Ph.D. DISSERTATION

Informatics techniques for associating key genes and phenotypes by network-based integration of multi-omics data

멀티오믹스 데이터의 통합 분석을 위한 네트워크 기반의 주요 유전자와 표현형을 연결하기 위한 정보학 기법

AUGUST 2019

INTERDISCIPLINARY PROGRAM IN BIOINFORMATICS
COLLEGE OF NATURAL SCIENCES
SEOUL NATIONAL UNIVERSITY

Chai-Jin Lee

Ph.D. DISSERTATION

Informatics techniques for associating key genes and phenotypes by network-based integration of multi-omics data

멀티오믹스 데이터의 통합 분석을 위한 네트워크 기반의 주요 유전자와 표현형을 연결하기 위한 정보학 기법

AUGUST 2019

INTERDISCIPLINARY PROGRAM IN BIOINFORMATICS
COLLEGE OF NATURAL SCIENCES
SEOUL NATIONAL UNIVERSITY

Chai-Jin Lee

Informatics techniques for associating key genes and phenotypes by network-based integration of multi-omics data

멀티오믹스 데이터의 통합 분석을 위한 네트워크 기반의 주요 유전자와 표현형을 연결하기 위한 정보학 기법

지도교수 김 선

이 논문을 이학박사 학위논문으로 제출함

2019 년 5 월

서울대학교 대학원

협동과정 생물정보학

이 채 진

이채진의 이학박사 학위논문을 인준함

2019 년 6 월

위 원 장	이병재
부위원장	김 선
위 원	손현석
위 원	황대희
위 원	김광수

Abstract

For the functional study of a gene, amplification or knock-out of a gene in the animal model is frequently performed. Experiments of this type are effective in associating a gene to a phenotype. To investigate further, measuring multi-omics data is a common practice. Analyzing such multi-omics data would explain how the gene of interest affects other genes, including regulatory mechanisms such as transcription factors, miRNA and epigenetic changes. However, analyzing multi-omics data is challenging since the integrated analysis of multi-omics data requires analyzing complex associations among genetic and epigenetic entities. To handle such a complex relationship, networks are the most effective tools. Thus, in my doctoral study, I developed network-based informatics techniques for associating key genes and phenotypes by analyzing multi-omics data.

In my first study, I investigated the genetic phenomenon caused by the knock-out gene EWS. MicroRNA data and mRNA expression data from the spinal cord of wildtype and EWS knock-out mice were analyzed and integrated. I used a negative-correlation network of miRNAs and target genes, and protein-protein interaction (PPI) network to investigate functional changes of DEGs. From the network analysis, I identified significantly down-regulated Gnai1 in the cholinergic synapse pathway. Gnai1 was suppressed by mmu-miR-381 and mmu-miR-181a/b/c, and inhibited by Rgs1 and Rgs19 in the spinal cord of EWS KO mice. In addition, the expression levels of Gnb1, Gnb2, and Gnb4, that are forming a G-protein complex with Gnai1 gene, were reduced.

In my second study, I investigated the effect of mutations in seven DNA methylation modifier genes on gene expression profiles on the genome scale in

cancer. Pan-cancer data were collected from TCGA, and 3865 samples having both transcriptome and methylome data were analyzed. In each carcinoma, samples were divided into two sample groups, one with mutations and the other without mutations in the seven DNA methylation modifier genes. First, genome-wide promoter methylation landscapes were significantly different between the two groups and differentially methylated regions (DMR) were identified. To investigate how DMRs affected genome-wide gene expression profiles, I first selected differentially expressed genes (DEG) between the two groups of samples. Then, DEGs were mapped to PPI and clusters of DEGs were computed to select gene sets in terms of biological functions. To associate DEG and DMR, I selected two cancers, AML and COAD, since the two cancers were most different in terms of mutation profiles of seven methylation modifier genes and methylation landscapes. Up-regulation of genes with hypomethylated promoter regions in AML and down-regulated genes with hypermethylated promoter regions in COAD was selected by graph-based sub-network clustering methods. To rule out expression changes of genes by a transcription factor (TF), I used the Transfac database to scan TF binding sites in the promoter regions, which compiled a list of TFs. If a TF that could bind to the promoter region of a gene that was expressed significantly different between the two sample groups, the gene was removed for further consideration to rule out the effect of TF. As a result, 42 up-regulated DEGs with hypomethylated promoter DMR in AML and 61 down-regulated DEGs with hypermethylated promoter DMR were identified. Many of these genes are known to be associated with either AML or COAD in the literature.

In the third study, I developed a computerized or *in silico* experimental system that can quickly test the relevance of a KO gene to disease using omics data. MicroRNA, PPI and TF network information were deployed for the *in*

silico testing. To transform a hypothesis to be tested into a target gene set, a literature-based search engine was used and the analysis results were evaluated by calculating the entropy of the number of target genes connected through the networks induced by the condition-specific gene expression levels. The *in silico* system was tested using E2f1 knock-out data. 11 out of 14 E2f1-related diseases showed to be highly associated with E2f1 while diseases that were not known to be related E2f1 failed in the *in silico* testing.

Although networks are effective tools for modeling complex interactions among biological entities, use of biological networks for analyzing multi-omics data is not straightforward. My doctoral study was to combine networks of PPI, miRNA, TF networks, and DNA methylation information to perform the integrated analysis of multi-omics data for mining new biological knowledge. *In silico* experiment tools using the integrated networks were developed for scientists to perform follow-up experiments.

Keywords: multi-omics data, network analysis, integration analysis, PPI network, miRNA network, TF network, methylation network

Student Number: 2011-30926

Contents

List of Figures

List of Tables

Chapter 1

Introduction

mRNA is transcribed from DNA and protein is translated from mRNA. Protein performs various biological functions, often leading to the phenotype of an organism. Mutations of DNA can result in the modification of protein and, in turn, a new phenotype such as disease. Due to the advances in instrument technologies, phenotypes have been investigated using transcriptome data by RNA sequencing. Mechanisms of producing or transcribing mature mRNA are complicated. Transcription is the first step of DNA based gene expression. First, transcription factors (TFs) bind to the promoter region of DNA and stabilize the binding of RNA polymerase to DNA. RNA polymerase bound to DNA produces mRNA. mRNA serves as a template for the protein's synthesis through translation. Some small RNAs, e.g. miRNA, bind to mRNA and interfere with protein translation to regulate gene expression. The complex flow of such genetic information is called central dogma (Crick, 1970) as shown in Figure 1.1.

Now sequencing technologies can be used to produce various omics data in addition to transcriptome data, and these various omics data are called multi-

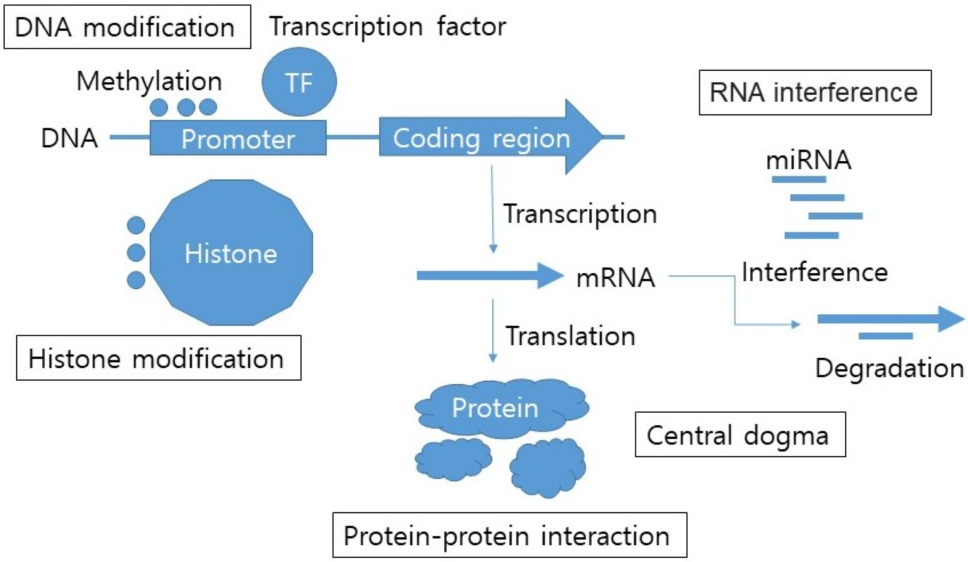


Figure 1.1 Complex mechanism of central dogma from DNA to protein expression. Central dogma consists of mRNA transcription from DNA and protein translations from mRNA, but the process has complex control mechanisms. DNA CpG site methylation and histone modification of the DNA promoter region regulate TF binding, and miRNAs have RNA interference that binds to and degrades mRNA. Expressed proteins interact with each other to perform their functions.

omics data. Analyzing multi-omics data can help characterize the regulatory relationship among multi-omics data and to explain phenotypes through the integrated data analysis.

1.1 Challenges

Analyzing multi-omics data is challenging due to complex interactions among various biological elements. In addition, different omics data are often in a

Table 1.1 **Difference of omics data.**

Data type	Identity	Number of variable
mRNA	Gene symbol	About 20,000
miRNA	miRNA symbol	About 2,000
methylation	CpG site ID number	28.3 million (450,000)

different format, so processing and interpreting omics data is not trivial. In the case of mRNA data, RNA-seq data processed and the quantities are measured in RPKM (Reads Per Kilobase Million), FPKM (Fragments Per Kilobase Million) or TPM (Transcripts Per Kilobase Million) (Mortazavi et al., 2008; Wagner et al., 2012). Since miRNA has short in sequence length, the data measurement method is different. In the case of methylation data, the value corresponding to the presence or absence of methylation on CpG site is measured and expressed as a ratio. The methylation value is measured only for some representative sites, not all sites. Thus, processing multi-omics data for the next step of the integrated analysis is complicated.

Integrating omics data requires handling large multi-dimensional feature space. There are more than 20,000 genes for mRNA and about 2,000 miRNAs. For methylation data, the whole number of CpG sites is 28.3 million (Babenko et al., 2017) (Table 1.1). On the other hand, there is a relatively small number of samples. Thus, analyzing multi-omics data is a high-dimensional, low sample problem, which is an unresolved machine learning problem. Since DNA methylation is usually measured from a bulk of cells, rather than a single cell. Thus, there is not a readily available computational method for the analysis of multi-omics data.

1.2 My approaches

In order to address the problem of combining multi-omics data, the core of my approach is to use biological networks.

In my doctoral thesis, three studies were conducted using an approach to integrate and analyze multi-omics data.

- A study using miRNA network and protein-protein interaction (PPI) network for integrated analysis of miRNA and mRNA data
- A study of cancer data through correlation clustering of DNA methylation network and gene expression correlation network
- Development of hypothesis verification system using miRNA network, PPI network and TF network

1.2.1 Using miRNA network and PPI network for integrated analysis of miRNA and mRNA data.

MicroRNAs consisting of about 22 nucleotide sequences were known to bind to mRNA and regulate gene expression. Due to the discovery of the regulatory mechanisms of miRNAs that are non-coding genes, the paradigm shifted from functional studies of protein-coding genes to novel gene regulation studies, and the importance of regulation by non-coding gene had increased.

It is not easy to analyze integrated miRNA and mRNA data to discover mRNA regulation by miRNA. The relationship between over 2,000 miRNAs and about 20,000 genes has over 40 million feature spaces, and analyzing all the relationships is a very difficult problem. A miRNA-mRNA network based on the complementary sequence was used a way to reduce the feature space in order to solve the challengeable problem in my doctoral study. In addition, domain knowledge about the negative correlation between miRNA and mRNA because of cleavage of double-stranded mRNAs bonded with miRNAs into two pieces was used to my doctoral thesis analysis. Furthermore, for the functional study of a set of genes affected by a specific gene, I studied the function of the gene by examining other gene set affected through the PPI network.

In my doctoral study, the integrated miRNA and mRNA data in EWS knockout mouse were analyzed by this method using the miRNA-mRNA network and PPI network. Through the analysis, some miRNAs and Gnai1 and neighbor genes that affect the phenotype were identified and validated with qRT-PCR.

1.2.2 Using DNA methylation network and gene expression correlation network clustering for integrated analysis of methylation and gene expression data.

The regulation of gene expression by epigenetics such as DNA methylation is a new field of genetic regulation. Studying the effects on cancer expression by DNA methylation is a very important issue to be able to create a new treatment method for cancer patients by recognizing the new effects that have not been understood in the past. However, studies on the effects of methylation are still in its infancy, and research is still lacking. Analyzing the impact between 28 million CpG sites and 20,000 gene expression is a near-impossible problem. Especially when comparing hundreds of samples, analysis becomes more difficult. The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015) provides 450,000 CpG sites, but it is still a large number and it is difficult to solve because it has a high dimensional feature space and it is combined with 20,000 genes and compared with more than 100 samples.

This difficult problem can also be solved by using a sub-network clustering method and domain knowledge to reduce the feature space. Among clusters of highly correlated networks, genes were identified by comparing gene expression levels in clusters and identifying genes with a negative correlation with methylation expression. In addition, using the TF network, genes associated with TF expression can be traced to remove genes with positive correlation, and finally, genes that are affected by methylation can be identified.

I used the above method to analyze the effects of seven methylated genes on 12 cancer types of TCGA data and analyzed a total of 3865 samples. Through analysis, 42 hypo-methylated promoter differentially methylated regions (DMRs) up-regulated differentially expressed genes (DEGs) in Acute Myeloid Leukemia patients and 61 hyper-methylated DMR down-regulated DEGs in Colon Adenocarcinoma patients were identified by methylation regardless of the expression of TF. And I had confirmed that several genes have been previously reported in other experimental papers.

1.2.3 Hypothesis test of a key gene with diseases using PPI, TF and miRNA networks.

Analysis of phenotypes about mutation effects and diseases of specific genes is a very necessary study to set the direction of biology research. Although it is necessary to confirm the various effects possible through experimentation, methods based on biological experiments are inefficient in terms of time and money, and there is no computer-based test method yet. Therefore, it was necessary to develop new computational experimental tools to analyze the impact of genes on the disease.

To investigate the effects of hypotheses and diseases, it was necessary to convert to a gene set that is highly related to disease or hypothesis by linking to the literature-based searching engine. I had constructed a database of various networks, such as public miRNA and PPI networks those are made by recent research and the TF network generated through Pearson's correlation. And then all the networks between the key gene and the target genes by converting the hypothesis were extracted. The final test result network selected the relevant networks considering the gene expression level and calculated the entropy of the resulting network.

The developed computable experimental analysis tools were verified using E2f1 data and related disease names to confirm a high association with related diseases by entropy value.

1.3 Outline of the thesis

My doctoral research consists of three studies that combine various network information and existing domain knowledge in data integration analysis to efficiently analyze difficult problems caused by mutual omics data integration. Chapters 2, 3, and 4 introduce the difficult challengeable problem of integrating different types of omics data, using a variety of networks to find out how each key gene has an effect on phenotype, disease, and cancer. Each of the studies uses network information and existing domain knowledge to solve the problems.

Chapter 2 explains the effect of genes on phenotype using the miRNA network, PPI network information, and miRNA-mRNA negative correlation domain knowledge.

Chapter 3 presents a study of the effects of methylated genes on cancer. I use a sub-network clustering method to select a highly correlated sub-network and select genes that affect cancer by 7 methylation genes, taking into account the gene expression level and the amount of methylation expression and TF gene expression level.

Chapter 4 describes the effects of genes on diseases by using miRNA, TF, and PPI network information to find related networks. Sub-networks are selected using gene expression information, and calculated the entropy value of the network to identify the impact.

Chapter 5 summarizes the studies. The bibliography of the cited references is at the end of this paper.

Chapter 2

Integrated analysis of omics data using microRNA-target mRNA network and PPI network reveals regulation of Gnai1 function in the spinal cord of Ews/Ewsr1 KO mice

2.1 Related works

Ewing sarcoma is the second most common bone and soft tissue tumor that predominantly afflicts children and adolescents (Meltzer, 2007; Barker et al., 2005; Miser et al., 2004). Understanding the biological mechanisms underlying this tumor is critical to the identification of new cancer therapy targets. The Ewing sarcoma gene (EWS)/EWS RNA-Binding Protein 1 (EWSR1), a transcription factor, encodes an RNA binding protein whose specific functional targets are still largely unknown (Bertolotti et al., 1999). In previous studies, fusion genes such as EWS-FLI-1, EWSR1-WT1, EWSR1-KLF17, EWSR1-ATF1, and

EWSR1-CREB3L1, are known to be produced by rearrangement of the EWSR1 gene with different gene fusion partners and these fusion genes have functions related to a variety of soft tissue tumors (May et al., 1993; Fisher, 2014; Huang et al., 2015; Rossi et al., 2007; Lau et al., 2013). To characterize the functions of EWS, I used RNA-seq gene expression data and miRNA expression data measured by using the spinal cord samples of Ews/Ewsr1 knock-out (KO) mouse and wild type.

2.2 Motivation

Multi-function genes interact with a number of coding and non-coding genes and perform a variety of functions depending on cell conditions and tissue types. Multi-function gene EWSR1 is known to regulate Drosha and microRNAs that inhibits RNA splicing (Kim et al., 2014; Chansky et al., 2001). However, it is still unknown which genes are regulated by and which biological functions are related to EWSR1. To characterize the functions of EWSR1, I used a well-known DEG set analysis. I performed functional analysis of top 200 up-regulated DEGs and top 200 down-regulated DEGs (2% of the whole genes) using gene ontology (GO) and KEGG pathway. From the GO analysis, I found 322 genes of 400 top DEGs were involved in 44 GO terms in the GOTERM_BP_FAT category which is the summarized version of Biological Processes in the Gene Ontology. Top three GO terms with the largest number of genes were ion transport, immune response, and homeostatic process. It is not clear how these three biological processes are related to EWS. In addition, I tried molecular function GO terms, which did not produce coherent biological functions related to EWS. From the KEGG pathway result, 93 of 400 genes hit 140 pathways. Only two pathways had more than 10 genes: metabolic pathway and cell adhesion molecules. Most of the pathways were not significant. Overall, GO and KEGG pathway analysis using DEGs did not produce meaningful clues on the role of EWS.

For the analysis of miRNA expression data, it is not clear how to perform an integrated analysis of gene expression data and miRNA expression data. In addition, a multifunction gene can play roles at various levels such as transcription, gene regulation, translation and protein activity level. To address this computational challenge, I developed a novel computational framework for the characterization of EWS multifunctional gene using gene expression data and miRNA expression data measured under a knockout condition of the multifunctional gene. The framework utilized microRNA-target gene network and PPI network and incorporates the two networks in a workflow. The workflow of the framework can be viewed as an effort to model the role of EWS at various levels, DEG analysis at the transcription level, the microRNA-target gene network analysis at the gene regulation level, and PPI network analysis at the translation and protein activity level.

2.3 Methods

I developed a three-step pipeline for the integrated analysis of omics data using the mRNA-microRNA network and protein-protein interaction network. I describe the workflow and computational methods used in each step in this section. The Figure 2.1 illustrates the workflow of the proposed omics data analysis pipeline. In the “Results” section, I discuss output from each step in detail.

2.3.1 Step 1 - MicroRNA-target gene regulation network analysis

To investigate the roles of EWS, I analyzed the translational regulatory network. The microRNA-target gene integrated network analysis was performed

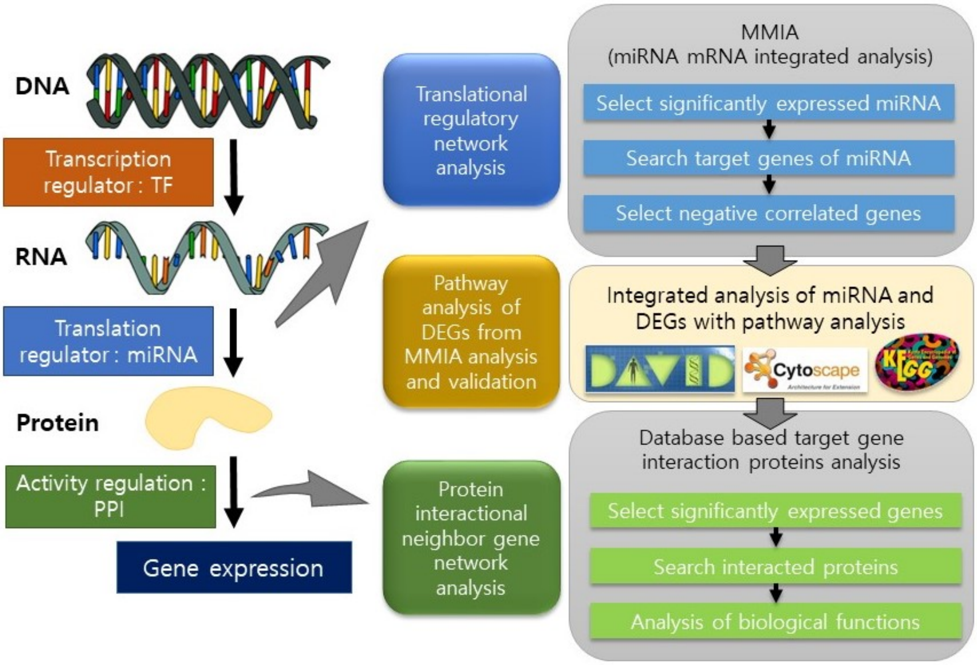

DNA
Transcription regulator : TF
RNA
Translation regulator : miRNA
Protein
Activity regulation : PPI
Gene expression
Translational regulatory network analysis
Pathway analysis of DEGs from MMIA analysis and validation
Protein interactional neighbor gene network analysis
MMIA (miRNA mRNA integrated analysis)
Select significantly expressed miRNA
Search target genes of miRNA
Select negative correlated genes
Integrated analysis of miRNA and DEGs with pathway analysis
DAVID
Cytoscape
KEGG
Database based target gene interaction proteins analysis
Select significantly expressed genes
Search interacted proteins
Analysis of biological functions

Figure 2.1 **Illustration of the workflow.** TF gene has multiple functions to regulate transcription. Generated mRNAs are regulated by microRNA and translated proteins have functions with interacted proteins and molecules. RNA sequencing data and microRNA (miRNA) microarray data are generated from spinal cord extraction in Ews/Ewsr1 knockout and wild type mice. SAM (Significance Analysis of Microarrays) is used for the selection of significantly expressed miRNA from miRNA microarrays. TargetScan and miRDB were used to predict the target genes of miRNAs. From RNA sequencing data, gene expression values are mapped to the reference genome data using Tophat. Then negative correlated DEGs are selected. Significantly expressed microRNA target genes have many interacting proteins. Specific target gene interactional neighbor proteins are searched in the STRING DB. PPI network analyzed with gene expression value. Analysis results of miRNA-mRNA network and PPI network are integrated by analyzing correlation in expression levels. Regulated genes further are analyzed and visualized with DAVID, KEGG, and Cytoscape.

following the strategy in MMIA (Nam et al., 2009).

- Input: gene expression data, miRNA expression data
- Output: differentially expressed miRNAs and their target genes

1.1. Selection significantly expressed microRNAs

I selected significantly up- or down-regulated microRNAs in the Ews/Ewsr1 KO condition compared to the wild type condition. To select significantly differentially expressed miRNAs from microarray data, I used the SAM tool package (Tusher et al., 2001). (More information in the detailed method section.)

1.2. Prediction of microRNAs target genes

After selecting significantly expressed microRNAs, I predicted regulatory target genes of the selected differentially expressed microRNA by TargetScan (Lewis et al., 2005) and miRDB (Wang and El Naqa, 2007; Wang, 2008).

1.3. Reselection target genes by correlation

I further investigated miRNA and gene target relationship by measuring negative correlation in expression levels between miRNAs and genes targeted by miRNAs since up-regulated microRNA inhibits translation of mRNA.

2.3.2 Step 2 - Pathway analysis of DEGs from MMIA analysis and validation

- Input: DEGs selected in Step 1
- Output: important pathways related to EWS and key genes in the pathways

2.1. DEG analysis

DEGs analysis of NGS RNA-seq was performed in the following steps. First, adaptor sequences of reads in raw data were trimmed. The Ensembl mouse reference genome sequence was downloaded for mapping short reads. Bowtie (Langmead et al., 2009) was used to build an index of the reference genome sequence for alignment. Trimmed reads were then mapped to the reference genome sequence using Tophat2 (Kim et al., 2013). Finally, Cufflinks was used to calculate gene expression levels. I compared gene expression values and selected DEGs by using Cuffdiff in the Cufflinks package (Trapnell et al., 2010).

2.2. Integrated analysis of miRNA and mRNA expression data

15 differentially expressed miRNAs were found to target 4342 genes based on TargetScan and miRDB. To further screen target genes, I integrated miRNAs target information and mRNA-seq based gene expression levels. The negative correlation analysis reduced the number of targets to 1338 genes. The negative correlation analysis is based on the techniques in (Xin et al., 2008; Marbach et al., 2012). The rationale for the negative correlation analysis is that if a miRNA targets a gene the expression levels of the miRNA and the gene should have negative correlation due to the regulatory effect of miRNA on the target gene. These DEGs were then analyzed by GSEA (Gene Set Enrichment Analysis) using DAVID (The Database for Annotation, Visualization and Integrated Discovery) (Dennis et al., 2003).

2.3. Pathway analysis

To characterize functions of selected target DEGs by negative correlation in the spinal cord of Ews/Ewsr1 KO mice, I performed biological pathway analysis using the KEGG Mapper (Kanehisa and Goto, 2000). KEGG Mapper highlighted DEGs with colors: up-regulated DEGs as red, down-regulated DEGs as blue, and other DEGs as light green. In addition, I performed additional pathway interpretation based on gene ontology by using ClueGO (Bindea et al., 2009), a Cytoscape (Shannon et al., 2003) plug-in, that analyzes biological pathway interpretation with KEGG ontology (2014 version) to integrate GO terms and KEGG/BioCarta pathways to generate a functionally organized GO/pathway term network.

2.4. Verification of Gnai1 expression by Quantitative real-time PCR (qRT-PCR)

To verify whether the expression of target genes is correlated with the analysis, I performed qRT-PCR using RNA isolated from the spinal cords of Ews/Ewsr1 WT and KO mice.

2.3.3 Step 3 - Protein-protein interaction network analysis

After selecting the key gene in Step 2, I investigated the biological functions of the genes by extending gene sets with neighboring genes of the key gene.

- Input: Key genes identified in Step 2
- Output: G protein complex genes and regulators

After selecting the key gene in Step 2, I investigated the biological functions of the genes by extending gene sets with neighboring genes of the key gene.

3.1. Selection significantly expressed gene

From gene set analysis (GSA) and pathway analysis (see the detailed methods section), I selected specific genes.

3.2. Search for proteins that interact with the selected gene

PPI network analysis of genes neighboring the key gene was performed by using STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) (Snel et al., 2000), the most widely used database of known and predicted protein interactions.

3.3. Analysis of biological functions

Relationship between the key gene and neighbor genes was investigated by performing the literature search. When I considered the relationship among genes, I also considered the regulatory roles of genes, i.e., activators or repressors, if applicable. For the regulatory relationship, I considered gene expression change information.

2.4 Results

2.4.1 Analysis of multifunctional EWS by using the network-based workflow

In this section, I present the result from each computational step of the workflow and discuss biological meanings if possible.

Part 1 - Translational regulatory network analysis: MicroRNA-mRNA network

Selection of differentially expressed miRNAs I selected 18 significantly expressed miRNAs from the total 1193 mouse miRNAs by SAM tool. 15 miRNAs expression levels were significantly up-regulated, and 3 miRNAs were down-regulated in the Ews/Ewsr1 KO mice against WT mice. In the order of the significance score by SAM, 15 up-regulated miRNAs are mmu-miR-127, mmu-miR-410, mmu-miR-433, mmu-miR-138, mmu-miR-181c, mmu-miR-382, mmu-miR-19b, mmu-miR-381, mmu-miR-666-3p, mmu-miR-376a, mmu-miR-873, mmu-miR-181a, mmu-miR-383, mmu-miR-181b, and mmu-miR-99b. Down-regulated 3 miRNAs were mmu-miR-1224, mmu-miR-9-3p, and mmu-miR-26a in the order of the significance score by SAM. Analysis of potential biological functions of these miRNAs was performed by using genes targeted by the miRNAs (see the DEG analysis from RNA-seq data result section).

Prediction of target mRNA regulated by selected miRNA To perform the integrated analysis of miRNA and their target genes, I need to predict the targets of miRNAs. Predicted target genes of miRNAs were collected by using TargetScan and miRDB. 5,779 and 5,448 genes were predicted by TargetScan and miRDB, respectively. 1,927 genes were targeted by multiple miRNAs in the prediction result of TargetScan, and 2,371 genes were multiply targeted according to miRDB. After discarding repeatedly predicted genes, a total of 4,342 genes were predicted as targets of 15 differentially expressed miRNAs. Only 36 percent (1,587 genes) of predicted target genes were predicted by both TargetScan and miRDB. In other words, the genes targeted by each miRNAs of prediction results by TargetScan and miRDB do not agree much. 4,342 target genes predicted by both TargetScan and miRDB were further analyzed by performing a negative correlation analysis to sort out potentially true miRNA-gene relationships (see the next section).

Negative correlation analysis of DEGs with DE microRNA Predicted target genes were further screened by considering negative correlations in expression levels between miRNA and each of its target genes. The rationale for

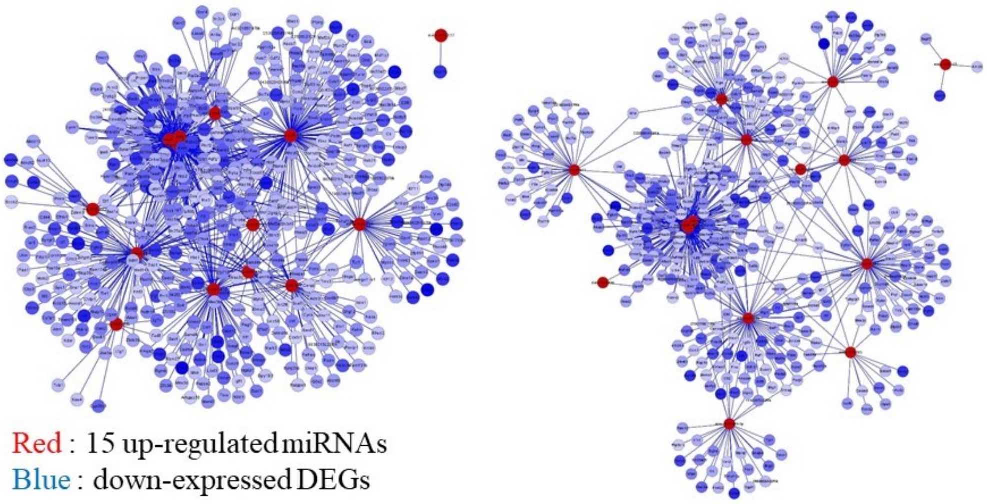


Figure 2.2 **Network of microRNAs and mRNAs.** Up-regulated miRNAs (Red nodes) are selected by SAM. Target genes (mRNAs, blue nodes) of selected miRNAs are predicted by TargetScan (left) and miRDB (right). Down-regulated genes targeted by up-regulated miRNA are selected from each predicted results. miRNA-mRNA interaction network is drawn by Cytoscape. Color intensity denotes the level of gene expression.

the negative correlation analysis is that miRNA degrades its target genes, thus a higher expression level of miRNA should result in a lower expression level of its target. I applied the same technique used in (Lewis et al., 2005; Wang and El Naqa, 2007). Negatively correlated miRNA-mRNA interaction network of miRNAs and their target DEGs were visualized by using Cytoscape (Figure 2.2). In Figure 2.2, significantly up-regulated 15 miRNAs are in red color, and negative correlated target DEGs are in blue color. Color intensity denoted the level of gene expression. As a result of the correlation analysis, 4,342 genes were reduced to 860 genes. Among the 860 DEGs, 339 target genes were targeted by multiple miRNAs.

Part 2 - Pathway analysis of DEGs from MMIA analysis and validation

KEGG pathway analysis of DEGs gene set targeted by miRNA I mapped the 860 negatively correlated DEGs to the KEGG pathway using the KEGG Mapper. 201 pathways were hit by the negatively correlated DEGs. I selected 13 pathways with eight or more gene hits. Metabolic pathways, calcium signaling pathway, PI3K-Akt signaling pathway, axon guidance, pathways in cancer, MAPK signaling pathway, tight junction, dilated cardiomyopathy, circadian entrainment, proteoglycans in cancer, regulation of actin cytoskeleton, cholinergic synapse and focal adhesion pathways were selected. Analysis of KEGG pathways of DEGs was highlighted in colors chosen by KEGG Mapper. Blue color genes were down-regulated genes, and red color genes were up-regulated genes in the pathways of Ews/Ewsr1 KO mice. Color intensity denoted the level of gene expression.

Gene ontology based network analysis Networks of negatively correlated target DEGs in terms of KEGG ontology were generated using ClueGO (Figure 2.3). "Cholinergic synapse pathway" term was highly clustered by down-regulated DEGs belonging pathways. ECM-receptor interaction pathway, focal adhesion pathway, tight junction pathway, and actin cytoskeleton regulation pathway were mostly correlated with selected down-regulated DEGs. Gnai1, which is most significantly down-regulated in the cholinergic synapse pathway, was selected for further investigation. More discussion on the biological functions of these pathways is presented in the Conclusion section.

qRT-PCR of Gnai1 qRT-PCR was performed to confirm the difference of Gnai1 expression in the spinal cords of Ews/Ewsr1 WT and KO mice. Average

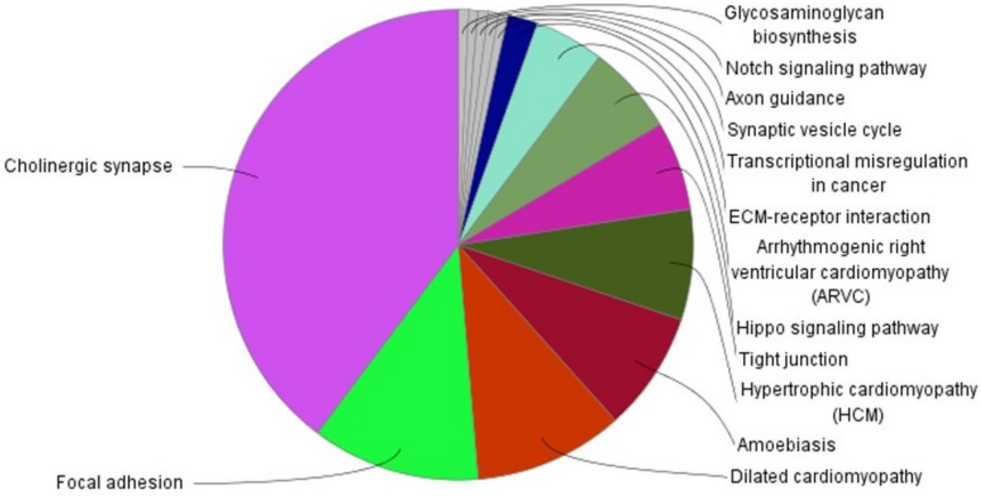


Figure 2.3 **Venn diagram generated by ClueGO.** ClueGO analyzes KEGG ontology of selected down-regulated genes which are targeted by up-regulated miRNA. Cholinergic synapse pathway is showed highly clustered by down-regulated gene pathways.

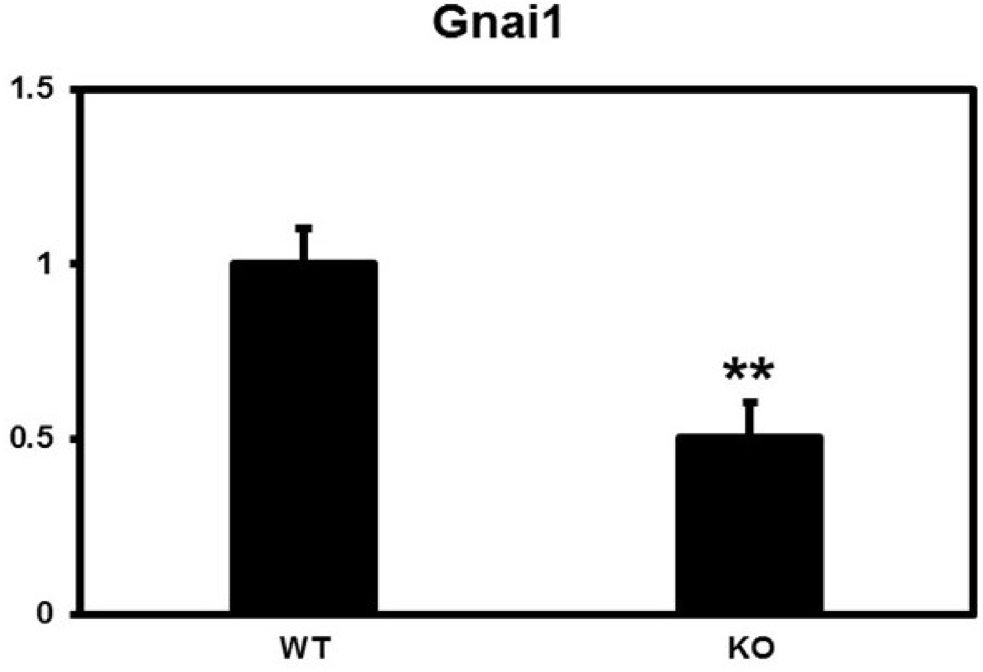


Figure 2.4 **Verification of altered Gnai1 expression in Ews/Ewsr1 WT and KO mice.** The gene expression level of Gnai1 was significantly lower in the spinal cords of Ews/Ewsr1 KO mice (n=6) compared to EWS WT mice (n=6). The bar graph represents the average ± standard error mean (SEM). **, Significantly different at $p<0.01$ by T-test.

gene expression levels of Gnai1 in Ews/Ewsr1 KO mice were significantly lower than those in Ews/Ewsr1 WT mice. This data validated that Gnai1 expression level was down-regulated in Ews/Ewsr1 KO mice (Figure 2.4).

Part 3 - Protein-protein interactions network analysis

I selected Gnai1 that is down-regulated in cholinergic synapse pathways and actin cytoskeleton regulation pathway. To investigate the effect of down-regulation of Gnai1, I used the STRING protein-protein interaction network DB. In the PPI network, genes neighboring Gnai1 were further investigated for their biological functions. Looking at gene expression values, I was able to confirm the relationship between G-protein genes and RGS genes. Genes neighboring Gnai1 were selected by using STRING (Figure 2.5). Top 20 interacted genes are shown in Table 2.4.1. Gnai1 and G-protein related genes, such as Gnb1, Gnb2, and Gnb4, had down-regulated gene expression level (Figure 2.6). In contrast, Rgs1 and Rgs19, Regulator of G-protein signaling genes, were up-regulated and inhibited with Gnai1 (Figure 2.6).

2.5 Discussion

2.5.1 A potential interaction map of EWS, RGS, and G-protein complex genes.

A growing body of evidence shows multifunctional roles of the EWS/EWSR1 fusion oncoproteins (May et al., 1993; Huang et al., 2015; Rossi et al., 2007; Lau et al., 2013). However, the role of wild-type (WT) EWS/EWSR1 is not fully understood yet. EWS/EWSR1 deficiency contributes to the failure of precursor B lymphocyte development and leads to the premature cellular senescence in mouse embryonic fibroblasts (MEFs) (Li et al., 2007; Cho et al., 2011). It seems likely that the WT EWS/EWSR1 protein exhibits many different cellular functions in a cell-type specific manner. In the spinal cord of Ews/Ewsr1 KO mice, microRNAs, such as mmu-miR-381 and mmu-miR-181a/b/c were up-regulated. These microRNAs suppressed the expression of Gnai1 (Gi Protein Alpha subunit). Concurrently, RGS (Regulator of G-protein Signaling) genes, Rgs1 and Rgs19, were up-regulated, which repressed Gnai1 activity. In addition, G Protein Beta subunit genes, Gnb1, Gnb2, and Gnb4 were down-regulated. Thus in

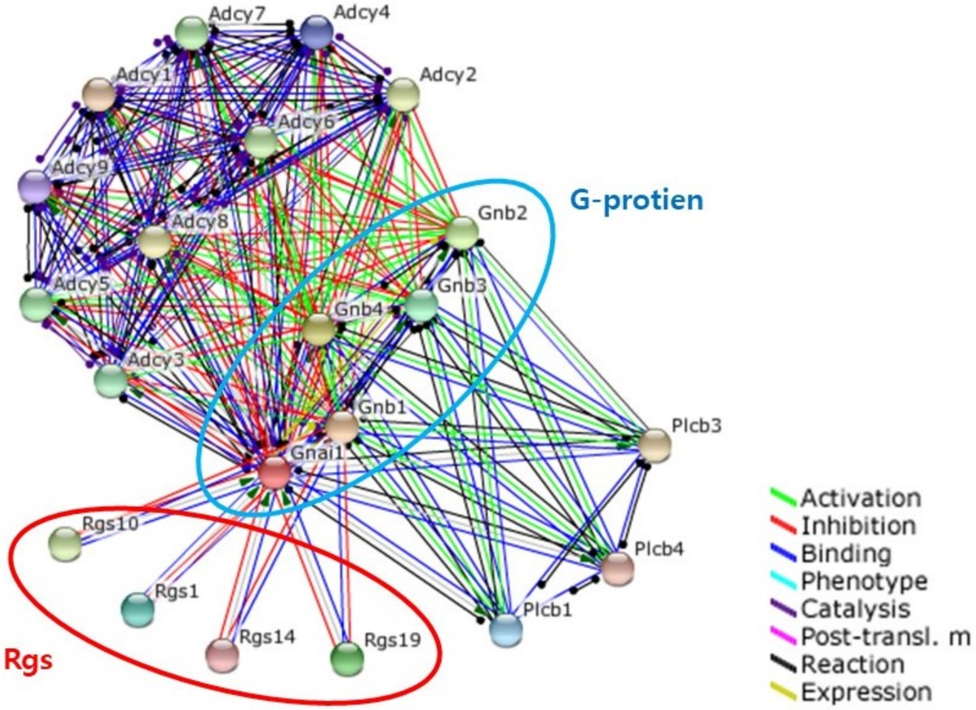
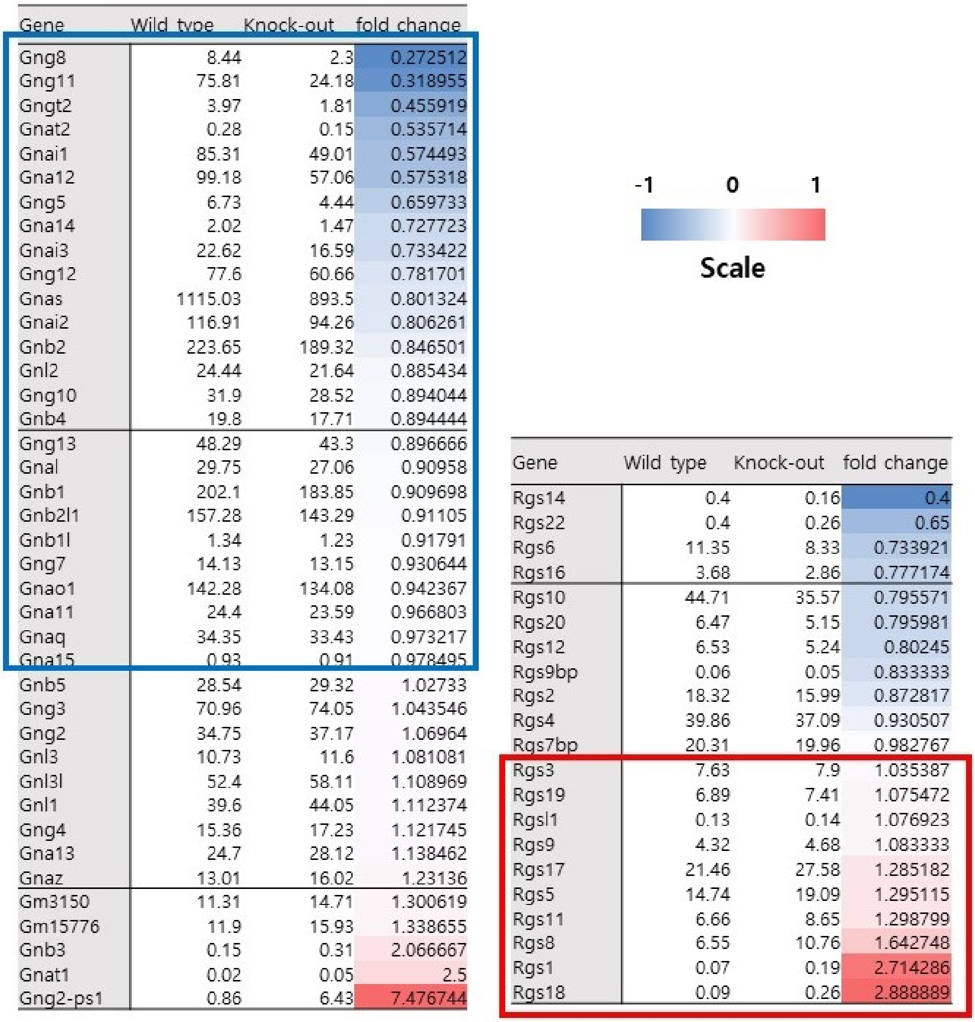


Figure 2.5 **PPI network of Gnai1 from the STRING DB.** The Gnai1 protein binds with neighbor protein Gnb1, Gnb2, Gnb3 and Gnb4 that are in G-protein family. Rgs1, Rgs10, Rgs14 and Rgs19 proteins inhibit the activity of Gnai1 protein.

Table 2.1 **Top 20 interacted genes with Gnai1 from the STRING DB.** These genes are sorted by prediction score. 13 genes are related to inhibition with Gnai1.

Gene Symbol	Prediction	Score	Inhibition
Gnb1	0.994	Yes	
Gnb4	0.98	Yes	
Gnb2	0.98	Yes	
Rgs19	0.979	Yes	Yes
Gnb3	0.978	Yes	
Rgs1	0.976	Yes	Yes
Plcb1	0.974	Yes	
Adcy4	0.973	Yes	Yes
Adcy9	0.973	Yes	Yes
Rgs14	0.972	Yes	Yes
Plcb4	0.97	Yes	
Adcy1	0.97	Yes	Yes
Plcb3	0.97	Yes	
Adcy8	0.969	Yes	Yes
Adcy2	0.969	Yes	Yes
Rgs10	0.969	Yes	Yes
Adcy6	0.967	Yes	Yes
Adcy7	0.967	Yes	Yes
Adcy5	0.966	Yes	Yes
Adcy3	0.966	Yes	Yes



Gene	Wild type	Knock-out	fold change
Gng8	8.44	2.3	0.272512
Gng11	75.81	24.18	0.318955
Gngt2	3.97	1.81	0.455919
Gnat2	0.28	0.15	0.535714
Gnai1	85.31	49.01	0.574493
Gna12	99.18	57.06	0.575318
Gng5	6.73	4.44	0.659733
Gna14	2.02	1.47	0.727723
Gnai3	22.62	16.59	0.733422
Gng12	77.6	60.66	0.781701
Gnas	1115.03	893.5	0.801324
Gnai2	116.91	94.26	0.806261
Gnb2	223.65	189.32	0.846501
Gnl2	24.44	21.64	0.885434
Gng10	31.9	28.52	0.894044
Gnb4	19.8	17.71	0.894444
Gng13	48.29	43.3	0.896666
Gnal	29.75	27.06	0.90958
Gnb1	202.1	183.85	0.909698
Gnb2l1	157.28	143.29	0.91105
Gnb1l	1.34	1.23	0.91791
Gng7	14.13	13.15	0.930644
Gnao1	142.28	134.08	0.942367
Gna11	24.4	23.59	0.966803
Gnaq	34.35	33.43	0.973217
Gna15	0.93	0.91	0.978495
Gnb5	28.54	29.32	1.02733
Gng3	70.96	74.05	1.043546
Gng2	34.75	37.17	1.06964
Gnl3	10.73	11.6	1.081081
Gnl3l	52.4	58.11	1.108969
Gnl1	39.6	44.05	1.112374
Gng4	15.36	17.23	1.121745
Gna13	24.7	28.12	1.138462
Gnaz	13.01	16.02	1.23136
Gm3150	11.31	14.71	1.300619
Gm15776	11.9	15.93	1.338655
Gnb3	0.15	0.31	2.066667
Gnat1	0.02	0.05	2.5
Gng2-ps1	0.86	6.43	7.476744

Gene	Wild type	Knock-out	fold change
Rgs14	0.4	0.16	0.4
Rgs22	0.4	0.26	0.65
Rgs6	11.35	8.33	0.733921
Rgs16	3.68	2.86	0.777174
Rgs10	44.71	35.57	0.795571
Rgs20	6.47	5.15	0.795981
Rgs12	6.53	5.24	0.80245
Rgs9bp	0.06	0.05	0.833333
Rgs2	18.32	15.99	0.872817
Rgs4	39.86	37.09	0.930507
Rgs7bp	20.31	19.96	0.982767
Rgs3	7.63	7.9	1.035387
Rgs19	6.89	7.41	1.075472
Rgsl1	0.13	0.14	1.076923
Rgs9	4.32	4.68	1.083333
Rgs17	21.46	27.58	1.285182
Rgs5	14.74	19.09	1.295115
Rgs11	6.66	8.65	1.298799
Rgs8	6.55	10.76	1.642748
Rgs1	0.07	0.19	2.714286
Rgs18	0.09	0.26	2.888889

Figure 2.6 **G-proteins and RGS (regulator of G-protein) expression level and log2 fold change value in Ews/Ewsr1 wild type and knock-out.**

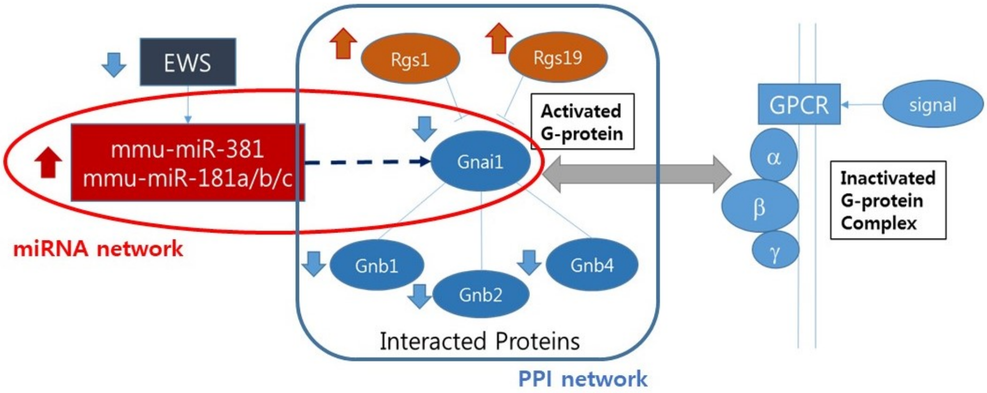


Figure 2.7 **Roles of G proteins and its regulatory mechanisms by miRNAs in the spinal cord of Ews/Ewsr1 KO mouse.** The direction of the arrow means with a change of gene expression level in Ews/Ewsr1 KO mice. Upper arrows are up-regulated gene expression level, and bottom arrows are the opposite.

the Ews/Ewsr1 KO condition, G protein complex was not formed (Figure 2.7).

Since Gnai1 was down-regulated, it is proposed that Gnai1 may be unable to inhibit downstream adenylate cyclase genes, such as Adcy9 and Adcy4, in cholinergic synapse pathway. Adenylate cyclase catalyzes the conversion of ATP to cAMP, and the cAMP regulates cAMPproteins, transcription factors, and cAMP-dependent kinases. Adenylate cyclase is an enzyme with key regulatory roles, and Adenylate cyclase regulator Gnai1 has important roles in the cholinergic synapse.

My study presents for the first time that Ews/Ewsr1 deficiency modulates microRNA processing in the spinal cord. Notably, increased levels of mmu-miR-381 and mmu-miR-181a/b/c were directly associated with the down-regulation of the G protein complex in the spinal cord of Ews/Ewsr1 KO mice. We have previously shown that Ews/Ewsr1 deficiency leads to abnormal microRNA processing and skin development via Drosha-dependent pathway (Kim et al., 2014). Furthermore, we found that Ews/Ewsr1 deficiency reduces the expression of Uvrag (UV radiation resistance associated) gene at the post-transcription level via mmu-miR-125a and mmu-miR-351 (Kim et al., 2015). Interestingly, the reduction of Uvrag by mmu-miR-125a and mmu-miR-351 impaired autophagy function in Ewsr1 KO MEFs and KO mice. Considering that G protein-coupled signaling transduction pathway is very complex, the Gnai1-dependent cellular function and mechanism in vitro and in vivo models of EWSR1 deficiency remain to be determined in future studies.

Chapter 3

Impact of mutations in DNA methylation genes on genome-wide methylation landscapes and downstream gene activations in pan-cancer

3.1 Related works

DNA mutation is one of the major causes of many diseases, thus understanding the impact of mutations in genes is an important research problem. For example, mutations in oncogenes and tumor suppressor genes have been extensively studied over the years (Wee et al., 2019; Kim and Kim, 2018; Bailey et al., 2018). Some class of genes, e.g., epigenetic genes, have roles in regulating gene expression, rather than being directly related to certain phenotypes. Epigenetic genes can be divided into functional groups: epigenetic modulators, modifiers, and mediators (Feinberg et al., 2016). Epigenetic mediators have corresponded to the tumor progenitor genes, epigenetic modifiers of the mediators are frequently mutated in cancer, and epigenetic modulators upstream of the modifiers are related to changes in the cellular environment. An epigenetic modifier gene that modifies DNA methylation status or chromatin structure is studied for interpretation of cancer. Among the epigenetic modifiers, DNA methylation regulatory genes, DNMT1, DNMT3A, MBD1, MBD4, TET1, TET2, and TET3, are known to be involved in cancer (Yan et al., 2011; Couronné et al., 2012; Grossmann et al., 2013; Abdel-Wahab et al., 2009; Langemeijer et al., 2009; Network et al., 2012; Imielinski et al., 2012; Stephens et al., 2012; Neumann et al., 2013; Delhommeau et al., 2009; Scourzic et al., 2015; Krauthammer et al., 2012). DNMT3A mutation was found at a high rate of 22.1 percent of LAML patients (Ley et al., 2010) and in at least one of DNA methylation modifiers, a mutation was found in about 13 percent (1,474/11,315) of 33 TCGA projects.

In general, mutations on a gene can affect the function of a gene, even loss or gain of a function. Many DNA methylation modification genes are enzymes. Thus, mutations on the epigenetic modifiers could affect the activity of epigenetic modifiers, which would result in the difference in genome-wide methylation profiles and in turn, activation of downstream genes. However, there is no systematic study of this important topic. In this paper, I investigated the effect of mutations on DNA methylation modification genes such as DNMT1, DNMT3A, MBD1, MBD4, TET1, TET2, and TET3 through a pan-cancer analysis. First, I investigated the effect of mutations in DNA methylation modification genes on genome-wide methylation profiles in 12 major cancer types in TCGA.

As a result, I found that genome-wide methylation landscapes were significantly different between two sample groups with mutations and without mutations in the DNA methylation modifier genes. Second, I investigated the effect of DNA methylations in the promoter regions on downstream genes in 12 cancer types. To investigate the effect of mutations on gene expression further, I chose an up-regulated gene cluster where DEGs were mostly hypomethylated promoter regions in Acute Myeloid Leukemia and another down-regulated gene cluster where DEGs had mostly hypermethylated promoter regions in Colon adenocarcinoma.

3.2 TCGA data of DNA methylome and transcriptome

To perform pan-cancer data analysis, I downloaded data for 12 major cancer types from TCGA: Acute Myeloid Leukemia (LAML), Bladder Urothelial Carcinoma (BLCA), Breast invasive carcinoma (BRCA), Colon adenocarcinoma (COAD), Glioblastoma multiforme (GBM), Head and Neck squamous cell carcinoma (HNSC), Kidney renal clear cell carcinoma (KIRC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Ovarian serous cystadenocarcinoma (OV), Rectum adenocarcinoma (READ) and Uterine Corpus Endometrial Carcinoma (UCEC). A total of 3,864 samples that had both methylome and transcriptome data were collected. Among 3864 samples, 598 samples had at least one mutation in seven DNA methylation modifier genes, thus samples were divided into two groups, one with mutations in DNA methylation modifiers (598 samples) and the other group without mutations (3,266 samples), excluding TCGA-OV that had only 9 samples. Thus, I analyzed 11 cancer types. (Table 3.1)

3.3 Workflow

The analysis of the mutation data of seven DNA methylation modifiers on the pan-cancer scale was performed in three phases and the analysis workflow is

Table 3.1 **Number of samples per 12 major cancer type in TCGA.** Each value represents the number of samples that have both methylome and transcriptome data and the number of samples that have mutations in seven DNA methylation modifier genes.

Cancer type	Total samples	Mutated samples
TCGA-BRCA	784	61
TCGA-HNSC	521	60
TCGA-LUAD	455	80
TCGA-BLCA	408	98
TCGA-LUSC	369	73
TCGA-KIRC	319	21
TCGA-COAD	279	94
TCGA-UCEC	173	59
TCGA-LAML	170	33
TCGA-READ	93	12
TCGA-GBM	64	6
TCGA-OV	9	1
Sum	3864	598

shown in a schematic diagram (Figure 3.1).

3.3.1 Part 1 - Impact of mutations in DNA methylation modifiers on genome-wide methylation landscape

First, I investigated the effect of mutations in DNA methylation modifiers on genome-wide methylation profiles.

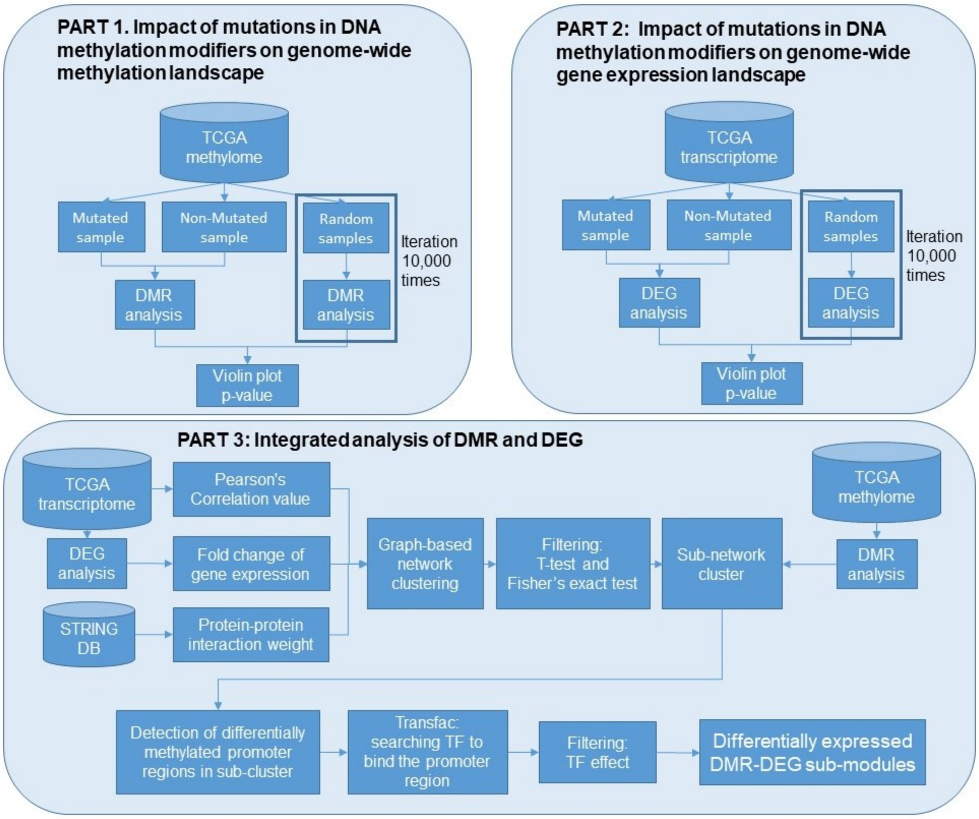

PART 1. Impact of mutations in DNA methylation modifiers on genome-wide methylation landscape
TCGA methylome
Mutated sample
Non-Mutated sample
Random samples
DMR analysis
DMR analysis
Iteration 10,000 times
Violin plot p-value
PART 2: Impact of mutations in DNA methylation modifiers on genome-wide gene expression landscape
TCGA transcriptome
Mutated sample
Non-Mutated sample
Random samples
DEG analysis
DEG analysis
Iteration 10,000 times
Violin plot p-value
PART 3: Integrated analysis of DMR and DEG
TCGA transcriptome
Pearson's Correlation value
DEG analysis
Fold change of gene expression
STRING DB
Protein-protein interaction weight
Graph-based network clustering
Filtering: T-test and Fisher's exact test
Sub-network cluster
TCGA methylome
DMR analysis
Detection of differentially methylated promoter regions in sub-cluster
Transfac: searching TF to bind the promoter region
Filtering: TF effect
Differentially expressed DMR-DEG sub-modules

Figure 3.1 **Workflow.** See the workflow section for more details.

1.1. Statistics on mutations in seven DNA methylation modifiers

Before investigating the genome-wide effects of seven DNA methylation modifiers, I analyzed the samples statistically. Mutation frequencies in DNA methylation modifiers were collected for each cancer.

1.2. Genome-wide methylation landscapes

To investigate the genome-wide effects of seven DNA methylation modifiers, I analyzed the difference in DNA methylation profiles in pan-cancer. To compare the difference in methylation of samples that were divided into DNA methylation modifiers mutation, mutated and non-mutated samples (3,266 vs. 598 samples) in terms of log ratios (See Methods section for the detail).

1.3. Statistics of the number of DMRs between two groups

To evaluate whether these differences are significant or not, I analyzed them statistically. I compared the number of DMRs in samples with mutations in the DNA methylation modifier with the number of DMRs in randomly selected samples. The analysis of DMR counts was performed with randomly sampled the same size as the number of mutation samples and repeated 10,000 times to calculate the p-value.

3.3.2 Part 2 - Impact of mutations in DNA methylation modifiers on genome-wide gene expression landscape

Since DNA methylation can have a significant effect on gene expression profiles, I compared gene expression profiles between the mutated and the non-mutated samples. In this part, I only compared gene expression profiles between two groups, without attempting to investigate the effect of DNA methylation on gene expression, which was reported in Part 3.

2.1. Statistics on gene expression profiles

DEG counts were collected from randomly chosen same size samples, repeating 10,000 times to calculate p-values.

2.2. Clustering analysis of transcriptome

To investigate the biological functions of DEGs, I divided DEGs into smaller gene sets based on network-based gene clustering analysis and then performed GO term enrichment test on each set of DEGs to compare the difference in functions of genes between the mutated and non-mutated groups. Before performing sub-network clustering, correlation values between genes were calculated. Pearson's correlation value was calculated for transcriptome data, and PPI score from STRING (Szklarczyk et al., 2016) database was multiplied by the weight. Using the log_2 fold change value obtained from the DEG analysis, I removed genes that had opposite interaction or the small change amount. Thus, I selected a set of genes with over 0.15 of the absolute value of log_2 fold change of gene expression and over 0.5 positive correlated genes network. I performed graph-based sub-network clustering using iCluster (see Methods section) with a fold change of gene expression using pre-processed gene-gene interaction score. To select meaningful clusters after clustering, I performed one sample t-test with gene expression levels and Fisher's exact test using the GO term enrichment test. Clusters with p-value under 10^{-9} were selected.

3.3.3 Part 3 - Integrated analysis of DMR and DEG

Now, I tried to associate DEGs and DMRs between the two groups as below.

3.1. Integration of gene expression and methylation expression

To investigate the effect of DMRs on DEGs, I focused on methylation difference in the promoter regions. First, I selected gene clusters with significantly enriched DEGs and DMRs using a Fisher's exact test for each of gene clusters. Then, gene sets were selected by considering the negative correlation between promoter methylation and the corresponding gene expression.

3.2. Transcription factor binding site search with Transfac

In addition to the negative correlation between promoter methylation and the corresponding gene expression, I considered expression levels of TFs that could bind to the promoter regions. Thus, I searched for all TF binding sequences in the DEG promoter region using Transfac (Matys et al., 2003).

3.3. Comparison without TF effect

The expression level of the TFs that had binding sites in the promoter regions was considered to remove cases where gene expression difference could result from TF expression difference. For example, if TF binding to the promoter of up-regulated DEG is not up-regulated, the up-regulated DEG can be determined by the effect of DMR regardless of the effect of TF. Thus, both up-regulated DEG with up-regulated TF and down-regulated DEG with down-regulated TF were removed.

3.4 Results and Discussions

3.4.1 Part 1 - Statistic analysis of mutation effect of seven DNA methylation modifier genes

To analyze the effects of seven DNA methylation modifier genes, I collected 3,865 TCGA methylome and transcriptome data. First, the number of mutation samples in DNA methylation modifier genes was found to be between 7% and 34% of the total sample for 12 major cancer types (Table 3.2). Excluding OV cancer that had only nine samples, 11 cancer types were analyzed.
The seven DNA methylation modifier genes that I studied were DNMT1, DNMT3A, MBD1, MBD4, TET1, TET2, and TET3. DNMT1 and DNMT3A function as DNA methyl-transfer and TET1, TET2, and TET3 have demethylation functions. Mutation statistics of the seven modifiers are summarized in Figure 3.2. Cancer types of BLCA, BRCA, COAD, LUAD, and LUSC were predominantly mutated in the TET genes that have de-methylation functions. In the case of LAML, DNMT3A mutation samples were high, while remaining GBM, HNSC, and KIRC, the ratio was similar. In the case of GBM, KIRC, and READ, the total mutation rate was less than 13%, and the number of mutations for each gene was 5 or less (Table 3.2).

Effect of mutations in seven DNA methylation modifier genes on genome-wide methylation landscapes

I compared genome-wide methylation landscapes between the mutated and the non-mutated groups. Since the comparison of genome-wide methylation landscapes between the two groups was difficult to interpret, I compared promoter regions instead.

Among the annotated 450,000 CpG sites, I selected the 140,040 sites as

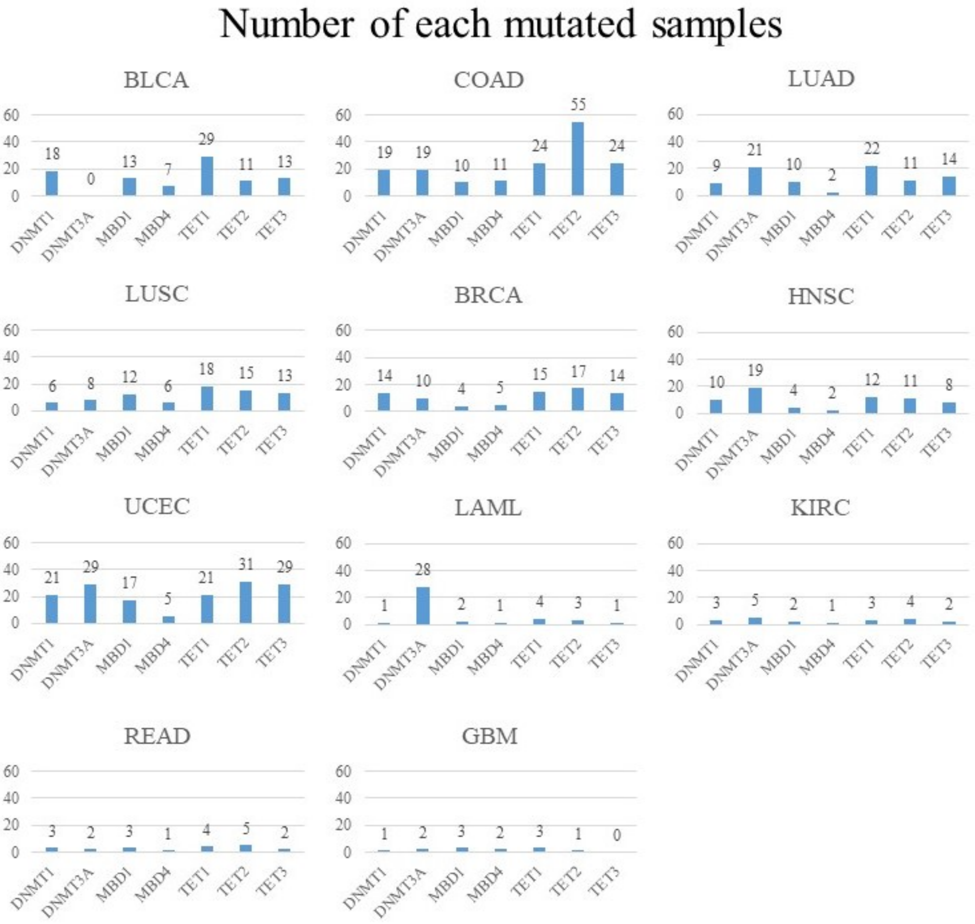


Figure 3.2 **The number of samples that each of the seven DNA methylation modifier genes is mutated.** DNMT3A mutation is dominant in LAML samples. In COAD, mutations in TET1, TET2, and TET3 are dominant.

Table 3.2 **Summary of the mutation status of seven DNA methylation modifier genes in each cancer.** Each value represents the number of samples that have both methylome and transcriptome data, the number of samples that have mutations in seven DNA methylation modifier genes, the number of samples that don't have mutations, ratio of the mutation samples per non-mutation samples, the number of DMRs and the number of DEGs that were selected by 0.05 false discovery rate.

Cancer type	Total samples	Mutated samples	Non-mutated samples	Mutation sample ratio	Number of DMRs	Number of DEGs
BRCA	784	61	723	8%	12,040	80
HNSC	521	60	461	12%	10,454	102
LUAD	455	80	375	18%	12,899	379
BLCA	408	98	310	24%	9,016	437
LUSC	369	73	296	20%	27,145	451
KIRC	319	21	298	7%	16,664	148
COAD	279	94	185	34%	43,982	904
UCEC	173	59	114	34%	54,956	2,079
LAML	170	33	137	19%	28,215	438
READ	93	12	81	13%	49,091	217
GBM	64	6	58	9%	79,204	173

promoters when the sites are annotated as TSS200 or TS1500; TSS200 is the region that covers zero to 200 bases upstream of the transcription start site (TSS) and TSS1500 covers 200 to 1500 bases upstream of the TSS. For each of 11 cancer types, methylation differences in 140,040 promoter regions of CpG sites were examined separately. I compared mutated and non-mutated samples of seven DNA methylation modifier genes, and the methylation values for each CpG site were expressed as log ratio values by comparing mean values. For the selected CpG sites, the average of DNA methylation of the mutation versus non-mutation samples was calculated as the log ratio and a heatmap was drawn by selecting 29,879 CpG sites with the log ratio value bigger than 1 or smaller than -1. In the heat map results, COAD showed the largest number of hypermethylation promoter regions, and LAML showed the lowest number of hypermethylation promoter regions. GBM showed the highest number of hypomethylation regions (Figure 3.3). The heatmap results showed that there was a change in methylation due to the mutation of seven DNA methylation modifier genes, and detailed analysis was conducted to investigate the CpG site of promoter region with methylation changes in 11 cancer types.

DMR analysis to investigate the mutation effects of seven DNA methylation modifiers.

Mutated samples of seven DNA methylation genes were compared with non-mutated samples using `bumperhunter` of the minfi package for DMR analysis. The significance of the number of DMRs potentially caused by the mutation of seven DNA methylation modifiers was compared with the number of DMRs in random samples. Random sampling DMR analysis was performed by repeatedly choosing samples of the same size for 10,000 times. P-value of the mutant sample was calculated from the distribution of DEG and DMR values obtained from 10,000 repeated tests. In the result of DMR test, 8 cancer types of 11, as BRCA, HNSC, LUAD, BLCA, LUSC, COAD, UCEC and LAML, showed significantly low p-value (Supplementary Figure). The other cancer type KIRC, READ, and GBM were not significant due to having few mutation samples (See Figure 3.2). Overall, it seemed that mutations of seven DNA methylation genes affected genome-wide promoter methylation differences.

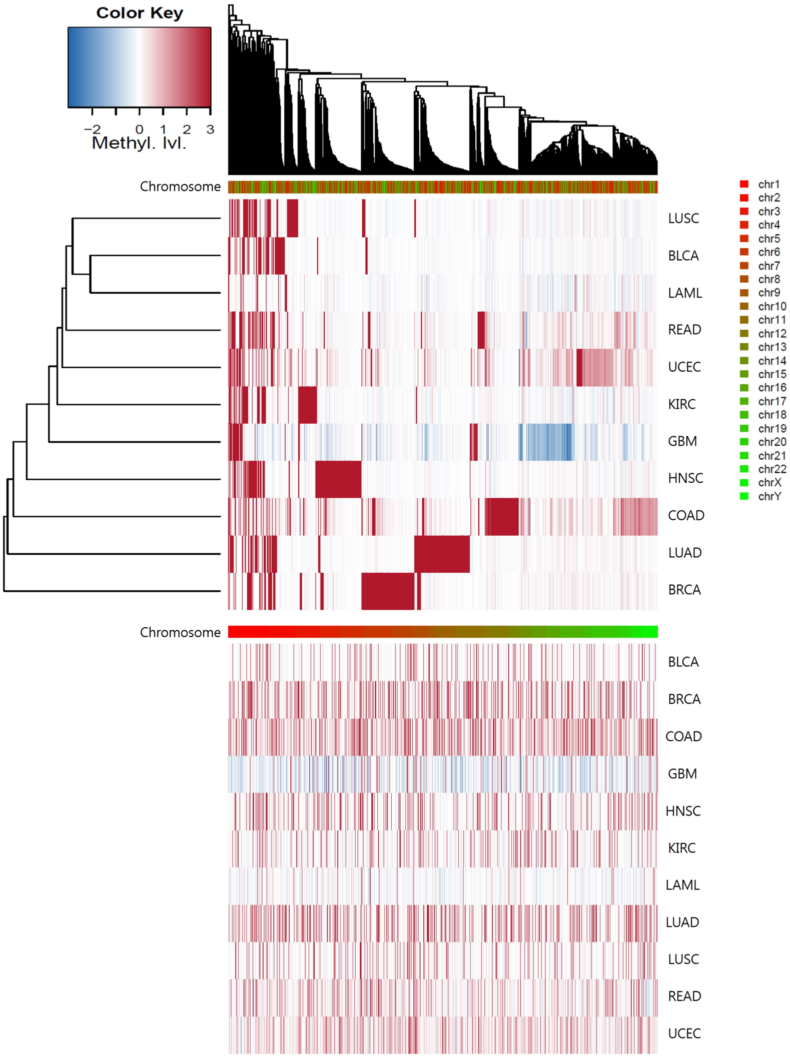


Figure 3.3 **Genome-wide landscape of promoter methylation.** Hypermethylated regions are colored in red and hypomethylated regions are colored in blue. The heatmap in the upper panel suggests that COAD shows a distinct sign of promoter hypermethylation while LAML shows no such tendency. The heatmap in the lower panel shows the methylation status in the order of chromosome and it is also observed that the promoters are hypermethylated in COAD and there is no strong methylation signal in LAML.

3.4.2 Part2 - Genome-wide association analysis of mutation effect of seven DNA methylation modifier genes

Sub-network clustering result in pan-cancer scale

I performed graph-based clustering of DEGs. First, I used the network topology of the STRING database and chose edges between two genes only when expression values of the two genes were highly correlated. Edges were weighted by the STRING database confidence scores. After that, the clustering was performed, and the clusters were filtered using the t-test.

The selected clusters were visualized using Cytoscape (Shannon et al., 2003) (Figure 3.4). Up-regulated DEG is displayed in a gradual red color and down-regulated DEG is displayed in a gradual blue color by the fold change value of gene expressions. Promoter DMR information was integrated into the DEG clusters and the case of DMR in the promoter of the up- and down-regulated DEG was marked in the cluster. DEGs with methylated promoter regions were colored in pink for hypermethylation and sky blue for hypomethylation.

Cluster selection for in-depth analysis

I performed Fisher's exact test with the number of DMR-DEGs (differentially expressed gene with differentially methylated promoter region) in each cluster to select statistically significant clusters. In the case of LAML, a cluster was selected in which mutated samples of DNMT3A were abundant and DEGs were up-regulated. In COAD cancer clusters, TET1/2/3 genes were mutated with promoter hypermethylated, so I selected a cluster that contained the largest number of down-regulated DEGs. For the functional analysis of DEGs in the clusters, I selected one cluster of up-regulated DEGs in LAML and another cluster of down-regulated DEGs in COAD (Figure 3.5).

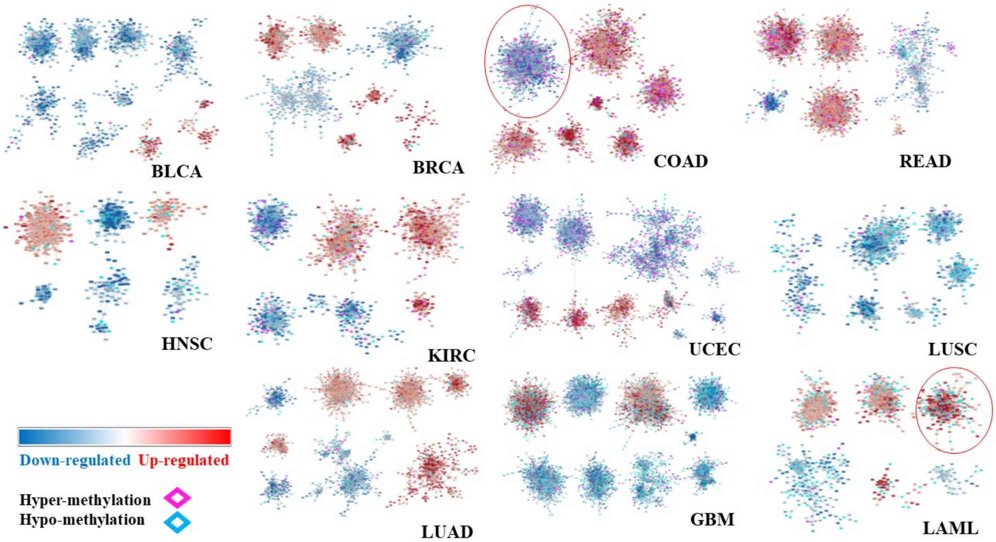


Figure 3.4 **Graph-based clustering results.** Up-regulated DEGs are colored in red and down-regulated DEGs are colored in blue. The red circles indicate the selected clusters in LAML and COAD.

TF selection related to DMR-DEGs

Among the genes in the clusters of COAD and LAML, I selected DEGs that the expression changes were not associated with TFs. To investigate TF-DNA-methylation interaction, I searched for all TF binding sites in the promoter regions using the Transfac (Matys et al., 2003) database. In COAD, there were 184 DMR-DEGs and I detected 381 TFs. In LAML, 86 DMR-DEGs were selected, and 254 TFs were detected by Transfac using a promoter sequence of DEGs.

3.4.3 Part 3 - DMR-DEGs in-depth analysis

Selection of cancers for in-depth analysis.

For the in-depth analysis to investigate the effect of mutations in DNA methylation modifiers, I first selected cancers based on the mutation profiles Figure 3.2. In COAD, the number of the samples of which the demethylation-related genes, TET1, TET2, and TET3, were mutated was bigger than that of the samples with mutations in the methylation-related genes. On the contrary, in LAML, mutations in the methylation-related genes, e.g., DNMT3A, were dominant. I also looked genome-wide promoter methylation landscape to see relations between the mutations in the methylation-related genes and the methylation status of the promoters of the genes. As shown in Figure 3.3, I was able to observe that there was a distinct signature of promoter hypermethylation in COAD (Figure 3.5). On the contrary, in LAML, the promoters were hypomethylated rather than hypermethylated. GBM also showed the promoter hypomethylation but the number of samples with mutations was too small to analyze the effect of mutations (Figure 3.2). Thus, I selected COAD and LAML for further analyses.

Selection of DMR-DEG possibly without TF-mediated regulation.

Before associating DMR-DEG, I excluded the DMR-DEGs that the expression changes were possibly affected by TFs. Among selected TFs that had binding sites in the promoter regions (see cluster selection in PART 2), if expression levels of TFs were different significantly between the mutated and non-mutated sample groups, TF expression difference could affect expression levels of downstream genes, thus I remove genes whose promoter regions had binding sites of such TFs. I set 0.2 and -0.2 as cutoff values for log_2 fold change to deter-

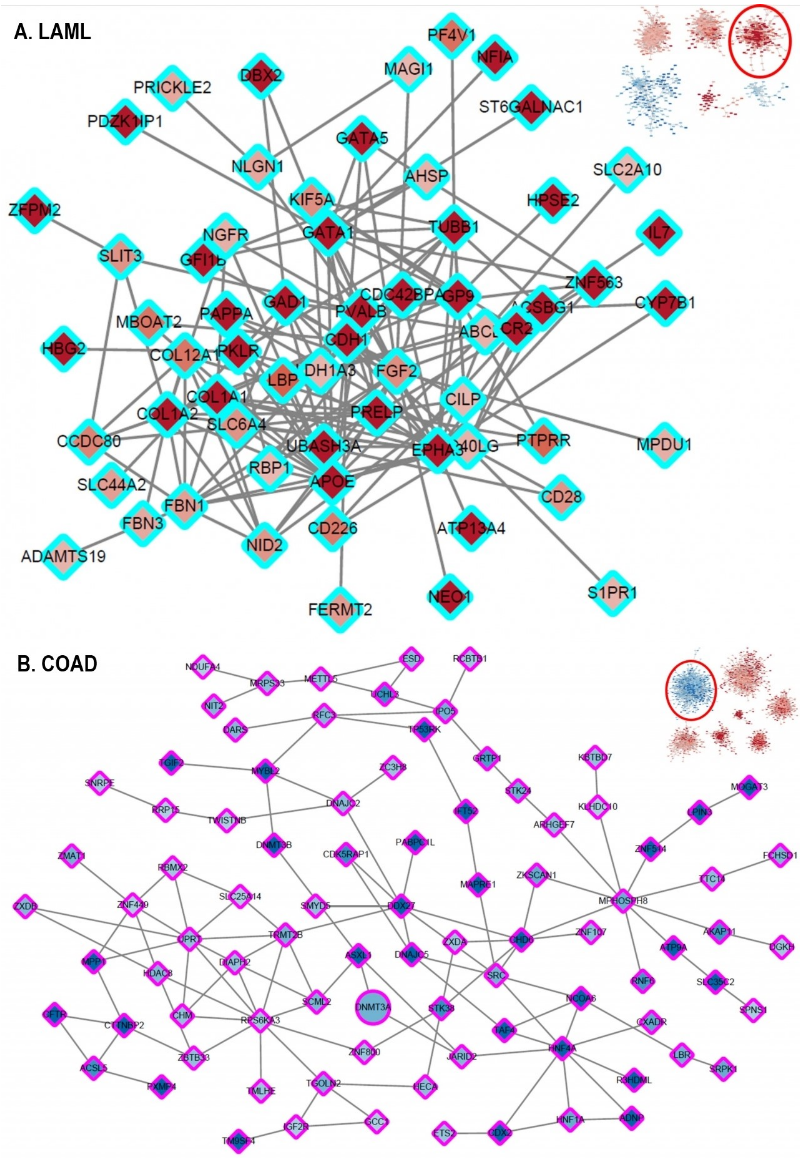


Figure 3.5 **Selected sub-network clusters in LAML and COAD.** Up-regulated genes were colored in red and down-regulated genes were colored in blue color according to the expression fold change level. The borders of the genes are colored in pink or sky blue when the promoters of the genes are either hypermethylated or hypomethylated, respectively.

mine if a gene or a TF is up-regulated or down-regulated. When a gene is up-regulated and a TF targeting the gene is up-regulated, the DEG was removed. Likewise, when a gene is down-regulated and a TF targeting the gene is also down-regulated, the DEG was removed. Finally, 42 DMR-DEGs in LAML and 61 DMR-DEGs in COAD were selected and studied for functional effects (Table 3.3).

Table 3.3 **List of 42 DMR-DEGs in LAML and 61 DMR-DEGs in COAD.**

Selected cluster	DMR-DEGs
The cluster of LAML	CD226, CACNA2D1, GP9, CD28, GATA5, GATA1, KIF5A, RNF182, ZNF563, NID2, TUBB1, FGF2, CR2, MINPP1, CD40LG, PF4V1, EPHA3, MBOAT2, TRIM58, ADAMTS19, PKLR, C7, NLGN1, CLCN4, IL7, COL1A2, COL1A1, SLC35D3, NCS1, SMARCA1, PRICKLE2, GFI1B, ATP13A4, NEO1, SLIT3, SLC44A2, FBN3, FBN1, MOV10L1, ST6GALNAC1, CILP, PDK3
The cluster of COAD	IMMP2L, UPRT, ZXDA, PRPF3, SMAP1, LBR, SAMD13, ZNF572, DNAJC15, MTMR6, MAPRE1, IFT52, CHM, POT1, TMLHE, ZNF449, ZKSCAN1, TTC14, ZNF775, NIT2, CDKN1B, ENAH, CHD6, LANCL2, GCC1, CEACAM6, HECA, MOGAT3, ZC3H8, ANKRD26, DNAJC5, DNMT3B, RPS7, SCML2, TP53RK, PABPC1L, AKAP8L, ARF5, REPS2, NDUFA4, ZNF800, CXADR, STAU2, PIPOX, EIF2AK1, ZNRF2, PHF20L1, ZMAT1, ELF1, CDK18, LPIN3, RCBTB1, MLLT3, HNF1A, USP11, PXMP4, ARL11, NCK2, RPL31, ATP6V1C1, ESD

42 up-regulated DEGs related to hypo-DMR in LAML

42 up-regulated DEGs with hypomethylated promoters were selected in LAML. To investigate the biological function of these genes, I searched the literature to find the relevance of these genes to LAML. For 42 DEGs in LAML, I searched with the terms ”methylation” or ”AML”. CD226, CACNA2D1, GATA1, EPHA3, IL7, GFI1B and SLIT3 genes are related to a disorder of methylation in Acute myeloid leukemia. CD226 (Cluster of Differentiation 226, DNAM-1 (DNAX Accessory Molecule-1)) is a 65 kDa glycoprotein expressed on the surface of natural killer cells, platelets, monocytes and a subset of T cells. TIGIT binding with CD226 has up-regulated on CD8(+) T cells in LAML (Sanchez-Correa et al., 2012). CACNA2D1 (Voltage-dependent calcium channel subunit alpha-2/delta-1) encodes a member of the alpha-2/delta subunit family, a protein in the voltage-dependent calcium channel complex. CACNA2D1 has DMR in oxytocin signaling pathway in LAML (Gao et al., 2018). GATA1 (GATA-binding factor 1) regulates the expression of an ensemble of genes that mediate the development of red blood cells and platelets. Its critical roles in red blood cell formation include promoting the maturation of precursor cells. GATA-1 binds to the PU.1 gene and inhibits expression in LAML (Burda et al., 2016). EPHA3 (ephrin type-A receptor 3) has been implicated in mediating developmental events, particularly in the nervous system. Receptors in the EPH subfamily typically have a single kinase domain and an extracellular region containing a Cys-rich domain and 2 fibronectin type III repeats. EphA3 was methylated in leukemia patients (Rush et al., 2004). IL7 (Interleukin 7) stimulates proliferation of all cells in the lymphoid lineage (B cells, T cells, and NK cells). IL-7 has abnormal methylation in peripheral blood of LAML patients (Li et al., 2019). GFI1B (Growth factor independent 1b, Zinc finger protein Gfi-

1b) are highly expressed in LAML (Vassen et al., 2009). SLIT3 (Slit homolog 3 protein) is a ligand-receptor SLIT-ROBO family. Low expression of SLIT and high expression of ROBO1 and ROBO2 suggests their participation in LAML pathogenesis (Gołos et al., 2019).

FGF2, SLC44A2, and PDK3 genes are related to LAML. FGF2 (basic fibroblast growth factor) is present in basement membranes and in the subendothelial extracellular matrix of blood vessels. FGF2 promotes resistance to FLT3 inhibitors in acute myeloid leukemia (Traer et al., 2016). SLC44A2 (Choline transporter-like protein 2) is located in a pathway controlling DNA damage and repair and affects the survival in LAML (Bruedigam et al., 2014). PDK3 (Pyruvate dehydrogenase lipoamide kinase isozyme 3) inhibits pyruvate dehydrogenase activity by phosphorylation of the E1 subunit PDHA1 and thereby regulates glucose metabolism and aerobic respiration. The overexpression of PDK3 conferred poor prognosis in LAML (Cui et al., 2018a).

In GO-term enrichment test with ”Molecular Function” category, the 42 genes in LAML were found to be related with ”platelet-derived growth factor binding”, ”integrin binding”, ”protease binding”, ”RNA polymerase II transcription factor binding”, ”voltage-gated calcium channel activity” and ”cytokine activity” (Table 3.4).

Table 3.4 **Enriched GO terms of 42 DMR-DEGs in LAML.**

GO term ID	Term description	P-value	Z-score
GO:0048407	platelet-derived growth factor binding	0.00028	-2.6767
GO:0005178	integrin binding	0.001042	-1.44853
GO:0002020	protease binding	0.001991	-1.69845
GO:0001085	RNA polymerase II transcription factor binding	0.002137	-1.13219
GO:0005245	voltage-gated calcium channel activity	0.002885	-1.8684
GO:0005125	cytokine activity	0.004276	-1.21879

61 down-regulated DEGs related to hyper-DMR in COAD

61 down-regulated DEGs with hypermethylated promoters were selected in COAD. To investigate the biological function of these genes, I searched the literature to find the relevance of these genes to COAD. For 61 DEGs selected in the cluster of COAD, I searched the literature with the terms ”methylation” or ”Colon adenocarcinoma”. CDKN1B, DNMT3B, and RPS7 genes are related to a disorder of methylation in Colon Adenocarcinoma. CDKN1B (Cyclin-dependent kinase inhibitor 1B, p27) is considered a tumor suppressor because of its function as a regulator of the cell cycle. In cancers, it is often inactivated via impaired synthesis, accelerated degradation, or mislocalization. Downregulation of CDKN1B is caused by increased ubiquitin-mediated proteasomal degradation in colorectal cancer(Ogino et al., 2007). DNMT3B (DNA (cytosine-5-)-methyltransferase 3 beta) encodes a DNA methyltransferase which is thought to function in de novo methylation, rather than maintenance methylation. The protein localizes primarily to the nucleus and its expression is developmentally regulated. DNMT3B expression contributes to CpG island methylator phenotype in colorectal cancer (Nosho et al., 2009). RPS7 (40S ribosomal protein S7) is a component of the 40S subunit. In eukaryotes, ribosomes, the organelles that catalyze protein synthesis, consist of a small 40S subunit and a large 60S subunit. Aberrant promoter hypermethylation of RPS7 inhibits colorectal cancer growth (Zhang et al., 2016).

UPRT, MAPRE1, NIT2, CEACAM, and NDUFA4 genes are related with COAD. UPRT (Uracil Phosphoribosyltransferase Homolog) modulate the sensitivity of the human colon cancer cells (Koyama et al., 2000). MAPRE1 (Microtubule-associated protein RP/EB family member 1) is often mutated in familial and sporadic forms of colorectal cancer (Ladd et al., 2012). NIT2 (Nitrilase Family

Member 2) has an omega-amidase activity to remove potentially toxic intermediates by converting alpha-ketoglutaramate and alpha-ketosuccinamate to biologically useful alpha-ketoglutarate and oxaloacetate. Downregulation of NIT2 inhibits colon cancer cell proliferation and induces cell cycle arrest (Zheng et al., 2015). CEACAM6 (Carcinoembryonic antigen-related cell adhesion molecule 6) is a member of the carcinoembryonic antigen (CEA) gene family. Expression of CEACAM6 in colorectal cancer is an independent prognostic factor that can subdivide patients into low-risk and high-risk groups (Jantscheff et al., 2003). NDUFA4 (mitochondrial complex associated) codes for a subunit of Complex I of the respiratory chain, which transfers electrons from NADH to ubiquinone. Up-regulated NDUFA4 facilitate the tumorigenesis of colorectal cancer (Cui et al., 2018b).

In GO-term enrichment test with "Molecular Function" category, the 61 genes in COAD were found to be related with "ubiquitin-protein transferase inhibitor activity", "RNA polymerase II intronic transcription regulatory region sequence-specific DNA binding", "hydrogen-exporting ATPase activity, phosphorylative mechanism", "chromo shadow domain binding", "G-rich strand telomeric DNA binding", "intronic transcription regulatory region sequence-specific DNA binding" and "C2H2 zinc finger domain binding" (table 3.5).

3.5 Methods

DEG analysis

A Bioconductor (version 3.8) EBSeq package (Leng et al., 2013) was used for the DEG analysis of RNA data. For each cancer type, I divided the samples into two groups into mutated versus non-mutated samples and performed DEG analysis. Number of DEGs was counted with a false discovery rate (FDR) less

Table 3.5 **Enriched GO terms of 61 DMR-DEGs in COAD.**

GO term ID	Term description	P-value	Z-score
GO:0055105	ubiquitin-protein transferase inhibitor activity	0.021159	-3.31913
GO:0001162	RNA polymerase II intronic transcription regulatory region sequence-specific DNA binding	0.021159	-3.28735
GO:0008553	hydrogen-exporting ATPase activity, phosphorylative mechanism	0.030091	-3.26128
GO:0070087	chromo shadow domain binding	0.021159	-3.00936
GO:0098505	G-rich strand telomeric DNA binding	0.033051	-2.87252
GO:0001161	intronic transcription regulatory region sequence-specific DNA binding	0.027123	-2.78111
GO:0070742	C2H2 zinc finger domain binding	0.038944	-2.77426

than 0.05. Fold change values of gene expression level were used in the following clustering analysis.

DMR analysis

For the methylation data analysis, the DMR was analyzed with an FDR of 0.05 using ”bumperhunter” in the minfi package (Aryee et al., 2014) of Bioconductor (version 3.8). For each cancer type, I divided the samples into two groups into mutated versus non-mutated samples as same as DEG analysis. The DMRs found were annotated using ”matchgene” to select the genes with DMR in the promoter.

Random sample test

Random sampling was performed to compare the seven DNA methylation modifier mutation samples of each cancer types. Random samples were selected with the same size as the seven DNA methylation modifiers mutation samples, and

DEG and DMR analysis were performed 10,000 times using the selected and remaining samples.

The log ratio of average methylation levels in promoter regions

To compare the methylation levels of each promoter region between the samples of which the seven DNA methylation modifier genes were mutated and the other samples, I calculated the average of methylation levels of each promoter region for the samples with mutation and the other samples, respectively. After that, the log ratio of the averaged methylation levels was calculated and the equation is shown below:

$$LR_{ij} = log_2 \frac{Avg\_mut_{ij} + pseudo}{Avg\_non_{ij} + pseudo}$$

where j indicates each probe, i is the index of cancer, Avg_mut_{ij} is the average of the methylation levels of probe j for the samples with mutation in cancer i, Avg_non_{ij} is the average of the methylation levels of probe j for the samples without mutation in cancer i and LR_{ij} is the log ratio of two average values of probe j in cancer i. Pseudo is the value of 0.001 I added to the averages to avoid the error caused by dividing by zero.

Gene expression correlation analysis

For transcriptome data, correlation values between genes were calculated using Pearson's correlation of "pearsonr" of scipy (Blanco-Silva, 2013) for each cancer type. The final correlation value between the final genes was calculated using the weight value of the PPI score of the STRING database. These correlation values are used in the following clustering analysis.

Graph-based clustering

I used the ”igraph” package (Csardi et al., 2006) of R to detect the multilevel community and perform sub-network clustering. For the graph-based clustering, I used the fold change value of the gene and correlation values between genes. Before clustering, I discard genes with fold change less than 0.2 and edge of correlation with less than 0.5. After clustering, I perform the GO enrichment test and one-sample t-test for each cluster.

Network visualization with Cytoscape

Visualization of the sub-network cluster is shown using Cytoscape (version 3.7.1).

Promoter binding TF search by Transfac

To search all TFs to bind the promoter sequence of DEG, I used Transfac.

Chapter 4

In silico experiment system for testing hypothesis on gene functions using three condition-specific biological networks

4.1 Related works

Important regulators such as TF genes have system-wide effects on many genes, often resulting in significant changes in phenotypes (Latchman, 1997). To understand the role of TF, it is a common practice to use model organisms, e.g., a mouse with the TF knocked out. Subsequently, sequencing technologies are used to measure changes in gene expression levels at the whole cell level. A common practice for the analysis of transcriptome data is to perform the DEG analysis to measure the system-wide effects of a TF. However, the DEG analysis has several limitations. First, there are too many DEGs, up to several thousand, depending on the criteria for beings DEGs. More importantly, the DEGs analysis do not explain how a TF affects DEGs since connections from the TF to DEGs are unknown. In addition, users do not have any control on the DEG analysis process except changing the cut-off values, even when the user has a good hypothesis on which biological mechanisms or pathways are likely to be affected by knocking out the TF.

Recently, there have been significant advances in bioinformatics technologies. Among them, a number of biological networks have been constructed using experimental data and/or computational methods. Thus, it is possible to use networks to investigate the system-wide effects of a TF by following edges of networks from the TF to all other genes. In addition, literature mining technologies have been advanced significantly and they were used in the recent research projects (Hur et al., 2016; Lee et al., 2016a; Oh et al., 2017). These literature mining technologies are now powerful enough to identify the relationship between the specific hypothesis of the user, e.g., disease names or certain biological pathways, and genes that are reported to be relevant to the hypothesis in the literature. By leveraging these recent advances, I developed a novel information system that can be used to perform *in silico* experiments for testing on functions of a TF.

4.2 Methods

An *in silico* experiment is performed as follows. Given a user-provided transcriptome and miRNA data from a knockout mouse experiment, the user can specify his/her hypothesis in English. The current system may not handle free-style sentences, thus a set of nouns are to be specified as input. Then, the hypothesis is translated to a set of genes using my literature knowledge mining system, BEST (Lee et al., 2016b). I call these genes *target genes*. Then, connections from the knockout gene to the target genes are constructed and evaluated by condition-specific networks that are instantiated by gene expression data. Three condition-specific networks are TF, miRNA and PPI networks. The connectivity between the regulator gene such as the knocked out TF and the target genes are determined by computing shortest paths. Intuitively, more target genes are reachable from the TF, an *in silico* experiment accepts or supports the user hypothesis while fewer connections would reject the hypothesis. Of course, my literature based experiment is not meant to use to determine the function of a TF since biological experiments should be performed to confirm the functions. However, my system allows the user to exploit his/her expert knowledge to explore potential functions of a TF, which, I expect, will reduce the burden of scientists significantly so that much smaller number of *in vivo* or *in vitro* experiments can confirm the function of a TF.

User input

- Transcriptome (mRNA expression) data
- microRNA expression data
- Regulator gene name (ex. knockout gene name)
- Hypothesis (ex. disease, pathway or gene) specified by a set of nouns

(My tool supports raw data of microarray, pre-processed data of microarray and pre-processed data of RNA-seq. I do not support raw data of RNA-seq.)

Output result

- DEG analysis result
- Candidate target genes related to regulator gene and hypothesis

- The network of regulator gene and target gene within TF, microRNA and PPI network
- Statistical and informational test results

4.3 Workflow

The workflow of my system is shown in Figure 4.1. Each step and the workflow is explained in detail in this section. To help understand readers the workflow, I will define genes in three categories. A *regulator gene* is a TF gene that is knocked out in the biological experiment. The *Mediator* or *network genes* are genes in the condition-specific TF network, miRNA network, and PPI network. *Target genes* are genes that are relevant to the hypothesis or pathway that the user specified. These genes are called as targets since my *in silico* experiment is to test how well a regulator gene is connected to the target genes via network genes.

4.3.1 Step 1 - Select target genes from hypothesis

1.1. DEG analysis of miRNA and mRNA.

The input mRNA and miRNA expression data are analyzed using `limma` (Ritchie et al., 2015) for microarray and `DEseq2` (Love et al., 2014) for RNA-sequencing data. DEGs are selected based on the log2 fold change value and the p-value are calculated for the expression level of each gene.

1.2. Search target genes related to the hypothesis from the literature.

The user needs to provide, as input, the regulator gene name and the hypothesis that is specified in English. The validity of an input gene name can be checked by clicking the check button on the web page. Instead of specifying a hypothesis,

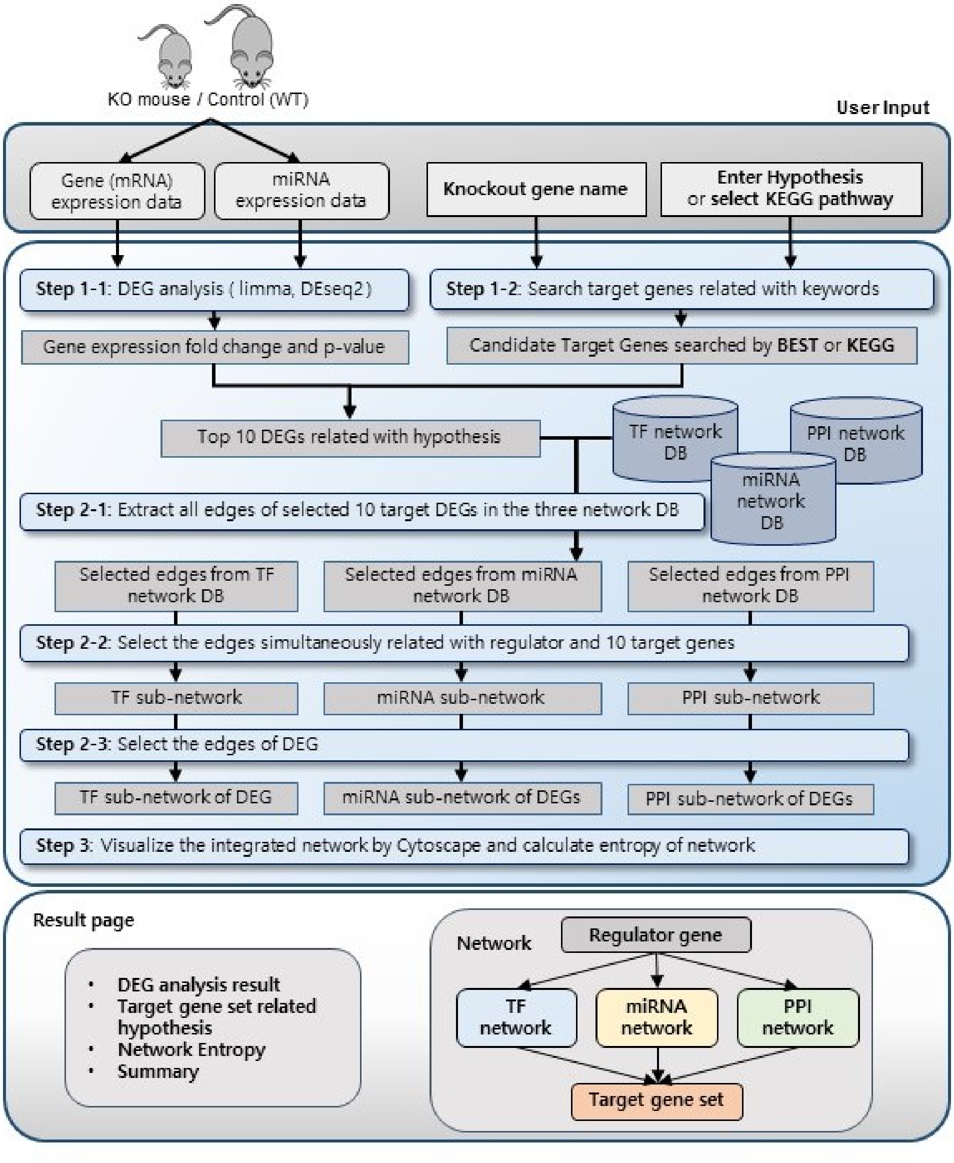

KO mouse / Control (WT)
User Input
Gene (mRNA) expression data
miRNA expression data
Knockout gene name
Enter Hypothesis or select KEGG pathway
Step 1-1: DEG analysis (limma, DEseq2)
Step 1-2: Search target genes related with keywords
Gene expression fold change and p-value
Candidate Target Genes searched by BEST or KEGG
Top 10 DEGs related with hypothesis
TF network DB
PPI network DB
miRNA network DB
Step 2-1: Extract all edges of selected 10 target DEGs in the three network DB
Selected edges from TF network DB
Selected edges from miRNA network DB
Selected edges from PPI network DB
Step 2-2: Select the edges simultaneously related with regulator and 10 target genes
TF sub-network
miRNA sub-network
PPI sub-network
Step 2-3: Select the edges of DEG
TF sub-network of DEG
miRNA sub-network of DEGs
PPI sub-network of DEGs
Step 3: Visualize the integrated network by Cytoscape and calculate entropy of network
Result page
DEG analysis result
Target gene set related hypothesis
Network Entropy
Summary
Network
Regulator gene
TF network
miRNA network
PPI network
Target gene set

Figure 4.1 **The workflow of my method.** The schematic diagram of the workflow.

the user can select a KEGG (Kanehisa and Goto, 2000) pathway from the list of KEGG pathways that are provided on the web page. With a user-specified hypothesis, the BEST system converts the hypothesis to a set of genes. When the user selects a pathway name, genes in the pathway is selected as candidate target genes by searching the KEGG pathway database.

1.3. Select top DEGs as target genes.

Using the DEG analysis (Step 1-1) results, top 10 target DEGs in terms of gene expression level changes are selected from the candidate target genes determined by BEST tool or KEGG pathway DB (Step 1-2) with a p-value. The top 10 target DEGs consist of 5 up-regulated DEGs and 5 down-regulated DEGs.

4.3.2 Step 2 - Condition-specific TF, miRNA and PPI network generation by the DEG set

2.1. Extract all edges of selected 10 target DEGs in the three network DB.

Template TF network, miRNA network, and PPI network are instantiated by the gene expression information from the microarray or RNA-seq experiments. However, these networks are too big to be displayed on a web page. Thus, genes that are directly or indirectly connected to the 10 target genes are chosen since genes that do not have a connection to the target genes are not relevant. A miRNA network database was obtained from TargetScan (Lewis et al., 2005), and STRING (Franceschini et al., 2013) was used as a PPI network. A mouse TF network database was created using NARROMI (Zhang et al., 2013).

2.2. Select the edges of DEGs.

Since I am interested in how the regulator gene affected the target genes, gene expression levels should be different in the expression value in the control vs. treated experiment. Thus, only edges incident to DEGs are selected and the others are removed from the network.

4.3.3 Step 3 - Performing *in silico* experiment

3.1. Compute the shortest path between the regulator gene and target genes in the networks.

My system computes the shortest path in the networks between the user provided regulator gene and 10 target genes from the hypothesis that is being tested.

3.2. Visualize networks in a graph.

The networks computed in the previous steps are visualized using `Cytoscape` (Shannon et al., 2003). The regulator gene is located at the top position, and the target genes are located at the bottom, and network genes are grouped into TF, miRNA, and PPI networks. Edges in the networks show how each of the genes is connected according to changes in expression level. The expression level of each gene was visualized in color according to the amount of change. Up-regulated genes are in red and down-regulated genes are in blue. In addition, a coding gene is denoted by a circle and a non-coding gene is denoted by a diamond shape. The blue colored edge represents the miRNA network, the purple colored edge represents the PPI network, and the yellow colored edge represents the TF network. Clicking a specific gene shows a list of connected genes.

4.3.4 Step 4 - Evaluation of the user hypothesis

4.1. A statistical evaluation by random permutation.

To evaluate network connections between the regulator gene and the target genes, a statistical evaluation is performed. While edges incident to the regulator gene and the target gene are fixed, random network networks are generated 1000 times to compute a p-value. The p-value is in how many times random networks the target genes are connected from the regulator genes. The significance of p-value means that the regulator gene has an effect on the target genes through mediator DEGs.

4.2. An information theoretic evaluation by measuring entropy.

In the network results, if there is a high correlation between the regulator gene and the target genes of hypothesis, there is a large number of mediator genes linking the regulator and the target genes. In addition, if the regulator gene relates to the given hypothesis, most of the target genes are connected via three-level networks. To quantitatively measure this connectivity concept, a normalized entropy of network connections is calculated by using degree information of target genes (Equation (4.1), (4.2), (4.3)).

$$\begin{aligned} P(tg_i) &= \frac{degree(tg_i) + \beta}{\sum_{j=1}^{10}(degree(tg_j) + \beta)} && (4.1) \\ H(TG) &= -\sum_{i=1}^{10} P(tg_i)\log_2 P(tg_i) && (4.2) \\ E_H &= H(TG)/H_{max} = H(TG)/\log_2 10 && (4.3) \end{aligned}$$

where:

- $degree(tg_i)$: degree of i-th target gene in the Target Gene Set $TG = \{tg_1, tg_2, ..., tg_{10}\}$
- $beta$: pseudo count ($beta = 0.00001$)
- H_{max} : maximum entropy of generated network

4.3.5 Optimization

To make an *in silico* experiment performed online, I need to speed up some computations, especially for statistical significance of connectivity between the regulator gene and target genes from the user provided hypothesis.

To compute a p-value, a network with 18 million edges and 31,897 nodes is randomly generated 1,000 times. This experiment takes too much time to provide the service online, so instead of rebuilding the entire network, I create a partial network where edges are generated randomly only for edges incident to nodes in the path between the regulator gene and the target genes according to the edge formation probability from the network density information. I also pre-upload the entire network at the *in silico* network server.

4.3.6 Explanation of the experiment result page

In the *in silico* experiment result, the user can test the hypothesis visually on the web page and a new *in silico* experiment can be retested again by simply clicking the button. It also shows a target gene list associated with hypothesis and the DEG analysis results of the mRNA and miRNA data and combines the results to visualize the top ten target genes associated with the regulator gene and hypothesis. In addition to the selected five up-regulated genes and the five down-regulated genes, additional genes of interest can be added as target genes. The network analysis results can be visualized as a graph. The networks can be zoomed in or out. Selecting a gene in the network shows only neighbors in a highlighted fashion. Genes in the networks can be reconfigured to arbitrary positions by clicking and dragging. Multiple genes can be selected for navigation by dragging a rectangle. A summary of the network information is presented along with p-value and entropy values.

4.4 Results & discussion

4.4.1 Test results of E2f1 and the hypothesis

I tested the *in silico* experiment system with a miRNA and mRNA dataset of GSE33902 from Gene Expression Omnibus (GEO) at NCBI and obtained the regulator gene name `E2f1` and hypothesis keyword `Lymphoma` from the original paper (Warg et al., 2012) that produced the data from an E2f1 knockout mouse. A null hypothesis keyword for E2f1 data, `Muscular Dystrophy` was selected in reference to the disease outcome of E2f1 in the MalaCards (Rappaport et al., 2016). Thus, the test was whether the hypothesis of `Lymphoma` is accepted and the hypothesis of `Muscular Dystrophy` was rejected in the two *in silico* experiments. The network results are shown in Figure 4.2. From the *in silico* experiment of E2f1 and `Lymphoma`, I found 127 genes associated with E2f1 and Lymphoma by BEST. Top five up-regulated genes, Cdkn1b, E2f2, H2afx, Bbc3, and E2f1, and top five down-regulated genes, E2f3, Anxa5, Rbl2, Casz1, and Ezh2, were used for network analysis. I found 55 DEGs connecting selected top 10 target genes, 4 in the miRNA network, 9 in the TF network and 41 in the PPI network. The total number of edges was 201, and the entropy value was 0.784. A higher entropy value means the well-connected network. E2f1 has a high correlation with the immune disease at 0.784 entropy score. Network connectivity maps by E2f1 and immune disease Lymphoma are shown

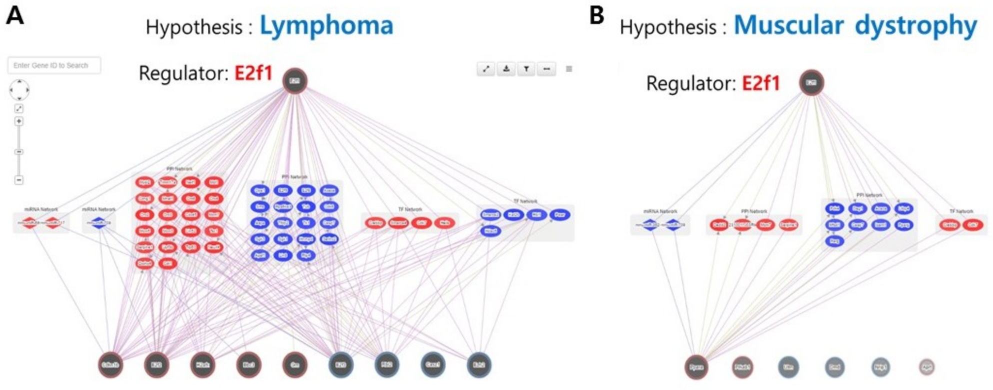


Figure 4.2 **Network results of E2f1 and hypotheses.** (A) (Accept case) Network result of "E2f1" (regulator gene) and "Lymphoma" (hypothesis). (B) (Reject case) Network result of "E2f1" and "Muscular Dystrophy".

in Figure 4.2 (A). For the `Muscular Dystrophy` hypothesis that is not related to E2f1 in the MalaCards (Figure 4.2 (B)), I found eight genes associated with E2f1 and Muscular Dystrophy. Among the eight genes, only E2f1 and Ppara genes were DEGs. I found 17 intermediated DEGs, 2 in the miRNA network, 2 in the TF network and 13 in the PPI network. This network graph had only 34 edges and the entropy value was 0.0138. Thus, E2f1 and Muscular Dystrophy did not seem to be related.

4.4.2 Test results of E2f1 and 62 diseases in the MalaCards as a hypothesis

I performed *in silico* experiments with 62 hypotheses of different disease names, selecting five diseases with a high MalaCards information score (MIFTS) in each of 18 categories of the MalaCards and excluding duplicate diseases. In the MalaCards DB, 14 diseases were related to E2f1 and the remaining 48 diseases were unrelated. As a result of the experiment, among 14 diseases associated with E2f1, 10 diseases such as Hepatitis, Hepatocellular Carcinoma, Colorectal Cancer, Lung Cancer, Retinoblastoma, Breast Cancer, Pancreatic Cancer, Prostate Cancer, Renal Cell Carcinoma and Esophageal Cancer had entropy values over 0.7 (Figure 4.3 (A)). For 14 diseases associated with E2f1, Table 4.1 summarizes entropy values for each experiment and reference to research papers that support the relevance of the disease to E2f1. 31 of 48 other diseases were not found to be associated with E2f1 because there was no target gene from the hypothesis or only a small number of genes are mapped to networks. Among the 48 diseases not associated with E2f1, four diseases, such as Obesity, Liver Disease, Ataxia-Telangiectasia, and Asthma, were accepted with an entropy value of over 0.8 in the experimental results, which could be new findings (Figure 4.3 (B)).

4.4.3 Test results of Lrrk2 and 23 diseases in the MalaCards as a hypothesis

I conducted additional tests using Lrrk2 data (GSE52584) that were downloaded from GEO. Leucine-rich repeat kinase 2 (Lrrk2) is an enzyme encoded by the PARK8 gene (Paisán-Ruız et al., 2004). I tested Lrrk2-related 23 diseases from the MalaCards database as a hypothesis. Seven diseases had an entropy value of 0.8 or more and 18 diseases had an entropy value of over 0.5, thus accepted in the *in silico* experiments. (Figure 4.4 (A)).

4.4.4 Test results of Dicer1 and 32 diseases in the MalaCards as a hypothesis

I conducted additional tests using Dicer1 data (GSE34910) that were downloaded from GEO. Dicer1 that is classified a Ribonuclease III, has a role of processing microRNA. 32 diseases that are known to be related with Dicer1 in

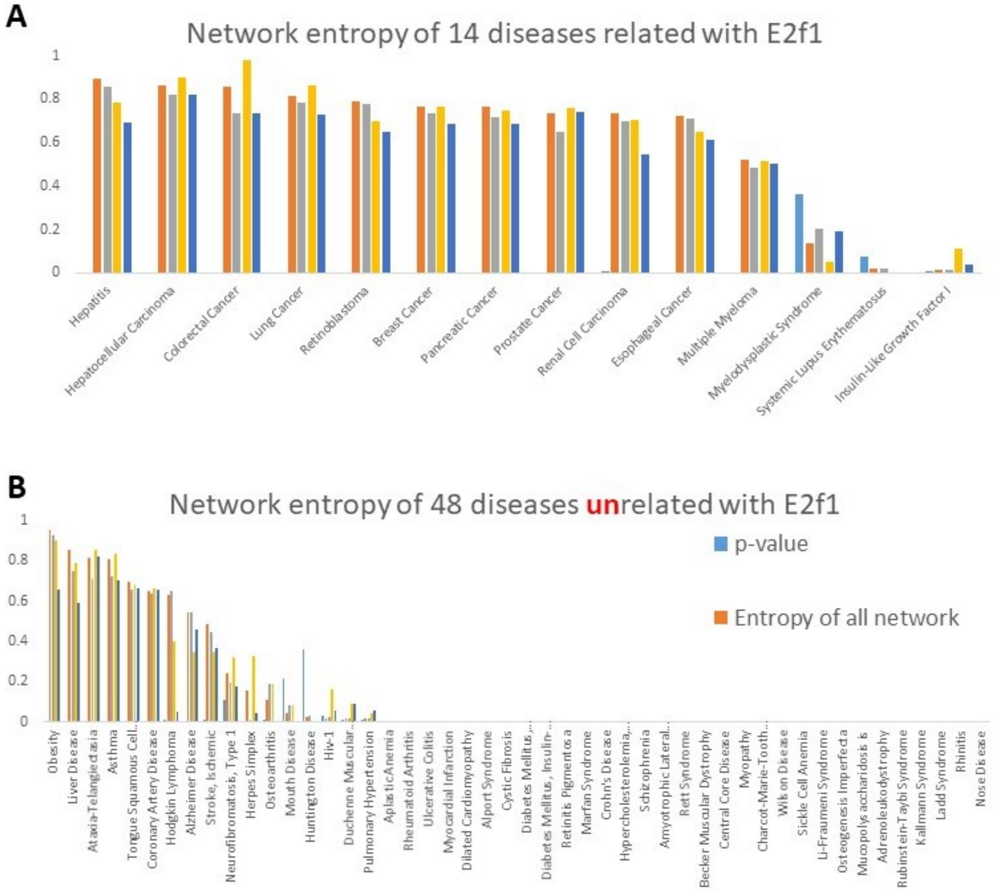


Figure 4.3 **Network entropy of E2f1 related diseases.** (A) Network entropy results of 14 diseases associated with E2f1. 10 of 14 had entropy values over 0.7. A high entropy value means that the association is high. (B) 31 of 48 other diseases were not found to be associated with E2f1 because there was no target gene or only a small number of genes are mapped to networks. 62 diseases were selected 5 diseases with a high MIFTS in each of 18 categories of MalaCards and excluding duplicate diseases.

Table 4.1 **A summary of *in silico* experiments with 14 diseases known to be relevant to E2f1 in MalaCards.** Shown are entropy value in the descending order and research papers that support the relevance of E2f1 to the disease. 10 of 14 diseases had over 0.7 entropy values, thus accepted in the experiments.

Disease name	Entropy (0∼1)	Reference
Hepatitis	0.8974	(Ghosh et al., 2016)
Hepatocellular Carcinoma	0.8628	(Ghosh et al., 2016)
Colorectal Cancer	0.8559	(Sulzyc-Bielicka et al., 2016)
Lung Cancer	0.8169	(Li et al., 2016)
Retinoblastoma	0.7921	(Pappas et al., 2017)
Breast Cancer	0.7676	(Cataldo et al., 2016)
Pancreatic Cancer	0.7671	(Chen et al., 2017)
Prostate Cancer	0.7371	(Liang et al., 2016)
Renal Cell Carcinoma	0.7369	(Gao et al., 2016)
Esophageal Cancer	0.7235	(Li et al., 2015)
Multiple Myeloma	0.5192	(Liu et al., 2013)
Myelodysplastic Syndrome	0.1358	(Saberwal et al., 2003)
Systemic Lupus Erythematosus	0.0177	(Aboelenein et al., 2013)
Insulin-Like Growth Factor I	0.0123	(Schayek et al., 2010)

the MalaCards database were tested. Breast cancer had 0.9341 entropy value, and 14 diseases including Blastoma, Lymphoma, and Ovarian cancer had an entropy value of over 0.7 and 21 of 32 diseases had an entropy value of over 0.5, thus accepted in the *in silico* experiments. (Figure 4.4 (B)).

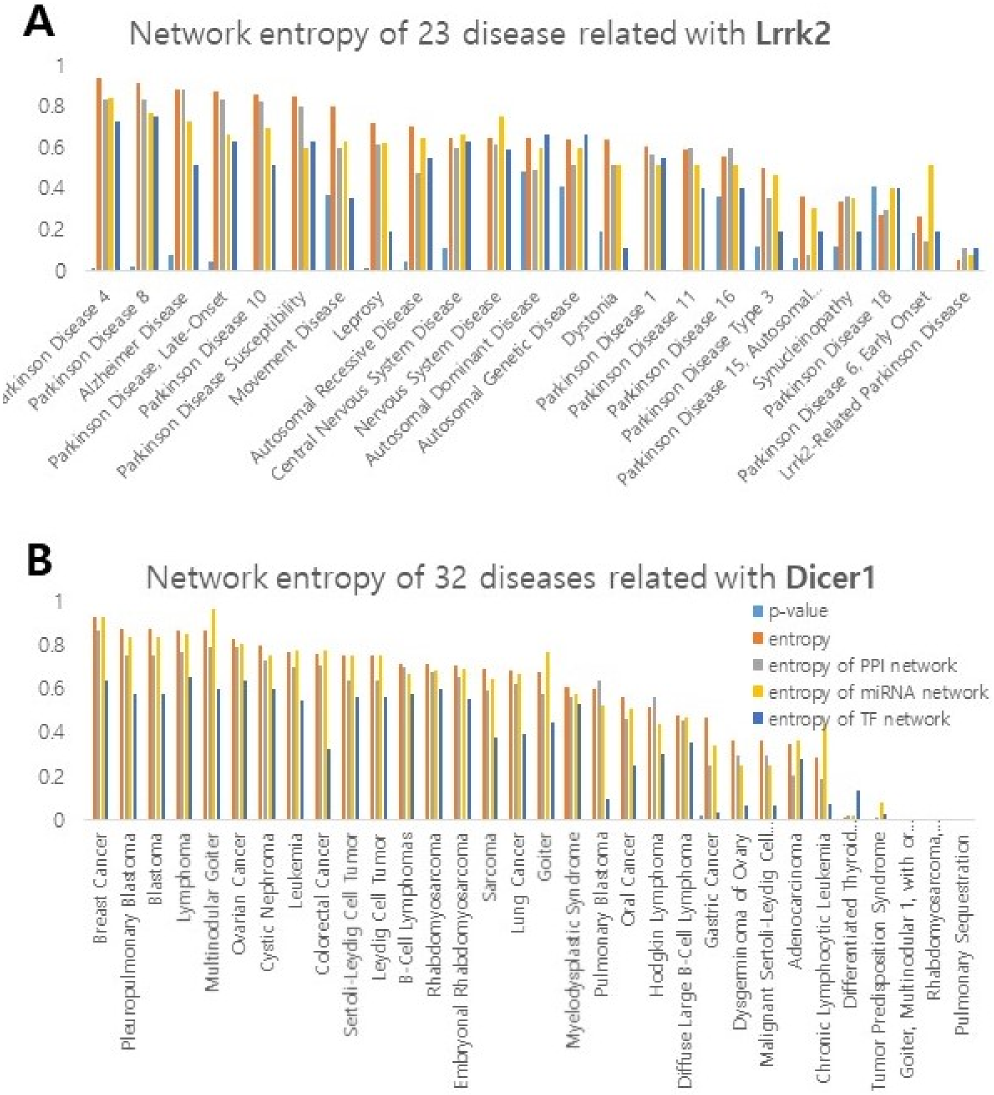


Figure 4.4 Network results of Lrrk2 and Dicer1. (A) Network entropy result of 23 diseases associated with Lrrk2. 7 of the 23 diseases had entropy values over 0.8 and 18 had over 0.5. (B) Network entropy result of 32 diseases associated with Dicer1. 14 of the 32 diseases had entropy values over 0.7 and 21 had over 0.5, thus accepted in the *in silico* experiments.

Chapter 5

Conclusion

The phenotype of an organism is manifested through the complex changes of various substances that make up the organism. To understand these biological phenomena, it is necessary to analyze various biological data in an integrated manner. Integrating and analyzing multi-omics, which are various data of biology, is a very difficult problem because it has a high dimensional large feature space. To solve the difficulty of an integrated analysis of multi-omics, it is very useful to use prior knowledge network information. This thesis presented three studies using prior knowledge network information for integration analysis of different omics data:

1. A study using miRNA networks and PPI networks for integrated analysis of miRNA and mRNA data.

2. A study using methylation network and gene expression correlation network clustering for integrated analysis of methylation and gene expression data.

3. A study on the development of hypothesis testing tools for key genes with the disease using PPI, TF and miRNA networks.

In the first study, integrated analysis of omics data using microRNA-target mRNA network and PPI network reveals regulation of Gnai1 function in the spinal cord of Ews/Ewsr1KO mice, I analyzed miRNA and mRNA data of EWS using miRNA network and PPI network. Both miRNA and mRNA data was integrated with the miRNA network. I found regulated miRNAs by the EWS gene and identified Gnai1 by miRNA network. I considered neighbor protein of Gnai1 in PPI network, analyzed the correlation of gene expression values. Gnai1 was suppressed by mmu-miR-381 and mmu-miR-181a/b/c and inhibited by Rgs1 and Rgs19 in the spinal cord of EWS KO mice, also reduced the expression levels of the expression of Gnb1, Gnb2, and Gnb4, which are complex with Gnai1 gene. It shows that an integrated analysis of miRNA and mRNA omics data are well analyzed in miRNA and PPI networks.

In the second study, the impact of mutations in DNA methylation genes on genome-wide methylation landscapes and down-stream gene activations in pan-cancer, I examined the effect of 7 DNA methylation modifier genes using sub-network clustering method in pan-cancer scare. Pan-cancer data were collected from TCGA, and 3865 samples with both transcriptome and methylation data were analyzed. In each carcinoma, samples were divided and analyzed for the presence of a mutation in 7 DNA methylation modifier genes. Up-regulated genes with hypomethylated promoter regions in AML and down-regulated genes with hypermethylated promoter regions in COAD were selected by graph-based sub-network clustering methods. Through analysis, 42 hypomethylated promoter DMRs up-regulated DEGs in AML and 61 hypermethylated DMR down-regulated DEGs in COAD was identified by methylation regardless of the expression of TF and showed that some of the genes found were previously reported in other experimental papers. Research of methylation data and gene expression data analysis using network clustering showed significant results of a gene set associated with DNA methylation genes.

In the last study, *in silico* experiment system for testing hypothesis on gene functions using three condition-specific biological networks, I developed a computerized experimental system that can quickly test the relevance of a key gene to disease from biological data. MicroRNA, PPI and TF network information were deployed for the *in silico* testing. To transform a user given gene or hypothesis into a gene set, a literature-based search engine was used and the analysis results were evaluated by calculating the entropy of the network combining the condition-specific gene expression levels. The network results with high complexity showed a high score of hypothesis verification. The constructed system was validated using E2f1 knock-out data. Eleven out of 14 E2f1-related diseases showed a high association and a low association for low-relational diseases. My development tool demonstrated high-level of hypothesis verification through simulation using miRNA, PPI, and TF networks. In conclusion, my doctoral study challenged to solve the difficulties of the integrated omics data analysis and successfully analyzed using the prior knowledge network. I contributed to bioinformatics by providing successful analysis cases and analysis tools for multi-omics integrated analysis using network-based analysis techniques. An integrated analysis of network-based multi-omics is an attempt to gain an integrated understanding of living things. A better understanding of living things requires a higher level of analysis and challenges to more complex problems.

Bibliography

Omar Abdel-Wahab, Ann Mullally, Cyrus Hedvat, Guillermo Garcia-Manero, Jay Patel, Martha Wadleigh, Sebastien Malinge, JinJuan Yao, Outi Kilpivaara, Rukhmi Bhat, et al. Genetic characterization of tet1, tet2, and tet3 alterations in myeloid malignancies. *Blood*, 114(1):144–147, 2009.

Heba Ragaee Abdelhakam Aboelenein, Samia Salah, Yasmine Adel Lashine, and Ahmed Ihab Abdelaziz. Dual downregulation of microrna 17-5p and e2f1 transcriptional factor in pediatric systemic lupus erythematosus patients. *Rheumatology international*, 33(5):1333–1338, 2013.

Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

Vladimir N Babenko, Irina V Chadaeva, and Yuriy L Orlov. Genomic landscape of cpg rich elements in human. *BMC evolutionary biology*, 17(1):19, 2017.

Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385, 2018.

Lisa M Barker, Thomas W Pendergrass, Jean E Sanders, and Douglas S Hawkins. Survival after recurrence of ewing's sarcoma family of tumors. *Journal of clinical oncology*, 23(19):4354–4362, 2005.

Anne Bertolotti, Brendan Bell, and Làszlò Tora. The n-terminal domain of human taf ii 68 displays transactivation and oncogenic properties. *Oncogene*, 18(56):8000, 1999.

Gabriela Bindea, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf-Herman Fridman, Franck Pagès, Zlatko Trajanoski, and Jérôme Galon. Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091–1093, 2009.

Francisco J Blanco-Silva. *Learning SciPy for Numerical and Scientific Computing*. Packt Publishing Ltd, 2013.

Claudia Bruedigam, Frederik O Bagger, Florian H Heidel, Catherine Paine Kuhn, Solene Guignes, Axia Song, Rebecca Austin, Therese Vu, Erwin Lee, Sarbjit Riyat, et al. Telomerase inhibition effectively targets mouse and human aml stem cells and delays relapse following chemotherapy. *Cell Stem Cell*, 15(6):775–790, 2014.

Pavel Burda, Jarmila Vargova, Nikola Curik, Cyril Salek, Giorgio Lucio Papadopoulos, John Strouboulis, and Tomas Stopka. Gata-1 inhibits pu. 1 gene via dna and histone h3k9 methylation of its distal enhancer in erythroleukemia. *PloS one*, 11(3):e0152234, 2016.

Alessandra Cataldo, Douglas G Cheung, Andrea Balsari, Elda Tagliabue, Vincenzo Coppola, Marilena V Iorio, Dario Palmieri, and Carlo M Croce. mir-302b enhances breast cancer cell sensitivity to cisplatin by regulating e2f1 and the cellular dna damage response. *Oncotarget*, 7(1):786, 2016.

Howard A Chansky, Ming Hu, Dennis D Hickstein, and Liu Yang. Oncogenic tls/erg and ews/fli-1 fusion proteins inhibit rna splicing mediated by yb-1 protein. *Cancer research*, 61(9):3586–3590, 2001.

Shi Chen, Jia-Qiang Zhang, Jiang-Zhi Chen, Hui-Xing Chen, Fu-Nan Qiu, Mao-Lin Yan, Yan-Ling Chen, Cheng-Hong Peng, Yi-Feng Tian, and Yao-Dong Wang. The over expression of long non-coding rna anril promotes epithelial-mesenchymal transition by activating the atm-e2f1 signaling pathway in pancreatic cancer: An in vivo and in vitro study. *International Journal of Biological Macromolecules*, 102:718–728, 2017.

Joonseok Cho, Hongmei Shen, Hui Yu, Hongjie Li, Tao Cheng, Sean Bong Lee, and Byeong Chel Lee. Ewing sarcoma gene ews regulates hematopoietic stem cell senescence. *Blood*, 117(4):1156–1166, 2011.

Lucile Couronné, Christian Bastard, and Olivier A Bernard. Tet2 and dnmt3a mutations in human t-cell lymphoma. *New England Journal of Medicine*, 366 (1):95–96, 2012.

Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561, 1970.

Gabor Csardi, Tamas Nepusz, et al. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

Longzhen Cui, Zhiheng Cheng, Yan Liu, Yifeng Dai, Yifan Pang, Yang Jiao, Xiaoyan Ke, Wei Cui, Qingyi Zhang, Jinlong Shi, et al. Overexpression of pdk2 and pdk3 reflects poor prognosis in acute myeloid leukemia. *Cancer gene therapy*, page 1, 2018a.

Shanshan Cui, Xi Yang, Lihong Zhang, Yi Zhao, and Weiqun Yan. Lncrna mafg-as1 promotes the progression of colorectal cancer by sponging mir-147b and activation of ndufa4. *Biochemical and Biophysical Research Communications*, 506(1):251–258, 2018b.

François Delhommeau, Sabrina Dupont, Véronique Della Valle, Chloe James, Severine Trannoy, Aline Masse, Olivier Kosmider, Jean-Pierre Le Couedic, Fabienne Robert, Antonio Alberdi, et al. Mutation in tet2 in myeloid cancers. *New England Journal of Medicine*, 360(22):2289–2301, 2009.

Glynn Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. David: database for annotation, visualization, and integrated discovery. *Genome biology*, 4(9):R60, 2003.

Andrew P Feinberg, Michael A Koldobskiy, and Anita Göndör. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews Genetics*, 17(5):284, 2016.

Cyril Fisher. The diversity of soft tissue tumours with ewsr 1 gene rearrangements: a review. *Histopathology*, 64(1):134–150, 2014.

Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian Von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1):D808–D815, 2013.

Chundi Gao, Jing Zhuang, Chao Zhou, Lijuan Liu, Cun Liu, Huayao Li, Minzhang Zhao, Gongxi Liu, and Changgang Sun. Developing dna methylation-based prognostic biomarkers of acute myeloid leukemia. *Journal of cellular biochemistry*, 119(12):10041–10050, 2018.

Yu Gao, Hongzhao Li, Xin Ma, Yang Fan, Dong Ni, Yu Zhang, Qingbo Huang, Kan Liu, Xin-Tao Li, Lei Wang, et al. Klf6 suppresses metastasis of clear cell renal cell carcinoma via transcriptional repression of e2f1. *Cancer Research*, pages canres–0348, 2016.

Alip Ghosh, Suchandrima Ghosh, Debanjali Dasgupta, Amit Ghosh, Somenath Datta, Nilabja Sikdar, Simanti Datta, Abhijit Chowdhury, and Soma Banerjee. Hepatitis b virus x protein upregulates helg1/atad5 expression through e2f1 in hepatocellular carcinoma. *International journal of biological sciences*, 12(1):30, 2016.

Aleksandra Gołos, Dorota Jesionek-Kupnicka, Lidia Gil, Marcin Braun, Mieczyslaw Komarnicki, Tadeusz Robak, and Agnieszka Wierzbowska. The expression of the slit–robo family in adult patients with acute myeloid leukemia. *Archivum immunologiae et therapiae experimentalis*, pages 1–15, 2019.

Vera Grossmann, Claudia Haferlach, Sandra Weissmann, Andreas Roller, Sonja Schindela, Franziska Poetzinger, Kathrin Stadler, Frauke Bellos, Wolfgang Kern, Torsten Haferlach, et al. The molecular profile of adult t-cell acute lymphoblastic leukemia: mutations in runx1 and dnmt3a are associated with poor prognosis in t-all. *Genes, Chromosomes and Cancer*, 52(4):410–422, 2013.

Shih-Chiang Huang, Hsiao-Wei Chen, Lei Zhang, Yun-Shao Sung, Narasimhan P Agaram, Mary Davis, Morris Edelman, Christopher DM Fletcher, and Cristina R Antonescu. Novel fus-klf17 and ewsr1-klf17 fusions in myoepithelial tumors. *Genes, Chromosomes and Cancer*, 54(5):267–275, 2015.

Benjamin Hur, Sangsoo Lim, Heejoon Chae, Seokjun Seo, Sunwon Lee, Jaewoo Kang, and Sun Kim. Clip-gene: a web service of the condition specific context-laid integrative analysis for gene prioritization in mouse tf knockout experiments. *Biology direct*, 11(1):57, 2016.

Marcin Imielinski, Alice H Berger, Peter S Hammerman, Bryan Hernandez, Trevor J Pugh, Eran Hodis, Jeonghee Cho, James Suh, Marzia Capelletti, Andrey Sivachenko, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6):1107–1120, 2012.

Peter Jantscheff, Luigi Terracciano, Adam Lowy, Katharina Glatz-Krieger, Fritz Grunert, Burkhard Micheel, Jens Brummer, Urs Laffer, Urs Metzger, Richard Herrmann, et al. Expression of ceacam6 in resectable colorectal cancer: a factor of independent prognostic significance. *Journal of clinical oncology*, 21 (19):3638–3646, 2003.

Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4): R36, 2013.

Hyeongmin Kim and Yong-Min Kim. Pan-cancer analysis of somatic mutations and transcriptomes reveals common functional gene clusters shared by multiple cancer types. *Scientific reports*, 8(1):6041, 2018.

KY Kim, YJ Hwang, MK Jung, J Choe, Y Kim, S Kim, CJ Lee, H Ahn, J Lee, NW Kowall, et al. A multifunctional protein ews regulates the expression of drosha and micrornas. *Cell death and differentiation*, 21(1):136, 2014.

Yunha Kim, Young-Sook Kang, Na-Young Lee, Ki Yoon Kim, Yu Jin Hwang, Hyun-Wook Kim, Im Joo Rhyu, Song Her, Min-Kyung Jung, Sun Kim, et al. Uvrag targeting by mir125a and mir351 modulates autophagy associated with ewsr1 deficiency. *Autophagy*, 11(5):796–811, 2015.

F Koyama, H Sawada, H Fujii, H Hamada, T Hirao, M Ueno, and H Nakano. Adenoviral-mediated transfer of escherichia coli uracil phosphoribosyltransferase (uprt) gene to modulate the sensitivity of the human colon cancer cells to 5-fluorouracil. *European journal of cancer*, 36(18):2403–2410, 2000.

Michael Krauthammer, Yong Kong, Byung Hak Ha, Perry Evans, Antonella Bacchiocchi, James P McCusker, Elaine Cheng, Matthew J Davis, Gerald Goh, Murim Choi, et al. Exome sequencing identifies recurrent somatic rac1 mutations in melanoma. *Nature genetics*, 44(9):1006, 2012.

Jon J Ladd, Tina Busald, Melissa M Johnson, Qing Zhang, Sharon J Pitteri, Hong Wang, Dean E Brenner, Paul D Lampe, Raju Kucherlapati, Ziding Feng, et al. Increased plasma levels of the apc-interacting protein mapre1, lrg1, and igfbp2 preceding a diagnosis of colorectal cancer in women. *Cancer prevention research*, 5(4):655–664, 2012.

Saskia MC Langemeijer, Roland P Kuiper, Marieke Berends, Ruth Knops, Mariam G Aslanyan, Marion Massop, Ellen Stevens-Linders, Patricia van

Hoogen, Ad Geurts van Kessel, Reinier AP Raymakers, et al. Acquired mutations in tet2 are common in myelodysplastic syndromes. *Nature genetics*, 41(7):838, 2009.

Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

David S Latchman. Transcription factors: an overview. *The international journal of biochemistry & cell biology*, 29(12):1305–1312, 1997.

Patrick PL Lau, Philip CW Lui, Gene TC Lau, Derek TW Yau, Elaine TY Cheung, and John KC Chan. Ewsr1-creb3l1 gene fusion: a novel alternative molecular aberration of low-grade fibromyxoid sarcoma. *The American journal of surgical pathology*, 37(5):734–738, 2013.

Jusang Lee, Kyuri Jo, Sunwon Lee, Jaewoo Kang, and Sun Kim. Prioritizing biological pathways by recognizing context in time-series gene expression data. *BMC bioinformatics*, 17(17):477, 2016a.

Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi, Sunkyu Kim, Aik-Choon Tan, et al. Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS one*, 11(10):e0164680, 2016b.

Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart MG Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–1043, 2013.

Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *cell*, 120(1):15–20, 2005.

Timothy J Ley, Li Ding, Matthew J Walter, Michael D McLellan, Tamara Lamprecht, David E Larson, Cyriac Kandoth, Jacqueline E Payton, Jack Baty, John Welch, et al. Dnmt3a mutations in acute myeloid leukemia. *New England Journal of Medicine*, 363(25):2424–2433, 2010.

Bin Li, Wen Wen Xu, Xin Yuan Guan, Yan Ru Qin, Simon Law, Nikki Pui Yue Lee, Kin Tak Chan, Pui Ying Tam, Yuk Yin Li, Kwok Wah Chan, et al. Competitive binding between id1 and e2f1 to cdc20 regulates e2f1 degradation and thymidylate synthase expression to promote esophageal cancer chemoresistance. *Clinical Cancer Research*, 2015.

Hongjie Li, Wendy Watford, Cuiling Li, Alissa Parmelee, Mark A Bryant, Chuxia Deng, John O'Shea, and Sean Bong Lee. Ewing sarcoma gene ews is essential for meiosis and b lymphocyte development. *The Journal of clinical investigation*, 117(5):1314–1323, 2007.

ZH Li, Y Liu, and SY Gao. Correlation between il-7 genomic protein methylation level and acute myeloid leukemia. *European review for medical and pharmacological sciences*, 23(3):1196–1202, 2019.

ZL Li, F Jiao, Y Ma, Z Yue, and LJ Kong. Target genes regulated by transcription factor e2f1 in small cell lung cancer. *Sheng li xue bao:[Acta physiologica Sinica]*, 68(3):276, 2016.

Yu-Xiang Liang, Jian-Ming Lu, Ru-Jun Mo, Hui-Chan He, Jian Xie, Fu-Neng Jiang, Zhuo-Yuan Lin, Yan-Ru Chen, Yong-Ding Wu, Hong-Wei Luo, et al. E2f1 promotes tumor cell invasion and migration through regulating cd147 in prostate cancer. *International journal of oncology*, 48(4):1650–1658, 2016.

Jing-Lei Liu, Guang-Zhi Zeng, Xiao-Li Liu, Yong-Qiang Liu, Zhong-Guo Hu, Ying Liu, Ning-Hua Tan, and Guang-Biao Zhou. Small compound bigelovin exerts inhibitory effects and triggers proteolysis of e2f1 in multiple myeloma cells. *Cancer science*, 104(12):1697–1704, 2013.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15 (12):1, 2014.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Andrej Aderhold, Richard Bonneau, Yukun Chen, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796, 2012.

Vea Matys, Ellen Fricke, R Geffers, Ellen Gößling, Martin Haubrock, Reinhard Hehl, Klaus Hornischer, Dagmar Karas, Alexander E Kel, Olga V Kel-Margoulis, et al. Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, 31(1):374–378, 2003.

William A May, Stephen L Lessnick, Benjamin S Braun, Michael Klemsz, Brian C Lewis, Lynn B Lunsford, Robert Hromas, and Christopher T Denny. The ewing's sarcoma ews/fli-1 fusion gene encodes a more potent transcriptional activator and is a more powerful transforming gene than fli-1. *Molecular and cellular biology*, 13(12):7393–7398, 1993.

Paul S Meltzer. Is ewing's sarcoma a stem cell tumor? *Cell Stem Cell*, 1(1): 13–15, 2007.

James S Miser, Mark D Krailo, Nancy J Tarbell, Michael P Link, Christopher JH Fryer, Douglas J Pritchard, Mark C Gebhardt, Paul S Dickman, Elizabeth J Perlman, Paul A Meyers, et al. Treatment of metastatic ewing's sarcoma or primitive neuroectodermal tumor of bone: evaluation of combination ifosfamide and etoposide—a children's cancer group and pediatric oncology group study. *Journal of Clinical Oncology*, 22(14):2873–2876, 2004.

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621, 2008.

Seungyoon Nam, Meng Li, Kwangmin Choi, Curtis Balch, Sun Kim, and Kenneth P Nephew. Microrna and mrna integrated analysis (mmia): a web tool for examining biological functions of microrna expression. *Nucleic acids research*, 37(suppl_2):W356–W362, 2009.

Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330, 2012.

Martin Neumann, Sandra Heesch, Cornelia Schlee, Stefan Schwartz, Nicola Gökbuget, Dieter Hoelzer, Nikola P Konstandin, Bianka Ksienzyk, Sebastian Vosberg, Alexander Graf, et al. Whole-exome sequencing in adult etp-all reveals a high rate of dnmt3a mutations. *Blood*, 121(23):4749–4752, 2013.

Katsuhiko Nosho, Kaori Shima, Natsumi Irahara, Shoko Kure, Yoshifumi Baba, Gregory J Kirkner, Li Chen, Sumita Gokhale, Aditi Hazra, Donna Spiegelman, et al. Dnmt3b expression might contribute to cpg island methylator phenotype in colorectal cancer. *Clinical cancer research*, 15(11):3663–3671, 2009.

Shuji Ogino, Takako Kawasaki, Gregory J Kirkner, Taiki Yamaji, Massimo Loda, and Charles S Fuchs. Loss of nuclear p27 (cdkn1b/kip1) in colorectal cancer is correlated with microsatellite instability and cimp. *Modern pathology*, 20(1):15, 2007.

Minsik Oh, Sungmin Rhee, Ji Hwan Moon, Heejoon Chae, Sunwon Lee, Jaewoo Kang, and Sun Kim. Literature-based condition-specific mirna-mrna target prediction. *PloS one*, 12(3):e0174999, 2017.

Coro Paisán-Ruız, Shushant Jain, E Whitney Evans, William P Gilks, Javier Simón, Marcel van der Brug, Adolfo López de Munain, Silvia Aparicio, Angel Martınez Gil, Naheed Khan, et al. Cloning of the gene containing mutations that cause park8-linked parkinson's disease. *Neuron*, 44(4):595–600, 2004.

Lara Pappas, Xiaoliang Leon Xu, David H Abramson, and Suresh C Jhanwar. Genomic instability and proliferation/survival pathways in rb1-deficient malignancies. *Advances in Biological Regulation*, 64:20–32, 2017.

Noa Rappaport, Michal Twik, Inbar Plaschkes, Ron Nudel, Tsippi Iny Stein, Jacob Levitt, Moran Gershoni, C Paul Morrey, Marilyn Safran, and Doron Lancet. Malacards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Research*, page gkw1012, 2016.

Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, page gkv007, 2015.

Sabrina Rossi, Kàroly Szuhai, Marije Ijszenga, Hans J Tanke, Lucia Zanatta, Raf Sciot, Christopher DM Fletcher, Angelo P Dei Tos, and Pancras CW

Hogendoorn. Ewsr1-creb1 and ewsr1-atf1 fusion genes in angiomatoid fibrous histiocytoma. *Clinical Cancer Research*, 13(24):7322–7328, 2007.

Laura J Rush, Aparna Raval, Pauline Funchain, Amy J Johnson, Lisa Smith, David M Lucas, Melania Bembea, Te-Hui Liu, Nyla A Heerema, Laura Rassenti, et al. Epigenetic profiling in chronic lymphocytic leukemia reveals novel methylation targets. *Cancer research*, 64(7):2424–2433, 2004.

Gurveen Saberwal, Eileen Broderick, Imke Janssen, Vilasini Shetty, Sairah Alvi, Laurie Lisak, Parameswaran Venugopal, Azra Raza, and Suneel D Mundle. Involvement of cyclin d1 and e2f1 in intramedullary apoptosis in myelodysplastic syndromes. *Journal of hematotherapy & stem cell research*, 12(4): 443–450, 2003.

Beatriz Sanchez-Correa, Inmaculada Gayoso, Juan M Bergua, Javier G Casado, Sara Morgado, Rafael Solana, and Raquel Tarazona. Decreased expression of dnam-1 on nk cells from acute myeloid leukemia patients. *Immunology and cell biology*, 90(1):109–115, 2012.

Hagit Schayek, Itay Bentov, Itay Rotem, Metsada Pasmanik-Chor, Doron Ginsberg, Stephen R Plymate, and Haim Werner. Transcription factor e2f1 is a potent transactivator of the insulin-like growth factor-i receptor (igf-ir) gene. *Growth Hormone & IGF Research*, 20(1):68–72, 2010.

Laurianne Scourzic, Enguerran Mouly, and Olivier A Bernard. Tet proteins and the control of cytosine demethylation in cancer. *Genome medicine*, 7(1):9, 2015.

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

Berend Snel, Gerrit Lehmann, Peer Bork, and Martijn A Huynen. String: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research*, 28(18):3442–3444, 2000.

Philip J Stephens, Patrick S Tarpey, Helen Davies, Peter Van Loo, Chris Greenman, David C Wedge, Serena Nik-Zainal, Sancha Martin, Ignacio Varela, Graham R Bignell, et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400, 2012.

Violetta Sulzyc-Bielicka, Pawel Domagala, Dariusz Bielicki, Krzysztof Safranow, Wojciech Rogowski, and Wenancjusz Domagala. E2f1/ts immunophenotype and survival of patients with colorectal cancer treated with 5fu-based adjuvant therapy. *Pathology & Oncology Research*, 22(3):601–608, 2016.

Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937, 2016.

Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.

Elie Traer, Jacqueline Martinez, Nathalie Javidi-Sharifi, Anupriya Agarwal, Jennifer Dunlap, Isabel English, Tibor Kovacsovics, Jeffrey W Tyner, Melissa Wong, and Brian J Druker. Fgf2 from marrow microenvironment promotes resistance to flt3 inhibitors in acute myeloid leukemia. *Cancer research*, 76 (22):6471–6482, 2016.

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511, 2010.

Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

Lothar Vassen, Cyrus Khandanpour, Peter Ebeling, Bert A van der Reijden, Joop H Jansen, Stefan Mahlmann, Ulrich Dührsen, and Tarik Möröy. Growth factor independent 1b (gfi1b) and a new splice variant of gfi1b are highly expressed in patients with acute and chronic leukemia. *International journal of hematology*, 89(4):422–430, 2009.

Günter P Wagner, Koryu Kin, and Vincent J Lynch. Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. *Theory in biosciences*, 131(4):281–285, 2012.

Xiaowei Wang. mirdb: a microrna target prediction and functional annotation database with a wiki interface. *Rna*, 14(6):1012–1017, 2008.

Xiaowei Wang and Issam M El Naqa. Prediction of both conserved and non-conserved microrna targets in animals. *Bioinformatics*, 24(3):325–332, 2007.

Laura A Warg, Judy L Oakes, Rachel Burton, Amanda J Neidermyer, Holly R Rutledge, Steve Groshong, David A Schwartz, and Ivana V Yang. The role of the e2f1 transcription factor in the innate immune response to systemic lps. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 303 (5):L391–L400, 2012.

YongKiat Wee, Yining Liu, Salma Begum Bhyan, Jiachun Lu, and Min Zhao. The pan-cancer analysis of gain-of-functional mutations to identify the common oncogenic signatures in multiple cancers. *Gene*, 697:57–66, 2019.

Fuxiao Xin, Meng Li, Curt Balch, Michael Thomson, Meiyun Fan, Yunlong Liu, Scott M Hammond, Sun Kim, and Kenneth P Nephew. Computational analysis of microrna profiles and their target genes suggests significant involvement in breast cancer antiestrogen resistance. *Bioinformatics*, 25(4): 430–434, 2008.

Xiao-Jing Yan, Jie Xu, Zhao-Hui Gu, Chun-Ming Pan, Gang Lu, Yang Shen, Jing-Yi Shi, Yong-Mei Zhu, Lin Tang, Xiao-Wei Zhang, et al. Exome sequencing identifies somatic mutations of dna methyltransferase gene dnmt3a in acute monocytic leukemia. *Nature genetics*, 43(4):309, 2011.

Wen Zhang, Duo Tong, Fei Liu, Dawei Li, Jiajia Li, Xi Cheng, and Ziliang Wang. Rps7 inhibits colorectal cancer growth via decreasing hif-1α-mediated glycolysis. *Oncotarget*, 7(5):5800, 2016.

Xiujun Zhang, Keqin Liu, Zhi-Ping Liu, Béatrice Duval, Jean-Michel Richer, Xing-Ming Zhao, Jin-Kao Hao, and Luonan Chen. Narromi: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics*, 29(1):106–113, 2013.

Bo'an Zheng, Rui Chai, and Xiaojun Yu. Downregulation of nit2 inhibits colon cancer cell proliferation and induces cell cycle arrest through the caspase-

3 and parp pathways. *International journal of molecular medicine*, 35(5): 1317–1322, 2015.

초록

특정 유전자의 특징을 알기 위해서 대상 유전자의 증폭 및 녹아웃 실험을 통해 표현형을 살펴보는 연구가 많이 수행되고 있다. 이러한 생물학 실험과 더불어 만들어지는 다양한 오믹스 데이터를 통합 분석하여 표현형을 나타내는 것은 여러 가지 어려운 문제가 있다. 오믹스 데이터들은 다른 형식을 사용하기 때문에 통합하는 것이 어려운 문제이며 오믹스 간의 복잡 관계를 고려해야 한다. 또한, 통합한 오믹스 데이터는 고차원의 자료이기 때문에 분석과 해석 또한 어렵다. 성격이 다른 데이터를 연계하고 분석에 쉬운 특징 공간으로 줄이는 문제를 해결하기 위해서 네트워크 정보를 사용하는 것은 매우 유용하다. 본 박사 연구에서는 네트워크 정보를 사용한 다중 오믹스 데이터 통합 분석의 세 가지 연구를 수행하였고 다중 오믹스 통합 분석의 어려운 문제의 해결은 네트워크를 이용한 분석이 매우 유용함을 보였다.

첫 번째 연구에서는, 녹아웃 유전자 EWS에 의해 발현되는 유전적 현상을 설명하기 위하여 야생형 쥐와 EWS 녹아웃 쥐의 척수로부터 얻어진 miRNA 데이터와 mRNA 데이터를 통합하여 분석하였다. miRNA에 의해 조절 받은 DEG의 기능적 변화를 조사하기 위해 miRNA와 표적 유전자 간의 음의 상관 네트워크와 단백질-단백질 (PPI) 네트워크를 사용했다. miRNA와 PPI 네트워크를 이용한 분석을 통해, 콜린성 시냅스 경로에서 유의미하게 하향 조절된 Gnai1을 찾아냈다. Gnai1의 발현량은 억제된 EWS에 의해 발현량이 증가한 mmu-miR-381 및 mmu-miR-181a/b/c에 의해 발현이 억제되는 것을 miRNA 네트워크를 이용하여 찾아 확인하였다. 또한, 단백질 네트워크를 이용하여 발현량이 증가한 Rgs1 및 Rgs19에 의해서 Gnai1이 억제되는 것을 보았으며, G 단백질 복합체를 이루는 이웃 단백질 Gnb1, Gnb2 및 Gnb4의 발현 수준도 감소한 것을 함께 확인하여 유의미한 발견임을 보였다.

두 번째 연구에서는, 전체 암 규모에서 DNA 메틸화에 필요한 7개의 유전자의 효과를 분석하고자 했다. TCGA의 12종의 암 데이터에서 유전자 발현 데이터와 메틸화 데이터를 갖는 3865개의 표본을 수집하여 분석하였다. 각 암종에서 해당 유전자의 돌연변이 유무로 표본을 나누고 서브 네트워크 클러스터링 방법을 이용하여 생물학적 의미가 있는 유전자 그룹으로 나누어 메틸화 영향을 분석하여 의미 있는 유전자를 찾고자 했다. 분석으로 찾아낸 클러스터 중에서 급성골수성백혈병 환자에서 하이포 메틸화된 프로모터를 가지는 유전자군과 대장암 환자에서 과메틸화된 프로모터를 가지는 유전자군을 선택하여 심화 분석하였다. TF에 의한 영향과는 무관하고 메틸화에 의한 영향으로 발현량이 변화한 유전자를 선별하였고, 급성골수성백혈병 환자의 42개의 유전자와 대장암 환자의 61개의 유전자를 유의미한 것으로 찾아내었다. 선별한 유전자 일부는 이전의 다른 실험 논문에서 보고된 것을 확인하여 유의미한 것을 검증하였다.

세 번째 연구에서는 생물학적 데이터를 통합 분석하여 질병에 대한 특정 유전자의 관련성을 신속하게 확인할 수 있는 컴퓨터 실험 시스템을 개발했다. 이 분석 실험 도구는 miRNA, PPI 및 TF 3가지 네트워크 정보를 데이터베이스로 구축하여 네트워크상에서 시뮬레이션 분석이 가능하도록 하였고, 주어진 유전자 또는 가설을 유전자 세트로 변환하기 위해 문헌 기반 검색 엔진을 이용하여 만들었다. 확인 분석된 네트워크 결과는 유전자 발현 수준을 고려하였고 네트워크의 정보 엔트로피값을 계산하여 분석 결과를 평가하였다. 많은 네트워크 정보를 가진 결과는 가설 검증에서 높은 점수를 가지도록 하였다. 구축한 시스템은 E2f1 유전자의 데이터와 Lrrk2, Dicer1 각각의 유전자 데이터를 사용하여 검증하였다. MalaCards의 인간 질병 데이터베이스를 이용하여 E2f1 관련된 14개의 질병과 유전자의 연관성을 검증하였고, 11개의 질병은 높은 연관성을 보였고, 그 외의 무관한 48개의 질병에 대해서는 낮은 연관성을 가지는 것을 보임으로써 검증하였다.

요약하자면, 필자의 박사 연구는 유전자와 표현형에 대한 연관성을 분석하기 위해 다중 오믹스 데이터의 통합하여 분석하였고, 통합 분석의 어려운 문제를 네트워크 정보를 사용하여 유의미한 결과를 보였다. 다중 오믹스 데이터의 성격에 따라

PPI, miRNA, TF 네트워크 및 DNA 메틸화 정보 네트워크를 결합하는 방법을 사용하였고 생물학적으로 유의미한 분석 결과를 보여 네트워크를 이용한 분석이 유용함을 보였다. 또한, 네트워크를 이용한 다중 오믹스 데이터 분석 실험 도구를 개발하여 생물정보학 연구에 기여 하고자 하였다.

주요어: 멀티 오믹스, 네트워크 분석, 네트워크 통합 분석, 단백질 네트워크, miRNA 네트워크, TF 네트워크, 메틸화 네트워크

학번: 2011-30926

Acknowledgements

저의 박사 학위 취득은 지금까지 살아오면서 느낄 수 있는 최고의 결실이었고, 가장 기쁜 경험이었습니다. 혼자서는 절대로 불가능한 일이었고, 많은 사람의 도움과 지원이 있어서 가능했던 일이기 때문에 모두에게 감사의 마음을 전하고 싶습니다.

무엇보다 먼저 긴 시간 동안 인내를 가지고 아낌없는 조언과 지도를 해주신 김선 교수님께 깊은 감사 인사를 전하고 싶습니다. 제가 가장 존경하는 생물정보학자이자 닮고 싶은 교수님으로서 학문에 대한 지식과 열정, 그룹을 이끌어가는 통솔력, 인생에 대해서 많은 것을 배울 수 있었습니다.

또한, 교수님 덕분에 만나게 된 Bio & Health Informatics 연구실의 동료들은 저의 연구에 큰 힘이 되어주었고 정말 감사를 드리고 싶습니다. 연구의 어려움이 있으며 함께 어울려서 생각을 논의하고 의견을 모아서 하나의 연구 목적으로 향해 나아갈 힘이 되어주었습니다. 특별히 항상 따뜻한 마음으로 대해 준 안홍렬 박사에게 깊은 감사를 표시하고 싶습니다. 세심하고 적극적인 태도의 조언은 연구에 많은 도움이 되었고 지치고 힘들 때 큰 위로와 힘이 되었습니다.

더불어 논문 심사 과정에서 아낌없는 조언을 해주신 이병재 교수님, 손현석 교수님, 황대희 교수님, 김광수 교수님께 감사를 표합니다.

그리고, 저를 낳아주시고 무한한 사랑과 신뢰를 주시고 모든 지원과 적극적인 지지를 보내 주신 부모님께 무한한 감사를 드리고 싶습니다. 부모님을 비롯한 제 동생과 친척 가족들 모두의 깊은 관심과 지지는 제가 학문에 몰두할 수 있는 큰

힘이 되었습니다.

마지막으로, 항상 저의 곁에서 진심으로 지지하고 격려해준 사랑하는 나의 아내에게 고마움을 전합니다.

저의 박사학위 논문의 성취는 모든 이들의 도움 없이는 불가능한 일이었습니다. 정말 감사드립니다.