Ph.D. DISSERTATION

Informatics systems using network approaches to prioritize genes in RNA-Seq data

RNA-Seq 데이터에서 유전자의 랭킹을 책정하기 위한 네트워크 접근법을 사용한 정보 과학 시스템

BY

Benjamin Hur

AUGUST 2019

INTERDISCIPLINARY PROGRAM IN BIOINFORMATICS
COLLEGE OF NATURAL SCIENCES
SEOUL NATIONAL UNIVERSITY

Ph.D. DISSERTATION

Informatics systems using network approaches to prioritize genes in RNA-Seq data

RNA-Seq 데이터에서 유전자의 랭킹을 책정하기 위한 네트워크 접근법을 사용한 정보 과학 시스템

BY

Benjamin Hur

AUGUST 2019

INTERDISCIPLINARY PROGRAM IN BIOINFORMATICS
COLLEGE OF NATURAL SCIENCES
SEOUL NATIONAL UNIVERSITY

Informatics systems using network approaches to prioritize genes in RNA-Seq data

RNA-Seq 데이터에서 유전자의 랭킹을 책정하기 위한 네트워크 접근법을 사용한 정보 과학 시스템

지도교수 김 선

이 논문을 이학박사 학위논문으로 제출함

2019 년 8 월

서울대학교 대학원

협동과정 생물정보학

허 영 회

허영회의 이학박사 학위논문을 인준함

2019 년 8 월

위 원 장	이병재
부위원장	김선
위 원	손현석
위 원	황대희
위 원	김광수

Abstract

Transcriptomic analysis, the measurement of transcripts on the genome scale, is now routinely performed in high resolution. Since the number of genes obtained in the transcriptome data is usually large, it is difficult for researchers to identify genes that are relevant to their research goals, without additional analysis. Analysis of transcriptome data is often performed utilizing heterogeneous resources such as biological networks, annotated gene information, and published literature. However, the relationship among heterogeneous resources is often too complicated to decipher which genes are relevant to the experimental design. Therefore, powerful computational methods should be coupled with these heterogeneous resources in order to effectively determine and illustrate key genes that are relevant to specific research goals. In my doctoral study, I have developed three bioinformatics systems that use network approaches to analyze transcriptome data and rank genes that are relevant to the experimental design.

The first study was conducted to develop a bioinformatics system that could be used to analyze RNA-Seq data of gene knockout (KO) mice, where the sample number is small. In this case, the main objectives were to investigate how the KO gene affects the expression of other genes and identify the key genes that contribute significantly to the phenotypic difference. To address these questions, I developed a gene prioritization system that utilizes the characteristics of RNA-Seq data. The system prioritizes genes by removing the less informative differentially expressed genes (DEGs) using gene regulatory network (GRN) and biological pathways. Next, it filters out genes that might be differ-

ent due to genetic differences between samples using single nucleotide variant (SNV) information. Consequently, this study demonstrated that the integration of networks and SNV information was able to increase the performance of gene prioritization.

The second study was conducted to develop a gene prioritization system that allows the user to specify the context of the experiment. This study was inspired by the fact that the currently available analysis methods for transcriptome data do not fully consider the experimental design of gene KO studies. Therefore, I envisaged that users would prefer an analysis method that took into consideration the characteristics of the KO experiments and could be guided by the context of the researcher who has designed and performed the biological experiment. Therefore, I developed CLIP-GENE, a web service of the condition-specific context-laid integrative analysis for prioritizing genes in mouse TF KO experiments. CLIP-GENE prioritizes genes of KO experiments by removing the less informative DEGs using GRN, discards genes that might have sample variance, using SNV information, and ranks genes that are related to the user's context using the text-mining technique, as well as considering the shortest path of protein-protein interaction (PPI) from the KO gene to the target genes.

The last study was conducted to develop an informative system that could be used to compare multiple RNA-Seq experiments using Venn diagrams. In general, RNA-Seq experiments are performed to compare samples between control and treated groups, producing a set of DEGs. Each region in a Venn diagram (a subset of DEGs) generally contains a large number of genes that could complicate the determination of the important and relevant genes. Moreover, simply comparing the list of DEGs from different experiments could be misleading because some of the DEG lists may have been measured using different controls. To address these issues, Venn-diaNet was developed, an analysis frame-

work that integrates Venn diagram and network propagation to prioritize genes for experiments that have multiple DEG lists. We demonstrated that Venn-diaNet was able to reproduce research findings reported in the original papers by comparing two, three, and eight biological experiments measured in different conditions. I believe that Venn-diaNet can be very useful for researchers to determine genes for their follow-up studies.

In summary, my doctoral study aimed to develop computational tools that can prioritize genes from transcriptome data. To achieve this goal, I combined network approaches with multiple heterogeneous resources in a single computational environment. All three informatics systems are deployed as software packages or web tools to support convenient access to researchers, eliminating the need for installation or learning any additional software packages.

Keywords: RNA-seq, Gene prioritization, Informatics system, Network-based approach

Student Number: 2012-23113

Contents

List of Figures

List of Tables

Chapter 1

Introduction

RNA-Seq is one of the popular technologies that estimates the abundance of global RNAs to identify genes that are relevant to the experimental design (Ozsolak and Milos, 2011; Li and Li, 2018). The data provides an unprecedented amount of information and details that cannot be handled in a single process. Therefore, the expression profile from RNA-Seq data is generally analyzed by using multiple databases and methods in order to obtain useful insight (Li and Li, 2018).

For more than a decade, studies have introduced a number of approaches and applications to analyze RNA-Seq data. However, it is largely difficult to pinpoint a gene with enough evidence to infer the relationship between the gene and the phenotype, in a single-step analysis. For example, differentially expressed gene (DEG) analysis is an analysis that finds genes that have statistically altered expressions, however, it does not provide enough information to explain why and how there is a phenotypic difference between samples. To overcome the limitations of single-step analysis, a number of studies have intro-

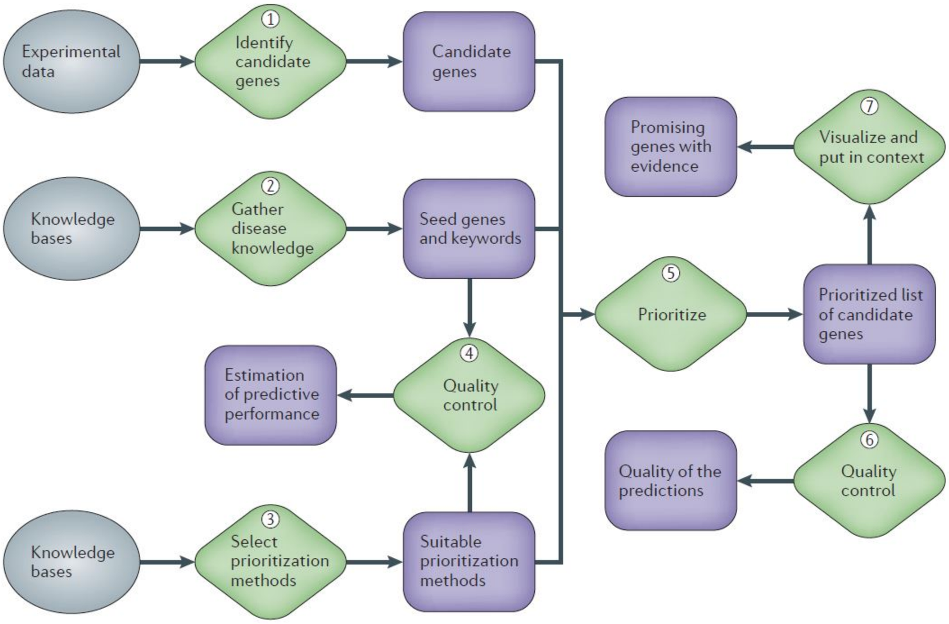

Experimental data
1
Identify candidate genes
Candidate genes
Knowledge bases
2
Gather disease knowledge
Seed genes and keywords
Promising genes with evidence
7
Visualize and put in context
5
Prioritize
Prioritized list of candidate genes
4
Quality control
Estimation of predictive performance
6
Quality control
Quality of the predictions
Knowledge bases
3
Select prioritization methods
Suitable prioritization methods

Figure 1.1: Work flow for prioritizing genes (Moreau and Tranchevent, 2012)

duced strategies that combine multiple data sources and methods to compensate for the insufficient information obtained from a single-step analysis (Figure 1.1)(Moreau and Tranchevent, 2012; Tranchevent *et al.*, 2010).

However, a number of databases and analysis methods, as well as strategies to combine these elements exist (Figure 1.2). Therefore, it is now a challenging task to select a strategy that is tailored for the goal of the experiment, with the appropriate combination of tools to analyze RNA-Seq data.

My doctoral study addresses the challenges in analyzing RNA-Seq data with three informatics systems that prioritize genes based on networks. The first study was conducted to analyze RNA-Seq data that have a small number of samples. The second study was conducted to overcome the challenges of gene ranking that does not reflect user interest. The third study was conducted

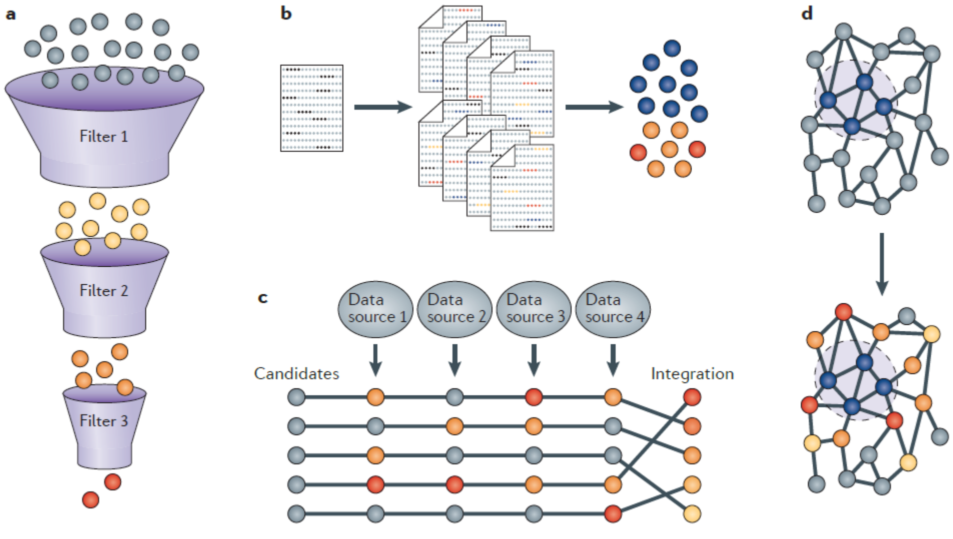


Figure 1.2: Various strategies to prioritize genes (Moreau and Tranchevent, 2012)

to analyze complicated RNA-Seq data that have multiple conditions, with an intuitive interpretation on.

1.1 Challenges of analyzing RNA-Seq data

The challenges to find phenotype-related genes with RNA-Seq data can be summarized into three reasons. *(i)* Large number of databases and methods make selection of the correct combination difficult and confusing *(ii)* Knowledge-bias that prioritize less relevant genes. *(iii)* Complicated experiment designs (i.e: multiple control/treatments, a small number of samples) that is difficult to analyze.

1.1.1 Excessive amount of databases and analysis methods

Intensive research has led to the creation of a large number and variety of databases and methods for the analysis of transcriptomic data. However, the vast array of options makes selection of the correct combination that would be ideal for one's research, difficult. To overcome this issue, studies have suggested various data combination strategies to identify promising genes (Moreau and Tranchevent, 2012)(Figure 1.2), however, these strategies have their limitations, which need to be addressed.

For example, filtering strategy, a strategy that uses multiple databases (or methods) as filters and removes the less significant candidate genes step wise (Figure 1.2a), is a straight forward strategy that strictly reduces the number of candidates that do not satisfy each criterion. However, if the filters do not have enough discrimination power, they will fail to screen out the less promising candidates. On the contrary, if the filters are too stringent, this strategy will give rise to a number of false negatives. Therefore, it is very challenging to adjust the level of discrimination power according to the combination of data sources.

Unlike filtering strategy, data fusion strategy has the advantage of eliminating false negatives caused by the stringent thresholds among filters by scoring the candidates at each data source and summarizes the overall ranks (Figure 1.2b). However, because the strategy combines heterogeneous data sources, the relationship between input and output data sources becomes complicated, and the complexities increase according to the number of data sources. Therefore, it is difficult to make an intuitive interpretation for the final results and this makes the analysis tools to have 'black-box' like characteristics (Moreau and Tranchevent, 2012).

Therefore, despite the availability of an abundance and variety of databases and methodologies, it is difficult to effectively combine these sources.

1.1.2 Knowledge bias that prioritizes less relevant genes

The knowledge bias between genes in databases can cause difficulties in identifying promising genes with data fusion strategy (Moreau and Tranchevent, 2012). Because the strategy evaluates candidate genes by utilizing multiple data sources, a well-studied gene is likely to be over-represented in multiple databases and cause biased rank even if it is not relevant to the experiment design. Therefore, it is necessary to address the biased-rank genes as well as prioritizing genes that are focused on the context of experimental design. Moreover, knowledge bias affects the network-based strategies that use network propagation, which is now one of the most powerful and common techniques to investigate the relationship between candidate genes and known genes (usually disease-related genes). However, this strategy largely relies on seeds that require prior knowledge to select proper seeds. If the prior information is insufficient for the selection of appropriate seed genes, the results of the network propagation will be less likely to reveal similarities between two different genes. Nevertheless, a few studies have used sequence features or topology features instead of prior knowledge to overcome these difficulties (López-Bigas and Ouzounis, 2004; Adie *et al.*, 2005; Chen *et al.*, 2009b). However, seed selection is difficult for transcriptome-based experiments that have poor prior knowledge.

1.1.3 Complicated experiment designs

DEGs are common elements used as initial candidates that are combined with other data sources for the identification of promising genes. However, discovering phenotype-dependent genes from complicated experimental designs (such as mouse gene KO experiments that have a small number of samples, or experiments that have multiple control/treated groups) is difficult with the currently available strategies. If the number of samples is small, the statistical evidence of DEGs can be weak, suggesting that differential gene expression (giving rise to DEGs) may have been caused by effects other than the conditional differences (i.e. genetic difference between biological replicates). If the false positives are considered during gene prioritization, identification of the true phenotype-related genes with the strategies that combine DEGs, is more challenging.

Studies that compare multiple DEG lists usually have experimental designs that include multiple control and treatment groups, increasing the complexity of sample comparison. Since each DEG list with different controls (or treatment groups) indicates different biological differences, simply adding or subtracting the entries between these lists might not be intuitive enough to decipher the biological meaning of the subset of genes. The complexity of the problem further increases with increasing number of DEG lists.

1.2 My approach to address the challenges for the analysis of RNA-Seq data

This thesis introduces three studies, each of the studies introduces an informatics system that uses a unique combination of network and data sources to solve the challenges prioritizing genes that are related to the phenotypic difference.

1. Combined analysis of gene regulatory network and SNV information enhances identification of potential gene markers in mouse gene KO studies with a small number of samples : a filtering strategy that addresses the challenge of complicated experimental design having a small number of samples by (i) removing less informative DEGs using gene regulatory network (GRN), biological pathways, and (ii) filtering out genes that differ between samples based on the presence of SNVs. As a result, this study was able to show that the integration of network and SNV information increases the performance of gene prioritization. The key idea of the method is to reconfirm that the DEGs have resulted as an effect of gene KO, rather than due to genetic differences between different samples.

2. CLIP-GENE: a web service of the condition-specific context-laid integrative analysis for gene prioritization in mouse TF KO experiments : a data fusion strategy that focuses on rank genes that are related to the experimental design. CLIP-GENE (i) removes less informative DEGs using GRN, (ii) discards genes that have sample variance with SNV, and (iii) ranks genes by using protein-protein interaction (PPI) network information and text-mining technique.

3. Venn-diaNet : Venn diagram based network propagation analysis framework for comparing multiple biological experiments a Venn diagram based network propagation analysis framework to prioritize genes that address the challenge of complicated experimental design, having multiple controls and treatment groups as well as seed selection for network propagation. Venn-diaNet was able to reproduce the original findings of experiments which comprised analysis and comparison of multiple biological transcriptomic data, measured in multiple conditions.

1.3 Background

1.3.1 Differentially expressed gene

Once the abundance of global mRNAs is measured, estimating the DEGs between samples is one of the great starting points to understand the characteristics of the phenotype differences (Marioni *et al.*, 2008). Measuring the expression differences that is statistically significant have proved to be a successful approach to find genes that is responible for the phenotypic differences (Hardcastle and Kelly, 2010; Robinson *et al.*, 2010; Anders and Huber, 2012; Trapnell *et al.*, 2013; Leng *et al.*, 2013; Li and Tibshirani, 2013; Tarazona *et al.*, 2015). The statistical approach to calculate DEG varies based on the distributional assumptions. Software such as DEGseq (Wang *et al.*, 2009), MyRNA (Langmead *et al.*, 2010), and PoissonSeq (Li *et al.*, 2012) use Poisson model for RNA-Seq count data while edgeR (Robinson *et al.*, 2010), DESeq (Anders and Huber, 2012), and DESeq2 (Love *et al.*, 2014) use negative binomial model. In addition, there are more DEG calculation software tools that use more other statistical models. However, it is important to understand the characteristics of the models and carefully apply to the data (Huang *et al.*, 2015).

1.3.2 Gene prioritization

Gene prioritization is a strategy that identifies the most promising genes from a large pool of candidates by integrating multiple data source for further downstream screens (Figure 1.2). The integration of the list of genes and external data sources allows increasing the data dimension from 1D (simple gene list) to a higher dimension that can have much more explanation to the data (Moreau and Tranchevent, 2012; Cowen *et al.*, 2017). Currently, the strategy of gene prioritization can be generally categorized into four types. *(i)* filtering strategy, *(ii)* profiling and data fusion, *(iii)* text-mining, and *(iv)* network analysis (Moreau and Tranchevent, 2012).

Filtering strategy is an approach that uses multiple data source (or methods) as filters while each filter removes less significant candidate genes step by step (Figure 1.2a). Unlike filtering strategy, data fusion strategy has the advantage to avoid the false negatives caused by the hard thresholds among filters by scoring the candidates at each data sources and summarizes the overall rank (Figure 1.2b). Text-mining is a data mining method that finds the associations between given keywords. In bioinformatics, text-mining is a strategy that finds the associations between candidate genes and knowledge (disease, phenotype or else) while the relationship between two elements is retrieved by information retrieval methods (Krallinger *et al.*, 2008; Winnenburg *et al.*, 2008)(Figure 1.2c). Network-based strategy uses various type of networks (biological pathway, PPI, GRN) to find the similarity between the candidate genes and networks. Network propagation is one of the popular technique to find the similarity between candidate genes and seed genes using networks while seed genes are often defined as disease genes or phenotype-relevant genes that requires prior knowledge (Figure 1.2d).

1.4 Outline of the thesis

Chapter 2 elaborates on the process of integration of network and SNV information, which was able to improve the statistical bias from mouse gene KO experiments that have a small number of samples. Chapter 3 describes a system that combines GRN, PPI, and text-mining techniques to prioritize genes that focus on the context of the experiment. Chapter 4 demonstrates that Venn diagram has a great advantage in prioritizing genes that can address the challenges of heterogeneous data and seed selection issues for network propagation. Chapter 5 summarizes and concludes the studies that are presented in this thesis. The bibliography of the cited references is organized at the end of the thesis.

Chapter 2

A filtering strategy that combines GRN, SNV information to enhances the gene prioritization in mouse KO studies with small number of samples

2.1 Background

DEGs from RNA-Seq data are often used for finding significant genes that can explain the phenotypic differences between control and cases (Oshlack *et al.*, 2010; Frazee *et al.*, 2014). However, in gene KO studies, discovering phenotype-dependent gene only with DEG can be difficult because distinguishing whether the expression alteration is resulted by the inactivation of the KO gene or by the genetic variations that were merely from differences in samples rather than phenotypic differences. And the problem becomes much more challenging when the number of samples is small, an issue that RNA-Seq experiments face frequently (Tarazona *et al.*, 2011). Various methods and models were proposed to overcome the difficulties of selecting phenotype related DEGs from a small number of samples such as the Poisson model (Marioni *et al.*, 2008), Bayesian approaches (Vreugdenhil *et al.*, 2008; Anders and Huber, 2010), or increased the sequencing depth of samples (Tarazona *et al.*, 2011).

Even if a number of studies have resolved the difficulties of DEG detection to some degree, addressing phenotype related DEGs from a small number of samples is still a challenging process. Studies suggested to increase the number of biological samples is the most critical factor have significant DEGs (Zhou *et al.*, 2013). However, increasing the number of biological samples is not easy for many reasons. Thus, a new approach that can detect significant gene markers in a small number of samples is necessary. This study proposes a new method that distinguishes genes that are relevant to the phenotypic differences in mouse gene KO experiments that have a small number of samples. The method uses the filter-out gene prioritization strategy that combines GRN, biological pathways and SNV information using DEGs as input (Hur *et al.*, 2015).

2.2 Methods

The gene prioritization method uses a reductionist approach by adding more filters at each step as described below (Figure 2.1).

1. The first filter is to use a method to identify DEGs between control and case samples. In this study, we used fold change, a classical DEG selection method.

2. The filter at the second step is to use GRN. GRN is constructed from a large volume of public data to represent the whole gene regulatory network. DEGs that are included in the network are selected as candidates.

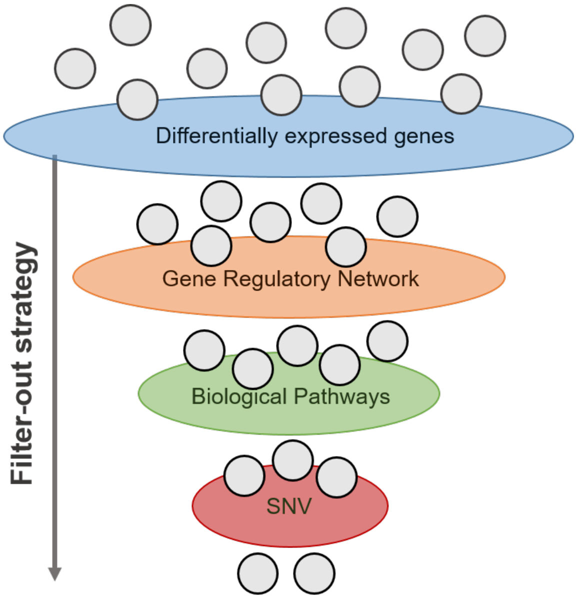


Figure 2.1: Filtering strategy combining networks and SNV

3. The third filter utilizes biological pathway information. Candidates that are not included in the pathways are discarded.

4. Finally, candidates that have higher than a certain rate of SNVs are discarded since the DEGs that have SNVs possibly resulted from genetic differences rather than phenotypic differences.

2.2.1 First filter : DEG

From the given expression profile, DEGs are considered as initial candidates. DEGs are used for the purpose of observing the alteration of expression patterns that could explain the phenotypic differences among samples. DEGs were selected by using fold change of the expression value (FPKM) between case and control. The study used multiple cutoffs in order to compare and observe differences in the number of selected genes. Note that this study used samples that does not have enough biological replicates to perform statistical testing to calculate DEGs. Therefore, expression fold change was used as a DEG estimation.

2.2.2 Second filter : GRN

The concept of reverse engineering the regulatory network from transcriptome data, GRN is a very effective method that can consider complex relationships of many genes (Basso *et al.*, 2005). GRN is used as the second filter of the gene prioritization process in order to discard genes that have less significant roles in the regulatory network.

In order to construct a GRN that is appropriate for mouse gene KO experiment data, public data (Microarray, RNA-Seq) of mouse were collected from NCBI GEO. For microarray, each series matrix files from GSE45929 (Ramsey *et al.*, 2013), GSE16741 (Yun *et al.*, 2010), GSE30906 (Shan *et al.*, 2012), GSE36780 (Bae *et al.*, 2012), GSE40375 (not published), GSE41380 (Nusinow *et al.*, 2012), GSE43663 (Ruan *et al.*, 2013) were used for GRN construction. These data contain gene expression value of multiple samples that differs in mouse's strain, genotype, and treatment (42 samples in total) and were created by the same microarray platform (Illumina MouseWG-6 v2.0 expression beadchip) and preprocessed by R bioconductor lumi package (Du *et al.*, 2008) (variance stabilizing transform, quantile normalization). The study integrates gene expression values of 7 series matrix files (GSE45929, GSE16741, GSE30906, GSE36780, GSE40375, GSE41380, GSE43663) into a single matrix and quantile normalized gene expression values of every sample and used it as an expression profile for construction of GRN.

GRN is constructed by using NARROMI (Zhang *et al.*, 2012) while a list of transcription factors and co-factors from the Animal Transcription Factor

Database (Zhang *et al.*, 2011) was used for regulatory information for NARROMI. For the gene list, we simply defined it as a list of whole genes that includes not only transcription factors and co-factors but also non-transcription factors. As a result, NARROMI constructed a network topology of 2950865 edges. The study supports a URL for the network topology file which was used in this study (epigenomics.snu.ac.kr/mouse_network/total_mouse.topology).

With the constructed GRN, the study discards candidates that have weak or no regulatory roles. The method filters out less significant DEGs that do not have any potential regulatory roles upon the calculated GRN. As a result, candidates that participate in a regulatory role remains.

2.2.3 Third filter : Biological Pathway

The combination of DEG and GRN information was used not only for reducing the number of candidates but also to select significant genes that have regulatory roles that could represent the phenotypic differences between WT and KO mouse. However, GRN is a hypothetical topology that gains regulatory information from the given data. Therefore, it is also important to ensure whether the candidates have biological evidence. In this study, KEGG pathway (Kanehisa and Goto, 2000) for confirming the candidates in terms of domain knowledge.

2.2.4 Final filter : SNV

Even if the study reduced the number of candidates by using multiple filtering methods, it is necessary to eliminate genes that have genetic differences that may not represent phenotypic differences. Since the statistical power is weak in a small number of samples, it is difficult to distinguish whether the genetic differences were caused by phenotypic differences or not. Therefore the study removed genes that have a certain or higher SNV rate. This process will remove

SNVs from the genetic differences but also by the phenotypic differences. A possibility to have false negative results. However, it will completely avoid the risk of selecting SNVs resulting from genetic differences.

2.3 Results and Discussion

GSE47851 were used for evaluating the filter-out method. The performance of the method is discussed by comparing between the genes reported from the original research article (Yagi *et al.*, 2014) and the genes prioritized by the filtering method.

RNA-Seq data of GSE47851 are from an experiment of Gata3 KO that have multiple SRA files. The study used 8 SRA files (SRR896215, SRR896216, SRR896217, SRR896218, SRR896219, SRR896220, SRR896221, SRR896222) that have two conditions where each of the conditions has 2 biological samples and 2 technical replicates of each biological sample.

The study reported that genes of TNF and TNFR superfamilies, members of NFkB and cell surface markers of ILC2s have expression alterations when Gata3 is not activated in ILC2 cells (Yagi *et al.*, 2014). The authors reported that when Gata3 is inactivated, many TNF and TNFR superfamily genes, such as Tnfrsf9 and Tnfsf21 and NFkB family members, including Nfkb2 and Relb, have altered expression patterns while cell-cycle inhibitor Cdkn2b was up-regulated. According to the authors we report, the reductionist approach was able to reproduce 4 out of 5 genes (except Tnfsf21). In addition, we were able to reconfirm the following facts by mapping the candidate genes to the KEGG pathway. Figure 2.2 represents expression alteration in NF-kappa B signaling pathway, showing down regulations of Nfkb2 (p100) and Relb when Gata3 is inactivated. Expression alteration was also detected in the TNF signaling pathway (Figure 2.2).

Filtering Steps	NONE	1st filter	2nd filter	3rd filter	Final filter
SelectedCandidates	12298	2153	1184	478	343
TruePositives	23	23	19	18	14
Accuracy	0.002	0.827	0.905	0.962	0.972
Precision	0.002	0.011	0.016	0.038	0.041
Recall	0.885	0.885	0.731	0.692	0.538
F-measure	0.004	0.021	0.032	0.073	0.076

Table 2.1: Performance comparison of filters

The table represents the remaining candidates, number of correctly predicted true positives, and the performance of each adapted filters.

TNF and TNFR superfamily genes, such as Tnf and Tnfrsf9, were successfully detected in the pathway as well as the statement (Figure 2.3).

The study also stated about the expression alterations in cell-surface markers of ILC2s. The study reported that 130 genes are positively regulated by GATA3 in ILC2s, and not in Th2 cells. Cell-surface markers of ILC2s, such as Icos, Il2ra, Kit, Il1r2, Cysltr1, Htr1b, and Tph1 were included. As a result, the reductionist approach was able to reproduce 4 genes among 7 were successfully matched (Figure 2.4C).

In addition, the study evaluated whether each filter had a significant role during the filter-out process (Table 2.1). Table 2.1 summarizes the performance of prioritizing candidates at each filtering step. Without no filter (NONE), it is obvious that there is a very few chances to prioritize genes reported in the original paper. However, when filters are gradually added, the number of false positives decreased rapidly. In addition, the recall has steadily decreased at each filtering steps, but the F-measure represents that the general performance of

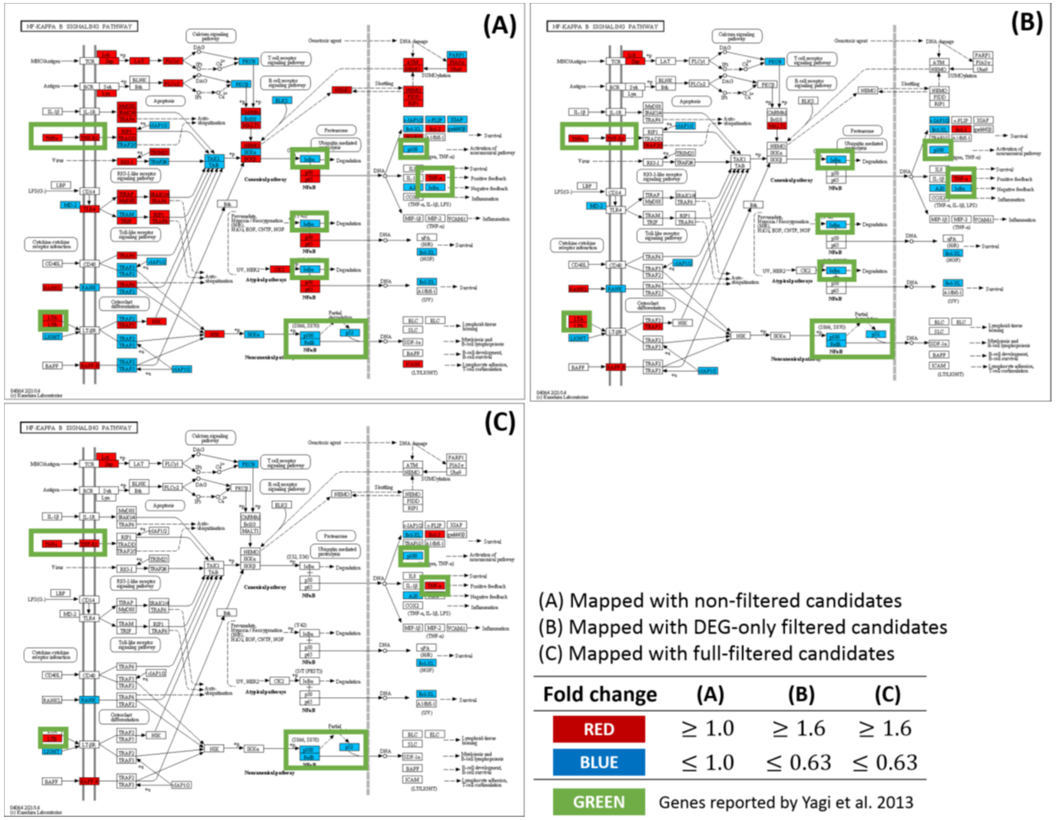
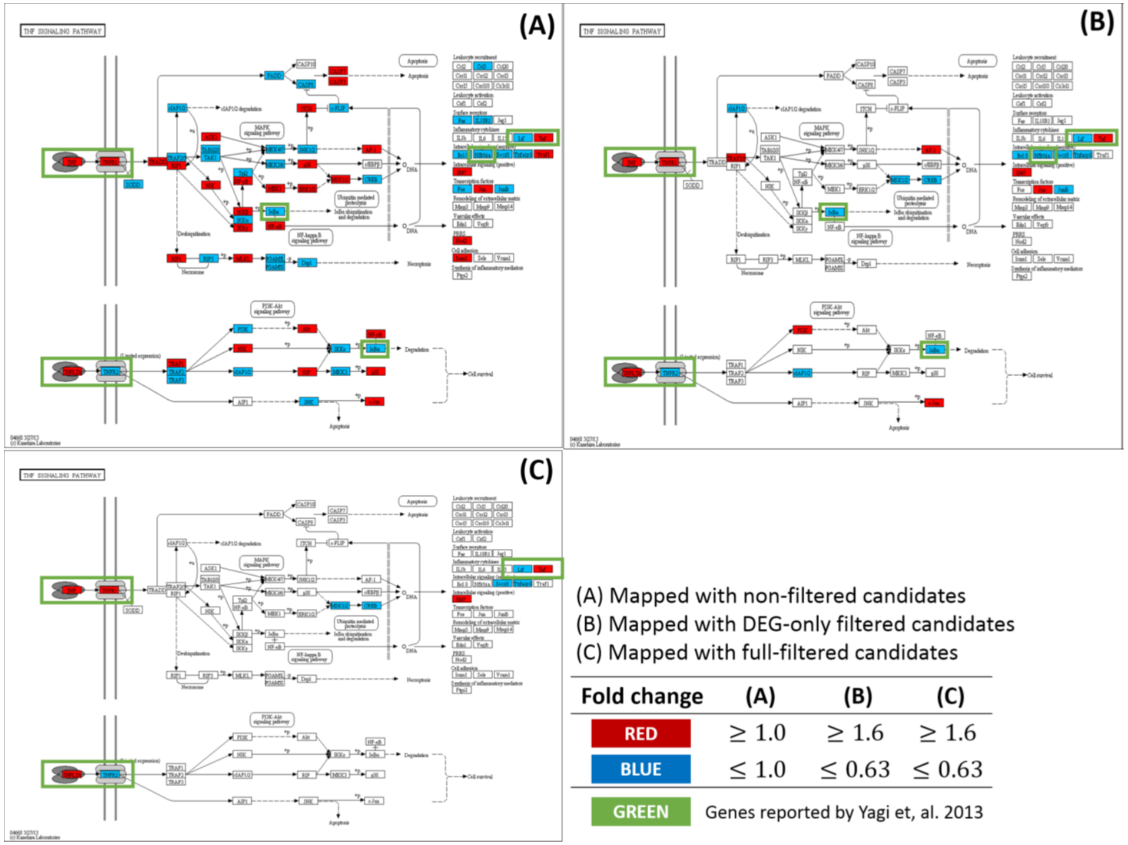

(A)
(B)
(C)
NF-KAPPA B SIGNALING PATHWAY
NF-KAPPA B SIGNALING PATHWAY
NF-KAPPA B SIGNALING PATHWAY
(A) Mapped with non-filtered candidates
(B) Mapped with DEG-only filtered candidates
(C) Mapped with full-filtered candidates
Fold change (A) (B) (C)
RED ≥ 1.0 ≥ 1.6 ≥ 1.6
BLUE ≤ 1.0 ≤ 0.63 ≤ 0.63
GREEN Genes reported by Yagi et al. 2013

Figure 2.2: Expression alteration in NF-kappa B signaling pathway (Hur *et al.*, 2015)

(A) NF-kappa B signaling pathway mapped with non-filtered candidates. With no filtering method, too many genes are shown in the pathway which makes it difficult to find an appropriate gene marker. (B) NF-kappa B signaling pathway mapped with candidates filtered by DEG. The number of genes is greatly reduced compared to the non-filtered method. However, difficulty exists in finding significant gene markers as the number of genes is still too great. (C) NF-kappa B signaling pathway mapped with full-filtered candidates. The number of genes was greatly reduced compared to non-filtered or DEG-only filtered candidates while keeping the genes reported by Yagi et al.(2013).


(A)
(B)
(C)
TNF SIGNALING PATHWAY
(A) Mapped with non-filtered candidates
(B) Mapped with DEG-only filtered candidates
(C) Mapped with full-filtered candidates

Fold change	(A)	(B)	(C)
RED	≥ 1.0	≥ 1.6	≥ 1.6
BLUE	≤ 1.0	≤ 0.63	≤ 0.63
GREEN	Genes reported by Yagi et, al. 2013		

Figure 2.3: Expression alteration in TNF signaling pathway (Hur *et al.*, 2015)

(A) TNF signaling pathway mapped with non-filtered candidates. With no filtering method, too many genes are shown in the pathway which makes it difficult to find an appropriate gene marker. (B) TNF signaling pathway mapped with candidates filtered by DEG. The number of genes is greatly reduced than non-filtered method. However, difficulty exists in finding significant gene marker as the number of genes is still too many. (C) TNF signaling pathway mapped with full-filtered candidates. The number of genes was greatly reduced than nonfiltered or DEG-only filtered candidates while keeping the genes reported by Yagi et al.(2013).

the filtering process was better than the previous steps.

2.4 Discussion

This study proposed a novel method that uses four filtering steps to distinguish phenotype-dependent genes from RNA-Seq data of mouse gene KO studies that have a small number of samples. The study demonstrated that the combination of DEG, GRN, biological pathways and SNV information was able to narrow down the significant genes that have regulatory roles and reduced the risk of including candidates that have genetic differences. However, several limitations of this study need to be addressed. First of all, there should be a more rigorous study of GRN construction. Using much omics data for GRN construction somehow preserves important relationships between transcription factors and their target genes, but how much data is needed for GRN construction is not rigorously studied. In this study, we had enough omics data for the network construction, therefore we were able to use a simple method using NARROMI (Zhang *et al.*, 2012). However, when the amount of omics data for network construction is not enough, special techniques such as low order partial correlation based methods (Zuo *et al.*, 2014) should be considered. Second, removing genes with genetic variation allows us to focus on genes that are relevant to the underlying biological mechanisms for the KO study. However, genetic variations do not always affect the transcription activity of genes, and it is possible that the suggested method might discard a number of SNVs that were affected by the KO gene. Thus, it is necessary to investigate the effect of genetic variations on transcription activities.

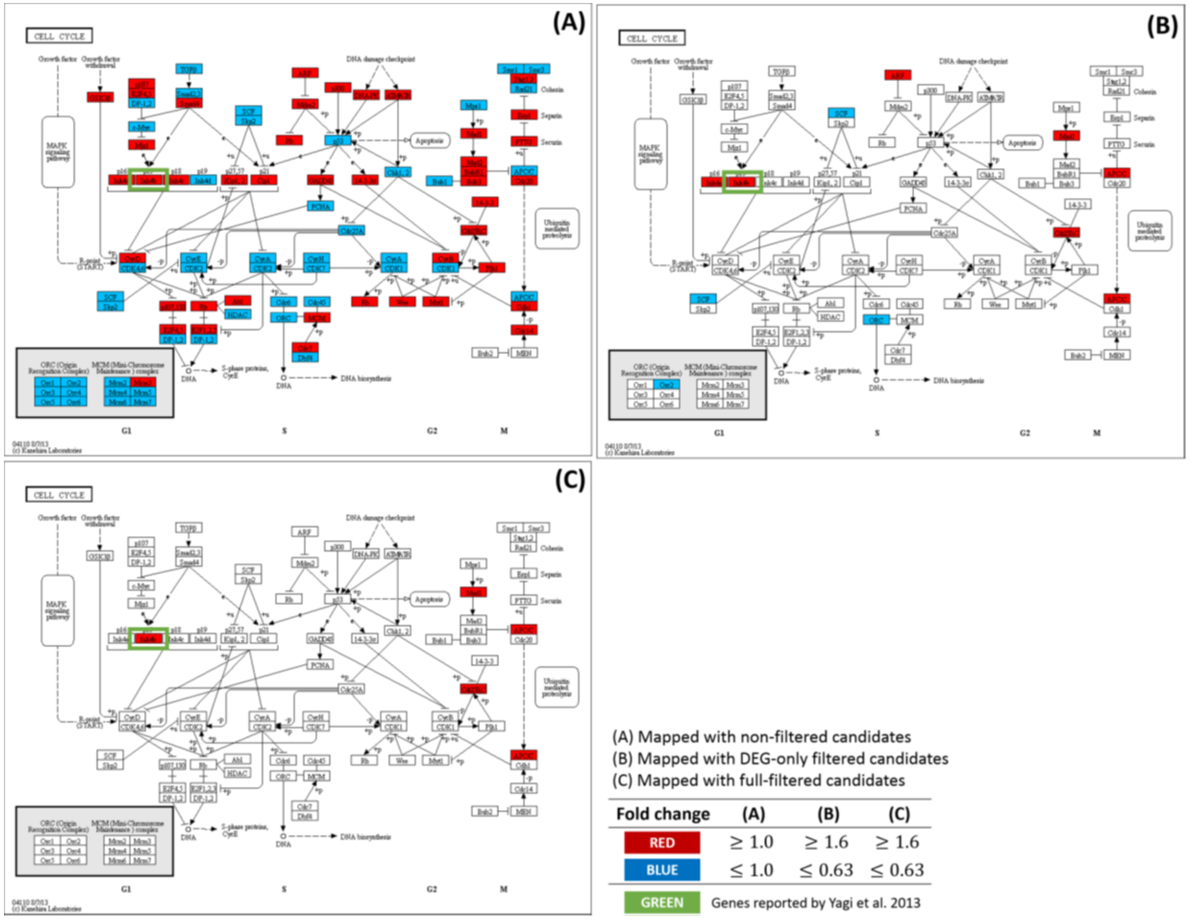

(A)
(B)
(C)
CELL CYCLE
(A) Mapped with non-filtered candidates
(B) Mapped with DEG-only filtered candidates
(C) Mapped with full-filtered candidates
Fold change (A) (B) (C)
RED ≥ 1.0 ≥ 1.6 ≥ 1.6
BLUE ≤ 1.0 ≤ 0.63 ≤ 0.63
GREEN Genes reported by Yagi et al. 2013

Figure 2.4: Expression alteration in Cell cycle pathway

(A) TNF signaling pathway mapped with non-filtered candidates. With no filtering method, too many genes are shown in the pathway which makes it difficult to find an appropriate gene marker. (B) TNF signaling pathway mapped with candidates filtered by DEG. The number of genes is greatly reduced than non-filtered method. However, difficulty exists in finding significant gene marker as the number of genes is still too many. (C) TNF signaling pathway mapped with full-filtered candidates. The number of genes was greatly reduced than non-filtered or DEG-only filtered candidates while keeping the genes reported by Yagi et al.(2013).

Chapter 3

An integration of data-fusion and text-mining strategy to prioritize context-laid genes in mouse TF KO experiments

3.1 Background

To overcome the limitations of the DEG methods, studies suggested data fusion techniques that utilize additional information to effectively identify genes that are related to the phenotypic differences. However, it is known that the integration of heterogeneous databases has several difficulties while prioritizing candidates for data of gene knock study that motivated this study. First, most of the existing gene prioritization tools are not appropriate for the condition-specific data such as mouse KO data. When a certain gene is knocked out, researchers have specific hypotheses that are related to the observed phenotypic differences. Thus, to select genes that are related to phenotypic differences, it is important to not only consider gene expression alteration but also to prioritize genes with the researcher's interest. Without considering the condition or the goal of the experiment, gene prioritization will likely to focus on genes that have enough supporting evidence instead of considering the intention of the experiment design. The best strategy is to provide information about the conditions of the experiment or a specific hypothesis that the user has. When the user provides such information, genes can be prioritized by consulting the literature database. Therefore, it is necessary to perform an integrative analysis of transcriptome data and literature data for the condition-specific gene selection and prioritization.

Second, complex relationships among genes should be considered in order to selected and prioritize genes that are related to the phenotype. Therefore, networks such as GRN and PPI are useful in explaining alteration among genes by considering gene-gene and regulatory relationships. Many KO experiments investigated transcription factors (TFs) that could result in the phenotypic differences by analyzing the GRN (Geier *et al.*, 2007; Madhamshettiwar *et al.*, 2012; Wang *et al.*, 2012; Ud-Dean and Gunawan, 2015).

Thus, considering GRN (to be specific, GRN) is essential to characterize the roles of TFs from KO data. In addition to GRNs, PPI networks also assist in explaining expression alteration among genes since PPI networks consist of more entities than other networks such as GRNs and biological pathway networks. Since we need to use both TF and PPI networks, an issue is how to utilize two different networks in a single computational framework. Our approach uses GRN to select candidate genes from the TF KO experiment and uses the combination of PPI and literature information to prioritize candidate genes in a condition-specific manner.

Third, existing computational methods for prioritizing genes are not designed for mouse KO data. Only 3 among 27 tools (listed in Gene Prioritization

Portal (Tranchevent *et al.*, 2010)) are designed for the mouse data (van Dam *et al.*, 2012; Tranchevent *et al.*, 2008; Nitsch *et al.*, 2011).

However, these tools are generally not applicable to evaluate RNA-Seq data of KO experiments. For example, even though PINTA (Nitsch *et al.*, 2011) and GeneFriends (van Dam *et al.*, 2012) can prioritize genes based on the concept of the guilt-by-association or network analysis, these tools require a pre-selected gene list of a certain size: up to 200 genes in PINTA and up to 500 genes in GeneFriends. Both tools are not applicable when the number of genes is large, such as DEG results. Although the use of a stringent cutoff value can reduce the number of candidate genes that can be used for the aforementioned tools, there may be too many false negatives. Therefore, the requirement of a pre-selected gene list in PINTA and GeneFriends is not easy to be resolved. In addition, PINTA is designed for microarray data and prioritizes genes by referring to the expression profiles of its neighbors from the PPI network, but it does not consider the influence of the KO gene. Likewise, GeneFriends prioritizes genes by considering co-expression of other genes but does not reflect the effect of the KO gene. Another tool, Endeavor (Tranchevent *et al.*, 2008), is able to prioritize genes from a large number of gene list that does not require pre-selection from a gene list. However, Endeavor requires a gene list from prior knowledge for a training dataset, and it is designed to select disease-related genes rather than KO related genes. To address the discussed issues, this study developed CLIP-GENE (Context Laid Integrative analysis to Prioritize genes) (Hur *et al.*, 2016). A web-based tool that takes a DEG list as input and uses GRN and SNV information to narrow down candidate genes and prioritizes genes with PPI information and literature information. In particular, CLIP-GENE allows researchers to specify the context of the experiment as a set of keywords input to a bio-medical entity search tool (BEST) (Lee *et al.*, 2016).

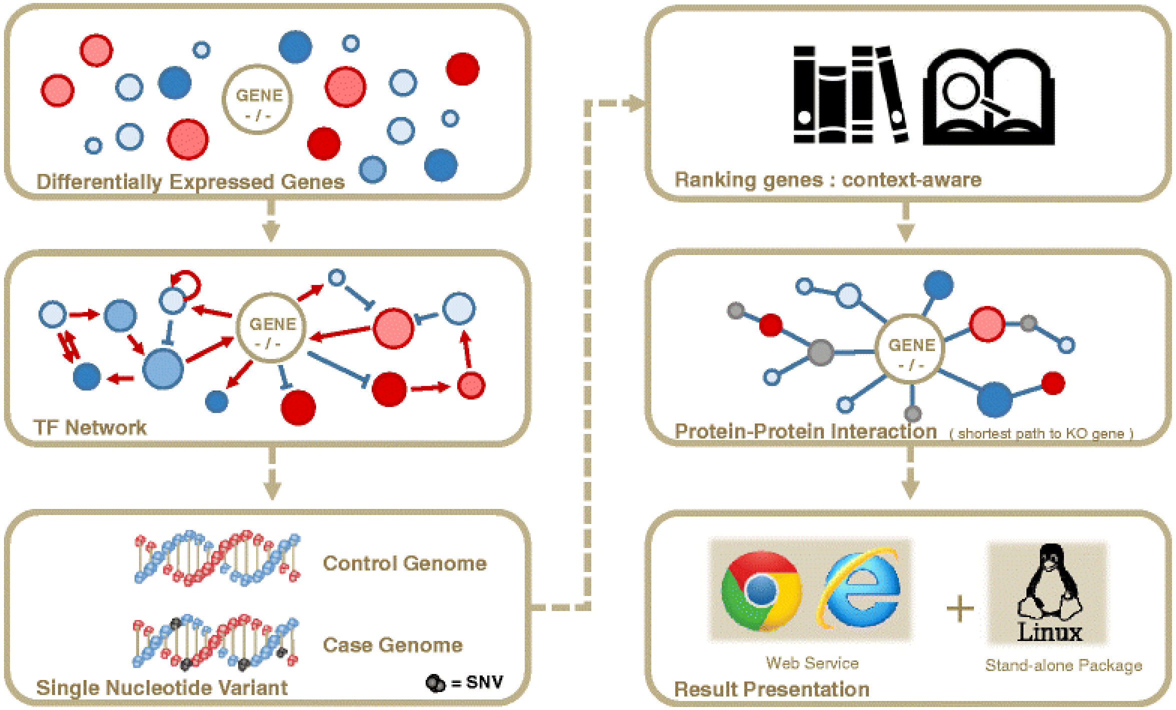

GENE
- / -
Differentially Expressed Genes
TF Network
Control Genome
Case Genome
Single Nucleotide Variant
= SNV
Ranking genes : context-aware
Protein-Protein Interaction (shortest path to KO gene)
Web Service
+
Linux
Stand-alone Package
Result Presentation

Figure 3.1: A workflow of CLIP-GENE (Hur *et al.*, 2016)

CLIP-GENE prioritize user-interested genes that are relevant to phenotypic/functional differences of KO mouse data. CLIP-GENE takes DEG as input and filters out genes by using GRN and SNV information. Then prioritize these genes by using BEST and PPI information

3.2 Methods

CLIP-GENE prioritizes genes with two major steps, selection, and ranking. For the selection step, GRN and SNV information are used to select candidate genes that are affected by the KO gene as well as expressed differentially between wild type and KO mouse. For the ranking step, BEST and PPI information are used to prioritize genes according to the researcher's context or hypothesis. With the assistance of a BEST (Lee *et al.*, 2016), it allows specifying certain context or hypothesis with a set of keywords by a user that is expected from the data. Afterward, PPI is used to consider the gene-gene relationship between the candidate genes and the KO gene. Workflow of CLIP-GENE is illustrated in Figure 3.1. Details of each step are described below.

3.2.1 Selection of initial candidate genes.

CLIP-GENE takes a DEG list from the KO experiment and investigates the regulatory role of the DEGs by referring to GRN. GRN is created using NARROMI (Zhang *et al.*, 2012) with data of normal inbred mouse data that varied in its strains, developmental stage, and tissues (150 samples of wild type mouse RNA-Seq data from 17 independent studies) (Yao *et al.*, 2014; Tena *et al.*, 2014; Stilling *et al.*, 2014; Srivastava *et al.*, 2015; Shen *et al.*, 2014; Roger *et al.*, 2014; Ntziachristos *et al.*, 2014; Moniot *et al.*, 2014; Mielcarek *et al.*, 2014; Liu *et al.*, 2014; Kayo *et al.*, 2014; Harmacek *et al.*, 2014; Gu *et al.*, 2014; Deng *et al.*, 2014; Bhatnagar *et al.*, 2014; Altboum *et al.*, 2014; Alpern *et al.*, 2014). while a list of transcription factors and co-factors from the Animal Transcription Factor Database (Zhang *et al.*, 2011) was used for the regulatory information for NARROMI.

CLIP-GENE takes a list of DEGs as input and uses them as initial candidates. Then, by referring to the mouse GRN that was constructed using 150 mouse expression profiles, DEGs that do not affect other DEGs or DEGs that are not affected by the KO gene are excluded. This step is performed to focus on the relationship between the regulator and its target genes that are significantly altered.

After CLIP-GENE selects candidate DEGs that take a part in the regulatory role, SNV information is used to filter out DEGs that might be caused by the genetic differences rather than the influence of the KO gene. It is well known that even if the inbred mouse are raised in a controlled environment, genetic differences are likely to be present (Eisener-Dorman *et al.*, 2009). If a large number of RNA-Seq experiments can be performed, it is possible to screen genes that may be expressed differentially due to the genetic difference. However, it is not practical to perform such a large number of RNA-Seq experiments that is enough to remove such genes. To compensate the low statistical power of the typical RNA-Seq data, candidate genes with over than a certain rate of SNVs in the KO mouse are discarded (Hur *et al.*, 2015).

3.2.2 Prioritizing genes with the user context and PPI

Candidate genes selected in the previous step are ranked in terms of the relevance to the phenotype in two different criteria: the user specified context and the PPI information.

Researchers can specify their hypothesis for the KO data as 'context' in a set of keywords. Specifically, context means a set of subjective words that describe the user's interest such as 'expected biological function when the gene is KO' or 'known function of the KO gene'. For example, a context for Gata3 KO data can be described as 'Immune response', 'Cell signaling', or 'Inflammatory response' (Yagi *et al.*, 2014; Wan, 2014). Then genes that are related to the user-specified keywords can be determined by looking for the relevance between keywords since certain keywords are documented in the literature in relation to a certain gene. Thus this can be viewed as a process to find a keyword-keyword relationship and keyword-gene relationship to prioritize genes. In order to find the relevance between two different keywords, literature search systems based on the named entity recognition (NER) are known to be effective (Spampinato *et al.*, 2011). For CLIP-GENE, BEST (Lee *et al.*, 2016) is used to find the relevance between the KO gene and candidate genes as well as the relationship between candidate genes and the user given context. With the user-specified keywords, BEST computes relevance between any pair of keywords from PubMed and returns a relevance score of genes with ranks. Once the relevance score of 'context to candidate gene' and 'KO gene to candidate gene' is calculated, the maximum of them is used to represent how the candidate gene is relevant to the user's interest or the KO gene. As a result, a candidate gene with a higher relevance score is ranked with higher priority.

PPI information is used to rank candidates by computing the shortest interaction path to the KO gene on the STRING PPI network (Szklarczyk *et al.*, 2010). Candidates that have a shorter interaction path to the KO gene are considered to be more relevant to the phenotypic/functional difference, hence they are ranked with a higher priority. Finally, CLIPGENE summarizes candidates with ranks by combining the BEST and PPI information with unweighted Borda count (Grazia, 1953). Figure 3.2 describes the overview of gene prioritization.

3.3 Results and Discussion

For the performance evaluation, we used datasets that come with publications reporting which genes are relevant to the functional difference when the gene is

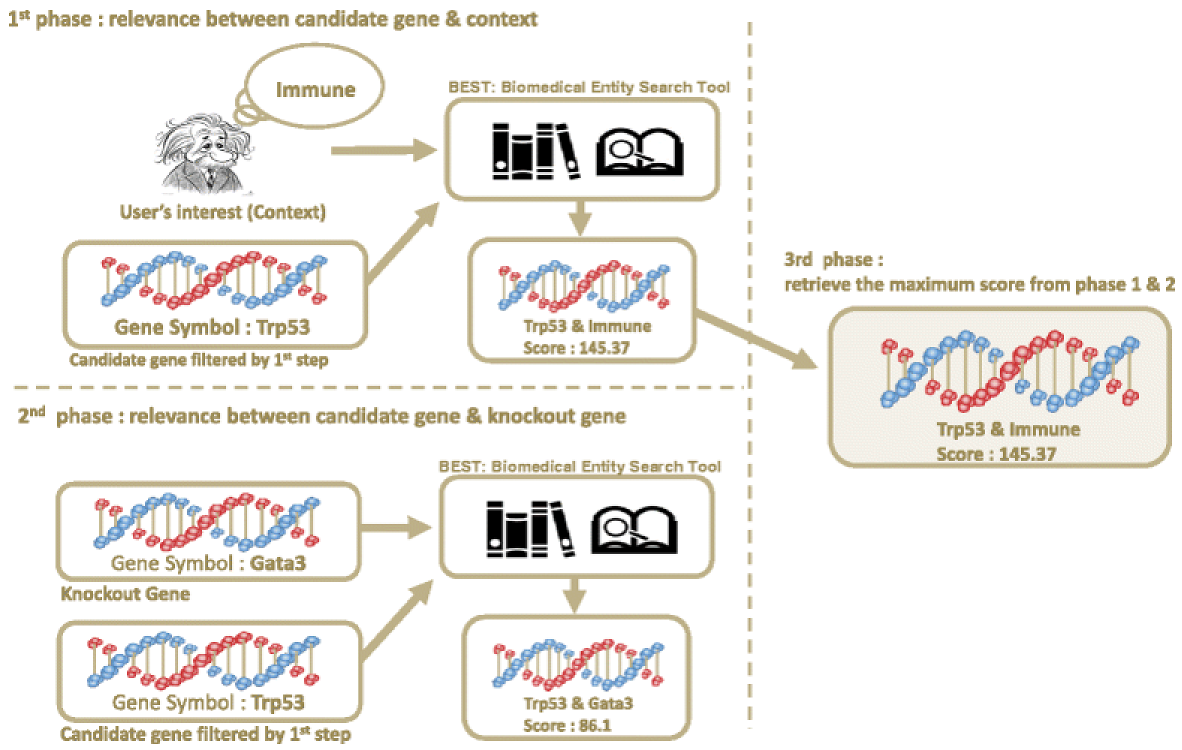

1st phase : relevance between candidate gene & context
Immune
User's interest (Context)
BEST: Biomedical Entity Search Tool
Gene Symbol : Trp53
Candidate gene filtered by 1st step
Trp53 & Immune
Score : 145.37
3rd phase :
retrieve the maximum score from phase 1 & 2
Trp53 & Immune
Score : 145.37
2nd phase : relevance between candidate gene & knockout gene
BEST: Biomedical Entity Search Tool
Gene Symbol : Gata3
Knockout Gene
Gene Symbol : Trp53
Candidate gene filtered by 1st step
Trp53 & Gata3
Score : 86.1

Figure 3.2: Gene selection and ranking process (Hur *et al.*, 2016)

Prioritizing genes with Biomedical Entity Search Tool (BEST). BEST is utilized to find the relevance between KO gene and candidate gene as well as the relationship between the candidate gene and given context. Then CLIP-GENE retrieves the maximum score to represent that the candidate gene is highly relevant to the user's interest or KO gene. As a result, the candidate gene with higher relevance score is ranked with high priority silenced. These genes are used as true positives to measure the precision, recall, and F-measure in terms of genes reported in the publications for data sets, GSE47851 (Yagi *et al.*, 2014), GSE54932 (Zhang *et al.*, 2014), and GSE53398 (Zhuang *et al.*, 2014). CLIP-GENE was compared with methods and tools that can be used for RNAseq mouse data. This study compared DEG-only method (DEG), integrative analysis method (IA) (Hur *et al.*, 2015), and GeneFriends (van Dam *et al.*, 2012) in terms of the predictive power. In addition, since the user can specify context with a set of keywords, the performance depends on the context that the user provides. In this experiment, four different sets of keywords are used as context. To compare the predictive power, the study designated the best case and the worst case in terms of the number of genes reproduced by CLIP-GENE. In addition, as BEST investigates the relationship between two given keywords by referring the abstract from PubMed, we chose keywords that were not mentioned in the abstract of the corresponding publications. This process is done to make sure that BEST did not consider the keywords from the publication that generated the data while calculating the relevance score.

Dataset GSE47851 is from a Gata3 KO mouse study that reported 25 genes were relevant to the functional difference between the wild type and the KO. For the performance evaluation, four different contexts: 'Inflammatory response', 'Immune regulation', 'Cell differentiation', 'Cell proliferation', the known functions of Gata3 (Yagi *et al.*, 2014; Wan, 2014). Dataset GSE54932 is from a Setd2 KO study, reporting 21 genes that are relevant to the phenotypic/functional differences between the wild type and the KO. 'Cell proliferation', 'DNAmismatch repair', 'Endodermal differentiation', and 'Histone modification' were used as the contexts for the Setd2 KO study since they are keywords representing well-known functions of Setd2 (Zhang *et al.*, 2014; Feng *et al.*, 2015). Dataset GSE53398 of Barx2 KO mouse, was used for the last evaluation. The study

Methods	Precision	Recall	F-measure
DEG	0.0105	1	0.0208
IA	0.0239	0.72	0.0463
GeneFriends	0.0038	0.92	0.0075
CLIP-GENE (Immune regulation*)	**0.0613**	**0.64**	**0.1122**
CLIP-GENE (Inflammatory response)	0.0354	0.76	0.0677
CLIP-GENE (Cell differentiation)	0.0294	0.72	0.0564
CLIP-GENE (Cell proliferation)	0.0201	0.72	0.0391

Table 3.1: Performance of CLIP-GENE while analyzing GSE47851 (Gata3 KO) The best performed measurement is marked with a star (*) with a bold context.

reported that 47 genes significantly differs when Barx2 is silenced. For the corresponding KO mouse data, we used ‘Myoblast progeny’, ‘Muscle maintenance’, ‘Chondrogenesis’, ‘Morphogenesis’ as the contexts for CLIP-GENE (Olguin and Olwin, 2004; Mi *et al.*, 2016; Zammit *et al.*, 2004; Meech *et al.*, 2012, 2005; Tsau *et al.*, 2011).

3.3.1 Performance with the best context

In terms of F-measure, CLIP-GENE achieved better performance in finding phenotypical/functional relevant (validated) genes than other methods 3.1,3.2, 3.3, as well as prioritizing phenotypic/functionally relevant genes with proper ranks (Hur *et al.*, 2016).

Context ‘Immune regulation’ achieved the best performance for the Gata3 KO data, which performed about 5.4 times better than DEG, 2.4 better than IA, and 15 times better than GeneFriends while ranking 4 genes in the top 10 gene list among 25 validated genes. For the Setd2 KO data, CLIP-GENE ranked 4 genes among 21 validated genes in top 10 with the context ‘Endodermal

Methods	Precision	Recall	F-measure
DEG	0.0099	0.5238	0.0195
IA	0.0183	0.1905	0.0333
GeneFriends	0.0015	0.5238	0.0031
CLIP-GENE (Endodermal differentiation*)	**0.2083**	**0.2381**	**0.2222**
CLIP-GENE (Cell proliferation)	0.0252	0.3333	0.0468
CLIP-GENE (DNA mismatch repair)	0.1304	0.1429	0.1364
CLIP-GENE (Histone modification)	0.0408	0.1905	0.0672

Table 3.2: Performance of CLIP-GENE while analyzing GSE54932 (Setd2 KO)
The best performed measurement is marked with a star (*) with a bold text.

Methods	Precision	Recall	F-measure
DEG	0.0071	0.7872	0.0142
IA	0.0111	0.3617	0.0215
GeneFriends	0.0036	0.617	0.0071
CLIP-GENE (Myoblast progeny)	**0.1818**	**0.0426**	**0.069**
CLIP-GENE (Muscle maintenance)	0.0476	0.0426	0.0449
CLIP-GENE (Chondrogensis)	0.1667	0.0426	0.0678
CLIP-GENE (Morphogenesis)	0.0217	0.4255	0.0412

Table 3.3: Performance of CLIP-GENE while analyzing GSE53398 (Barx2 KO)
The best performed measurement is marked with a star (*) with a bold text.

differentiation', achieving 11 times better than DEG, 6.7 times better than IA, and 72 times better than GeneFriends. For the Barx2 KO data, context 'Myoblast progeny' achieved the best performance, achieving 4.8 times better than the DEG, 3.2 times better than IA method, and 9.7 times better than Gene Friends. In addition, CLIP-GENE was able to prioritize 2 genes among 47 validated genes in the top 10 from Barx2 KO data.

3.3.2 Performance with the worst context

In terms of F-measure, even with the worst performed context, CLIP-GENE achieved better performance in predicting phenotypic/functionally relevant genes. For the Gata3 KO data, context 'Cell proliferation' performed 1.9 times better than DEG and 5.2 times better than GeneFriends, and slightly poor than IA. CLIP-GENE ranked one gene in the top 10 among 25 validated genes. The context 'Cell proliferation' performed the worst case for the Setd2 KO data, which still performed better than DEG, IA, and GeneFriends while reporting one gene among 21 validated genes in the top 10. 'Morphogenesis' was the worst context for the Barx2 KO dataset. However, CLIP-GENE still performs better than other methods while ranking 2 genes from the 47 validated genes in top 10, which again suggests that CLIP-GENE promises significant results than other compared methods even with the worst context.

3.4 Discussion

The performance of CLIP-GENE depends on the context that the user provided. However, in terms of candidate selection and prioritization, even with the context that performed worst, CLIP-GENE was consistently superior to DEG, IA, and GeneFriends. Transcriptome data from mouse models with certain genes knocked out are widely used to investigate gene functions in terms of phenotypes. In order to determine genes that are affected by the knocked out TF, both selecting candidate genes and prioritizing genes are necessary. Only three tools are available for the mouse data, but none of these tools was appropriate to prioritize genes of user's interest from KO data. This study presents a novel web service that select and prioritizes the candidate genes in terms of the user's experimental context. Two major contributions are: (*i*) CLIP-GENE allows researchers to specify the experimental conditions in a set of keywords. Our system automatically determines relevance between the keywords and genes so that we can provide rankings of the candidate genes in the users' context. (*ii*) CLIP-GENE provides a comprehensive web service for the mouse KO experiments by integrating multiple resources into a single framework: mouse GRN, SNV information, PPI network, and literature information.

Chapter 4

Integrating Venn diagram to the network-based strategy for comparing multiple biological experiments

4.1 Background

Before performing advanced analysis (i.e. network analysis, gene set analysis, or more) in transcriptome data, identifying DEGs is the very first step to understand the characteristics of the experiment. Since the number of DEGs can be hundreds or thousands, understanding the difference between samples with a list (or lists) of DEGs is not easy. An effective method to summarize a large number of DEGs is to use Venn diagram. A simple, yet a powerful tool that can illustrate the portion of each gene sets. The intuitive diagram helps researchers to understand the common and distinctive characteristics of the experiments that assist the decision for further investigation. However, there are several issues when Venn diagram tries to compare and analyze multiple experiments.

First of all, current Venn diagram tools are difficult to find genes that are responsible for the phenotype differences. Most of the current Venn diagram applications are developed with the purpose of visualizing the correct appearance of the diagram or to compare gene sets that aid researchers' brief understanding by giving additional knowledge such as enriched sub-network or gene sets (Kestler *et al.*, 2004; Martin *et al.*, 2012; Kestler *et al.*, 2008; Chen and Boutros, 2011; Heberle *et al.*, 2015; Hulsen *et al.*, 2008; Wang *et al.*, 2014; Jeggari *et al.*, 2018). The provided information may be useful but it is difficult to design a follow-up experiment with a simple list of gene sets.

Moreover, elucidating the phenotypic difference for the experiment designs that have different controls is also an issue. For example, when a dataset of two experiments that focus to find the differences of gene KO (KO) effect between liver and muscle, the DEG of each experiment represents the tissue-specific phenotypic difference. Thus, comparing the gene sets and the number of genes of the two experiments is not informative enough to pinpoint whether the genes are affected by the gene knock out effect or the tissue effect.

If it is possible to rank DEGs in a region of Venn diagram, then the researcher can make a more informed decision and overcome the difficulties that are described. To rank DEGs, this study combined the gene prioritization method into Venn diagram. Gene prioritization is a widely used method to rank genes by combining multiple database and methods to maximize the biological relevance to answer a difficult question that cannot be easily solved in a single data. Network propagation is one of the widely used technique that computes the influence of initial nodes (or seeds) to other nodes (Cowen *et al.*, 2017), and prioritize genes in the context of biological networks (Li and Patra, 2010; Smedley *et al.*, 2014; Köhler *et al.*, 2008; Vanunu *et al.*, 2010; Lee *et al.*, 2011; Chen *et al.*, 2009a, 2006). However, selection of seed genes is one of the critical factors for the network propagation analysis and becomes more important when prior knowledge is not available or is not enough. This paper suggests that the seed selection issue can be handled by allowing the user to select seed genes freely in arbitrary combinations of regions in a Venn diagram. We present Venn-diaNet: a web-based Venn diagram based network analysis framework that can prioritize genes to compare multiple biological experiments of transcriptome data. A convenient web-based user interface is provided to generate Venn diagrams of DEGs dynamically and to perform network propagation experiments to investigate which genes are relevant to certain phenotypes. This study suggests that Venn diagram, coupled with analytic methods such as network propagation, can be a very useful tool for comparing multiple biological experiments with different controls.

4.2 Methods

4.2.1 Taking input data

Venn-diaNet takes multiple DEG lists as input while each DEG list is resulted by the comparison of treatment/control or treatment/treatment experiment (Figure 4.1: Step 1). Each file must include one DEG list from one experiment. For example, if a researcher wants to compare three different experiments, three independent files of DEG list must be provided. The format of the file is as follows. Each input file requires gene ID (transcript ID) for the first column and gene symbol for the second column. We provide an example data on the web page of Venn-diaNet for better understanding.

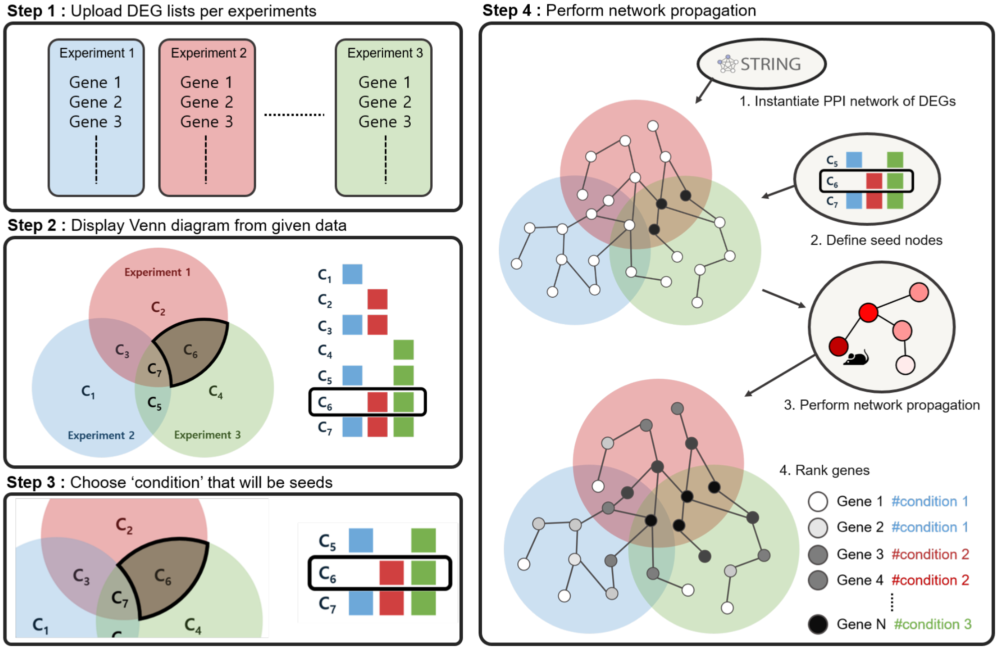


Figure 4.1: Venn-diaNet work flow

Step 1 : Venn-diaNet receives DEG lists per experiments from user. Step 2 : Uploaded DEGs from step 1 are interpreted with a Venn diagram as well as organized as sets with table. Step 3 : Define specific or multiple C_i as seeds for further network propagation analysis. Step 4 : Once the seed is defined, Venn-diaNet instantiates a PPI network of DEGs from STRING DB. Network propagation with given seeds from the previous steps. As a result, DEGs are ranked by the probability score calculated during the Markov Random Walk.

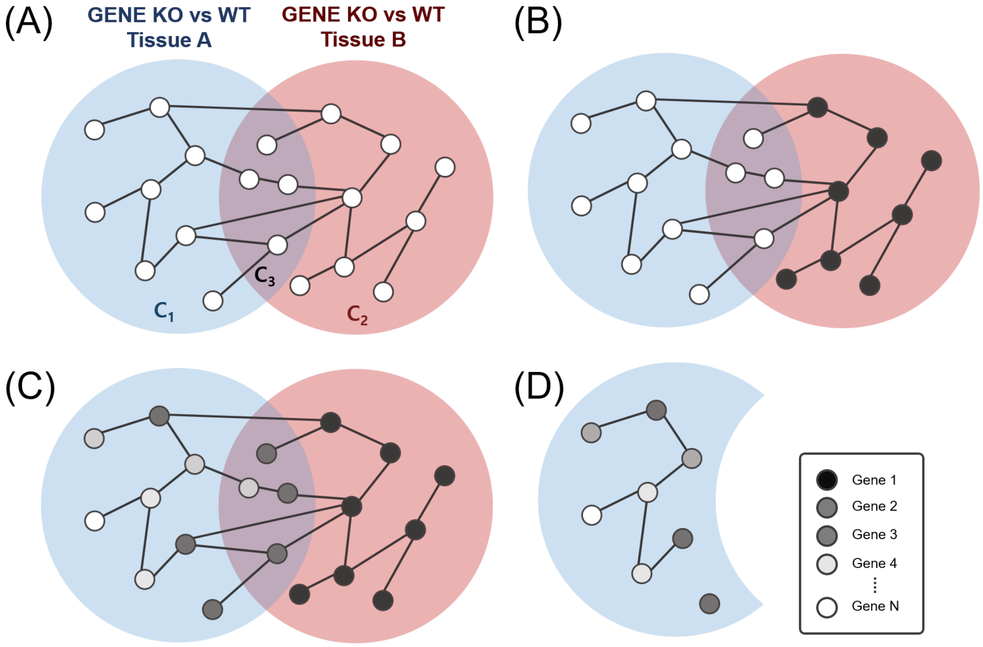


Figure 4.2: Key concept of Venn-diaNet

(A) Instantiate a PPI network with the DEGs from the multiple experiments. (B) When we are interested in C_1 that has similar function as C_2, we can define C_2 as seeds. (C) Performing network propagation with Markov Random Walk. (D) Discard C_3 genes (as well as seed genes) in order to focus on C_1 genes. Remaining genes are ranked by the probability score calculated from the previous step.

4.2.2 Generating Venn diagram of DEG sets

Venn-diaNet considers each experiment as a set for the diagram. Therefore, With given number (=n) of experiments E, Venn-diaNet generates a diagram of n circles that have $2^n - 1$ regions. Each region is denoted as C_i ($1 \leq di \leq 2^n - 1$) while each C_i contains genes of

$$\mathbf{C}_i = \{\mathbf{g} : \mathbf{g} \in \textstyle\bigcap_{j=1}^{N} \mathbf{G}(\mathbf{b}_j)\}, \qquad \mathbf{G}(b_j) = \begin{cases} E_j & if \;\; j = 1 \\ E_j^c & if \;\; j = 0 \end{cases}$$

b represents the binary number of C_i (i.e. $C_1 = 001$) while b_j indicates the position of digits (i.e. $b_1 = 1$, $b_2 = 0$, $b_3 = 0$). If Venn-diaNet receives DEG lists from 3 experiments, Venn-diaNet illustrates a Venn diagram of 3 sets (E_1,E_2,E_3) that have 7 regions (C_1,C_2,C_3, $\cdots$ C_7), where C_7 contains genes of $E_1 \cap E_2 \cap E_3$. C_i represents specific DEGs to certain region that could be considered as 'condition specific genes'.

Seed selection

This step is the most important part of Venn-diaNet. A user can select multiple (or a single) C_i as seeds for network propagation to measure the global influence of the seed DEGs. Thus, the results will vary depending on the selected seeds. Network propagation methods generally use informative genes as seeds. Such as 'disease-related genes', 'phenotype-related genes', or else. The idea of network propagation in Venn-diaNet is very similar but does not need to select genes that require prior knowledge. As the DEG in each region of the Venn diagram can be considered as condition-specific DEGs, the DEG in C_i can be a guide to find the similarities or dissimilarities to other C_j ($j \neq i$) that researchers are interested in. Because the selection is crucial, this study provides three possible seed selection scenarios to help to understand the seed selection.

The first scenario is to consider *'condition-specific function'* as seeds. Again, DEGs in specific region can be considered as condition-specific DEGs. If the researcher uses these genes as seeds, it can prioritize DEGs belonging to other conditions in terms of functional similarity to the seed DEGs. For example, if a user wants to prioritize tissue A-specific DEGs (Figure 4.2A: C_1) that have a similar function to the tissue B-specific DEGs when the same gene is KO, tissue B specific-DEGs (Figure 4.2A: C_2) can be used as seeds.

The second scenario is to consider *'common function'* as seeds. In some cases, a user might be interested in condition-specific DEGs that have a common function in different experiments. For instance, if the user is interested in tissue A-specific DEGs (Figure 4.2A: C_1) that have similar function between two different tissues, C_3 can be seeds. Similarly, if the common KO effect in different tissues are in interest (C_3), C_1+C_2 can be seeds.

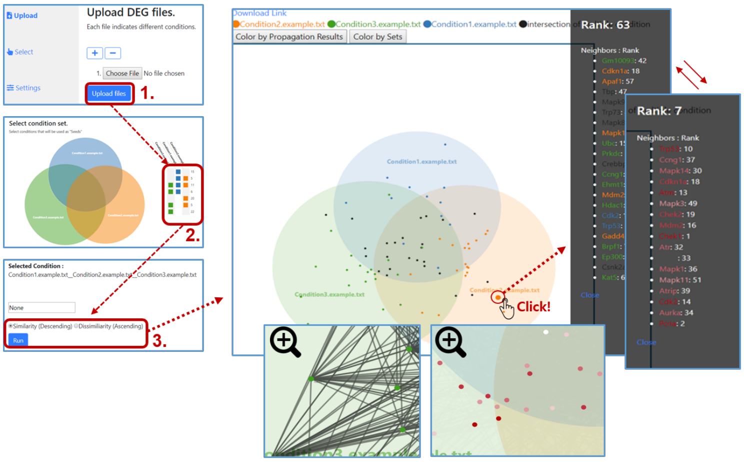
The last scenario is to consider seeds that have *'Functional similarity'*. Distinct from the two scenarios stated above, this study assumed a case that there is no sufficient knowledge to select a certain condition as seeds. In this case, a 'minimum guideline' to choose certain conditions as seeds to rank the genes of interest. If the user has multiple experiments and expects some DEGs in the condition of interest (C_i) to have functional similarity to other condition DEGs (C_j), the condition that has functional similarity to the condition of interest will be appropriate to be as seeds. This guideline is suggested to prioritize genes for experiments that study compound effects of multiple treatments which will be introduced later.

4.2.3 Network propagation and gene ranking

When a set of seed DEGs are selected, Venn-diaNet instantiates a protein-protein interaction (PPI) network of DEGs from STRING DB (Szklarczyk *et al.*, 2014). In the instantiated network, nodes are DEGs and an edge between two DEGs is defined when the corresponding edge in the original PPI network is of high-confidence (combined score $>$ 700). Then, Markov Random Walk (MRW) (Dirmeier, 2018) is performed using the seeds selected in the previous step (Figure 4.1: Step 4). The goal of network propagation is to quantify the influence of seed DEGs to the remaining DEGs. The selected seed DEGs can be considered as the hypothesis that a user wants to test. Thus, by performing a network propagation analysis, the user can obtain the DEGs pertaining to the hypothesis. For the network propagation, an R package `diffusr`, the implementation of MRW, is used. The equation is shown below:

$$p^{t+1} = (1-r)A'p^t + rp^0$$

where p^0 is the vector of initialized nodes, t is a time step, p^t is the vector at the current time step t, p^{t+1} is the vector at the next time step, A' is column-normalized matrix of adjacency matrix A, and r is the restart rate. p^0 is initialized in 1 or 0, to represent the assigned seed DEGs and target DEGs, and normalized so the sum of the elements in p^0 becomes 1. The adjacency matrix A is a matrix consists with 0 or 1 that represents a graph with no weighted edges. 0.5 is used for r and network propagation stops when $L1$ norm difference between p^t and p^{t+1} is smaller than 10^{-4}, which are the default progress of the `diffusr` package. When the algorithm stops, Venn-diaNet returns ranked gene sets based on the network propagation result.



Upload

Select

Settings

Upload DEG files.

Each file indicates different conditions.

+ −

1. Choose File No file chosen

Upload files

1.

Select condition set.

Select conditions that will be used as "Seeds"

2.

Selected Condition :
Condition1.example.txt__Condition2.example.txt__Condition3.example.txt

None

◉Similarity (Descending) ○Dissimiliarity (Ascending)

Run

3.

Download Link

●Condition2.example.txt ●Condition3.example.txt ●Condition1.example.txt ●intersection

Color by Propagation Results | Color by Sets

Rank: 63

Neighbors : Rank

- Gm10093: 42
- Cdkn1a: 18
- Apaf1: 57
- Tbp: 47
- Mapk9
- Trp73
- Mapk8
- Mapk1
- Ubc: 1
- Prkdc
- Crebbp
- Ccng1
- Ehmt1
- Mdm2
- Hdac1
- Cdk2
- Trp53
- Gadd4
- Brpf1
- Ep300
- Csnk2
- Kat5: 6

Close

Rank: 7

Neighbors : Rank

- Trp53: 10
- Ccng1: 37
- Mapk14: 30
- Cdkn1a: 18
- Atm: 13
- Mapk3: 49
- Chek2: 19
- Mdm2: 16
- Chek1: 1
- Atr: 32
- : 33
- Mapk1: 36
- Mapk11: 51
- Atrip: 39
- Cdk2: 14
- Aurka: 34
- Pcna: 2

Close

Figure 4.3: Venn-diaNet (web) work flow

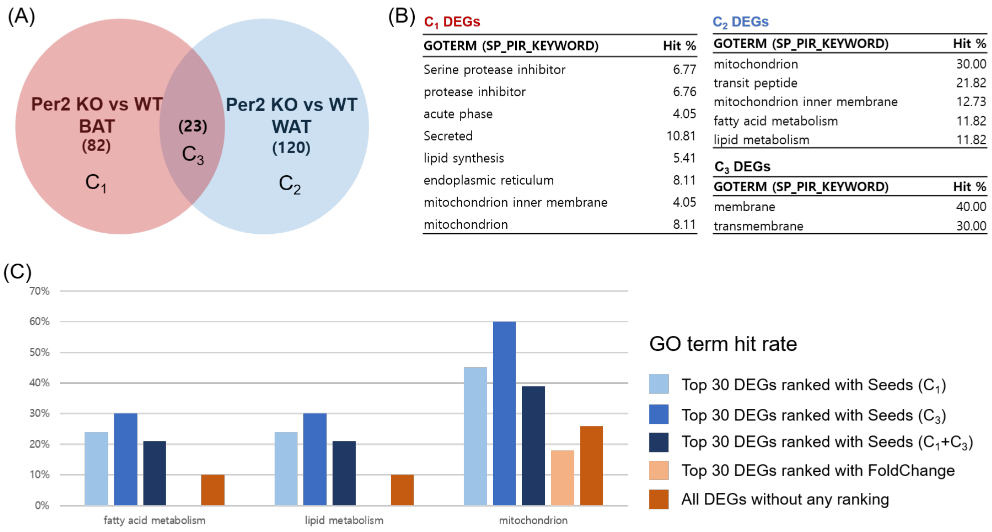
A work flow of Venn-diaNet (web). Step 1: Upload DEG lists per experiment. Step 2: Select seed condition C_i Step 3: Perform analysis. Venn-diaNet gives user (1) list of ranked genes, (2) gene's neighbor nodes information (when the node is clicked). (3) Venn diagram with PPI network (when the Venn diagram is zoomed in).

4.3 Results and Discussion

This study evaluated the performance of Venn-diaNet using three datasets downloaded from the Gene Expression Omnibus (GEO) (Edgar *et al.*, 2002) or from the supplementary data of the corresponding published paper. Three datasets were selected to show how to perform network propagation analysis with different seed gene selections.

4.3.1 Venn-diaNet for two experiments

The dataset is from a study of Per2 KO mouse with two different tissues (Grimaldi *et al.*, 2010): (*i*) Per2 KO vs WT in white adipose tissue (WAT Per2 KO), and (*ii*) Per2 KO vs WT in brown adipose tissue (BAT Per2 KO). The authors used these DEGs and reported that several WAT specific expressed



C_1 DEGs

GOTERM (SP_PIR_KEYWORD)	Hit %
Serine protease inhibitor	6.77
protease inhibitor	6.76
acute phase	4.05
Secreted	10.81
lipid synthesis	5.41
endoplasmic reticulum	8.11
mitochondrion inner membrane	4.05
mitochondrion	8.11

C_2 DEGs

GOTERM (SP_PIR_KEYWORD)	Hit %
mitochondrion	30.00
transit peptide	21.82
mitochondrion inner membrane	12.73
fatty acid metabolism	11.82
lipid metabolism	11.82

C_3 DEGs

GOTERM (SP_PIR_KEYWORD)	Hit %
membrane	40.00
transmembrane	30.00

Figure 4.4: Venn-diaNet Per2 GO term Comparison

(A) Venn-diagram of GSE20165 experiment. C_1 represents Per2 KO vs WT DEGs that is specific to BAT while C_2 represents WAT specific Per2 KO vs WT DEGs. (B) Enriched GO terms by DAVID gene functional clustering analysis. Gene functional clustering was performed for each specific condition (C_i). (C) Enriched GO terms of Top 30 genes prioritized by corresponding seeds.

Gene	FC	C_1	C_3	C_1+C_3
Ucp1	2	18	16	14
Cidea	4	26	18	25
Acsm3	47	30	39	35
Pdk4	20	71	61	74
Cpt1b	11	6	20	6
Acads	129	58	27	61
Acadm	119	14	15	11
Acadl	95	52	28	58
Acadvl	67	37	12	34
Hadha	111	5	10	3
Hadhb	54	8	13	5
Cox7a1	14	62	66	67
Cox8b	12	22	43	28
PredictedCandidates	120	100	100	100

Table 4.1: Comparing ranking results of the Per2 KO experiment performed by (Grimaldi et al., 2010)

genes have similar behavior also BAT when Per2 is KO.

Venn-diaNet used these two experiments from this study to evaluate how well Venn-diaNet could reproduce the effects of the corresponding study. For convenience, this study denoted BAT Per2 KO specific DEGs as C_1, WAT Per2 KO as C_2, and the intersection DEGs of BAT Per2 KO and WAT Per2 KO as C_3 (Figure 4.4A). Venn-diaNet used this data to show that Venn-diaNet can reproduce the authors' results by following the authors' inputs, interest, and approach. The original paper reported that Per2 KO caused BAT-specific genes to express in WAT by controlling PPARγ-dependent genes. Therefore, the aim of this study is to find promising C_2 DEGs that have the similar characteristic in BAT tissue. Three suggested seed scenarios can be used to address the authors interests. For each seed scenarios, the study compared (*i*) how the GO terms of ranked top 10% genes match the GO terms reported in the original paper, and (*ii*) how many genes match to the genes that were reported in the original paper. Note that the authors used only fold change to rank genes and did not use any gene prioritization method.

Condition specific function (C_1) & common function as seeds (C_3)

BAT Per2 KO specific DEGs (C_1), can be used as seeds in order to prioritize genes of WAT Per2 KO specific DEGs (C_2). This scenario is to investigate that some of the unknown PPARγ-dependent genes that express exclusively in BAT somehow seems to be expressed in WAT when Per2 is KO. The phenomenon indicates that there might be a functional similarity between these two different conditions. Likewise, common DEGs between two experiments (C_3) can also be considered as seeds. Activation of BAT-specific PPARγ-dependent genes in WAT also means that BAT and WAT have common functions. Thus, the common function of these genes (C_3) might be a guideline to prioritize WAT-specific genes with the context of 'functional similarity' between two different tissues. It is interesting that Venn-diaNet could prioritize genes in top 30 (about 10% of total candidates) as well as prioritizing genes that are related to the functions that the authors reported (Figure 4.4C, Table 4.1).

Analysis scenario with functional similarity as seeds (C_1)

As discussed in the previous section, researchers might encounter a situation where the user does not have sufficient knowledge to select seeds. In this case, the suggested 'minimum guideline' to choose a certain condition as seeds to rank genes in a condition of interest. For this, the study defined 'The condition that has functional similarity to the condition of interest will be appropriate to be as seeds' as a 'minimum guideline' to find seeds.

The process is very straight-forward. (*i*) Find the major GO terms of each C_i, and (*ii*) use genes in C_i if the GO terms are similar to the condition C_j ($j \neq i$) that we want to prioritize. As a result, the study found that GO term (mitochondrion) in C_1 was similar to the condition of interest (C_2) (Figure 4.4B). Thus, C_1 becomes an appropriate seed for this scenario and the results share the same which we discussed in the previous subsection.

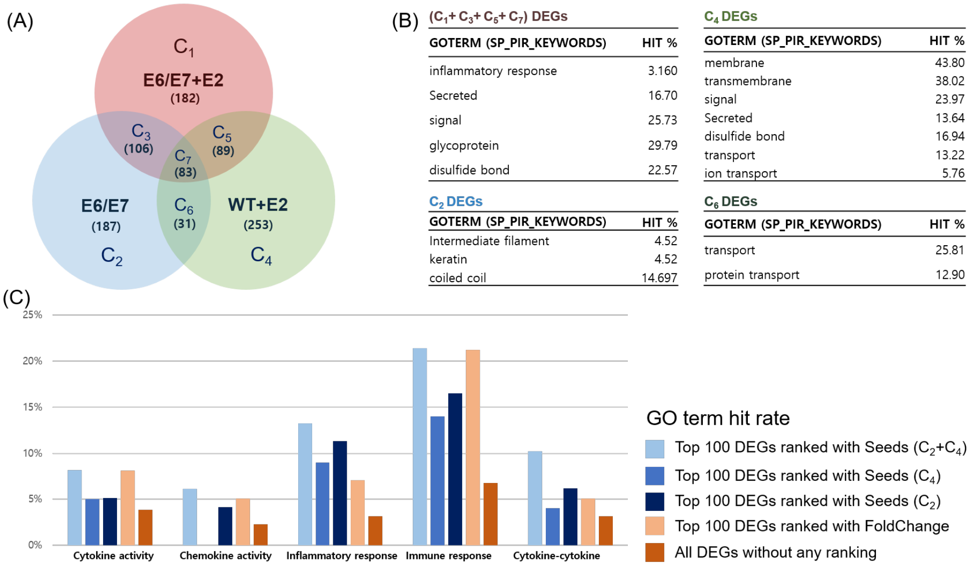
Venn-diaNet is also tested with other possible seed scenarios ($C_1 + C_3$) to confirm whether Venn-diaNet performs better than random seeds.

4.3.2 Venn-diaNet for three experiments

Data from a study of human papillomavirus oncogenes (Spurgeon *et al.*, 2017) is used for Venn-diaNet validation to consider the case of more complicated experiment designs. The study observes the independent, synergistic effects of two treatments: (*i*) K14E6/E7 bitransgenic mouse vs WT mouse (E6/E7), (*ii*) estrogen treated mouse vs WT mouse (E2), and (*iii*) K14E6/E7 bitransgenic

Gene	FC	C_2	C_4	C_6	C_2+C_4	C_2+C_6	C_4+C_6	C_2+C_4+C_6
Ccl3	284	236	27	159	78	246	35	91
Ccl6	107	15	188	171	53	24	206	62
Ccl28	257	36	220	242	94	54	240	107
Cd14	39	174	21	107	44	186	25	51
Cxcl1	12	125	131	142	144	143	148	156
Cxcl2	5	132	9	139	17	151	11	16
Cxcl3	9	139	166	202	232	161	184	248
Cxcl5	1	179	43	196	63	202	52	74
Cxcl16	268	121	139	129	159	141	156	169
Ecm1	207	238	312	282	320	258	319	324
Enpp3	346	14	230	122	71	21	241	77
Il1a	45	232	179	118	269	242	191	278
Il1b	111	117	34	20	32	114	19	21
Il1f6	213	378	378	373	385	380	380	387
Il23a	389	62	177	164	88	73	192	101
Il33	104	364	104	358	208	366	123	223
Met	211	127	60	21	52	118	27	34
Pglyrp1	16	73	341	303	179	86	347	199
Pycard	226	50	241	172	115	60	250	134
S100a8	7	248	65	121	130	257	77	143
S100a9	3	227	238	39	296	188	218	286
Spp1	99	80	189	155	118	92	205	135

Table 4.2: Comparing ranking results of the E6/E7 experiment performed by (Spurgeonet al., 2017)



(C_1+ C_3+ C_5+ C_7) DEGs

GOTERM (SP_PIR_KEYWORDS)	HIT %
inflammatory response	3.160
Secreted	16.70
signal	25.73
glycoprotein	29.79
disulfide bond	22.57

C_4 DEGs

GOTERM (SP_PIR_KEYWORDS)	HIT %
membrane	43.80
transmembrane	38.02
signal	23.97
Secreted	13.64
disulfide bond	16.94
transport	13.22
ion transport	5.76

C_2 DEGs

GOTERM (SP_PIR_KEYWORDS)	HIT %
Intermediate filament	4.52
keratin	4.52
coiled coil	14.697

C_6 DEGs

GOTERM (SP_PIR_KEYWORDS)	HIT %
transport	25.81
protein transport	12.90

Figure 4.5: Venn-diaNet HPV experiment GO term Comparison

(A) Venn-diagram of the experiment by (Spurgeonet al., 2017). C_1 C_2, and C_4 represents E6/E7+E2 specific DEGs, E6/E7 specific DEGs, and E2 specific DEGs, respectively. (B) Enriched GO terms by DAVID gene functional clustering analysis. Gene functional clustering was performed for each specific condition (C_i). (C) Enriched GO terms of Top 100 genes prioritized by corresponding seeds.

mouse treated with estrogen mouse vs WT mouse (E6/E7+E2) (Figure 4.5).

The study focused on E6E7+E2 DEGs (C_1 + C_3 + C_5 + C_7) to determine the synergistic effect of E6/E7 and E2. E6/E7 specific DEGs and E2 specific DEGs (C_2 + C_4) were selected for the seed scenario of 'condition specific function'. The seed scenario represents that the independent effect of each treatment as a guideline to find the effect of the combined factors. The goal for this experiment is to reproduce GO terms and genes that the authors reported.

Condition specific function as seeds (C_2 + C_4)

As a result, Venn-diaNet could prioritize genes and GO terms that were reported in the original paper by using the combination of independent effects of two factors as seeds (C_2 + C_4) (Figure 4.5C and 4.2). However, several careful consideration remains to be discussed. When Venn-dianet considers the prioritized top 20% genes, Venn-diaNet was not superior to the authors approach, but it could prioritize genes that are related to the GO terms where the original paper focused. In addition, Venn-diaNet could prioritize other genes that were related to the function of interest (immune response & inflammatory response) that are responsible for the HPV associated cervical cancer while the authors did not.

For example, Tlr2, a gene that is known to be related to having a significant role in HPV associated cervical cancer (Woodby *et al.*, 2016; Zom *et al.*, 2016; Halec *et al.*, 2018; Yang *et al.*, 2018), was distinctively overexpressed in E6/E7+E2. The results support that Tlr2 might be one of the significant gene that is enhanced by the combined effect of E6/E7 and E2, which achieves the condition of 'inflammatory response are increased by epithelial E6/E7 expression and further enhanced by estrogen'. The study conjectured that Tlr2 was not included in the original paper because the fold change of Tlr2 is not significant (ranked 332th in terms of fold change rankings). However, our gene prioritization analysis ranked Tlr2 much higher in the 33rd place.

Likewise, CD74 has been reported that it may play an important role in the pathogenesis and angiogenesis of cervical cancer (Cheng *et al.*, 2011) as well as the influence of the HPV (Klymenko *et al.*, 2017). Venn-diaNet placed this gene in the 76th position while fold change could only rank them as 182th. Icam1 was ranked 76th in foldchange but had the 3rd position in Venn-diaNet which also might have a E6/E7+E2 specific expression while Icam1 was also reported to have a role with HPV related cervical carcinoma (Viac *et al.*, 1992) The comparison of Top 100 ranked genes related to 'inflammatory response' & 'immune response' is summarized in 4.2.

Functional similarity as seeds (C_4)

C_4 was selected by following the 'minimum guideline' to select seeds. Unlike 'Condition specific function as seeds', seeds chosen by functional similarity performed weaker than the previous seeds. This is probably because the seed scenario does not reflect the effect of E6/E7. E6/E7 is well known to change the activity of cytokine and chemokine, and Venn-diaNet could not prioritize those genes with not considering those effects in seeds (Figure 4.5C). The study emphasized that this seed scenario reflects that using seed genes from a singular treatment is not effective to rank genes that is under the influence of multiple treatments. However, Venn-diaNet could still prioritize 7 genes in top 100 with seeds of 'functional similarity' (4.2). In addition, Venn-diaNet also tested every other possible seeds, and the results indicate other seeds are less effective than the suggested seed scenarios.

4.3.3 Venn-diaNet for eight experiments

Case 3 uses a dataset from a study that designed the experiments with three treatments and four tissues (Julien *et al.*, 2017): (*i*) narciclasine (ncls), (*ii*) high-fat diet (HFD), (*iii*) normal chow diet (NCD), (*iv*) WAT, (*v*) BAT, (*vi*) liver, and (*vii*) muscle. The initial number of sets of this study were extremely complicated that makes almost impossible to interpret the DEG list at once. Thus, the authors used a step-by-step filtering method to find promising genes for these multi-conditioned data. The authors searched the relation between treatments and tissues using hierarchical clustering and narrowed down to compare two DEG lists (HFD-ncls/HFD-veh, NCD-veh/HFD-veh) of muscle. The study reported genes that have low expression level in HFD, changes to have a high expression level when ncls was given. The results indicate that a natural compound ncls can attenuate diet-induced obesity and the associated genes can enhance the energy expenditure.

To reproduce the results what the authors made, we planned two different scenarios. The first scenario is to follow the story of the authors: using two DEG lists. The authors compared the expression profile of treatments and tissues using hierarchical clustering as a very first step. They discovered that muscle had partial mutual exclusive expression pattern to other tissues and made a hypothesis of 'ncls might accelerate genes to be expressed again while the genes were suppressed in HFD environment in muscle'. The study assumed to reached this step and use Venn-diaNet for the DEGs of HFD-ncls/HFD-veh and NCD-veh/HFD-veh. Venn-diaNet will mimic this story with the concept of `'Case 1: Venn-diaNet for two experiments'` analysis of Venn-diaNet.

Another scenario is to find promising genes purely by Venn-diaNet, using eight DEG lists. The goal of this scenario is to check whether Venn-diaNet can

		2 DEG list		8 DEG list
	FC	C_1+ C_2	C_2	C_3+C_5+C_{192}
Actc1	31	1	7	-
Tnni1	8	6	26	12
Myl2	10	23	25	9
Myh7	6	28	31	11
Tnnt1	13	5	28	10
Myl3	11	4	3	5
Tnnc1	9	20	24	15

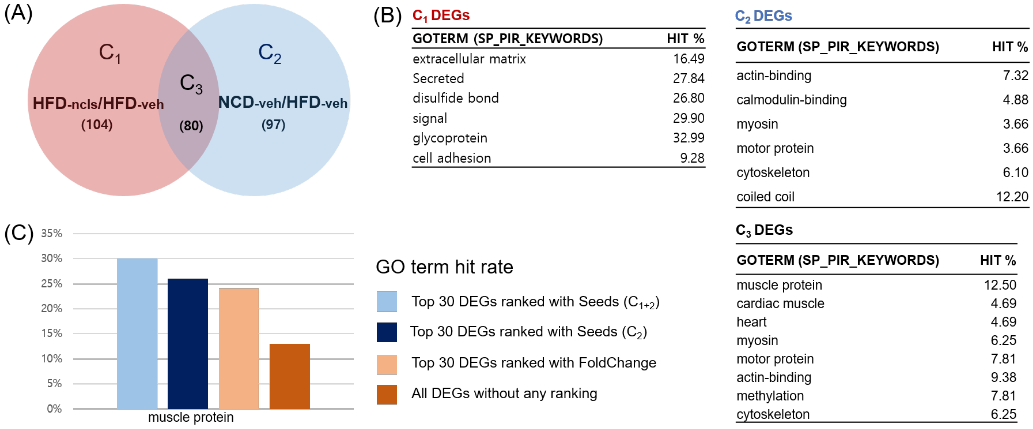
Table 4.3: Comparing ranking results of the HFD experiment performed by (Julien et al., 2017)

track down the reported genes, with a reasonable story.

Authors' approach : two DEG list

As described in the previous section, the study also assumed to performed hierarchical clustering and focus to find certain genes in C_3 (Figure 4.6A) that have the common characteristics of up-regulation when ncls is induced and up-regulated in NCD without any treatments .

In order to prioritize genes in C_3, the study used the seed scenario of `Condition specific function as seeds`. DEGs that are common in both experiments can be prioritized using the independent effects of each factor. Therefore, C_1+C_2, the independent effect of each treatment was selected as seeds to observe the influence to the genes that have same activity alteration in HFD-ncls/HFD-veh and NCD-veh/HFD-veh (C_3). The study found that VenndiaNet could prioritize and reproduce the genes where the authors reported



C_1 DEGs

GOTERM (SP_PIR_KEYWORDS)	HIT %
extracellular matrix	16.49
Secreted	27.84
disulfide bond	26.80
signal	29.90
glycoprotein	32.99
cell adhesion	9.28

C_2 DEGs

GOTERM (SP_PIR_KEYWORDS)	HIT %
actin-binding	7.32
calmodulin-binding	4.88
myosin	3.66
motor protein	3.66
cytoskeleton	6.10
coiled coil	12.20

C_3 DEGs

GOTERM (SP_PIR_KEYWORDS)	HIT %
muscle protein	12.50
cardiac muscle	4.69
heart	4.69
myosin	6.25
motor protein	7.81
actin-binding	9.38
methylation	7.81
cytoskeleton	6.25

Figure 4.6: Venn-diaNet HFD GO term Comparison

(A) Venn-diagram of GSE63268 experiment. C_1 represents HFD (ncls/veh) specific DEGs while C_2 shows veh (NCD/HFD) specific DEGs. (B) Enriched GO terms by DAVID gene functional clustering analysis. Gene functional clustering was performed for each specific region. (C) Enriched GO terms of Top 100 genes prioritized by corresponding seeds

(Table 4.3) as well as prioritizing GO terms of the authors' interest with better hit ratio (Figure 4.6C). The minimum guideline, 'Functional similarity as seeds' (C_2) showed weaker gene prioritization but still had a better focus on GO terms (Figure 4.6C and Table 4.3). In addition, this study is designed to find the common effect from independent conditions, meaning that the condition of interest is closely related to each other condition. Therefore, it is natural to have poor performance with the same reason that is discussed in the previous section.

Venn-diaNet approach: All (eight) DEG list

The study assumed that the researcher does not have enough knowledge of the corresponding data, and try whether Venn-diaNet could reach to the authors' conclusion. Venn-diaNet simply with all DEG lists (that contains up and down-regulation) from eight different experiments at once (Figure 4.7A). The Venn

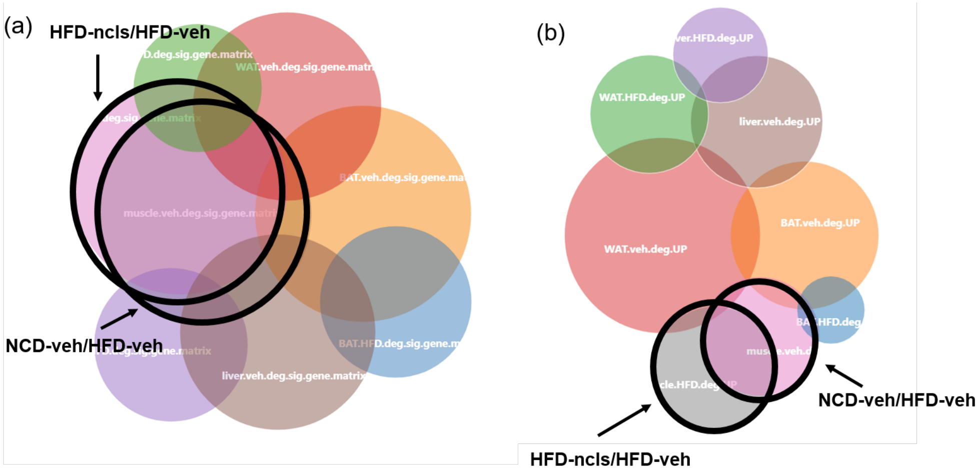


Figure 4.7: Venn-diaNet using 8 different DEG list

(A) Using up and down-regulated DEG list to Venn-diaNet (web). The Venn diagram directly shows muscle DEGs in HFD-ncls/HFD-veh C_48, and NCD-veh/HFD-veh are similar to each other while other tissues are not similar to each other. (B) Using up-regulated DEG list to Venn-diaNet. The Venn diagram shows that up-regulated muscle DEGs in HFD-ncls/HFD-veh, and NCD-veh/HFD-veh are very similar to each other while other tissues are not similar to each other.

diagram shows that the intersection of HFD-ncls/HFD-veh and NCD-veh/HFD-veh shared many DEGs in muscle (C_{48}) than any other tissues (C_3, C_{12}, C_{192}).

The findings of Venn diagram reaffirms the authors' hierarchical clustering results and leads to the idea that the intersection of HFD-ncls/HFD-veh and NCD-veh/HFD-veh in muscle have common functions than other tissues, and needs to be analyzed in detail. To start the detailed search, up-regulated DEG list is used to examine whether Venn-diaNet can answer for the hypothesis of 'ncls might accelerate genes to be expressed again while the genes were suppressed in HFD environment in muscle'. As a result, This study discovered that the condition of interest was much more distinct to other conditions (Figure 4.7B: C_{48}) and the portion of common genes between HFD-ncls/HFD-veh and NCD-veh/HFD-veh in muscle was bigger than any other tissue (C_{48}, C_3, C_{12}, C_{192}). The findings indicate that up-regulation of C_{48} is likely to be more specific and distinct to other tissues. To prioritize genes in C_{48}, 'common functions as seeds' is chosen for the seed scenario. This study selected the intersection of HFD-ncls/HFD-veh and NCD-veh/HFD-veh of other tissues as seeds (C_3, C_{12}, C_{192}) to represent that the function of 'ncls might accelerate genes to be expressed again while the genes were suppressed in HFD environment' in other tissues can assist to prioritize genes in muscle. As a result, this study was able to reproduce genes that the authors reported in their original paper (Table 4.3).

In addition to seed selection, the minimum guideline cannot be used for this complex condition data. The data is composed of 255 conditions that make it difficult to compare and analyze the GO terms of all these conditions.

4.3.4 Venn-diaNet performance with different PPI network

Currently, there are multiple databases that contain PPI information while Venn-diaNet performed network propagation to the network topology of STRING

Gene	FC	C_1	C_3	C_1+C_3
Ucp1	2	-	-	-
Cidea	4	7	17	8
Acsm3	47	-	-	-
Pdk4	20	28	8	30
Cpt1b	11	51	49	52
Acads	129	60	59	60
Acadm	119	22	51	24
Acadl	95	53	52	54
Acadvl	67	56	56	57
Hadha	111	19	44	21
Hadhb	54	21	46	23
Cox7a1	14	-	-	-
Cox8b	12	-	-	-
PredictedCandidates	120	62	62	62

Table 4.4: Comparing ranking results of the Per2 KO experiment performed by (Grimaldi et al., 2010) using PPI network from BioGRID

Gene	FC	C_2	C_4	C_6	C_2+C_4	C_2+C_6	C_4+C_6	C_2+C_4+C_6
Ccl3	284	124	122	103	145	126	124	147
Ccl6	107	-	-	-	-	-	-	-
Ccl28	257	-	-	-	-	-	-	-
Cd14	39	38	165	154	78	41	166	82
Cxcl1	12	-	-	-	-	-	-	-
Cxcl2	5	139	138	123	156	141	140	158
Cxcl3	9	-	-	-	-	-	-	-
Cxcl5	1	-	-	-	-	-	-	-
Cxcl16	268	-	-	-	-	-	-	-
Ecm1	207	-	-	-	-	-	-	-
Enpp3	346	-	-	-	-	-	-	-
Il1a	45	28	134	118	62	30	136	65
Il1b	111	47	72	28	99	53	76	104
Il1f6	213	-	-	-	-	-	-	-
Il23a	389	-	-	-	-	-	-	-
Il33	104	172	172	165	180	173	173	181
Met	211	72	9	36	19	79	11	19
Pglyrp1	16	-	-	-	-		-	-
Pycard	226	176	176	169	184	177	177	185
S100a8	7	123	121	102	144	125	123	146
S100a9	3	-	-	-	-	-	-	-
Spp1	99	127	125	106	148	129	127	150

Table 4.5: Comparing ranking results of the E6/E7 experiment performed by (Spurgeonet al., 2017) using PPI network from BioGRID

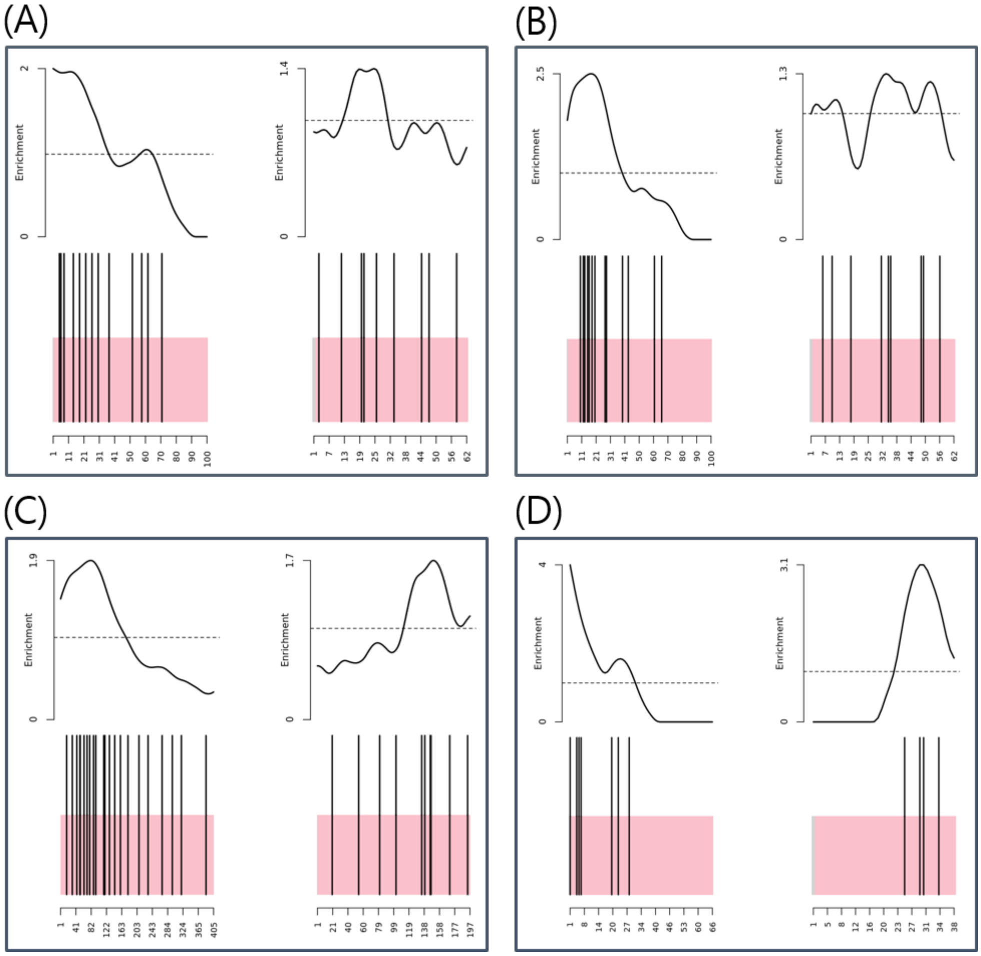


Figure 4.8: Venn-diaNet performance comparison between STRING and BioGrid

A barcode plot to distinguish the prioritized results. (A) Per2 KO experiment with Seed C_1. left; SRINGDB, right; BioGrid. (B) Per2 KO experiment with Seed C_C. left; SRINGDB, right; BioGrid. (C) E6/E7 experiment with Seed C_2+C_4. left; SRINGDB, right; BioGrid. (B) HFD experiment with Seed C_1+C_2. left; SRINGDB, right; BioGrid.

		2 DEG list		8 DEG list
	FC	C_1+ C_2	C_2	C_3+C_5+C_{192}
Actc1	31	36	35	-
Tnni1	8	-	-	-
Myl2	10	-	-	-
Myh7	6	11	7	13
Tnnt1	13	-	-	-
Myl3	11	33	26	20
Tnnc1	9	35	34	27

Table 4.6: Comparing ranking results of the HFD experiment performed by (Julien et al., 2017) using PPI network from BioGRID

DB (combined score > 0.4). We considered that the Venn-diaNet might have different gene prioritization results with different network topology and the network propagation from STRING DB could contain a number of false positives due to the nature of the STRING DB (several PPI information are not biologically validated). Therefore it is important to compare whether the results varies by using the different network that contains higher biological evidence. We additionally performed the same process and compared the ranks between results between STRING to BioGRID (Stark *et al.*, 2006). As a result, we confirmed that network propagation using network from STRING DB was overall more effective than the network from the BioGRID that can prioritize the reported genes in higher ranks (Table 4.4, 4.5 and 4.6)(Figure 4.8).

4.4 Discussion

This study presented Venn-diaNet, a web-based software that does not require any additional installment or registration. Venn-diaNet draws a Venn diagram from a given input and prioritizes genes by network propagation. This study suggested that a Venn diagram can support selecting seeds for network propagation and introduced several examples to show the idea can effectively prioritize genes that are related to the function of interests. Venn-diaNet is designed not only to avoid the 'black-box' issue in gene prioritization which is caused by the integration of heterogeneous databases but also to address a logical approach for seed selection of network propagation. Venn-diaNet supports gene list with ranking and additional features that explains how the specific gene is influential to other genes. Venn-diaNet is available at: biohealth.snu.ac.kr/software/venndianet.

Chapter 5

Conclusion

Identifying promising genes from a large pool of candidates that represent the phenotypic differences, is one of the common goals during RNA-Seq analysis. Although, a number of studies have demonstrated various approaches for prioritizing genes, the challenges in deciding the most appropriate combination of analysis methods for a complicated experimental design, as well as the difficulties in handling strategies that rank irrelevant genes, still persist. This thesis summarized three studies to address these difficulties:

1. A filtering strategy that combines DEG, GRN, pathways, and SNVs to handle the statistical bias caused by a small number of samples in mouse gene KO data.

2. A data fusion strategy that combines text-mining and PPI network to rank genes filtered by DEG, GRN, and SNVs to focus on the context of the experiment design.

3. A network strategy that uses Venn diagram to have an advantage in seed

selection and interpretation, and perform network propagation to rank candidate genes.

The first study analyzes RNA-Seq data by (i) removing the less informative DEGs using GRN, biological pathways, and (ii) filtering out genes that differ among samples on the basis of SNVs. This study demonstrated that the integration of multiple filters enabled better gene prioritization and refined the candidates by eliminating the genetic differences among samples. The second study developed an informatics system to avoid ranking irrelevant genes by allowing the user to specify the context of the experiment. CLIP-GENE prioritizes genes of KO experiments by (i) removing the less informative DEGs using GRN, (ii) discarding genes that vary among samples on the basis of SNVs, and (iii) ranking genes that are related to the user's context using text-mining technique, as well as considering the shortest path of PPI to the KO gene. The last study addressed the seed selection issue by integrating Venn diagrams into the network-based strategy. The study developed an informative gene prioritization system that can compare multiple biological experiments in Venn diagram and select seed genes that are free from the pressure of prior knowledge. The study demonstrated that Venn-diaNet was able to reproduce the findings of the original papers that have reported complicated experiments with an intuitive interpretation.

In conclusion, this thesis summarizes the difficulties of RNA-Seq analysis methods and has created three different informatics systems that combine network approaches with other methods to prioritize phenotype-specific genes from RNA-Seq data. For each approach, we have developed software packages and web tools for researchers to have convenient access to the methods, and hope that these methods will provide a good starting point for RNA-seq analysis.

Bibliography

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., and Pickard, B. S. (2005). Speeding disease gene discovery by sequence based candidate prioritization. *BMC bioinformatics*, **6**(1), 55.

Alpern, D., Langer, D., Ballester, B., Le Gras, S., Romier, C., Mengus, G., and Davidson, I. (2014). Taf4, a subunit of transcription factor ii d, directs promoter occupancy of nuclear receptor hnf4a during post-natal hepatocyte differentiation. *Elife*, **3**, e03613.

Altboum, Z., Steuerman, Y., David, E., Barnett-Itzhaki, Z., Valadarsky, L., Keren-Shaul, H., Meningher, T., Mendelson, E., Mandelboim, M., Gat-Viks, I., *et al.* (2014). Digital cell quantification identifies global immune cell dynamics during influenza infection. *Molecular systems biology*, **10**(2), 720.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, **11**(10), R106.

Anders, S. and Huber, W. (2012). Differential expression of rna-seq data at the gene level–the deseq package. *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL).*

Bae, Y., Yang, T., Zeng, H.-C., Campeau, P. M., Chen, Y., Bertin, T., Dawson, B. C., Munivez, E., Tao, J., and Lee, B. H. (2012). mirna-34c regulates notch signaling during bone development. *Human molecular genetics*, **21**(13), 2991–3000.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human b cells. *Nature genetics*, **37**(4), 382.

Bhatnagar, S., Zhu, X., Ou, J., Lin, L., Chamberlain, L., Zhu, L. J., Wajapeyee, N., and Green, M. R. (2014). Genetic and pharmacological reactivation of the mammalian inactive x chromosome. *Proceedings of the National Academy of Sciences*, **111**(35), 12591–12598.

Chen, H. and Boutros, P. C. (2011). Venndiagram: a package for the generation of highly-customizable venn and euler diagrams in r. *BMC bioinformatics*, **12**(1), 35.

Chen, J., Aronow, B. J., and Jegga, A. G. (2009a). Disease candidate gene identification and prioritization using protein interaction networks. *BMC bioinformatics*, **10**(1), 73.

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009b). Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, **37**(suppl_2), W305–W311.

Chen, J. Y., Shen, C., and Sivachenko, A. Y. (2006). Mining alzheimer disease relevant proteins from integrated protein interactome data. In *Biocomputing 2006*, pages 367–378. World Scientific.

Cheng, R.-j., Deng, W.-g., Niu, C.-b., Li, Y.-y., and Fu, Y. (2011). Expression of macrophage migration inhibitory factor and cd74 in cervical squamous cell carcinoma. *International Journal of Gynecological Cancer*, **21**(6), 1004–1012.

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, **18**(9), 551.

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**(6167), 193–196.

Dirmeier, S. (2018). *diffusr: Network Diffusion Algorithms*. R package version 0.1.4.

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing illumina microarray. *Bioinformatics*, **24**(13), 1547–1548.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, **30**(1), 207–210.

Eisener-Dorman, A. F., Lawrence, D. A., and Bolivar, V. J. (2009). Cautionary insights on knockout mouse studies: The gene or not the gene? *Brain, Behavior, and Immunity*, **23**(3), 318 – 324.

Feng, C., Ding, G., Jiang, H., Ding, Q., and Wen, H. (2015). Loss of mlh1 confers resistance to pi3kβ inhibitors in renal clear cell carcinoma with setd2 mutation. *Tumor Biology*, **36**(5), 3457–3464.

Frazee, A. C., Sabunciyan, S., Hansen, K. D., Irizarry, R. A., and Leek, J. T. (2014). Differential expression analysis of rna-seq data at single-base resolution. *Biostatistics*, **15**(3), 413–426.

Geier, F., Timmer, J., and Fleck, C. (2007). Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC systems biology*, **1**(1), 11.

Grazia, A. D. (1953). Mathematical derivation of an election system. *Isis*, **44**(1/2), 42–51.

Grimaldi, B., Bellet, M. M., Katada, S., Astarita, G., Hirayama, J., Amin, R. H., Granneman, J. G., Piomelli, D., Leff, T., and Sassone-Corsi, P. (2010). Per2 controls lipid metabolism by direct regulation of pparγ. *Cell metabolism*, **12**(5), 509–520.

Gu, S., Zhang, Y., Jin, L., Huang, Y., Zhang, F., Bassik, M. C., Kampmann, M., and Kay, M. A. (2014). Weak base pairing in both seed and $3'$ regions reduces rnai off-targets and enhances si/shrna designs. *Nucleic acids research*, **42**(19), 12169–12176.

Halec, G., Scott, M. E., Farhat, S., Darragh, T. M., and Moscicki, A.-B. (2018). Toll-like receptors: Important immune checkpoints in the regression of cervical intra-epithelial neoplasia 2. *International journal of cancer*, **143**(11), 2884–2891.

Hardcastle, T. J. and Kelly, K. A. (2010). bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, **11**(1), 422.

Harmacek, L., Watkins-Chow, D. E., Chen, J., Jones, K. L., Pavan, W. J., Salbaum, J. M., and Niswander, L. (2014). A unique missense allele of baf155, a core baf chromatin remodeling complex protein, causes neural tube closure defects in mice. *Developmental neurobiology*, **74**(5), 483–497.

Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., and Minghim, R. (2015). Interactivenn: a web-based tool for the analysis of sets through venn diagrams. *BMC bioinformatics*, **16**(1), 169.

Huang, H.-C., Niu, Y., and Qin, L.-X. (2015). Differential expression analysis for rna-seq: an overview of statistical methods and computational software: supplementary issue: sequencing platform modeling and analysis. *Cancer informatics*, **14**, CIN–S21631.

Hulsen, T., de Vlieg, J., and Alkema, W. (2008). Biovenn–a web application for the comparison and visualization of biological lists using area-proportional venn diagrams. *BMC genomics*, **9**(1), 488.

Hur, B., Chae, H., and Kim, S. (2015). Combined analysis of gene regulatory network and snv information enhances identification of potential gene markers in mouse knockout studies with small number of samples. *BMC medical genomics*, **8**(2), S10.

Hur, B., Lim, S., Chae, H., Seo, S., Lee, S., Kang, J., and Kim, S. (2016). 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.

Jeggari, A., Alekseenko, Z., Petrov, I., Dias, J. M., Ericson, J., and Alexeyenko, A. (2018). Evinet: a web platform for network enrichment analysis with flexible definition of gene sets. *Nucleic acids research*, **46**(W1), W163–W170.

Julien, S. G., Kim, S.-Y., Brunmeir, R., Sinnakannu, J. R., Ge, X., Li, H., Ma, W., Yaligar, J., KN, B. P., Velan, S. S., *et al.* (2017). Narciclasine attenuates diet-induced obesity by promoting oxidative metabolism in skeletal muscle. *PLoS biology*, **15**(2), e1002597.

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

Kayo, H., Kiga, K., Fukuda-Yuzawa, Y., Hedlund, S., Murakami, K., De La Rosa-Velazquez, I. A., Kimura, T., Shimoda, K., Tanabe, M., and Fukao, T. (2014). mir-212 and mir-132 are dispensable for mouse mammary gland development. *Nature genetics*, **46**(8), 802.

Kestler, H. A., Müller, A., Gress, T. M., and Buchholz, M. (2004). Generalized venn diagrams: a new method of visualizing complex genetic set relations. *Bioinformatics*, **21**(8), 1592–1595.

Kestler, H. A., Müller, A., Kraus, J. M., Buchholz, M., Gress, T. M., Liu, H., Kane, D. W., Zeeberg, B. R., and Weinstein, J. N. (2008). Vennmaster: area-proportional euler diagrams for functional go analysis of microarrays. *BMC bioinformatics*, **9**(1), 67.

Klymenko, T., Gu, Q., Herbert, I., Stevenson, A., Iliev, V., Watkins, G., Pollock, C., Bhatia, R., Cuschieri, K., Herzyk, P., *et al.* (2017). Rnaseq analysis of differentiated keratinocytes reveals a massive response to late events during human papillomavirus type 16 infection, including loss of epithelial barrier function. *Journal of virology*, pages JVI–01001.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, **82**(4), 949–958.

Krallinger, M., Valencia, A., and Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome biology*, **9**(2), S8.

Langmead, B., Hansen, K. D., and Leek, J. T. (2010). Cloud-scale rna-sequencing differential expression analysis with myrna. *Genome biology*, **11**(8), R83.

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., and Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research*, pages gr–118992.

Lee, S., Kim, D., Lee, K., Choi, J., Kim, S., Jeon, M., Lim, S., Choi, D., Kim, S., Tan, A.-C., and Kang, J. (2016). Best: Next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PLoS ONE*, **11**(10), e0164680.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M., and Kendziorski, C. (2013). Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, **29**(8), 1035–1043.

Li, J. and Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in rna-seq data. *Statistical methods in medical research*, **22**(5), 519–536.

Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, **13**(3), 523–538.

Li, W. V. and Li, J. J. (2018). Modeling and analysis of rna-seq data: a review from a statistical perspective. *Quantitative Biology*, **6**(3), 195–209.

Li, Y. and Patra, J. C. (2010). Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, **26**(9), 1219–1224.

Liu, G. J., Cimmino, L., Jude, J. G., Hu, Y., Witkowski, M. T., McKenzie, M. D., Kartal-Kaess, M., Best, S. A., Tuohey, L., Liao, Y., *et al.* (2014). Pax5 loss imposes a reversible differentiation block in b-progenitor acute lymphoblastic leukemia. *Genes & development*, **28**(12), 1337–1350.

López-Bigas, N. and Ouzounis, C. A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic acids research*, **32**(10), 3108–3114.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, **15**(12), 550.

Madhamshettiwar, P. B., Maetschke, S. R., Davis, M. J., Reverter, A., and Ragan, M. A. (2012). Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome medicine*, **4**(5), 41.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, **18**(9), 1509–1517.

Martin, B., Chadwick, W., Yi, T., Park, S.-S., Lu, D., Ni, B., Gadkaree, S., Farhang, K., Becker, K. G., and Maudsley, S. (2012). Vennture–a novel venn diagram investigational tool for multiple pharmacological dataset analysis. *Plos one*, **7**(5), e36911.

Meech, R., Edelman, D. B., Jones, F. S., and Makarenkova, H. P. (2005). The homeobox transcription factor barx2 regulates chondrogenesis during limb development. *Development*, **132**(9), 2135–2146.

Meech, R., Gonzalez, K. N., Barro, M., Gromova, A., Zhuang, L., Hulin, J.-A., and Makarenkova, H. P. (2012). Barx2 is expressed in satellite cells and is required for normal muscle growth and regeneration. *Stem cells*, **30**(2), 253–265.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P. D. (2016). Panther version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic acids research*, **45**(D1), D183–D189.

Mielcarek, M., Inuabasi, L., Bondulich, M. K., Muller, T., Osborne, G. F., Franklin, S. A., Smith, D. L., Neueder, A., Rosinski, J., Rattray, I., *et al.* (2014). Dysfunction of the cns-heart axis in mouse models of huntington's disease. *PLoS genetics*, **10**(8), e1004550.

Moniot, B., Ujjan, S., Champagne, J., Hirai, H., Aritake, K., Nagata, K., Dubois, E., Nidelet, S., Nakamura, M., Urade, Y., *et al.* (2014). Prostaglandin d2 acts through the dp2 receptor to influence male germ cell differentiation in the foetal mouse testis. *Development*, **141**(18), 3561–3571.

Moreau, Y. and Tranchevent, L.-C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, **13**(8), 523.

Nitsch, D., Tranchevent, L.-C., Goncalves, J. P., Vogt, J. K., Madeira, S. C., and Moreau, Y. (2011). Pinta: a web server for network-based gene prioritization from expression data. *Nucleic acids research*, **39**(suppl_2), W334–W338.

Ntziachristos, P., Tsirigos, A., Welstead, G. G., Trimarchi, T., Bakogianni, S., Xu, L., Loizou, E., Holmfeldt, L., Strikoudis, A., King, B., *et al.* (2014). Contrasting roles of histone 3 lysine 27 demethylases in acute lymphoblastic leukaemia. *Nature*, **514**(7523), 513.

Nusinow, D. P., Kiezun, A., O'Connell, D. J., Chick, J. M., Yue, Y., Maas, R. L., Gygi, S. P., and Sunyaev, S. R. (2012). Network-based inference from complex proteomic mixtures using snipe. *Bioinformatics*, **28**(23), 3115–3122.

Olguin, H. C. and Olwin, B. B. (2004). Pax-7 up-regulation inhibits myogenesis and cell cycle progression in satellite cells: a potential mechanism for self-renewal. *Developmental Biology*, **275**(2), 375 – 388.

Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From rna-seq reads to differential expression results. *Genome biology*, **11**(12), 220.

Ozsolak, F. and Milos, P. M. (2011). Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, **12**(2), 87.

Ramsey, M. R., Wilson, C., Ory, B., Rothenberg, S. M., Faquin, W., Mills, A. A., and Ellisen, L. W. (2013). Fgfr2 signaling underlies p63 oncogenic function in squamous cell carcinoma. *The Journal of clinical investigation*, **123**(8), 3525–3538.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Roger, J. E., Hiriyanna, A., Gotoh, N., Hao, H., Cheng, D. F., Ratnapriya, R., Kautzmann, M.-A. I., Chang, B., and Swaroop, A. (2014). Otx2 loss causes rod differentiation defect in crx-associated congenital blindness. *The Journal of clinical investigation*, **124**(2), 631–643.

Ruan, M. Z., Erez, A., Guse, K., Dawson, B., Bertin, T., Chen, Y., Jiang, M.-M., Yustein, J., Gannon, F., and Lee, B. H. (2013). Proteoglycan 4 expression protects against the development of osteoarthritis. *Science translational medicine*, **5**(176), 176ra34–176ra34.

Shan, M., Yuan, X., Song, L.-z., Roberts, L., Zarinkamar, N., Seryshev, A., Zhang, Y., Hilsenbeck, S., Chang, S.-H., Dong, C., *et al.* (2012). Cigarette smoke induction of osteopontin (spp1) mediates th17 inflammation in human and experimental emphysema. *Science translational medicine*, **4**(117), 117ra9–117ra9.

Shen, L., Inoue, A., He, J., Liu, Y., Lu, F., and Zhang, Y. (2014). Tet3 and dna replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. *Cell stem cell*, **15**(4), 459–471.

Smedley, D., Köhler, S., Czeschik, J. C., Amberger, J., Bocchini, C., Hamosh, A., Veldboer, J., Zemojtel, T., and Robinson, P. N. (2014). Walking the interactome for candidate prioritization in exome sequencing studies of mendelian diseases. *Bioinformatics*, **30**(22), 3215–3222.

Spampinato, C., Giordano, D., and Faro, A. (2011). Combining literature text mining with microarray data: advances for system biology modeling. *Briefings in Bioinformatics*, **13**(1), 61–82.

Spurgeon, M. E., den Boon, J. A., Horswill, M., Barthakur, S., Forouzan, O., Rader, J. S., Beebe, D. J., Roopra, A., Ahlquist, P., and Lambert, P. F. (2017). Human papillomavirus oncogenes reprogram the cervical cancer microenvironment independently of and synergistically with estrogen. *Proceedings of the National Academy of Sciences*, page 201712018.

Srivastava, J., Siddiq, A., Gredler, R., Shen, X.-N., Rajasekaran, D., Robertson, C. L., Subler, M. A., Windle, J. J., Dumur, C. I., Mukhopadhyay, N. D., *et al.* (2015). Astrocyte elevated gene-1 and c-myc cooperate to promote hepatocarcinogenesis in mice. *Hepatology*, **61**(3), 915–929.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic acids research*, **34**(suppl_1), D535–D539.

Stilling, R. M., Rönicke, R., Benito, E., Urbanke, H., Capece, V., Burkhardt, S., Bahari-Javan, S., Barth, J., Sananbenesi, F., Schütz, A. L., *et al.* (2014). K-lysine acetyltransferase 2a regulates a hippocampal gene expression network linked to memory formation. *The EMBO journal*, **33**(17), 1912–1927.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., *et al.* (2010). The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, **39**(suppl_1), D561–D568.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., *et al.* (2014). String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, **43**(D1), D447–D452.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in rna-seq: a matter of depth. *Genome research*, **21**(12), 2213–2223.

Tarazona, S., Furió-Tarí, P., Turra, D., Pietro, A. D., Nueda, M. J., Ferrer, A., and Conesa, A. (2015). Data quality aware analysis of differential expression in rna-seq with noiseq r/bioc package. *Nucleic acids research*, **43**(21), e140–e140.

Tena, J. J., González-Aguilera, C., Fernández-Miñán, A., Vázquez-Marín, J., Parra-Acero, H., Cross, J. W., Rigby, P. W., Carvajal, J. J., Wittbrodt, J., Gómez-Skarmeta, J. L., *et al.* (2014). Comparative epigenomics in distantly related teleost species identifies conserved cis-regulatory nodes active during the vertebrate phylotypic period. *Genome research*, **24**(7), 1075–1085.

Tranchevent, L.-C., Barriot, R., Yu, S., Van Vooren, S., Van Loo, P., Coessens, B., De Moor, B., Aerts, S., and Moreau, Y. (2008). Endeavour update: a web resource for gene prioritization in multiple species. *Nucleic acids research*, **36**(suppl_2), W377–W384.

Tranchevent, L.-C., Capdevila, F. B., Nitsch, D., De Moor, B., De Causmaecker, P., and Moreau, Y. (2010). A guide to web tools to prioritize candidate genes. *Briefings in bioinformatics*, **12**(1), 22–32.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, **31**(1), 46.

Tsau, C., Ito, M., Gromova, A., Hoffman, M. P., Meech, R., and Makarenkova, H. P. (2011). Barx2 and fgf10 regulate ocular glands branching morphogenesis by controlling extracellular matrix remodeling. *Development*, **138**(15), 3307–3317.

Ud-Dean, S. M. and Gunawan, R. (2015). Optimal design of gene knockout experiments for gene regulatory network inference. *Bioinformatics*, **32**(6), 875–883.

van Dam, S., Cordeiro, R., Craig, T., van Dam, J., Wood, S. H., and de Magalhães, J. P. (2012). Genefriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC genomics*, **13**(1), 535.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, **6**(1), e1000641.

Viac, J., Chardonnet, Y., Euvrard, S., and Schmitt, D. (1992). Epidermotropism of t cells correlates with intercellular adhesion molecule (icami) expression in human papillomavirus (hpv)-induced lesions. *The Journal of pathology*, **168**(3), 301–306.

Vreugdenhil, E., van Ommen, G.-J. B., Thygesen, H. H., den Dunnen, J. T., Boer, J. M., de Menezes, R. X., Vossen, R. H. A. M., Ariyurek, Y., and 't Hoen, P. A. C. (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research*, **36**(21), e141–e141.

Wan, Y. Y. (2014). Gata3: a master of many trades in immune regulation. *Trends in Immunology*, **35**(6), 233 – 242.

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2009). Degseq: an r package for identifying differentially expressed genes from rna-seq data. *Bioinformatics*, **26**(1), 136–138.

Wang, L., Wang, X., Arkin, A. P., and Samoilov, M. S. (2012). Inference of gene regulatory networks from genome-wide knockout fitness data. *Bioinformatics*, **29**(3), 338–346.

Wang, Y., Thilmony, R., and Gu, Y. Q. (2014). Netvenn: an integrated network analysis web platform for gene lists. *Nucleic acids research*, **42**(W1), W161–W166.

Winnenburg, R., Wächter, T., Plake, C., Doms, A., and Schroeder, M. (2008). Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Briefings in bioinformatics*, **9**(6), 466–478.

Woodby, B., Scott, M., and Bodily, J. (2016). The interaction between human papillomaviruses and the stromal microenvironment. In *Progress in molecular biology and translational science*, volume 144, pages 169–238. Elsevier.

Yagi, R., Zhong, C., Northrup, D. L., Yu, F., Bouladoux, N., Spencer, S., Hu, G., Barron, L., Sharma, S., Nakayama, T., *et al.* (2014). The transcription factor gata3 is critical for the development of all il-7rα-expressing innate lymphoid cells. *Immunity*, **40**(3), 378–388.

Yang, S., Liu, L., Xu, D., and Li, X. (2018). The relationship of the tlr9 and tlr2 genetic polymorphisms with cervical cancer risk: a meta-analysis of case-control studies. *Pathology & Oncology Research*, pages 1–9.

Yao, H., Goldman, D. C., Nechiporuk, T., Kawane, S., McWeeney, S. K., Tyner, J. W., Fan, G., Kerenyi, M. A., Orkin, S. H., Fleming, W. H., *et al.* (2014). Corepressor rcor1 is essential for murine erythropoiesis. *Blood*, **123**(20), 3175–3184.

Yun, B., Anderegg, A., Menichella, D., Wrabetz, L., Feltri, M. L., and Awatramani, R. (2010). Microrna-deficient schwann cells display congenital hypomyelination. *Journal of Neuroscience*, **30**(22), 7722–7728.

Zammit, P. S., Golding, J. P., Nagata, Y., Hudon, V., Partridge, T. A., and Beauchamp, J. R. (2004). Muscle satellite cells adopt divergent fates. *The Journal of Cell Biology*, **166**(3), 347–357.

Zhang, H.-M., Chen, H., Liu, W., Liu, H., Gong, J., Wang, H., and Guo, A.-Y. (2011). Animaltfdb: a comprehensive animal transcription factor database. *Nucleic acids research*, **40**(D1), D144–D149.

Zhang, X., Liu, K., Liu, Z.-P., Duval, B., Richer, J.-M., Zhao, X.-M., Hao, J.-K., and Chen, L. (2012). Narromi: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics*, **29**(1), 106–113.

Zhang, Y., Xie, S., Zhou, Y., Xie, Y., Liu, P., Sun, M., Xiao, H., Jin, Y., Sun, X., Chen, Z., Huang, Q., and Chen, S. (2014). H3k36 histone methyltransferase setd2 is required for murine embryonic stem cell differentiation toward endoderm. *Cell Reports*, **8**(6), 1989 – 2002.

Zhou, J., White, K. P., and Liu, Y. (2013). RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, **30**(3), 301–304.

Zhuang, L., Hulin, J.-A., Gromova, A., Nguyen, T. D. T., Ruth, T. Y., Liddle, C., Downes, M., Evans, R. M., Makarenkova, H. P., and Meech, R. (2014). Barx2 and pax7 have antagonistic functions in regulation of wnt signaling and satellite cell differentiation. *Stem cells*, **32**(6), 1661–1673.

Zom, G. G., Welters, M. J., Loof, N. M., Goedemans, R., Lougheed, S., Valentijn, R. R., Zandvliet, M. L., Meeuwenoord, N. J., Melief, C. J., de Gruijl, T. D., *et al.* (2016). Tlr2 ligand-synthetic long peptide conjugates effectively stimulate tumor-draining lymph node t cells of cervical cancer patients. *Oncotarget*, **7**(41), 67087.

Zuo, Y., Yu, G., Tadesse, M. G., and Ressom, H. W. (2014). Biological network inference using low order partial correlation. *Methods*, **69**(3), 266–273.

초록

RNA-seq 기술은 게놈 규모의 전사체를 고해상도로 분석 가능하게 만들었으나, 일반적으로 전사체 데이터에서 나타나는 유전자의 수는 많기 때문에 추가 분석 없이 연구 목표와 관련된 유전자를 식별하기가 어렵다. 따라서 전사체 데이터 분석은 종종 생물 네트워크, 유전자 정보 데이터베이스, 문헌 정보 같이 서로 다른 자원을 활용하여 분석하게 된다. 그러나 자원들 간의 관계는 이질적인 부분이 존재하여 서로 직접적으로 연결하여 해석하기 어려우며 어떠한 유전자가 실험 목표와 관련이 있는지를 구체적으로 이해하기 힘들다. 따라서 특정 연구 목표와 관련 있는 핵심 유전자를 효과적으로 결정하고 설명하기 위해서는 이러한 이질적인 자원을 효과적으로 통합할 강력한 전산 기법이 필요하다. 본 논문에서는 네트워크 기반 접근법을 사용하여 전사체 데이터를 분석하고 실험 목표와 관련 있는 유전자를 찾기 위한 세 가지 생물 정보 시스템을 개발했다.

첫 번째 연구는 RNA-Seq 데이터의 특성을 활용하여 샘플 수가 적은 유전자 녹아웃 (KO) 마우스 실험에서 중요한 유전자를 찾기 위한 정보학 시스템을 개발하였다. 이 시스템은 유전자 조절 네트워크 (GRN)와 패스웨이 정보를 사용하여 유의함이 적은 Differentially Expressed Gene (DEG)를 제거하고 단일 염기 변이 (SNV) 정보를 사용하여 샘플 간 유전적 차이로 인해 다를 수 있는 유전자를 제거한다. 이 연구는 네트워크와 SNV 정보의 통합을 통해서 후보 유전자의 수를 유의미하게 줄일 수 있음을 보여주었다.

두 번째 연구는 사용자의 실험 목표를 반영할 수 있는 유전자 랭킹 시스템인 CLIP-GENE을 개발하였다. CLIP-GENE은 쥐의 전사인자 KO 실험에서 유전자를 랭킹하기 위한 통합 분석 웹 서비스이다. CLIP-GENE은 후보 유전자에 랭킹을 부여하기 위해 GRN, SNV 정보를 이용하여 샘플 개체 간의 차이가 있고 덜 유의미한 후보 유전자를 제거하고 텍스트 마이닝 기술과 단백질-단백질 상호작용

네트워크 정보를 이용하여 사용자의 실험 목표와 관련된 유전자를 랭킹한다.

마지막 연구는 벤 다이어그램을 사용하여 다수의 RNA-Seq 실험을 비교분석할수 있는 정보 시스템을 개발하였다. RNA-Seq 실험은 일반적으로 비교 및 대조군의 샘플을 비교하여 DEG를 생성하고 벤 다이어그램을 통하여 샘플 간의 차이를 분석한다. 그러나 벤 다이어그램 상에서의 각 영역은 다양한 비율의 DEG를 포함하고 있으며, 특정 영역의 DEG는 서로 다른 비교군(혹은 대조군)에 의한 DEG이기에 단순히 유전자 목록 간의 차이를 비교하는 것은 적절하지 못하다. 이러한 문제를 해결하기 위해 벤 다이어그램과 네트워크 전파(Network Propagation)를 사용한 통합 분석 프레임워크인 Venn-diaNet이 개발했다. Venn-diaNet은 다수의 DEG 목록이 있는 실험의 유전자를 랭킹할 수 있는 정보 시스템이다. 우리는 Venn-diaNet이 서로 다른 조건에서 생물학적 실험을 비교함으로써 원본 논문에 보고된 연구 결과를 재현 할 수 있음을 보여주었다.

정리하면 이 논문은 전사체 데이터로부터 유전자를 랭킹할 수있는 정보 시스템을 개발하기 위해 네트워크 기반 분석법을 다양한 자원들과 결합하였으며, 다른 연구자의 편리한 사용 경험을 위해 친화적인 UI를 가진 웹도구 또는 소프트웨어 패키지로 제작 및 배포하였다.

주요어: RNA-seq, 유전자 랭킹, 정보 분석 시스템, 네트워크 분석

학번: 2012-23113