for many human cancers, including CML (Perrotti and Neviani, 2007). BCR-ABL oncoprotein was reported to affect the basal mRNA translation machinery by regulating the function of translation factors eukaryotic translation initiation factor 4E and its binding protein (Perrotti and Neviani, 2007). Other researchers found that the metabolic patterns of untreated CML patients are different from healthy controls, indicating the metabolic dysregulation in CML patients (A *et al.*, 2010). As a result, dysregulation of metabolism may be associated with the CML disease.

6.3 Conclusion

In this study, we explored gene co-expression from two aspects: i) functionally related genes tend to be co-expressed: NMP1-associated gene set; ii) genes regulated by the same regulators are more likely to be co-expressed: genes regulated by TFs or miRNAs. Starting from these two points, we identified the potentially dysregulated ribosome synthesis and translation process in the CML state by analyzing the co-expression pattern for the NMP1-associated genes. TFs and miRNAs are two important gene regulatory factors. E2F1–3 and MYC are TFs related to cell proliferation and apoptosis. Exploring the co-expression pattern of their target genes,

we found the potentially dysregulated cell proliferation, adhesion and angiogenesis properties due to the transcriptional regulation of E2F1–3 and MYC in the CML state. The potentially dysregulated metabolism processes were further identified due to the post-transcriptional regulation of miR-17-92 cluster in CML compared to the normal state. Hence, NPM1, E2F1–3, MYC and miR-17-92 cluster can be regarded as the potential drug targets for the CML treatment.

6.4 Future direction

In this study, we statistically identified the co-expression pattern differences between the normal and CML groups using our developed method. The function annotation with literature supports was used to validate the significant results. Future studies should be focused on the improvements based on these findings. The laboratory experiments and other potential areas will be performed to achieve better accuracy and higher performance of our developed method in the future.

6.4.1 Laboratory experimental validation

The identified strongly co-expressed gene pairs in each group will be validated by cell line experiment. K562 cell line is from bone marrow, which was established by Lozzio and Lozzio from the pleural effusion of the patient with CML in terminal blastic phase (Lozzio and Lozzio, 1975). Cells will be cultured in RPMI1640 medium supplemented with 10% fetal bovine serum (FBS). In order to compare the normal and the CML groups, the K562 cell line is perturbed by applying resveratrol treatment as stimulus. K562 cells are treated for 24 hours with 30 μM Resveratrol (Res) or with DMSO as a vehicle control. Cultures are incubated at 37 °C in 5% CO_2.

The cells will be examined for quality and absence of contamination by light microscope everyday. If big clones are found, we will count the cell number and passage the cells to two flasks. The cells will be collected and harvested for the total RNA extraction.

TRIzol Reagent (Life Technologies, Carlsbad, USA) is a ready to use reagent, which is designed to extract high quality total RNAs from cell and tissue samples. It is a mixture of phenol, guanidine isothiocyanate, red dye and other proprietary components. The red dye is used to detect the organic phase, which is not interactive with nucleic acids. TRIzol will be chosen to isolate the total RNAs according to the manufactory's protocol. The determination of mRNA expression levels will follow the traditional reverse transcription polymerase chain reaction (RT-PCR). After the RNA is transcribed to cDNA, the next step is to quantify the cDNA molecules using Real-time PCR. Then, we will compare the treatment and the control groups to validate the significantly co-expressed gene pairs.

6.4.2 Other potential areas

6.4.2.1 Master transcription factor and super enhancer

Transcription factors are proteins that bind to specific sites on DNA sequences, which control the transcription of genetic information from DNA to mRNA. Enhancers are DNA segments with a few hundred base pairs in length, which are usually occupied by multiple transcription factors (Whyte *et al.*, 2013). The activity of transcription-factor-bound enhancers can determine cell-type-specific patterns of gene expression (Whyte *et al.*, 2013). Oct4, Sox2 and Nanog are master transcription

factors to activate the gene expression program of pluripotent embryonic stem cells (Whyte *et al.*, 2013). Super-enhancers that consist of clusters of enhancers are densely bound by the transcription factors, which are different from typical enhancers in size, density and content of transcription factor, as well as other abilities (Whyte *et al.*, 2013). The super-enhancers at oncogenes and other important genes responsible for tumor pathogenesis have been found in cancer cells (Hnisz *et al.*, 2013). Moreover, genes with super-enhancers for cell-type-specific master transcription factors have been identified in differentiated cells to define cell identity (Whyte *et al.*, 2013). As a result, super-enhancers play vital roles in mammalian cell identity.

For the co-expression analysis of candidate target genes regulated directly and concurrently by E2F1–3 and MYC, we would like to perform further studies. In the future, we plan to explore if the super-enhancers are found in the loci of the normal-specific gene pairs and CML-specific gene pairs.

6.4.2.2 Combination of co-expression analysis and multiple regression analysis

The co-expressed genes are more likely to be regulated by the same mechanism(s), e.g. the same regulator. In this study, we explored the differences in the co-expression patterns of those genes regulated by the same regulators (TFs or miRNAs). Since the co-expressed genes may share the same regulators, in the future, we plan to explore the common TFs and miRNAs that regulating the identified strongly co-expressed gene pairs. After identifying the common regulators, we will further investigate the relationship between each gene from the strongly co-expressed gene pairs and the common TFs and miRNAs using the multiple regression analysis.

Since multiple TFs/miRNAs targeting the same mRNA may function together during gene expression process, multiple linear regression analysis is more suitable for studying their relationship and mimicking the real situation *in vivo*. The regression model will be adopted to identify TFs/miRNAs that have significant effects on the target genes (Formula 6) (Wang *et al*., 2014).

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n \quad (6)$$

where y stands for the expression profile of an mRNA (a gene); x_1, x_2, ... and x_n are the expression levels of the corresponding TFs/miRNAs targeting the mRNA; a_0 is the regression constant; a_1, a_2, ... and a_n represent the regression coefficients for each TF/miRNA.

After identifying the TFs/miRNAs that have significant effects on each gene of the strongly and specifically co-expressed gene pairs, we will infer which TF/miRNA has the key function in the normal state or the CML state. These significant TFs/miRNAs may be regarded as the drug targets for the diagnosis and treatment of cancer.

Appendix

A.1 NPM1-associated genes found in GSE5550

Official gene symbol	Gene title
ACTG1	actin, gamma 1
APEX1	APEX nuclease (multifunctional DNA repair enzyme) 1
APRT	adenine phosphoribosyltransferase
ATP5B	ATP synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide
ATP5G3	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit C3 (subunit 9)
CLIC1	chloride intracellular channel 1
COL16A1	collagen, type XVI, alpha 1
DUT	deoxyuridine triphosphatase
EEF1B2	eukaryotic translation elongation factor 1 beta 2
EEF2	eukaryotic translation elongation factor 2
EIF3F	eukaryotic translation initiation factor 3, subunit F
EIF4B	eukaryotic translation initiation factor 4B
FXR1	fragile X mental retardation, autosomal homolog 1
GNB2L1	guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1
H2AFZ	H2A histone family, member Z
HINT1	histidine triad nucleotide binding protein 1
HNRNPA1	heterogeneous nuclear ribonucleoprotein A1
HNRNPH1	heterogeneous nuclear ribonucleoprotein H1 (H)
HNRNPM	heterogeneous nuclear ribonucleoprotein M
HNRNPU	heterogeneous nuclear ribonucleoprotein U (scaffold attachment factor A)
HSP90AB1	heat shock protein 90kDa alpha (cytosolic), class B member 1
IK	IK cytokine, down-regulator of HLA II
ILF2	interleukin enhancer binding factor 2
KHDRBS1	KH domain containing, RNA binding, signal transduction associated 1
NCL	nucleolin
NONO	non-POU domain containing, octamer-binding
NPM1	nucleophosmin (nucleolar phosphoprotein B23, numatrin)
PNN	pinin, desmosome associated protein
POLG	polymerase (DNA directed), gamma

POLR2G	polymerase (RNA) II (DNA directed) polypeptide G
PPIA	peptidylprolyl isomerase A (cyclophilin A)
PPP1CC	protein phosphatase 1, catalytic subunit, gamma isozyme
PRMT1	protein arginine methyltransferase 1
PSMA1	proteasome (prosome, macropain) subunit, alpha type, 1
PSMA2	proteasome (prosome, macropain) subunit, alpha type, 2
PSMA5	proteasome (prosome, macropain) subunit, alpha type, 5
PSMB4	proteasome (prosome, macropain) subunit, beta type, 4
PSME1	proteasome (prosome, macropain) activator subunit 1 (PA28 alpha)
PTMA	prothymosin, alpha
PWP1	PWP1 homolog (S. cerevisiae)
QARS	glutaminyl-tRNA synthetase
RABGGTB	Rab geranylgeranyltransferase, beta subunit
RPL10A	ribosomal protein L10a
RPL11	ribosomal protein L11
RPL13A	ribosomal protein L13a
RPL17	ribosomal protein L17
RPL18	ribosomal protein L18
RPL19	ribosomal protein L19
RPL23	ribosomal protein L23
RPL27	ribosomal protein L27
RPL28	ribosomal protein L28
RPL31	ribosomal protein L31
RPL32	ribosomal protein L32
RPL34	ribosomal protein L34
RPL35	ribosomal protein L35
RPL36A	ribosomal protein L36a
RPL4	ribosomal protein L4
RPL6	ribosomal protein L6
RPL7	ribosomal protein L7
RPL7A	ribosomal protein L7a
RPL9	ribosomal protein L9
RPS11	ribosomal protein S11
RPS14	ribosomal protein S14
RPS16	ribosomal protein S16
RPS18	ribosomal protein S18
RPS19	ribosomal protein S19
RPS23	ribosomal protein S23
RPS24	ribosomal protein S24
RPS28	ribosomal protein S28

RPS3	ribosomal protein S3
RPS3A	ribosomal protein S3A
RPS5	ribosomal protein S5
RPS6	ribosomal protein S6
RPS8	ribosomal protein S8
RPS9	ribosomal protein S9
SAP18	Sin3A-associated protein, 18kDa
SET	SET nuclear proto-oncogene
SLC25A3	solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 3
SNRPA1	small nuclear ribonucleoprotein polypeptide A'
SNRPD2	small nuclear ribonucleoprotein D2 polypeptide 16.5kDa
SNRPD3	small nuclear ribonucleoprotein D3 polypeptide 18kDa
SNRPE	small nuclear ribonucleoprotein polypeptide E
SNRPF	small nuclear ribonucleoprotein polypeptide F
SRSF9	serine/arginine-rich splicing factor 9
SUMO2	small ubiquitin-like modifier 2
TAF9	TAF9 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 32kDa
TBCB	tubulin folding cofactor B
TCP1	t-complex 1
TIAL1	TIA1 cytotoxic granule-associated RNA binding protein-like 1
TPT1	tumor protein, translationally-controlled 1
TRIM28	tripartite motif containing 28
UBE2L3	ubiquitin-conjugating enzyme E2L 3
UQCRH	ubiquinol-cytochrome c reductase hinge protein

A.2 Candidate target genes regulated directly and concurrently by E2F1–3 and MYC extracted from GSE5550

Official gene symbol	Gene title
ACVR2A	activin A receptor, type IIA
ADCY7	adenylate cyclase 7
ADCY8	adenylate cyclase 8 (brain)
ADK	adenosine kinase
AKTIP	AKT interacting protein
ANP32A	acidic (leucine-rich) nuclear phosphoprotein 32 family, member A
AP3B1	adaptor-related protein complex 3, beta 1 subunit
API5	apoptosis inhibitor 5
APOC3	apolipoprotein C-III
ATP5G2	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit C2 (subunit 9)
ATP6AP2	ATPase, H+ transporting, lysosomal accessory protein 2
AZI2	5-azacytidine induced 2
BACH2	BTB and CNC homology 1, basic leucine zipper transcription factor 2
BARD1	BRCA1 associated RING domain 1
BMP6	bone morphogenetic protein 6
C1D	C1D nuclear receptor corepressor
C8A	complement component 8, alpha polypeptide
C8B	complement component 8, beta polypeptide
CACNA2D1	calcium channel, voltage-dependent, alpha 2/delta subunit 1
CALCA	calcitonin-related polypeptide alpha
CALCB	calcitonin-related polypeptide beta
CASK	calcium/calmodulin-dependent serine protein kinase (MAGUK family)
CAST	calpastatin
CD180	CD180 molecule
CDH8	cadherin 8, type 2
CDK17	cyclin-dependent kinase 17
CDKN2B	cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4)
CDKN2C	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)
CDX1	caudal type homeobox 1
CETN3	centrin, EF-hand protein, 3
CNTRL	centrosomal protein 110kDa
COL19A1	collagen, type XIX, alpha 1
COX11	COX11 cytochrome c oxidase assembly homolog (yeast)

CRIM1	cysteine rich transmembrane BMP regulator 1 (chordin-like)
CSNK1G3	casein kinase 1, gamma 3
CTBP2	C-terminal binding protein 2
CYP24A1	cytochrome P450, family 24, subfamily A, polypeptide 1
CYP7B1	cytochrome P450, family 7, subfamily B, polypeptide 1
DACH1	dachshund homolog 1 (Drosophila)
DACT1	dapper, antagonist of beta-catenin, homolog 1 (Xenopus laevis)
DBC1	deleted in bladder cancer 1
DCUN1D1	DCN1, defective in cullin neddylation 1, domain containing 1 (S. cerevisiae)
DLK1	delta-like 1 homolog (Drosophila)
DMXL1	Dmx-like 1
DNAH9	dynein, axonemal, heavy chain 9
DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha
DUSP6	dual specificity phosphatase 6
EBF2	early B-cell factor 2
EDN2	endothelin 2
EEF1B2	eukaryotic translation elongation factor 1 beta 2
EFNA5	ephrin-A5
EFNB2	ephrin-B2
EGR3	early growth response 3
ELAVL2	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 2 (Hu antigen B)
EPHA4	EPH receptor A4
ERCC5	excision repair cross-complementing rodent repair deficiency, complementation group 5
ESRRG	estrogen-related receptor gamma
ETNK1	ethanolamine kinase 1
FDX1	ferredoxin 1
FEZ2	fasciculation and elongation protein zeta 2 (zygin II)
FHIT	fragile histidine triad gene
FRZB	frizzled-related protein
FST	follistatin
FXR1	fragile X mental retardation, autosomal homolog 1
GAP43	growth associated protein 43
GATA3	GATA binding protein 3
GBA3	glucosidase, beta, acid 3 (cytosolic)
GBF1	golgi brefeldin A resistant guanine nucleotide exchange factor 1
GCK	glucokinase (hexokinase 4)
GCNT1	glucosaminyl (N-acetyl) transferase 1, core 2

GJA9	gap junction protein, alpha 9, 59kDa
GLRX3	glutaredoxin 3
GOT2	glutamic-oxaloacetic transaminase 2, mitochondrial (aspartate aminotransferase 2)
GPR63	G protein-coupled receptor 63
GRB14	growth factor receptor-bound protein 14
GTF2B	general transcription factor IIB
HABP2	hyaluronan binding protein 2
HIVEP3	human immunodeficiency virus type I enhancer binding protein 3
HNRNPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1
HNRNPAB	heterogeneous nuclear ribonucleoprotein A/B
HOXD4	homeobox D4
HS2ST1	heparan sulfate 2-O-sulfotransferase 1
HSF2	heat shock transcription factor 2
HSPA14	heat shock 70kDa protein 14
HTATIP2	HIV-1 Tat interactive protein 2, 30kDa
IGFBP5	insulin-like growth factor binding protein 5
INHBA	inhibin, beta A
ISL1	ISL LIM homeobox 1
ITGA2	integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)
JAG1	jagged 1
JARID2	jumonji, AT rich interactive domain 2
KCNA5	potassium voltage-gated channel, shaker-related subfamily, member 5
KCNJ2	potassium inwardly-rectifying channel, subfamily J, member 2
KCNMA1	potassium large conductance calcium-activated channel, subfamily M, alpha member 1
KIAA1967	KIAA1967
KIFAP3	kinesin-associated protein 3
KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
KITLG	KIT ligand
KLF3	Kruppel-like factor 3 (basic)
KLF6	Kruppel-like factor 6
KLF7	Kruppel-like factor 7 (ubiquitous)
LAMTOR3	MAPK scaffold protein 1
LDB2	LIM domain binding 2
LPHN2	latrophilin 2
LRP12	low density lipoprotein receptor-related protein 12
LRP1B	low density lipoprotein receptor-related protein 1B
LSAMP	limbic system-associated membrane protein
MAF	v-maf musculoaponeurotic fibrosarcoma oncogene homolog (avian)

MAP2K5	mitogen-activated protein kinase kinase 5
MAP3K8	mitogen-activated protein kinase kinase kinase 8
MAPK6	mitogen-activated protein kinase 6
MBP	myelin basic protein
MECOM	MDS1 and EVI1 complex locus
MEF2C	myocyte enhancer factor 2C
MEOX2	mesenchyme homeobox 2
MITF	microphthalmia-associated transcription factor
MKI67	antigen identified by monoclonal antibody Ki-67
MKLN1	muskelin 1, intracellular mediator containing kelch motifs
MPDZ	multiple PDZ domain protein
MVK	mevalonate kinase
NFIB	nuclear factor I/B
NOMO3	NODAL modulator 3
NOX3	NADPH oxidase 3
NPAS3	neuronal PAS domain protein 3
NR2F1	nuclear receptor subfamily 2, group F, member 1
NRP2	neuropilin 2
NRXN3	neurexin 3
NTM	neurotrimin
OPCML	opioid binding protein/cell adhesion molecule-like
ORC4	origin recognition complex, subunit 4
PAK3	p21 protein (Cdc42/Rac)-activated kinase 3
PARVA	parvin, alpha
PAWR	PRKC, apoptosis, WT1, regulator
PAX7	paired box 7
PBX3	pre-B-cell leukemia homeobox 3
PCDH9	protocadherin 9
PDE1A	phosphodiesterase 1A, calmodulin-dependent
PDGFRB	platelet-derived growth factor receptor, beta polypeptide
PEX2	peroxisomal biogenesis factor 2
PFKP	phosphofructokinase, platelet
PIAS1	protein inhibitor of activated STAT, 1
PIK3C3	phosphoinositide-3-kinase, class 3
PIP4K2A	phosphatidylinositol-5-phosphate 4-kinase, type II, alpha
PITX1	paired-like homeodomain 1
PITX3	paired-like homeodomain 3
PKIA	protein kinase (cAMP-dependent, catalytic) inhibitor alpha
PKN2	protein kinase N2
PNMA2	paraneoplastic antigen MA2

PNO1	partner of NOB1 homolog (S. cerevisiae)
PODXL	podocalyxin-like
POLR3G	polymerase (RNA) III (DNA directed) polypeptide G (32kD)
PPIC	peptidylprolyl isomerase C (cyclophilin C)
PPP1R12B	protein phosphatase 1, regulatory (inhibitor) subunit 12B
PPP2R3A	protein phosphatase 2, regulatory subunit B, alpha
PRDM1	PR domain containing 1, with ZNF domain
PRDM12	PR domain containing 12
PREP	prolyl endopeptidase
PRKACB	protein kinase, cAMP-dependent, catalytic, beta
PSAT1	phosphoserine aminotransferase 1
PSIP1	PC4 and SFRS1 interacting protein 1
PTDSS1	phosphatidylserine synthase 1
PTPRD	protein tyrosine phosphatase, receptor type, D
PTPRE	protein tyrosine phosphatase, receptor type, E
PTPRG	protein tyrosine phosphatase, receptor type, G
PTTG2	pituitary tumor-transforming 2
RAB28	RAB28, member RAS oncogene family
RASAL2	RAS protein activator like 2
RASGRF1	Ras protein-specific guanine nucleotide-releasing factor 1
RBMS3	RNA binding motif, single stranded interacting protein 3
RDX	radixin
RFC3	replication factor C (activator 1) 3, 38kDa
RIT2	Ras-like without CAAX 2
RORA	RAR-related orphan receptor A
RPS17	ribosomal protein S17
RPS24	ribosomal protein S24
RUNX1T1	runt-related transcription factor 1; translocated to, 1 (cyclin D-related)
RYBP	RING1 and YY1 binding protein
SATB1	SATB homeobox 1
SCGB1A1	secretoglobin, family 1A, member 1 (uteroglobin)
SCML1	sex comb on midleg-like 1 (Drosophila)
SH3GL2	SH3-domain GRB2-like 2
SIM1	single-minded homolog 1 (Drosophila)
SKAP2	src kinase associated phosphoprotein 2
SLC10A2	solute carrier family 10 (sodium/bile acid cotransporter family), member 2
SLC14A2	solute carrier family 14 (urea transporter), member 2
SLC30A4	solute carrier family 30 (zinc transporter), member 4
SLC39A1	solute carrier family 39 (zinc transporter), member 1
SLC4A3	solute carrier family 4, anion exchanger, member 3

SMARCA1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 1
SMNDC1	survival motor neuron domain containing 1
SNRPB2	small nuclear ribonucleoprotein polypeptide B
SOX14	SRY (sex determining region Y)-box 14
SOX3	SRY (sex determining region Y)-box 3
SOX9	SRY (sex determining region Y)-box 9
SQRDL	sulfide quinone reductase-like (yeast)
SRP54	signal recognition particle 54kDa
SRSF7	serine/arginine-rich splicing factor 7
SS18	synovial sarcoma translocation, chromosome 18
SSTR2	somatostatin receptor 2
STAP1	signal transducing adaptor family member 1
STX11	syntaxin 11
TCF4	transcription factor 4
TEAD1	TEA domain family member 1 (SV40 transcriptional enhancer factor)
TGFBR2	transforming growth factor, beta receptor II (70/80kDa)
TIMM9	translocase of inner mitochondrial membrane 9 homolog (yeast)
TRIM32	tripartite motif-containing 32
TXNL1	thioredoxin-like 1
USP2	ubiquitin specific peptidase 2
USP3	ubiquitin specific peptidase 3
UST	uronyl-2-sulfotransferase
UTRN	utrophin
WNT5A	wingless-type MMTV integration site family, member 5A
YKT6	YKT6 v-SNARE homolog (S. cerevisiae)
ZEB2	zinc finger E-box binding homeobox 2
ZFHX3	zinc finger homeobox 3
ZIC1	Zic family member 1 (odd-paired homolog, Drosophila)
ZMYM5	zinc finger, MYM-type 5

A.3 Candidate target genes directly regulated by miR-17-92 cluster extracted from GSE5550

Official gene symbol	Gene title
ABCA1	ATP-binding cassette, sub-family A (ABC1), member 1
ABCB7	ATP-binding cassette, sub-family B (MDR/TAP), member 7
ABR	active BCR-related
ACTC1	actin, alpha, cardiac muscle 1
ACTN1	actinin, alpha 1
ADCY3	adenylate cyclase 3
ADD3	adducin 3 (gamma)
ADM	adrenomedullin
ADRB1	adrenoceptor beta 1
ADSS	adenylosuccinate synthase
AFF1	AF4/FMR2 family, member 1
AKAP11	A kinase (PRKA) anchor protein 11
ANXA7	annexin A7
ARFGEF1	ADP-ribosylation factor guanine nucleotide-exchange factor 1 (brefeldin A-inhibited)
ARFIP1	ADP-ribosylation factor interacting protein 1
ARID4B	AT rich interactive domain 4B (RBP1-like)
ASAP2	ArfGAP with SH3 domain, ankyrin repeat and PH domain 2
ASNA1	arsA arsenite transporter, ATP-binding, homolog 1 (bacterial)
ATP2B2	ATPase, Ca++ transporting, plasma membrane 2
ATP2C1	ATPase, Ca++ transporting, type 2C, member 1
ATP6V1B2	ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B2
ATXN1	ataxin 1
BAMBI	BMP and activin membrane-bound inhibitor
BAZ2B	bromodomain adjacent to zinc finger domain, 2B
BCAT2	branched chain amino-acid transaminase 2, mitochondrial
BCL3	B-cell CLL/lymphoma 3
BCL7A	B-cell CLL/lymphoma 7A
BNIP2	BCL2/adenovirus E1B 19kDa interacting protein 2
BTG1	B-cell translocation gene 1, anti-proliferative
BTG3	BTG family, member 3
CACNA1C	calcium channel, voltage-dependent, L type, alpha 1C subunit
CAST	calpastatin
CCL1	chemokine (C-C motif) ligand 1

CCNL1	cyclin L1
CCNT2	cyclin T2
CD69	CD69 molecule
CDKN1C	cyclin-dependent kinase inhibitor 1C (p57, Kip2)
CDS1	CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 1
CHAF1A	chromatin assembly factor 1, subunit A (p150)
CHKA	choline kinase alpha
CHST1	carbohydrate (keratan sulfate Gal-6) sulfotransferase 1
CHST7	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 7
CLIP1	CAP-GLY domain containing linker protein 1
CLOCK	clock circadian regulator
CLTC	clathrin, heavy chain (Hc)
CNN1	calponin 1, basic, smooth muscle
CNOT7	CCR4-NOT transcription complex, subunit 7
CNTFR	ciliary neurotrophic factor receptor
COL1A2	collagen, type I, alpha 2
COL4A3	collagen, type IV, alpha 3 (Goodpasture antigen)
CR2	complement component (3d/Epstein Barr virus) receptor 2
CROT	carnitine O-octanoyltransferase
CTGF	connective tissue growth factor
CTSA	cathepsin A
DBN1	drebrin 1
DCUN1D1	DCN1, defective in cullin neddylation 1, domain containing 1
DDX3X	DEAD (Asp-Glu-Ala-Asp) box helicase 3, X-linked
DDX3Y	DEAD (Asp-Glu-Ala-Asp) box helicase 3, Y-linked
DDX5	DEAD (Asp-Glu-Ala-Asp) box helicase 5
DHRS3	dehydrogenase/reductase (SDR family) member 3
DICER1	dicer 1, ribonuclease type III
DLC1	DLC1 Rho GTPase activating protein
DNAI1	dynein, axonemal, intermediate chain 1
DNAJB1	DnaJ (Hsp40) homolog, subfamily B, member 1
DRD1	dopamine receptor D1
EIF4G2	eukaryotic translation initiation factor 4 gamma, 2
ELK3	ELK3, ETS-domain protein (SRF accessory protein 2)
ENC1	ectodermal-neural cortex 1 (with BTB domain)
EPHB3	EPH receptor B3
EPS15	epidermal growth factor receptor pathway substrate 15
ERBB3	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 3
ETV5	ets variant 5
F3	coagulation factor III (thromboplastin, tissue factor)

FASTK	Fas-activated serine/threonine kinase
FBN1	fibrillin 1
FGL2	fibrinogen-like 2
FHL2	four and a half LIM domains 2
FLNC	filamin C, gamma
FOSL1	FOS-like antigen 1
FOXF2	forkhead box F2
FRY	furry homolog (Drosophila)
GABRB2	gamma-aminobutyric acid (GABA) A receptor, beta 2
GALNT6	polypeptide N-acetylgalactosaminyltransferase 6
GAP43	growth associated protein 43
GATA2	GATA binding protein 2
GFPT1	glutamine--fructose-6-phosphate transaminase 1
GJA1	gap junction protein, alpha 1, 43kDa
GNB5	guanine nucleotide binding protein (G protein), beta 5
GOLGA4	golgin A4
GPR137B	G protein-coupled receptor 137B
GPR6	G protein-coupled receptor 6
GRK6	G protein-coupled receptor kinase 6
GRSF1	G-rich RNA sequence binding factor 1
GTF2H5	general transcription factor IIH, polypeptide 5
HBP1	HMG-box transcription factor 1
HIF1A	hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)
HIP1	huntingtin interacting protein 1
HIPK3	homeodomain interacting protein kinase 3
HNF1B	HNF1 homeobox B
HOXD1	homeobox D1
HPRT1	hypoxanthine phosphoribosyltransferase 1
HSPA2	heat shock 70kDa protein 2
ID2	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein
IDH1	isocitrate dehydrogenase 1 (NADP+), soluble
IGF1	insulin-like growth factor 1 (somatomedin C)
IGF2R	insulin-like growth factor 2 receptor
IGFBP3	insulin-like growth factor binding protein 3
IGSF3	immunoglobulin superfamily, member 3
IMPDH1	IMP (inosine 5'-monophosphate) dehydrogenase 1
INADL	InaD-like (Drosophila)
INSIG1	insulin induced gene 1
INTS6	integrator complex subunit 6

IPP	intracisternal A particle-promoted polypeptide
IRF2	interferon regulatory factor 2
ITGB8	integrin, beta 8
ITPR1	inositol 1,4,5-trisphosphate receptor, type 1
IVNS1ABP	influenza virus NS1A binding protein
KCNA4	potassium voltage-gated channel, shaker-related subfamily, member 4
KCNJ2	potassium inwardly-rectifying channel, subfamily J, member 2
KIF23	kinesin family member 23
KLF10	Kruppel-like factor 10
KLF13	Kruppel-like factor 13
KLF2	Kruppel-like factor 2
KPNA2	karyopherin alpha 2 (RAG cohort 1, importin alpha 1)
LBX1	ladybird homeobox 1
LDLR	low density lipoprotein receptor
LRP12	low density lipoprotein receptor-related protein 12
LRP2	low density lipoprotein receptor-related protein 2
MACF1	microtubule-actin crosslinking factor 1
MAGI2	membrane associated guanylate kinase, WW and PDZ domain containing 2
MAN1C1	mannosidase, alpha, class 1C, member 1
MAN2A1	mannosidase, alpha, class 2A, member 1
MAP2K3	mitogen-activated protein kinase kinase 3
MAP3K8	mitogen-activated protein kinase kinase kinase 8
MAP4K3	mitogen-activated protein kinase kinase kinase kinase 3
MAPK4	mitogen-activated protein kinase 4
MAPK6	mitogen-activated protein kinase 6
MED17	mediator complex subunit 17
MFHAS1	malignant fibrous histiocytoma amplified sequence 1
MID1	midline 1
MORF4L1	mortality factor 4 like 1
MPPED2	metallophosphoesterase domain containing 2
MRPL17	mitochondrial ribosomal protein L17
MSMO1	methylsterol monooxygenase 1
MYCN	v-myc avian myelocytomatosis viral oncogene neuroblastoma derived homolog
MYLIP	myosin regulatory light chain interacting protein
NAGK	N-acetylglucosamine kinase
NCOA1	nuclear receptor coactivator 1
NCOA3	nuclear receptor coactivator 3
NEFL	neurofilament, light polypeptide

NEFM	neurofilament, medium polypeptide
NEUROD1	neuronal differentiation 1
NME7	NME/NM23 family member 7
NOTCH2	notch 2
NOX4	NADPH oxidase 4
NPAS2	neuronal PAS domain protein 2
NR3C2	nuclear receptor subfamily 3, group C, member 2
NRBP1	nuclear receptor binding protein 1
NSF	N-ethylmaleimide-sensitive factor
NUP54	nucleoporin 54kDa
PAK6	p21 protein (Cdc42/Rac)-activated kinase 6
PCDHA10	protocadherin alpha 10
PCDHA2	protocadherin alpha 2
PCDHA6	protocadherin alpha 6
PCOLCE2	procollagen C-endopeptidase enhancer 2
PDCD1LG2	programmed cell death 1 ligand 2
PDE4D	phosphodiesterase 4D, cAMP-specific
PFKP	phosphofructokinase, platelet
PHF20	PHD finger protein 20
PI15	peptidase inhibitor 15
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha
PLA2G10	phospholipase A2, group X
PLAA	phospholipase A2-activating protein
PLXNC1	plexin C1
PNRC1	proline-rich nuclear receptor coactivator 1
PON2	paraoxonase 2
POSTN	periostin, osteoblast specific factor
PPP1R12A	protein phosphatase 1, regulatory subunit 12A
PPP6C	protein phosphatase 6, catalytic subunit
PRC1	protein regulator of cytokinesis 1
PRKAA1	protein kinase, AMP-activated, alpha 1 catalytic subunit
PRRG1	proline rich Gla (G-carboxyglutamic acid) 1
PRRG4	proline rich Gla (G-carboxyglutamic acid) 4 (transmembrane)
PSAP	prosaposin
PSD	pleckstrin and Sec7 domain containing
PTEN	phosphatase and tensin homolog
PTPN4	protein tyrosine phosphatase, non-receptor type 4 (megakaryocyte)
PTPRD	protein tyrosine phosphatase, receptor type, D
PTPRG	protein tyrosine phosphatase, receptor type, G
RAB23	RAB23, member RAS oncogene family

RAB8B	RAB8B, member RAS oncogene family
RAD21	RAD21 homolog (S. pombe)
RAF1	Raf-1 proto-oncogene, serine/threonine kinase
RAP1A	RAP1A, member of RAS oncogene family
RAP2C	RAP2C, member of RAS oncogene family
RAPGEF2	Rap guanine nucleotide exchange factor (GEF) 2
RAPGEF4	Rap guanine nucleotide exchange factor (GEF) 4
RASA1	RAS p21 protein activator (GTPase activating protein) 1
RBBP8	retinoblastoma binding protein 8
RBMS1	RNA binding motif, single stranded interacting protein 1
RFX1	regulatory factor X, 1 (influences HLA class II expression)
RFX4	regulatory factor X, 4 (influences HLA class II expression)
RGS3	regulator of G-protein signaling 3
RHOB	ras homolog family member B
RNF19A	ring finger protein 19A, RBR E3 ubiquitin protein ligase
RNF2	ring finger protein 2
RNF4	ring finger protein 4
RNF6	ring finger protein (C3H2C3 type) 6
RPS6KA2	ribosomal protein S6 kinase, 90kDa, polypeptide 2
RSRC2	arginine/serine-rich coiled-coil 2
RTN1	reticulon 1
S1PR1	sphingosine-1-phosphate receptor 1
SAR1B	secretion associated, Ras related GTPase 1B
SCAF11	SR-related CTD-associated factor 11
SCN1B	sodium channel, voltage-gated, type I, beta subunit
SDC1	syndecan 1
SEC11A	SEC11 homolog A (S. cerevisiae)
SEC22A	SEC22 vesicle trafficking protein homolog A (S. cerevisiae)
SERINC3	serine incorporator 3
SERTAD3	SERTA domain containing 3
SGK1	serum/glucocorticoid regulated kinase 1
SIRT5	sirtuin 5
SLC24A3	solute carrier family 24 (sodium/potassium/calcium exchanger), member 3
SLC25A12	solute carrier family 25 (aspartate/glutamate carrier), member 12
SLC26A4	solute carrier family 26 (anion exchanger), member 4
SLC31A2	solute carrier family 31 (copper transporter), member 2
SLC6A8	solute carrier family 6 (neurotransmitter transporter), member 8
SLC9A6	solute carrier family 9, subfamily A (NHE6, cation proton antiporter 6), member 6
SMARCA2	SWI/SNF related, matrix associated, actin dependent regulator of

	chromatin, subfamily a, member 2
SMARCD2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 2
SNX3	sorting nexin 3
SOCS1	suppressor of cytokine signaling 1
SOCS3	suppressor of cytokine signaling 3
SOX21	SRY (sex determining region Y)-box 21
SOX4	SRY (sex determining region Y)-box 4
SPHK2	sphingosine kinase 2
SPOCK1	sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 1
SRPK2	SRSF protein kinase 2
SRPR	signal recognition particle receptor (docking protein)
ST3GAL5	ST3 beta-galactoside alpha-2,3-sialyltransferase 5
STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)
SYT1	synaptotagmin I
TACC2	transforming, acidic coiled-coil containing protein 2
TAF4	TAF4 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 135kDa
TBK1	TANK-binding kinase 1
TBR1	T-box, brain, 1
TCFL5	transcription factor-like 5 (basic helix-loop-helix)
TFDP2	transcription factor Dp-2 (E2F dimerization partner 2)
TGFBR2	transforming growth factor, beta receptor II (70/80kDa)
TGFBR3	transforming growth factor, beta receptor III
TGIF1	TGFB-induced factor homeobox 1
TGM3	transglutaminase 3
THBS1	thrombospondin 1
TMEM2	transmembrane protein 2
TNFAIP3	tumor necrosis factor, alpha-induced protein 3
TNFRSF12A	tumor necrosis factor receptor superfamily, member 12A
TNKS	tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase
TNKS2	tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase 2
TRAK2	trafficking protein, kinesin binding 2
TRIM23	tripartite motif containing 23
TRIM3	tripartite motif containing 3
TRIM8	tripartite motif containing 8
TRIP10	thyroid hormone receptor interactor 10
TRIP11	thyroid hormone receptor interactor 11
TRPS1	trichorhinophalangeal syndrome I

TRPV6	transient receptor potential cation channel, subfamily V, member 6
TSG101	tumor susceptibility 101
TUB	tubby bipartite transcription factor
TXK	TXK tyrosine kinase
UBE2B	ubiquitin-conjugating enzyme E2B
UBL3	ubiquitin-like 3
UGP2	UDP-glucose pyrophosphorylase 2
URI1	URI1, prefoldin-like chaperone
USP6	ubiquitin specific peptidase 6
VLDLR	very low density lipoprotein receptor
VPS4B	vacuolar protein sorting 4 homolog B (S. cerevisiae)
VSX1	visual system homeobox 1
WBP4	WW domain binding protein 4
WDR1	WD repeat domain 1
WDR44	WD repeat domain 44
WNT3	wingless-type MMTV integration site family, member 3
WRNIP1	Werner helicase interacting protein 1
YTHDF2	YTH domain family, member 2
ZFPM2	zinc finger protein, FOG family member 2
ZFYVE9	zinc finger, FYVE domain containing 9
ZMYND11	zinc finger, MYND-type containing 11
ZNF217	zinc finger protein 217
ZNF287	zinc finger protein 287

References

A,J. *et al*. (2010) Chronic myeloid leukemia patients sensitive and resistant to imatinib treatment show different metabolic responses. *PLoS One*, **5**, e13186.

Abe,Y. *et al*. (2004) Mini review activin receptor signaling. *Growth Factors*, **22**, 105–110.

Aguayo,A. *et al*. (2000) Angiogenesis in acute and chronic leukemias and myelodysplastic syndromes. *Blood*, **96**, 2240–2245.

Aguda,B.D. *et al*. (2008) MicroRNA regulation of a cancer network: Consequences of the feedback loops involving miR-17-92, E2F, and Myc. *Proc. Natl. Acad. Sci. U S A*, **105**, 19678–19683.

Albano,F. *et al*. (2013) Gene expression profiling of chronic myeloid leukemia with variant t(9;22) reveals a different signature from cases with classic translocation. *Mol. Cancer*, **12**, 36.

Allocco,D.J. *et al*. (2004) Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, **5**, 18.

Altman,R.B. and Raychaudhuri,S. (2001) Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.*, **11**, 340–347.

Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Attisano,L. *et al*. (1996) Activation of signalling by the activin receptor complex. *Mol. Cell. Biol.*, **16**, 1066–1073.

Baek,D. *et al*. (2008) The impact of microRNAs on protein output. *Nature*, **455**, 64–71.

Bagga,S. *et al.* (2005) Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell*, **122**, 553–563.

Bakker,W.J. *et al.* (2013) HIF proteins connect the RB-E2F factors to angiogenesis. *Transcription*, **4**, 62–66.

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

Bernstein,E. *et al.* (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, **409**, 363–366.

Bhagavathi,S. and Czader,M. (2010) MicroRNAs in benign and malignant hematopoiesis. *Arch. Pathol. Lab. Med.*, **134**, 1276–1281.

Bhatia,R. *et al*. (1996) Interferon-alpha restores normal beta 1 integrin-mediated inhibition of hematopoietic progenitor proliferation by the marrow microenvironment in chronic myelogenous leukemia. *Blood*, **87**, 3883–3891.

Blanchette,M. *et al.* (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.*, **16**, 656–668.

Brady,S.N. *et al*. (2004) ARF impedes NPM/B23 shuttling in an Mdm2-sensitive tumor suppressor pathway. *Mol. Cell. Biol.*, **24**, 9327–9338.

Brazma,A. *et al.* (1998) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, **8**, 1202–1215.

Brentani,H. *et al*. (2003) The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc. Natl. Acad. Sci. U S A*, **100**, 13418–13423.

Breslin,T. *et al.* (2005) Signal transduction pathway profiling of individual tumor samples. *BMC Bioinformatics*, **6**, 163.

Burdette,J.E. *et al*. (2005) Activin A mediates growth inhibition and cell cycle arrest through Smads in human breast cancer cells. *Cancer Res.*, **65**, 7968–7975.

Calin,G.A. *et al*. (2004) MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proc. Natl. Acad. Sci. U S A*, **101**, 11755–11760.

Cammarata,G. *et al*. (2010) Differential expression of specific microRNA and their targets in acute myeloid leukemia. *Am. J. Hematol.*, **85**, 331–339.

Carmell,M.A. and Hannon,G.J. (2004) RNase III enzymes and the initiation of gene silencing. *Nat. Struct. Mol. Biol.*, **11**, 214–218.

Chan,L.W. *et al*. (2012) Genomic sequence analysis of EGFR regulation by microRNAs in lung cancer. *Curr. Top. Med. Chem.*, **12**, 920–926.

Chan,W.Y. *et al*. (1989) Characterization of the cDNA encoding human nucleophosmin and studies of its role in normal and abnormal growth. *Biochemistry*, **28**, 1033–1039.

Chen,C.Y. *et al*. (2011) Coregulation of transcription factors and microRNAs in human transcriptional regulatory network. *BMC Bioinformatics*, **12**, S41.

Cho,W.C. (2007) OncomiRs: the discovery and progress of microRNAs in cancers. *Mol. Cancer*, **6**, 60.

Chong,J.L. *et al*. (2009) E2f1-3 switch from activators in progenitor cells to repressors in differentiating cells. *Nature*, **462**, 930–934.

Chou,Y.T. *et al*. (2010) EGFR promotes lung tumorigenesis by activating miR-7 through a Ras/ERK/Myc pathway that targets the Ets2 transcriptional repressor ERF. *Cancer Res.*, **70**, 8822–8831.

Cines,D.B. *et al*. (1998) Endothelial cells in physiology and in the pathophysiology of vascular disorders. *Blood*, **91**, 3527–3561.

Coller,H.A. *et al*. (2007) ‘‘Myc’ed messages’’: myc induces transcription of E2F1 while inhibiting its translation via a microRNA polycistron. *PLoS Genet.*, **3**, 1319–1324.

Cui,Q. *et al*. (2007) MicroRNAs preferentially target the genes with high transcriptional regulation complexity. *Biochem. Biophys. Res. Commun.*, **352**, 733–738.

Dalenc,F. *et al*. (2002) Increased expression of a COOH-truncated nucleophosmin resulting from alternative splicing is associated with cellular resistance to ionizing radiation in HeLa cells. *Int. J. Cancer*, **100**, 662–668.

Deli,A. *et al*. (2008) Activins and activin antagonists in hepatocellular carcinoma. *World J. Gastroenterol.*, **14**, 1699–1709.

Denkert,C. *et al*. (2008) Metabolite profiling of human colon carcinoma-deregulation of TCA cycle and amino acid turnover. *Mol. Cancer*, **7**, 72.

Denli,A.M. *et al*. (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature*, **432**, 231–235.

Diaz-Blanco,E. *et al*. (2007) Molecular signature of CD34(+) hematopoietic stem and progenitor cells of patients with CML in chronic phase. *Leukemia*, **21**, 494–504.

Dudoit,S. *et al*. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U S A*, **95**, 14863–14868.

Elo,L.L. *et al*. (2007) Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. *Bioinformatics*, **23**, 2096–2103.

Faderl,S. *et al*. (1999) The biology of chronic myeloid leukemia. *N. Engl. J. Med.*, **341**, 164–172.

Falini,B. *et al*. (2005) Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N. Engl. J. Med.*, **352**, 254–266.

Farkas,I.J. *et al*. (2006) Topological basis of signal integration in the transcriptional-regulatory network of the yeast, Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 478.

Federici,L and Falini,B. (2013) Nucleophosmin mutations in acute myeloid leukemia: a tale of protein unfolding and mislocalization. *Protein Sci.*, **22**, 545–556.

Ferretti,V. *et al*. (2007) PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res.*, **35**, D122–D126.

Frazer,R. *et al*. (2007) Chronic myeloid leukaemia in the 21st century. *Ulster Med. J.*, **76**, 8–17.

Frehlick,L.J. *et al*. (2007) New insights into the nucleophosmin/nucleoplasmin family of nuclear chaperones. *Bioessays*, **29**, 49–59.

Fu,J. *et al*. (2012) Identifying microRNA-mRNA regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis. *BMC Syst. Biol.*, **6**, 68.

Fuller,T.F. *et al*. (2007) Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm. Genome*, **18**, 463–472.

Garzon,R. *et al*. (2008) MicroRNA signatures associated with cytogenetics and prognosis in acute myeloid leukemia. *Blood*, **111**, 3183–3189.

Gene Ontology Consortium (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.

Gordon,J.E. *et al*. (2013) MicroRNAs in myeloid malignancies. *Br. J. Haematol.*, **162**, 162–176.

Grisendi,S. *et al*. (2006) Nucleophosmin and cancer. *Nat. Rev. Cancer*, **6**, 493–505.

Gupta,A. *et al*. (2006) Elucidation of directionality for co-expressed genes: predicting intra-operon termination sites. *Bioinformatics*, **22**, 209–214.

Haferlach,T. *et al*. (2005) Global approach to the diagnosis of leukemia using gene expression profiling. *Blood*, **106**, 1189–1198.

Hagner,P.R. *et al*. (2010) Targeting the translational machinery as a novel treatment strategy for hematologic malignancies. *Blood*, **115**, 2127–2135.

Hammond,S.M. (2006) MicroRNAs as oncogenes. *Curr. Opin. Genet. Dev.*, **16**, 4–9.

Han,J. *et al*. (2004) The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev.*, **18**, 3016–3027.

Haskins,W.E. *et al*. (2011) MRCQuant– an accurate LC-MS relative isotopic quantification algorithm on TOF instruments. *BMC Bioinformatics*, **12**, 74.

Hazar-Rethinam,M. *et al*. (2011) The role of the E2F transcription factor family in UV-induced apoptosis. *Int. J. Mol. Sci.*, **12**, 8947–8960.

He,T.C. *et al*. (1998) Identification of c-MYC as a target of the APC pathway. *Science*, **281**, 1509–1512.

Himanen,J.P. and Nikolov,D.B. (2003) Eph receptors and ephrins. *Int. J. Biochem. Cell Biol.*, **35**, 130–134.

Himanen,J.P. *et al*. (2007) Cell-cell signaling via Eph receptors and ephrins. *Curr. Opin. Cell Biol.*, **19**, 534–542.

Hnisz,D. *et al*. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.

Horvath,S. and Dong,J. (2008) Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.*, **4**, e1000117.

Hosack,D.A. *et al*. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol*., 4, R70.

Huang,D.W. *et al*. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Huang,K. *et al*. (2010) MicroRNA roles in beta-catenin pathway. *Mol. Cancer*, **9**, 252.

Hussong,J.W. *et al*. (2000) Evidence of increased angiogenesis in patients with acute myeloid leukemia. *Blood*, **95**, 309–313.

Ideker,T. *et al*. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Ideker,T. *et al*. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**, S233–S240.

Irmer,D. *et al*. (2007) EGFR kinase domain mutations – functional impact and relevance for lung cancer therapy. *Oncogene*, **26**, 5693–5701.

Itahana,K. *et al.* (2003) Tumor suppressor ARF degrades B23, a nucleolar protein involved in ribosome biogenesis and cell proliferation. *Mol. Cell*, **12**, 1151–1164.

Kalidas,M. *et al*. (2001) Chronic myelogenous leukemia. *JAMA*, **286**, 895–898.

Kapp,A.V. *et al*. (2006) Discovery and validation of breast cancer subtypes. *BMC Genomics*, **7**, 231.

Kumar,M.S. *et al*. (2007) Impaired microRNA processing enhances cellular transformation and tumorigenesis. *Nat. Genet.*, **39**, 673–677.

Leto,G. (2010) Activin A and bone metastasis. *J. Cell. Physiol.*, **225**, 302–309.

Li,L. *et al*. (2010a) Computational approaches for microRNA studies: a review. *Mamm. Genome*, **21**, 1–12.

Li,W. *et al*. (2010b) Characterization of E2F3a function in HepG2 liver cancer cells. *J. Cell. Biochem.*, **111**, 1244–1251.

Lindström,M.S. (2011) NPM1/B23: A multifunctional chaperone in ribosome biogenesis and chromatin remodeling. *Biochem. Res. Int.*, **2011**, 195209.

Lozzio,C.B. and Lozzio,B.B. (1975) Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood*, **45**, 321–334.

Lu,M. *et al.* (2012) Combined effects of E2F1 and E2F2 polymorphisms on risk and early onset of squamous cell carcinoma of the head and neck. *Mol. Carcinog.*, **51**, E132–E141.

Ly,C. *et al*. (2003) Bcr-Abl kinase modulates the translation regulators ribosomal protein S6 and 4E-BP1 in chronic myelogenous leukemia cells via the mammalian target of rapamycin. *Cancer Res.*, **63**, 5716–5722.

Maggi,L.B.Jr. and Weber,J.D. (2005) Nucleolar adaptation in human cancer. *Cancer Invest.*, **23**, 599–608.

Maggi,L.B.Jr. *et al.* (2008) Nucleophosmin serves as a rate-limiting nuclear export chaperone for the mammalian ribosome. *Mol. Cell. Biol.*, **28**, 7050–7065.

Marcucci,G. *et al.* (2011) Molecular genetics of adult acute myeloid leukemia: prognostic and therapeutic implications. *J. Clin. Oncol.*, **29**, 475–486.

Matsuo,S.E. *et al.* (2006) Transforming growth factor-β1 and activin A generate antiproliferative signaling in thyroid cancer cells. *J. Endocrinol.*, **190**, 141–150.

Medina,P.P. and Slack,F.J. (2008) microRNAs and cancer: an overview. *Cell Cycle*, **7**, 2485–2492.

Meister,G. and Tuschl,T. (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature*, **431**, 343–349.

Melo,J.V. and Barnes,D.J. (2007) Chronic myeloid leukaemia as a model of disease evolution in human cancer. *Nat. Rev. Cancer*, **7**, 441–453.

Mendell,J.T. (2008) miRiad roles for the miR-17-92 cluster in development and disease. *Cell*, **133**, 217–222.

Naoe,T. *et al.* (2006) A versatile molecule associated with hematological malignancies. *Cancer Sci.*, **97**, 963–969.

Nowell,P.C. and Hungerford,D.A. (1960) A minute chromosome in human chronic granulocytic leukemia. *Science*, **32**, 1497–1501.

O'Connell,B.C. *et al.* (2003) A large scale genetic analysis of c-Myc-regulated gene expression patterns. *J. Biol. Chem.*, **278**, 12563–12573.

Palaiologou,M. *et al.* (2012) E2F-1 is overexpressed and pro-apoptotic in human hepatocellular carcinoma. *Virchows Arch.*, **460**, 439–446.

Panopoulou,E. *et al.* (2005) Activin A suppresses neuroblastoma xenograft tumor growth via antimitotic and antiangiogenic mechanisms. *Cancer Res.*, **65**, 1877–1886.

Pelengaris,S. and Khan,M. (2003) The many faces of c-MYC. *Arch. Biochem. Biophys.*, **416**, 129–136.

Pelengaris,S. *et al*. (2000) Action of Myc in vivo - proliferation and apoptosis. *Curr. Opin. Genet. Dev.*, **10**, 100–105.

Pelicano,H. *et al*. (2006) Glycolysis inhibition for anticancer treatment. *Oncogene*, **25**, 4633–4646.

Pelletier,C.L. *et al*. (2007) TSC1 sets the rate of ribosome export and protein synthesis through nucleophosmin translation. *Cancer Res.*, **67**, 1609–1617.

Perrotti,D. and Neviani,P. (2007) From mRNA metabolism to cancer therapy: chronic myelogenous leukemia shows the way. *Clin. Cancer Res.*, **13**, 1638–1642.

Piccaluga,P.P. *et al*. (2009) Cytoplasmic mutated nucleophosmin (NPM1) in blast crisis of chronic myeloid leukaemia. *Leukemia*, **23**, 1370–1371.

Press,R.D. *et al*. (2006) BCR-ABL mRNA levels at and after the time of a complete cytogenetic response (CCR) predict the duration of CCR in imatinib mesylate–treated patients with CML. *Blood*, **107**, 4250–4256.

Rowley,J.D. (1973) Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, **243**, 290–293.

Salesse,S. and Verfaillie,C.M. (2002) Mechanisms underlying abnormal tracking and expansion of malignant progenitors in CML: BCR/ABL-induced defects in integrin function in CML. *Oncogene*, **21**, 8605–8611.

Salvatori,B. *et al*. (2012) The microRNA-26a target E2F7 sustains cell proliferation and inhibits monocytic differentiation of acute myeloid leukemia cells. *Cell Death Dis.*, **3**, e413.

Scagliotti,G.V. *et al*. (2004) The biology of epidermal growth factor receptor in lung cancer. *Clin. Cancer Res.*, **10**, 4227s–4232s.

Schulze,A. and Downward,J. (2001) Navigating gene expression using microarrays–a technology review. *Nat. Cell Biol.*, **3**, E190–E195.

Schwemmle,S. and Pfeifer,G.P. (2000) Genomic structure and mutation screening of the E2F4 gene in human tumors. *Int. J. Cancer*, **86**, 672–677.

Seggerson,K. *et al*. (2002) Two genetic circuits repress the Caenorhabditis elegans heterochronic gene lin-28 after translation initiation. *Dev. Biol.*, **243**, 215–225.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Series B Stat. Methodol.*, **64**, 479–498.

Stuart,J.M. *et al*. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U S A*, **102**, 15545–15550.

Szebeni,A. and Olson,M.O. (1999) Nucleolar protein B23 has molecular chaperone activities. *Protein Sci.*, **8**, 905–912.

Tavazoie,S. *et al*. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Tetsu,O. and McCormick,F. (1999) Beta-catenin regulates expression of cyclin D1 in colon carcinoma cells. *Nature*, **398**, 422–426.

Timmers,C. *et al*. (2007) E2f1, E2f2, and E2f3 control E2F target expression and cellular proliferation via a p53-dependent negative feedback loop. *Mol. Cell. Biol.*, **27**, 65–78.

Topisirovic,I. *et al*. (2003) Aberrant eukaryotic translation initiation factor 4E-dependent mRNA transport impedes hematopoietic differentiation and contributes to leukemogenesis. *Mol. Cell. Biol.*, **23**, 8992–9002.

Torkamani,A. *et al*. (2010) Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. *Genome Res.*, **20**, 403–412.

Tukey,J.W. (1958) Bias and confidence in not quite large samples. *Ann. Math. Stat.*, **29**, 614.

Valenta,T. *et al*. (2003) HMG box transcription factor TCF-4's interaction with CtBP1 controls the expression of the Wnt target Axin2/Conductin in human embryonic kidney cells. *Nucleic Acids Res.*, **31**, 2369–2380.

Varambally,S. *et al*. (2008) Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer. *Science*, **322**, 1695–1699.

Vardiman,J.W. *et al*. (2002) The World Health Organization (WHO) classification of the myeloid neoplasms. *Blood*, **100**, 2292–2302.

Venturini,L. *et al*. (2007) Expression of the miR-17-92 polycistron in chronic myeloid leukemia (CML) CD34+ cells. *Blood*, **109**, 4399–4405.

Wadlow,R. and Ramaswamy,S. (2005) DNA microarrays in clinical cancer research. *Curr. Mol. Med.*, **5**, 111–120.

Wang,F. *et al*. (2014) Multiple regression analysis of mRNA-miRNA associations in colorectal cancer pathway. *Biomed Res. Int.*, **2014**, 676724.

Warner,J.R. (1990) The nucleolus and ribosome formation. *Curr. Opin. Cell Biol.*, **2**, 521–527.

Watkins,D.B. *et al.* (2013) NPM1 mutations occur rarely or not at all in chronic myeloid leukaemia patients in chronic phase or blast crisis. *Leukemia*, **27**, 489–490.

Whyte,W.A. *et al*. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307–319.

Wolfsberg,T.G. *et al*. (1999) Candidate regulatory sequence elements for cell cycle-dependent transcription in Saccharomyces cerevisiae. *Genome Res.*, **9**, 775–792.

Wu,L. and Belasco,J.G. (2005) Micro-RNA regulation of the mammalian lin-28 gene during neuronal differentiation of embryonal carcinoma cells. *Mol. Cell. Biol.*, **25**, 9198–9208.

Wu,L. *et al*. (2001) The E2F1-3 transcription factors are essential for cellular proliferation. *Nature*, **414**, 457–462.

Yao,Z. *et al*. (2010). B23 acts as a nucleolar stress sensor and promotes cell survival through its dynamic interaction with hnRNPU and hnRNPA1. *Oncogene*, **29**, 1821–1834.

Yi,R. *et al*. (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short Hairpin RNAs. *Genes Dev.*, **17**, 3011–3016.

Yu,H. *et al*. (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet.*, **19**, 422–427.

Yu,J. *et al*. (2006a) Human microRNA clusters: genomic organization and expression profile in leukemia cell lines. *Biochem. Biophys. Res. Commun.*, **349**, 59–68.

Yu,J. *et al*. (2007) Integrative genomics analysis reveals silencing of β-adrenergic signaling by polycomb in prostate cancer. *Cancer Cell*, **12**, 419–431.

Yu,Y. *et al*. (2006b) Nucleophosmin is essential for ribosomal protein L5 nuclear export. *Mol. Cell. Biol.*, **26**, 3798–3809.

Zhan,L. *et al*. (2014) Promising roles of mammalian E2Fs in hepatocellular carcinoma. *Cell. Signal.*, **26**, 1075–1081.

Zhang,M. *et al*. (2008) Inhibition of polysome assembly enhances imatinib activity against chronic myelogenous leukemia and overcomes imatinib resistance. *Mol. Cell. Biol.*, **28**, 6496–6509.

Zhu,D. *et al*. (2005) Network constrained clustering for gene microarray data. *Bioinformatics*, **21**, 4014–4020.