이 학 박 사 학 위 논 문

Hierarchical Structural Component Models for Integrative Analysis of miRNA and mRNA expression data

계층적 구조 모형을 이용한 miRNA, mRNA 발현 자료의 통합분석

2018 년 8 월

서울대학교 대학원

통 계 학 과

김 용 강

Hierarchical Structural Component Models for Integrative Analysis of miRNA and mRNA expression data

by

Yongkang Kim

A thesis
submitted in fulfillment of the requirement
for the degree of Doctor of Philosophy
in
Statistics

Department of Statistics
College of Natural Sciences
Seoul National University
Aug, 2018

Hierarchical Structural Component Models for Integrative Analysis of miRNA and mRNA expression data

지도교수 박 태 성

이 논문을 이학박사 학위논문으로 제출함

2018 년 8 월

서울대학교 대학원

통계학과 통계학 전공

김 용 강

김용강의 이학박사 학위논문을 인준함

2018 년 8 월

위 원 장 **Myunghee Cho Paik (인)**

부위원장 박 태 성 (인)

위 원 임 요 한 (인)

위 원 이 승 연 (인)

위 원 **Heungsun Hwang (인)**

Abstract

Hierarchical Structural Component Models for Integrative Analysis of miRNA and mRNA expression data

Yongkang Kim

Department of Statistics

The Graduate School

Seoul National University

Identification of multi-markers is one of the most challenging issues in this new era of "personalized medicine." Although many methods have been developed to identify candidate markers for each type of omics data, few can facilitate multi-marker identification. It is well known that microRNAs (miRNAs) affect phenotypes only indirectly, through regulating mRNA expression and/or protein translation. Toward addressing this issue, we suggest a hierarchical structured component analysis of microRNA-mRNA integration ("HisCoM-mimi") model that accounts for this biological relationship, to efficiently study and identify such integrated markers.

In this thesis, we suggest two types of HisCoM-mimi. First type of HisCoM-mimi is used for discriminant analysis. In simulation studies,

HisCoM-mimi showed the better performance than the other three methods. Also, in real data analysis, HisCoM-mimi successfully identified more gives more informative miRNA-mRNA integration sets relationships for pancreatic ductal adenocarcinoma (PDAC) diagnosis, compared to the other methods.

Second type of HisCoM-mimi is used for survival analysis (mimi-surv). As the result of comparison study of HisCoM-mimi for discriminant analysis, we found the statistical power of mimi-surv to be better than other models in simulated comparisons. In analysis of real clinical data, mimi-surv successfully identified miRNA-mRNA integrations sets associated with progression-free survival of PDAC patients. Interestingly, miR-93, a previously unidentified PDAC-related miRNA, was found by mimi-surv, both in patient data from Seoul National University Hospital and The Cancer Genome Atlas (TCGA). Also, methods that use known structure for miRNA-mRNA regularization, found more PDAC related miRNAs than others.

As exemplified by an application to pancreatic cancer data, our proposed model effectively identified integrated miRNA/target mRNA pairs as markers for diagnosis or prognosis of cancer, providing a much broader biological interpretation

Key words: microRNA, mRNA, omics integration, discriminant analysis, survival analysis

Student number: 2013-30080

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Biological background on omics data analysis

Presently, numerous types of "omics" data are generated by many accurate and cost-effective methods. For instance, next-generation sequencing (NGS) technology is used to find DNA or RNA variations, bisulfite sequencing is used to find DNA-methylated variants, and multiple reaction monitoring (MRM) is applied to measure protein abundances (1-3). These efficient omics data platforms allow researchers to use multi-omics data, obtained from the same subjects, for analyzing huge numbers of variants. As a result, efficient multi-omics data analysis is becoming more important in

integrating large-scale data sets, making it possible to interpret fundamental biological systems (4).

With the development of generating and handling genomic data, a genome-wide association (GWA) study has become a common approach for testing association between a single nucleotide polymorphism (SNP) and a complex disease of interest. There have been many successful results from GWAS. However, SNPs that were identified by GWAS have been shown to explain only a small fraction of disease etiology, because the relatedness between complex diseases and multiple genes and/or their interactions are ignored. For this reason, integration analysis of gene-protein or gene-environment has been emphasized as a new alternative for understanding the etiology of common complex traits. However, these omics integration analyses are hard to detect and characterize by using traditional parametric statistical methods, for the following reasons. First, high dimensional omics data may be of a sparse nature. The issue of data sparseness can be addressed by using exponentially large sample sizes when parametric statistical methods are used for omics integration. Second, traditional statistical model is hard to reflect biological relationships between omics data (5, 6). Following statements show why omics integration is important for biological interpretation, and traditional statistical methods cannot deal with.

1.1.1 Central dogma in biological procedure

The Central Dogma states that once 'information' has passed into protein it cannot get out again (7). In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein. The dogma is a framework for understanding the transfer of sequence information between information-carrying biopolymers, in the most common or general case, in living organisms. There are three major classes of such biopolymers: DNA and RNA (both nucleic acids), and protein. There are nine conceivable direct transfers of information that can occur between these. The dogma classes these into three groups of 3: three general transfers (believed to occur normally in most cells), three special transfers (known to occur, but only under specific conditions in case of some viruses or in a laboratory), and three unknown transfers (believed never to occur). The general transfers describe the normal flow of biological information: DNA can be copied to DNA (DNA replication), DNA information can be copied into mRNA (transcription), and proteins can be synthesized using the information in mRNA as a template (translation). The special transfers describe: RNA being copied from RNA (RNA replication), DNA being synthesized using an RNA template (reverse transcription), and proteins being synthesized directly from a DNA template without the use of mRNA. The unknown transfers describe: a protein being copied from a protein, synthesis of RNA using the primary structure of a

protein as a template, and DNA synthesis using the primary structure of a protein as a template - these are not thought to naturally occur.

1.1.2 Definition of miRNA inhibition process

MicroRNAs (miRNAs) are noncoding RNAs having a length less than 25 base pairs, regulating the expression of specific genes by mRNA degradation or blocking translation by binding to the 3′ regions of their ″target″ mRNAs. By downregulating their target mRNAs, miRNAs control nearly all developmental and pathological processes in animals, particularly in cell development, and many cancer types are affected by miRNA regulation (8).

Since miRNA has a well-known regulation mechanism, many studies focused on finding their target mRNAs, and explaining biological context by showing significant negative correlation between miRNA and target mRNA, and how such relationships affect phenotypes. For example, Enerly et al. performed hierarchical clustering on miRNA expression profiles, and found that specific clusters associated with specific expression levels of the tumor suppressor gene, TP53 (9). Although this approach is effective when the number of target genes is small, it more difficult for identifying novel miRNA-target gene integration sets.

Many recent studies have now implicated miRNAs in the pathogenesis of cancer, including triggering cancer initiation and progression. MiRNAs have been shown to have tissue-specific and disease-specific expression patterns (10-13). Intensive investigation is now underway for using applying miRNAs′

inhibitory information to mRNAs. For example, Nam et al. developed ″miRNA and mRNA integrated analysis″ (MMIA) to examine biological functions of miRNA expression (14). Moreover, Buffa et al. used pathway information to independently validate miRNAs significant for breast cancer (15), while Cho et al. performed network analysis, and hierarchical clustering, to find biological ″signatures″ of interstitial lung diseases (16).

Most miRNA and mRNA integration analyses focus on first identifying miRNAs significantly associated with the phenotype of interest, and then experimentally validating those miRNAs′ phenotype involvement by inhibiting or ectopically overregulating their expression (14-16). Although these approaches are effective at validating significant miRNAs, they do not provide information on how they regulate expression of their target mRNAs, as relevant to the pathway level.

To perform mRNA and miRNA integration analysis, based on information on large numbers of targets, many studies use a two-step analysis. First, miRNAs associated with specific phenotypes are chosen. Second, expression levels of known target mRNAs, which negatively correlate to each miRNA, are investigated further (17). Therefore, this approach integrates two different omics data efficiently, but only when focused on phenotypes and miRNAs. However, this approach cannot provide information about how miRNAs and their inhibited mRNAs affect to observed phenotype together.

1.1.3 Review of omics measuring techniques

Fig. 1.1 illustrates technical progress of omics data. Reverse transcription polymerase chain reaction (RT-PCR) is a technique commonly used in molecular biology to detect RNA expression (18).

RT-PCR is used to clone expressed genes by reverse transcribing the RNA of interest into its DNA complement through the use of reverse transcriptase. Subsequently, the newly synthesized cDNA is amplified using traditional PCR. Although this method could detect target mRNA expression or DNA mutants, it cannot detect multiple mRNAs or DNA mutants at once.

A microarray is a multiplex lab-on-a-chip. It is a 2D array on a glass that stores large amounts of biological material using high-throughput screening miniaturized, multiplexed and parallel processing and detection methods. The concept and methodology of microarrays was first introduced and illustrated in antibody microarrays by Tse Wen Chang (19). The "gene chip" industry started to grow significantly by the Ron Davis and Pat Brown (20). The technology of DNA microarrays has become the most sophisticated and the most widely used, while the use of protein, peptide and carbohydrate microarrays is expanding (21). With development of microarray, we could detect multiple mRNAs or DNA mutants with a one experiment. However, it uses known sequence to detect targets, unknown genes or DNA mutants cannot be obtained by microarray.

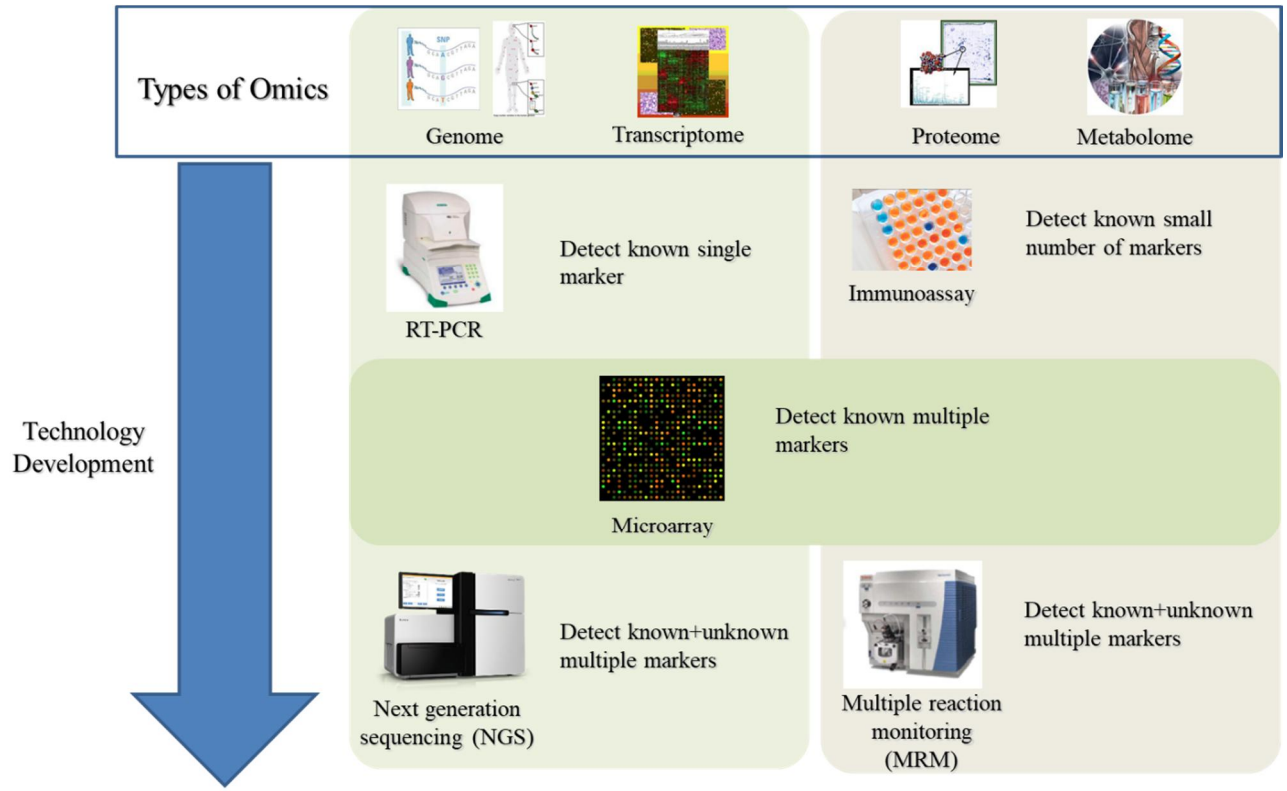

Fig. 2.1 Types of Omics data and Technologies for detecting omics data.

Microarrays are hybridization experiments involving comparison of relative amounts of cellular mRNA from two tissue samples. The terms ″hybridize″ and ″hybridization″ mean that a single strand of DNA or RNA consisting of unpaired nucleotide bases bonds to a respective complementary nucleotide strand of DNA or RNA. Genomic DNA is usually first transcribed into mRNA in the cell nucleus and subsequently translated into proteins in the cell cytoplasm. The amount of mRNA in the cell is thought to represent the transcription of the gene. Hence, the extraction, stabilization, and purification of total RNA that includes mRNA are important factors affecting the quality of the microarray results. Total RNA is extracted from the tissue, and the quality of the total RNA is verified by electrophoresis and spectrophotometry. The mRNA is labeled and hybridized to the array for quantification. This is achieved by introducing a fluorescent marker during the preparation of mRNA that can be detected and quantified by a laser scanner.

The exploitation of hybridization in microarray analyses sharply accelerated the search for defective genes. Hybridization is based on the Watson-Crick model of base pairing of nucleic acids such that adenine (A) binds to thymine (T) (or uracil [U], in the case of RNA), and cytosine (C) binds to guanine (G). Each probe on a microarray is designed to hybridize with unknown target mRNA. In this review, the term ″probe″ refers to known oligonucleotides or complementary DNA (cDNA) fragments immobilized on microarray slides. When samples labeled with fluorescence are applied to microarrays,

hybridization or binding reactions take place between each probe and the target mRNA. Each microarray probe recognizes cDNA sequences by base pairing (hybridization). After a series of washes to eliminate unbound nucleotides and nonspecific bindings, only the target probe complexes remain bound. Intensity of the fluorescent signal for each probe reflects the abundance of the target RNA in the RNA sample.

An immunoassay is a biochemical test that measures the presence or concentration of a macromolecule or a small molecule in a solution through the use of an antibody. The molecule detected by the immunoassay is often referred to as an "analyte" and is in many cases a protein, although it may be other kinds of molecules, of different size and types, as long as the proper antibodies that have the adequate properties for the assay are developed. Analytes in biological liquids such as serum or urine are frequently measured using immunoassays for medical and research purposes (22). Immunoassay could detect the small number of proteins or metabolites which have commercial antibodies. Thus, it cannot used to generate large dimensional data. Selected reaction monitoring (SRM) is a method used in tandem mass spectrometry in which an ion of a particular mass is selected in the first stage of a tandem mass spectrometer and an ion product of a fragmentation reaction of the precursor ion is selected in the second mass spectrometer stage for detection (23). SRM can be used for targeted quantitative proteomics by mass spectrometry (24). Multiple reaction monitoring (MRM) is the application of selected reaction monitoring to multiple product ions from one or more

precursor ions (3, 25, 26). Thus, MRM could detect multiple protein markers easily than other protein detection methods.

The primary technology for the detection of rare SNPs is sequencing, which may target regions of interest, or may examine the whole genome. Next-generation sequencing technologies, which process millions of sequence reads in parallel, provide monumental increases in speed and volume of generated data free of the cloning biases and arduous sample preparation characteristic of capillary sequencing. RNA is less stable in the cell, and also more prone to nuclease attack experimentally. As RNA is generated by transcription from DNA, the information is already present in the cell's DNA. However, it is sometimes desirable to sequence RNA molecules. While sequencing DNA gives a genetic profile of an organism, sequencing RNA reflects only the sequences that are actively expressed in the cells. To sequence RNA, the usual method is first to reverse transcribe the RNA extracted from the sample to generate cDNA fragments.

1.2 Statistical procedure to analyze omics data

Second, detecting integrated markers using traditional procedures leads to an increase in type II errors and a decrease in power. As a result, detecting interactions among variables is a well-known challenge in statistics and data mining (Freitas, 2001).

1.2.1 Quality control and normalization of omics data

Quality control (QC) is the most important data preprocessing procedure used to find any potential undesirable factors during the production of data. In bioinformatics, QC for omics data has been widely used through various statistical methods. These approaches check whether or not the data is well produced, by not violating the experimental protocol and whether or not any experiment batch effect exists, in different environments. In detail, QC procedures include detecting outliers caused by data contamination. For example, finding structural mistakes when grouping the data and mismatching the data labels. In the area of high-dimensional data, when tens of thousands of genes or proteins are frequently generated from a large number of samples via high-throughput technology (e.g., microarray and next generation sequencing), it becomes more difficult and time-consuming to conduct QC effectively.

The MicroArray Quality Control (MAQC) project showed the existence of inter- and intra-platform reproducibility of gene expression through QC experiments. Additionally, the MAQC project also showed several measurement errors and poor experiment qualities in some microarrays (27-31). Several processing tools for QC of bioinformatics data have now been well developed. For example, Wang et al. developed RSeQC for QC of RNA-Seq experiments, Wilson et al. made Simple affy package for QC specifically for Affymetrix microarrays, and Bock et al. created a BiQ Analyzer to perform QC on DNA methylation data from bisulfite sequencing (2, 32, 33).

ArrayQCplot, developed by Lee et al., provides a simple graphical tool to depict QC of microarray data (34).

Most current QC measures and plots play supporting roles, and do not provide an objective guideline to determine the quality of data. To provide a more objective guideline, based on p-values from distances, we introduced a high-dimensional data quality control (HidQC) plot (35). The HidQC plot uses distance measures, such as correlation and harmonically summed distance. We showed that the HidQC plot offers simple and easily interpretable QC results for high-dimensional omics data. Through these distance-based procedures, the HidQC plot checks data quality by investigating the consistency of each group. HidQC plot also provides permutation-based p-values for each sample, to determine the inclusion or exclusion, of the specific sample, in the main analysis. Through a HidQC plot, it is possible to simply detect outliers in high-dimensional omics data, and to easily check any instrumental or human-related errors, such as sample contaminations, mislabeling, or misgrouping the entities of data. The hypothesis underlying microarray analysis is that the measured intensities for each arrayed gene represent its relative expression level. Biologically relevant patterns of expression are typically identified by comparing measured expression levels between different states on a gene-by-gene basis. But before the levels can be compared appropriately, a number of transformations must be carried out on the data to eliminate questionable or low-quality measurements, to adjust the measured intensities to facilitate comparisons,

and to select genes that are significantly differentially expressed between classes of samples.

After QC process with HidQC plot, we could apply the transformation to expression data, referred to as normalization, adjusts the individual hybridization intensities to balance them appropriately so that meaningful biological comparisons can be made. With a normalization, we could obtain fair comparison of intensities which is originally affected by quantities of starting RNA, differences in labeling or detection efficiencies between the fluorescent dyes used, and systematic biases in the measured expression levels.

1.2.2 Statistical methods for finding significant features

DNA microarrays contain oligonucleotide or cDNA probes for measuring the expression of thousands of genes in a single hybridization experiment. Although massive amounts of data are generated, methods are needed to determine whether changes in gene expression are experimentally significant. Cluster analysis of microarray data can find coherent patterns of gene expression but provides little information about statistical significance. Methods based on conventional t tests provide the probability (P) that a difference in gene expression occurred by chance (2, 3). Although $P < 0.01$ is significant in the context of experiments designed to evaluate small numbers of genes, a microarray experiment for 10,000 genes would identify 100 genes by chance. To resolve this problem, Significance Analysis of Microarrays

(SAM) was developed. SAM identifies genes with statistically significant changes in expression by assimilating a set of gene-specific t tests. Each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene. Genes with scores greater than a threshold are deemed potentially significant.

Although SAM worked well in DEG analysis, many microarray studies could not be replicated by other researches. This limitation is caused by unwanted variation. It may lead to high rates of false discoveries, high rates of missed discoveries, or both. The causes of unwanted variation are often partially or entirely unknown. In some cases, factors that cause unwanted variation are known, but cannot be easily or precisely measured. In other cases, only proxies of the true unwanted factors may be known. For example, in a microarray study, ″batch effects″ may be created if samples in one batch are processed at a higher temperature than in another batch. A researcher may know that a batch effect exists, and even know which samples were processed in which batch, but not know the cause of the batch effect (temperature). The researcher may try to model the unwanted variation using a dummy variable for batch. However, the batch variable is only a proxy for temperature.

Methods to adjust for unwanted variation can be divided into 2 broad categories. In the first category are methods that can be used quite generally and provide a global adjustment. An example would be quantile normalization (QN), which is generally regarded as a self-contained step and plays no role in

the downstream analysis of the data. In the second category are application specific methods that incorporate the batch adjustment directly into the main analysis of interest.

To adjust unwanted variation, Gagnon-Bartsch et al., suggested removing unwanted variation method (RUV). RUV uses two types control genes. Negative control genes are genes whose expression levels are known a priori to be truly unassociated with the biological factor of interest. Conversely, positive control genes are genes whose expression levels are known a priori to be truly associated with the factor of interest. RUV was performed with factor analysis on just the negative control genes and incorporate the resulting factors into a linear regression model. The idea is that since the negative control genes are known to be unassociated with the factor of interest, there is no danger in picking up any of the relevant biology in the factor analysis step. RUV is widely applicable in omics analysis based on sequencing, microarray, mass spectrometry, and so on.

1.2.3 Multiple testing problems on Omics data analysis

The biological question of differential expression can be restated as a problem in multiple hypothesis testing: the simultaneous test for each gene j of the null hypothesis H_j of no association between the expression measure X_j and the response or covariate Y.

A standard approach to the multiple testing problem consists of two aspects:[1] computing a test statistic T_j for each gene j, and [2] applying a multiple testing procedure to determine which hypotheses to reject while controlling a suitably defined Type I error rate. To solve the multiple testing problem of microarray, many researchers used false discovery rate (FDR). Since, FDR controlled the proportion of false positive genes, it could reduce the chance of throwing out true signal genes comparing to family wise error rate (FWER).

Table 1.1. shows the possible outcomes when testing multiple null hypotheses. Let m be the total number of hypotheses, m_0 be the number of true null hypotheses, $m - m_0$ be the number of true alternative hypotheses. Then V denotes the number of false positives, S denotes the number of true positives, T denotes the number of false negatives, and U denotes the number of true negatives.

FWER is the probability of making at leat one type I error in the family of test.

$$\text{FWER} = P(V \geq 1) = 1 - P(V = 0)$$

Thus, $\text{FWER} \leq \alpha$ means the probability of at least one or more type I errors being occurred is less than α. To control FWER, we could consider Bonferroni procedure, Sidak procedure, Tukey′s procedure, Holm′s step-down procedure and so on.

Bonferroni procedure is the easiest and most common method in controlling FWER. Let p_i be the p-value for testing H_i. When Bonferroni procedure is performed, we reject H_i if $p_i \leq \frac{\alpha}{m}$.

Sidak procedure tests each hypothesis at level $\alpha_{SID} = 1 - (1 - \alpha)^{\frac{1}{m}}$. This procedure is more powerful than Bonferroni, however, can fail to control the FWER when the tests are negatively dependent.

Tukey′s procedure assumes independence of the observations being tested, as well as equal variation across observations. It uses Student′s t-test to corrects FWER.

Holm′s step-down procedure uses order statistics of p-values from tests: $P_{(1)}, P_{(2)}, \ldots, P_{(m)}$, and $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$. Let k be the minimal index such that $P_{(k)} > \frac{\alpha}{m+1-k}$. Reject the null hypothesis $H_{(1)}, H_{(2)}, \ldots, H_{(k-1)}$. If $k = 1$ then none of the hypotheses are rejected. This procedure is uniformly more powerful than the Bonferroni procedure.

Table 1.1 The possible outcomes when testing multiple null hypotheses

	Null hypothesis is true (H_0)	**Alternative hypothesis is true (H_1)**	**Total**
Reject null hypothesis	V	S	R
Accept null hypothesis	U	T	$m - R$
Total	m_0	$m - m_0$	m

1.2.4 Review of traditional data integration methods

As omics data gain technologies are advanced, the types of omics data and the number of markers for each omics are expanded. Thus, omics data integration with large-scale dataset is very important. Omics data integration analysis refers to the combination of at least two different types of omics data. Relationships between two sets of omics parameters such as the expression quantitative trait loci (eQTL) or the methylation-QTL (methQTL, have been recently reported. The approach most commonly used for this type of pairwise analysis has been performed calculating correlation coefficients or simple linear regression models. Since these approaches assumed that the changes in gene expression are only affected by one variable, it is hard to interpret complex associations between omics data.

Recently, many studies used dimension reduction approaches first, and then find the significant omics relation. Principal Component Analysis (PCA) or Canonical Correlation Analysis (CCA) is used to reduce data dimensionality and investigate the overall correlation between two sets of variables. However, these methods are descriptive or exploratory techniques rather than hypothesis-testing tools.

Lasso proposed by Tibshirani in 1996 and the Elastic Net (EN) proposed by Hui Zou and Trevor Hastie in 2005 are penalized regression methods which can model more than one type of omics data. More importantly, both methods simultaneously execute variable selection and parameter estimation, thus reducing the computation time, while the traditional methods work on the two

problems separately, first selecting the relevant parameters and then computing the estimates. LASSO and EN have already been applied to GWAS studies as well as in the context of integrative studies. One limitation of penalized regression techniques is that the penalty produces biased estimators; consequently, standard errors are not meaningful and cannot provide p-values to assess significance.

1.3 The purpose of this study

In this study, we propose a structured component-based analysis, for integrating omics data to identify multiple accurate biomarkers. It is well known that miRNAs affect phenotypes indirectly, by regulating mRNA expression or protein translation (13). Herein, we propose hierarchical structured component analysis of miRNA-mRNA integration (HisCoM-mimi) analysis, which models biological relationships as structured components, to efficiently yield integrated markers.

1.4 Outline of the thesis

This thesis is organized as follows. Chapter 1 is an introduction to this study to describe the purpose of this study. Chapter 2 is a review of the component-based structural equation models. Chapter 3 gives motivate examples. Chapter 4 introduces our proposed method, HisCoM-mimi with

binary phenotype. Chapter 5 introduces Hiscom-mimi with survival time of cancer patients. Finally, the summary and conclusions are presented in Chapter 6.

Chapter 2

Review of component-based structural equation models

3.1 Partial least square path modeling (PLS-PM)

Partial Least Squares is a family of regression-based methods designed for the analysis of high dimensional data. Manifest variables, or indicators, are observable variables who are supposed to analyze information about the behavior of latent variables, theoretical concepts, which are not directly observable. In the social sciences factor models are most commonly used for the analysis of the interplay between latent and manifest variables. Model

construction and estimation used to be focused mainly on the specification, validation and interpretation of factor loadings and latent variables.

Basic design of PLS-PM is as Fig. 2.1. Let f_1, f_2 be atent variables, w_1, w_2, w_3, w_4 be weights between latent variables and observations x_1, x_2, x_3, x_4. Conventionally, both indicators and latent variables are assumed to be standardized in PLS-PM. The individual scores of standardized latent variables can always be obtained by multiplying their normalized scores by the square root of sample size.

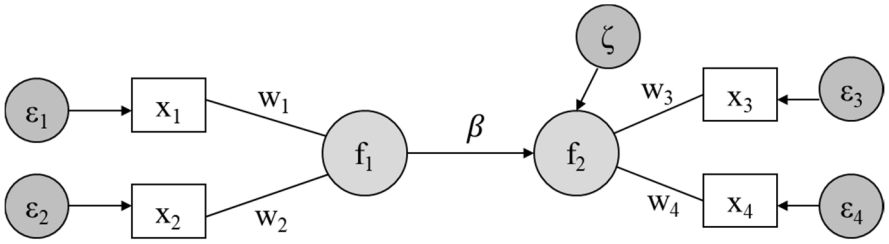

Fig. 2.1 The structure of PLS-PM

We provide a description of the Lohmöller′s (1989) algorithm for partial least squares path modeling. This algorithm carries out two main stages sequentially. The first stage estimates weights and latent variables iteratively. Inner estimates for latent η_j is updated by $\eta_j = \sum_{q=1}^{Q_j} \gamma_{jq} f_q$, where e_{jq} means a scalar value, called the inner weight. γ_{jq} is estimated by the signs of the correlations between η_q and η_j or the regression coefficients of η_j on η_q. The second estimates weights and loadings in closed form. w_j is updated with regression. $w_j = \left(X_j' X_j\right)^{-1} X_j' \eta_j$. β is upadated by $\beta = X_j w_j$ and normalized with $\beta'\beta = w_j' X_j' X_j w_j = 1$. These steps are repeated until w_j are converged.

3.2 Generalized structured component analysis (GSCA)

Hwang and Takane propose generalized structured component analysis (GSCA) as an alternative to SEM (36). GSCA maximizes the average or sum of explained variances of linear composites. GSCA consists of three defining elements: (1) a way to specify linear models, (2) an optimization criterion, and (3) an algorithm to obtain estimates. We illustrate all three elements of the GSCA approach next.

2.1.2.1. Model specification

GSCA is a component-based method whereby each component represents a latent variable. It involves three sub-models as found in Fig. 2.2:

measurement, structural, and weighted relation models. The measurement model is used to define the loading of each latent to observed variable. The structural model is used to define the relationship between latent variables. The weighted relation model is used to define a latent variable as a weighted composite or component of indicators. These sub-models can be written in matrix form as follows:

- Measurement model: $\boldsymbol{x}_i = \boldsymbol{C}\boldsymbol{f}_i + \boldsymbol{\epsilon}_i$

- Structural model: $\boldsymbol{f}_i = \boldsymbol{B}\boldsymbol{f}_i + \boldsymbol{\zeta}_i$

- Weighted relation model: $\boldsymbol{f}_i = \boldsymbol{W}\boldsymbol{x}_i$

where $\boldsymbol{x}_i$ is observed variables of *i*th individual, $\boldsymbol{f}_i$ is a vector of latent variables, C is a $P \times J$ matrix of loadings, B is a $J \times I$ matrix of path coefficients β, W is a $J \times P$ matrix of component weights, ε is a vector of the errors for observed variables, and ζ is a vector of errors for latent variables.

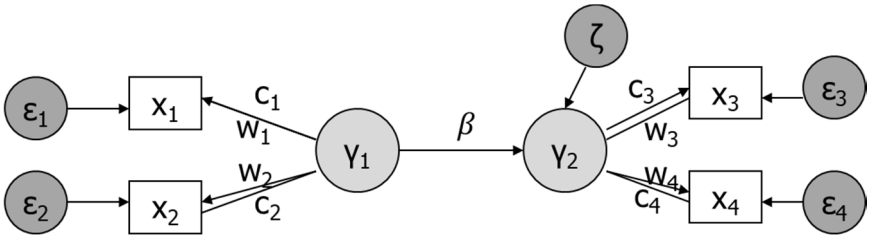

Fig. 2.2 The structure of GSCA

2.1.2.2. Model estimation and extension

To fit GSCA, we could summarize three sub-model as follows:

$$\begin{vmatrix} x_i \\ f_i \end{vmatrix} = \begin{vmatrix} C \\ B_i \end{vmatrix} f_i + \begin{vmatrix} \boldsymbol{\epsilon}_i \\ \boldsymbol{\zeta}_i \end{vmatrix} \quad (1)$$

$$\begin{vmatrix} I \\ W_i \end{vmatrix} x_i = V x_i = \begin{vmatrix} C \\ B_i \end{vmatrix} W x_i + \begin{vmatrix} \boldsymbol{\epsilon}_i \\ \boldsymbol{\zeta}_i \end{vmatrix} = AW x_i + E_i \quad (2)$$

GSCA estimates model parameters, including weights (W), path coefficients (B), and loadings (C), by minimizing the sum of the squares of E_i, or $\phi(\mathrm{A}, \mathrm{W})$ defined by:

$$\phi(\mathrm{A}, \mathrm{W}) = \sum_{i=1}^{n} \mathrm{E_i'} E_i = \sum_{i=1}^{n} \big((V - AW)x_i\big)' \big((V - AW)x_i\big) \quad (3)$$

Since we cannot estimate A and W simultaneously, alternating least squares (ALS) algorithm is used to minimize $\phi(\mathrm{A}, \mathrm{W})$ as follows.

1) Fix A and update $\mathrm{W} = \mathrm{argmin}_{\mathrm{W}}(\phi(\mathrm{A}, \mathrm{W}))$

2) Fix W, and update $\mathrm{A} = \mathrm{argmin}_{\mathrm{A}}(\phi(\mathrm{A}, \mathrm{W}))$

Iteratively perform 1) and 2) until converged.

In the traditional approach of GSCA, the standard errors and p-values for components are estimated by the bootstrap method.

3.3 Extended Redundancy Analysis (ERA)

Redundancy analysis (RA), also called reduced rank regression or principal components of instrumental variables is a useful technique for analyzing a

directional relationship between two sets of multivariate data (37). RA aims to extract a series of components from a set of exogenous variables in such a way that they are mutually orthogonal and successively account for the maximum variance of a set of endogenous variables. However, they are limited to model and fit a particular type of relationship among three sets of variables. Thus, RA was extended to specify and fit a variety of relationships among multiple sets of variables. The method proposed is called an extended redundancy analysis (ERA).

Let Z_1 be an $\mathrm{n} \times \mathrm{r}$ matrix of dependent or endogenous variables and Z_1 be an $\mathrm{n} \times \mathrm{t}$ matrix of independent or exogenous variables. RA could defined by $\mathrm{Z}_1 = Z_2WA + E = FA + E$, where W is a $\mathrm{t} \times \mathrm{d}$ matrix of component weights, A is a $\mathrm{d} \times \mathrm{r}$ matrix of component loadings, E is an $\mathrm{n} \times \mathrm{t}$ matrix of errors, and F is component scores. To estimate parameters in RA, we minimize the following least squares criterion,

$$\mathrm{f} = \mathrm{SS}(\mathrm{Z}_1 - Z_2WA),$$

with respect to W and A, where $\mathrm{SS}(\mathrm{X}) = \mathrm{trace}(\mathrm{X}'\mathrm{X})$.

In ERA, RA was extended to accommodate more diverse relationships among multiple sets of variables. In ERA, Z_1 is an endogenous variable, and Z_2 was an exogenous variable. ERA uses latent variables with separate observations. Thus, ERA can fit more flexible models than RA. To fit ERA, alternative least squares were used. Since RA used fixed form of models, canonical correlation analysis (CCA) or principle component analysis (PCA)

were used to fit RA. Alternative least squares also could be used for fitting RA, however, ERA cannot be fitted by CCA or PCA.

3.4 Pathway based approach using hierarchical components of collapsed rare variants (PHARAOH)

Extending GSCA and PLS-PM, pathway based approach using hierarchical components of collapsed rare variants (PHARAOH) model was developed for pathway-based analysis (38). PHARAOH uses a hierarchical structure of rare variants, genes, and pathways. The advantage of such hierarchical structural component models is their generation of latent variables, such as genes and pathways, which are inferred by observed variables, such as rare variants. Using latent variables, we can collapse unstructured data into a structured form, providing less ambiguous biological explanations of the results. In PHARAOH approach, we obtained p-values for components by permutation.

To estimate the parameters for PHARAOH, we used alternating least squares algorithm for the penalized log-likelihood function, with ridge parameters. Then, the objective function to maximize is given as follows:

$$logit(\pi_i) = \beta_0 + \sum_{j=1}^{J}\left[\sum_{k=1}^{G_j} X_{ijk} w_{jk}\right]\beta_j = \beta_0 + \sum_{j=1}^{J} f_{ij}\beta_j, \quad (4)$$

$$\varphi_1 = \sum_{i=1}^{n} \log p(y_i; \beta_j, \delta) - \frac{1}{2}\lambda_G \sum_{j=1}^{J}\sum_{k=1}^{G_j} w_{jk}^2 - \frac{1}{2}\lambda_P \sum_{j=0}^{J} \beta_j^2 \quad (5)$$

where $p(y_i;\gamma_i,\delta)$ is the probability distribution for the phenotype of the ith individual. λ_G and λ_P are ridge parameters for pathway of interest, representing the integrated latent components. X_{ijk} is measure of collapsed rare variants for the gene.

To maximize the objective function, φ_1, the iterative reweighted least squares (IRWLS) algorithm is used. Note that when using IRWLS, maximizing φ_1 is equivalent to minimizing the object function φ_2.

$$\begin{aligned}\varphi_2 &= \sum_{i=1}^{n} \mathrm{v_i}(z_i - \sum_{j=1}^{J} f_{ij}\beta_j)^2 - \frac{1}{2}\lambda_G \sum_{j=1}^{J}\sum_{k=1}^{G_j} w_{jk}^2 - \frac{1}{2}\lambda_P \sum_{j=0}^{J} \beta_j^2 \\ &= (\mathrm{z} - \mathrm{F}\beta)'\mathrm{V}(\mathrm{z} - \mathrm{F}\beta) + \lambda_{\mathrm{G}} \sum_{j=1}^{J} \left(w_j' w_j\right) + \lambda_P (B'B)\end{aligned} \tag{6}$$

To maximize the objective function, φ_2, an alternating least squares algorithm is used as following steps.

1) initialize $\hat{B}, \widehat{W}$=0.

2) Minimize φ_2 with fixed $\widehat{w}_{jk}$ subject to $\sum|\widehat{w}_{jk}| \le s_W$.

3) Minimize φ_2 with updated $\hat{B}$ subject to $\sum|B_i| \le s_B$.

4) Repeat steps 3,4 until both $\widehat{W}$ and $\hat{B}$ do not change.

Chapter 3

Motivating Example

3.1 Pancreatic ductal adenocarcinoma (PDAC)

Note that PC is one of the most fatal diseases in the world, having an all-stage 5-year survival rate of about 6%, worldwide (39), a mere 8% five-year survival rate in the USA and a 9.4% survival rate in the Republic of Korea (40-42). In particular, the tumor heterogeneity in PC patients' tumors makes early diagnosis harder than cancers of most other organs (43). To adjust for heterogeneity among tumor cells, we need a more robust and complex statistical model which can interpret and integrate several causes of cancer altogether. Although many bioinformatics research studies have been performed to find diagnostic markers for PC, to date, no clinically approved

prognostic markers exist (44). Although there were few approved prognostic markers, many researchers did not give up to find prognostic or diagnostic markers for PDAC. One of candidates is miRNA. One previous meta-analysis showed that PDAC patients with high expression of miR-21 had significant shorter overall survival time (OS) than other patients (45). Also, many studies have identified cell-free miRNAs as prognostic markers of PDAC (46).

3.2 Seoul National University Hospital (SNUH) PDAC samples

To analyze real microarray data, 95 PDAC samples were collected from SNUH. The enrolled PDAC patients′ average age was 65.2 years (standard deviation (SD): 9.4 years). 46 patients were males, and 49 females. Fig. 3.1 shows a Kaplan-Meier curve for OS of SNUH PDAC patients after surgery. The median survival time after surgery of this group of PDAC patients was 795 days.

To analyze mRNA expression, the Affy HuGene 1.0 array (Affymetrix, Santa Clara, CA, USA) was used. mRNA expression values were normalized by robust multi-array averaging (RMA), using the Affymetrix console. After RMA, we also performed quantile normalization. To analyze miRNA expression, the Affy GeneChip miRNA 3.0 array (Affymetrix) was used. miRNA expression values were also normalized by RMA. Since the Affy GeneChip miRNA 3.0 array has probes for the expression of human miRNAs,

and also those of other species, we only considered human-derived miRNA targets, but used other species′ probes as background intensities. Thus, we normalized out the background intensities of the jth human probe, of the ith sample, x_{ij}, by the following Eq. (1).

$$x_{ij(norm)} = x_{ij} - median(x_{ij}, j \in non\ human\ miRNA) \quad (1)$$

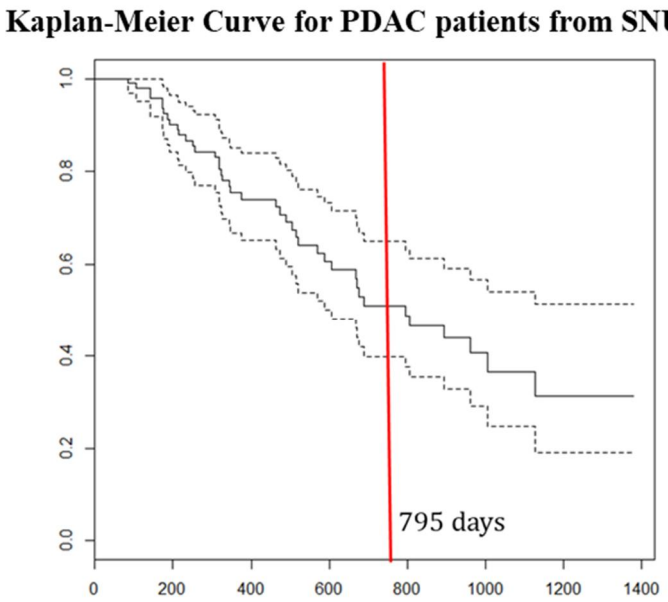

Fig. 3.3 Kaplan-Meier Curve for PDAC SNUH patients There were 75 patients enrolled in this analysis. Median survival time of SNUH dataset was 795 days.

3.3 The Cancer Genome Atlas (TCGA) PDAC samples

We downloaded a TCGA PDAC dataset from the Genomic Data Commons (GDC) data portal of the U.S. National Cancer Institute (https://portal.gdc.cancer.gov/) (47). For mRNA and miRNA expression profiling, Illumina Hiseq was used. To normalize mRNA-seq and miRNA-seq datasets, Fragments Per Kilobase Million (FPKM) was measured for each read count. We first collected 185 TCGA PDAC data samples for analysis. In sample quality control procedure, we excluded 25 non-PDAC samples, and 47 PDAC samples whose follow-up time was less than 3 months, because the cause of death was hard to determine. The PDAC patients′ average age was 63.9 years (standard deviation (SD): 11.1 years), 48 patients were male, and 64 were female. Fig. 3.2 shows the Kaplan-Meier curve for survival time of TCGA PDAC patients after surgery. The median survival time was 585 days

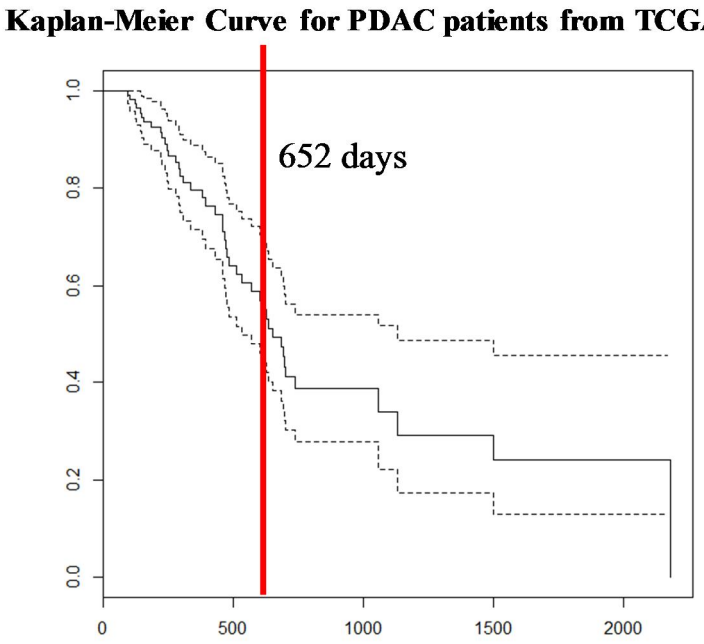

Fig. 3.2 Kaplan-Meier Curve for PDAC patients from TCGA dataset There were 112 patients enrolled in this analysis. Median survival time of TCGA dataset was 652 days

Chapter 4

Hierarchical structural component modeling of microRNA-mRNA integration model for binary phenotype

4.1 Introduction

In this chapter, we propose a structured component-based analysis, for integrating omics data for identifying multiple accurate biomarkers. It is well known that miRNAs affect phenotypes indirectly, by regulating mRNA expression or protein translation (13). Herein, we propose hierarchical structured component analysis of miRNA-mRNA integration (HisCoM-mimi)

analysis, which models biological relationships as structured components, to efficiently yield integrated markers. Our proposed model is based on generalized structured component analysis (GSCA), which tests hypothesized relationships between observed and latent variables (36).

Accordingly, our proposed HisCoM-mimi model can efficiently account for biological relationships between miRNA and mRNA, in the structured component, and effectively provide integrated (e.g., miRNA-to-target-mRNA) markers. As an illustration, we tried HisCoM-mimi for identifying biomarkers for the early diagnosis of PDAC.

Here, we applied HisCoM-mimi to computationally identify diagnostic markers of PDAC, the most common type of pancreatic cancer. By applying the HisCoM-mimi approach to miRNA and mRNA microarray data from PDAC patients, at Seoul National University Hospital (SNUH), we identified numerous cognate miRNA-mRNA partners, as markers for diagnosis of PDAC. Finally, our HisCoM-mimi provided integrated marker sets, with more biological and intuitive interpretation, than other existing methods.

4.2 Methods

4.2.1 HisCoM-mimi model for binary

To perform the integration analysis of miRNA and mRNA data, we developed and implemented our HisCoM-mimi approach. This model analyzes multiple subnetworks simultaneously, with specific regard to inverse

correlations between mRNA and miRNA. Fig. 4.1 shows the flowchart of the method. First, for a given miRNA, a miRNA-mRNA subnetwork, consisting of one miRNA and multiple potential target mRNAs, is constructed if the following two conditions are satisfied: (i) the mRNAs are reported as target of the miRNA by TargetScan 7.1 (targetscan.org) (48), and the negative correlation coefficients between the mRNA and miRNAs are significant (p-value <0.05). Second, for all entities deemed significant, we derived our hierarchical structural component model by using all miRNA-mRNA subnetworks.

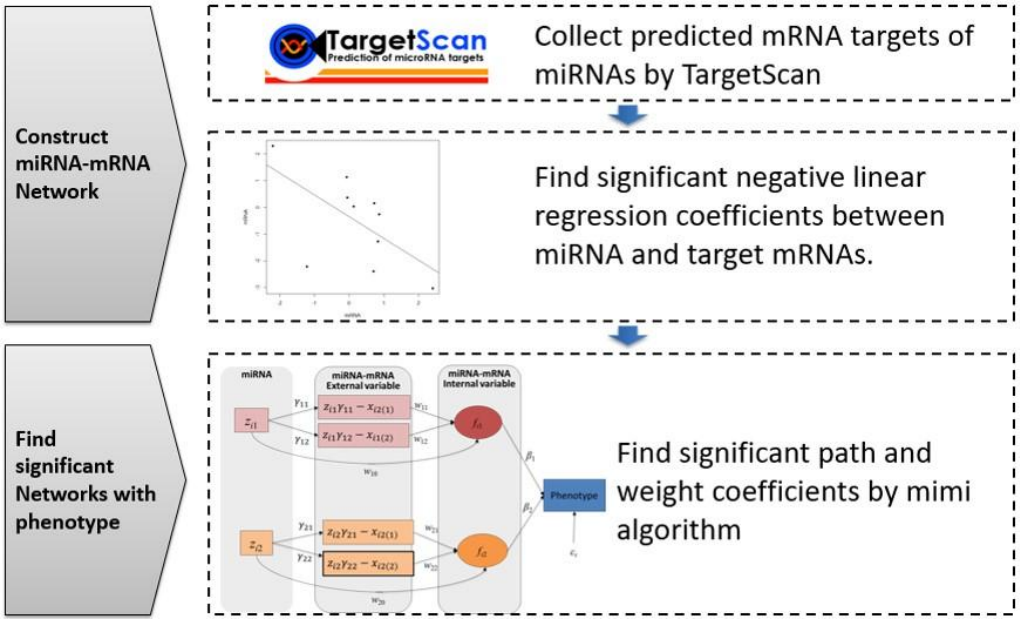

Fig. 4.1 Flow chart for analyzing mRNA-miRNA integration

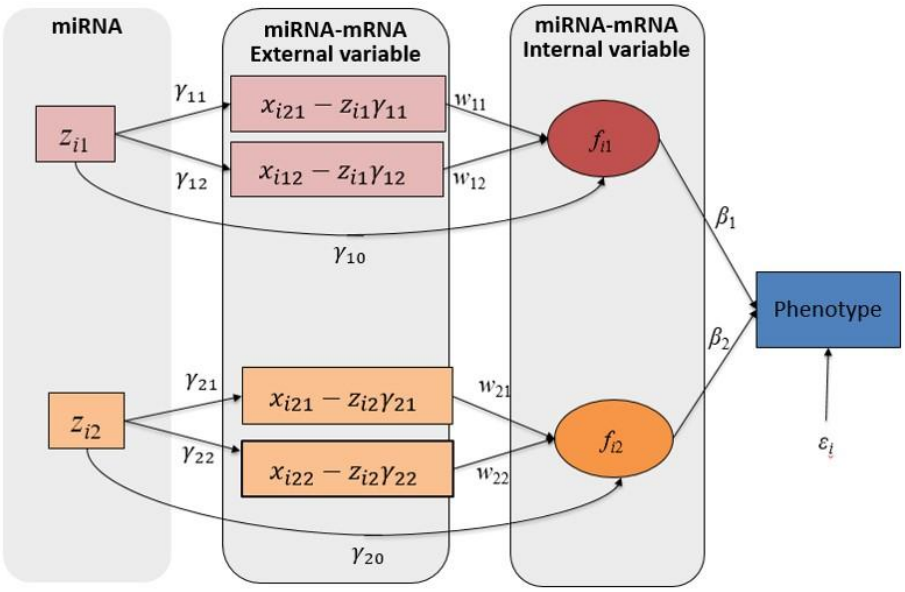

Fig. 4.2 Network Diagram for HisCoM-mimi model

As shown in Fig. 4.2, there are three structures to consider: miRNA-mRNA structure, miRNA integration latent structure, and phenotype-latent structure. Each structure can be represented as a generalized linear model, similar to PHARAOH (38).

$$\hat{X}_{ijk} = x_{ijk} - \gamma_{jk} z_{ij},\ j = 1, \ldots, G_j, \tag{4}$$

Above Eq. (4) shows how to obtain mRNA expression before inhibition by miRNA, subscript i means i th individual, x_{ijk} represents the mRNA expression of the kth gene related with j th miRNA, z_j the j th miRNA expression, γ_{jk} the inhibition coefficient for the j th miRNA for the k th gene, and G_j is the number of inhibited mRNAs by the j th miRNA. By estimating the coefficients γ_{jk}, mRNA expression after removing the inhibition effect of miRNA can be obtained.

$$f_{ij} = \gamma_{j0} z_j + \sum_{k=1}^{G_j} \hat{X}_{ijk} w_{jk} \tag{5}$$

The miRNA latent variable is defined in Eq. (5). The miRNA latent variable is built by linearly combining miRNA expression values. while γ_{j0} denotes the direct effect of the miRNA on the phenotype. Then, the latent variable f_{ij} represents the global effect of the miRNA′s activity through its inhibited mRNAs.

$$logit(\pi_i) = \beta_0 + \sum_{j=1}^{J} \left[\gamma_{j0} z_j + \sum_{k=1}^{G_j} \hat{X}_{ijk} w_{jk} \right] \beta_j = \beta_0 + \sum_{j=1}^{J} f_{ij} \beta_j \tag{6}$$

Let the phenotype variable y_i be a binary variable, distinguishing PDAC from normal tissues. Let π_i be the probability of $y_i = 1$ (PDAC).

$logit(\pi_i)$ is the logit link function, β_j represents the effect of f_{ij} on the phenotype, as interpreted as a log-odds ratio.

4.2.2 Fitting the HisCoM-mimi model for binary

To estimate the parameters for HisCoM-mimi, we adopted our previously developed PHARAOH algorithm (38), which is based on the alternating least squares algorithm for the penalized log-likelihood function, with ridge parameters. Then, the objective function to maximize is given as follows:

$$logit(\pi_i) = \beta_0 + \sum_{j=1}^{J}\left[\gamma_{j0}z_j + \sum_{k=1}^{G_j}\hat{X}_{ijk}w_{jk}\right]\beta_j = \beta_0 + \sum_{j=1}^{J} f_{ij}\beta_j, \quad (7)$$

$$\varphi_1 = \sum_{i=1}^{n} \log p(y_i; \beta_j, \delta) - \frac{1}{2}\lambda_m \sum_{j=1}^{J}\sum_{k=1}^{G_j} w_{jk}^2 - \frac{1}{2}\lambda_{mm}\sum_{j=0}^{J}\beta_j^2 \quad (8)$$

where $p(y_i; \gamma_i, \delta)$ is the probability distribution for the phenotype of the ith individual. λ_m and λ_{mm} are ridge parameters for miRNA-mRNA pairs of interest, representing the integrated latent components.

To maximize the objective function, φ_1, the iterative reweighted least squares (IRWLS) algorithm is used. Note that when using IRWLS, maximizing φ_1 is equivalent to minimizing the object function φ_2.

$$\varphi_2 = \sum_{i=1}^{n} \mathrm{v_i}(z_i - \sum_{j=1}^{J} f_{ij}\beta_j)^2 - \frac{1}{2}\lambda_m \sum_{j=1}^{J}\sum_{k=1}^{G_j} w_{jk}^2 - \frac{1}{2}\lambda_{mm}\sum_{j=0}^{J}\beta_j^2 \quad (9)$$

4.2.3 Comparative models for binary phenotype

To compare the results of HisCoM-mimi with other methods, we considered several alternative regression-based methods.

$$logit(\pi_i) = \beta_0 + \sum_{j=1}^{J} \theta_j z_{ij} + \sum_{k=1}^{K} \rho_k x_{ijk}\,,\ j = 1, \dots, J \tag{10}$$

$$\varphi_{LR}(\beta_0, \theta, \rho, \delta; X, Z) = \sum_{i=1}^{n} \log p(y_i; \beta_0, \theta, \rho) - \delta P_\alpha(\theta, \rho),\ j = 1, \dots, J \tag{10}$$

Firstly, we considered the ordinary penalized logistic regression (LR) methods such as lasso or elastic-net (EN) (49, 50). Eq. (9) shows the LR model, where θ_j and ρ_k represent the effect of the jth miRNA and the kth mRNA, respectively. Eq. (11) is the objective function to maximize for finding optimal parameters with the penalty function $P_\alpha(\theta, \rho)$. When lasso is used, $P_\alpha(\theta, \rho)$=$\sum_{\mathrm{k}} |\rho_k| + \sum_{\mathrm{j}} |\theta_j|$. If EN is used, $P_\alpha(\theta, \rho) = \alpha\left(\sum_{\mathrm{k}} |\rho_k| + \sum_{\mathrm{j}} |\theta_j|\right) + (1-\alpha)\left(\sum_{\mathrm{k}} \rho_k^2 + \sum_j \theta_j^2\right)$. Lasso or EN can then select the miRNAs and/or mRNAs of interest. However, these methods cannot use group information. Thus, ordinarily penalized LR methods cannot adequately account for the biological structure of miRNA-mRNA.

Secondly, we considered LR with a group lasso penalty (GL) (51), which has the benefit of using group information among the miRNAs and mRNAs of interest. In our analysis, a group can be defined as a set of one miRNA and its corresponding inhibited target mRNAs. GL uses the same likelihood in Eq. (11) with a different penalty function $P(\theta, \rho) = \sum_{j=1}^{J} \sqrt{\theta_j^2 + \sum_{k=1}^{G_j} |\rho_k|}$. Via this penalty function, miRNA integration set can be selected together. However, the GL approach does not easily provide p-values for each set of independent variables.

To fit the penalized LR models, we first performed 3-fold cross-validation to find the optimal tuning parameter, δ. after which we fitted the models with all the data sets.

4.2.4 Simulation Study

To compare HisCoM-mimi to the other three methods, we performed simulation studies and computed type I errors and power, simulating data from the same miRNA and mRNA data structure in our pancreatic cancer dataset. That is, we selected miRNA and mRNA data from the pancreatic cancer dataset, and then generated phenotype data iteratively from the LR model. We then considered two simulation scenarios. Scenario 1 assumed that a true causal integration set contains two mRNAs, with the same effect size. Scenario 2 assumed that a true causal integration set contains five mRNAs, with the same effect size. For each scenario, we randomly selected one causal miRNA-mRNA subnetwork, and then randomly selected another 9 miRNA-mRNA subnetworks, for which the number of inhibited mRNAs was less than 10. The selected miRNA-mRNA subnetworks for Scenario 1 are summarized in Table 4.1 and for Scenario 2 are in Table 4.2.

For Scenario 1, we used miR-217 as a true causal miRNA. To generate phenotypes, we considered the following LR model.

$$logit(\pi) = \beta_{miRNA} z_1 + \beta_1 x_1 + \beta_2 x_2, \quad (12)$$

where π is the probability of observing a disease $(Y = 1)$, z_1 represents the true causal miRNA expression, and x_1 and x_2 represent two

causal mRNA expression values. For type I error evaluation, we assumed $\beta_{miRNA} = \beta_1 = \beta_2 = 0$. For power comparison, we generated simulation data sets under the assumption that $\beta_{miRNA} = \beta_1 = 0.2,\ 0.25, 0.3, 0.35$. For the given 114 (97 PDAC and 17 normal tissues) values of $(z_1,\ x_1,\ x_2)$, from our pancreatic cancer dataset, we simulated 1,000 datasets.

For Scenario 2, we assumed that a true causal integration set contains five mRNAs, with the same effect size. In our dataset, miR-381 was the only miRNA having five inhibited target mRNAs. To generate phenotypes, we considered the following LR model:

$$logit(\pi) = \beta_{miRNA} z_1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5, \qquad (13)$$

where $x_1, \ldots, x_5$ represent five causal mRNA expression values. As in Scenario 1, we assumed $\beta_{miRNA} = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$, for type I error evaluation, and $\beta_{miRNA} = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0.2, 0.25, 0.3, 0.35$, for power comparison. For the given 114 values of $(z_1,\ x_1,\ x_2,\ x_3,\ x_4,\ x_5)$ from the pancreatic cancer dataset, 1,000 simulation datasets were generated. We used the significance level $\alpha = 0.05$ for HisCoM-mimi, as an FPR criterion. For lasso, EN, and group-lasso, we selected a threshold T which provides a comparable FPR to the type I error 0.05. T was determined by calculating the FPR for simulation settings such that a miRNA-mRNA subnetwork is selected when $\beta_{miRNA} \neq 0$ and $K(= \sum_{l=1}^{L} I(\beta_l \neq 0))$ exceeded the threshold T. Here, L is the number of inhibited mRNAs for true causal miRNA for each scenario: L = 2 for Scenario 1, and L = 5 for Scenario 2.

Table 4.1 List of used miRNAs and mRNAs for simulation Scenario 1

miRNA	Role in simulation	Inhibited mRNA
miR-217*	Causal	ITGBL1, ATP10A
miR-215	Non-Causal	CDC6, CTH, DNAJC19, DPP10, ELP4, FUNDC2,GLP1R, B3GALNT2, SLC39A8
miR-485	Non-Causal	CDX1, CTDNEP1, GPR3, HDAC5, KCNJ11, RASL10A,SLC39A14
miR-195	Non-Causal	CNDP2, SLC45A2, SLC7A2
miR-381	Non-Causal	DKK3, IGFBP5, LAMA4, OSBPL3, BAMBI
miR-132	Non-Causal	GLRB, GMPR, ARX, SALL3
miR-363	Non-Causal	SOSTDC1
miR-1	Non-Causal	FAM150B
miR-28	Non-Causal	SRPRB
miR-200	Non-Causal	NRG3

Table 4.2 List of used miRNAs and mRNAs for simulation Scenario 2

miRNA	Role in simulation	Inhibited mRNA
miR-381	Causal	DKK3, IGFBP5, LAMA4, OSBPL3, BAMBI
miR-215	Non-Causal	CDC6, CTH, DNAJC19, DPP10, ELP4, FUNDC2, GLP1, B3GALNT2, SLC39A8
miR-32	Non-Causal	COL1A2, BGN
miR-195	Non-Causal	CNDP2, SLC45A2, SLC7A2
miR-501	Non-Causal	PARM1, SLC32A1
miR-1	Non-Causal	FAM150B
miR-212	Non-Causal	KCNK2
miR-204	Non-Causal	CDH11
miR-200	Non-Causal	NRG3
miR-363	Non-Causal	SOSTDC1

4.3 Results

4.3.1 Simulation results

For our analyses, we first determined the false positive error rates (FPRs) of each method and chose the threshold values of T to make each penalized method provide (hold) FPRs close to 0.05. In Scenario 1, the type I error rate of HisCoM-mimi was 0.048 when $\alpha = 0.05$. The FPRs of lasso were 0.054, when T was 1, and that of EN was 0.064, when T was 1. Since type I error rates of lasso and EN were nearly 0.05 when T=1, we set T=1 to evaluate power of those two methods. The FPR of GL, when choosing a causal miRNA integration set, 0.064.

For Scenario 2, Table 4.3 shows the FPRs for lasso and EN, when varying the threshold T. For this result, we found that the type I error of lasso and EN were similar to 0.05, when T=1 and 2, respectively. The type I error rate of HisCoM-mimi was 0.054. On the other hand, GL did not select a causal miRNA integration set at all, such that the type I error rate was 0. Secondly, we compared the powers of each method for Scenarios 1 and 2. Fig. 4.3 shows bar plots of powers for scenario 1, where the x-axis shows the effect sizes (i.e., beta coefficients), and the y-axis shows the power. HisCoM-mimi showed the highest power, while EN was second, Lasso was third, and GL was last. The same tendency is shown in Fig. 4.4, for Scenario 2. Fig. 4.5 shows that the differences of power between HisCoM-mimi and the others were much larger than those of Scenario 1. Consequently, GL could not find

any significant miRNA-mRNA integration sets under Scenario 1, due to its GL′s penalty being too strict for many mRNAs, whose beta values were small.

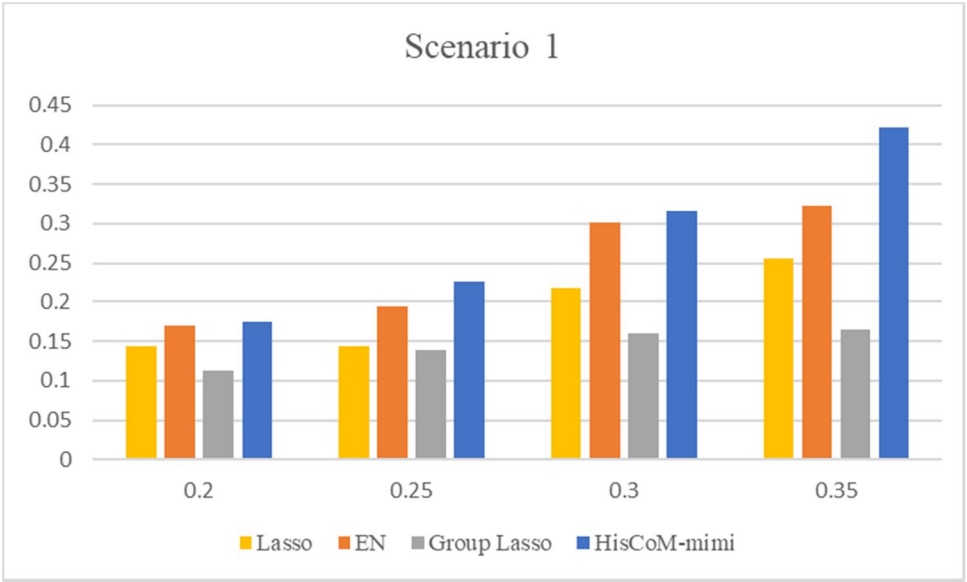

Fig. 4.3 Power comparison for scenario 1

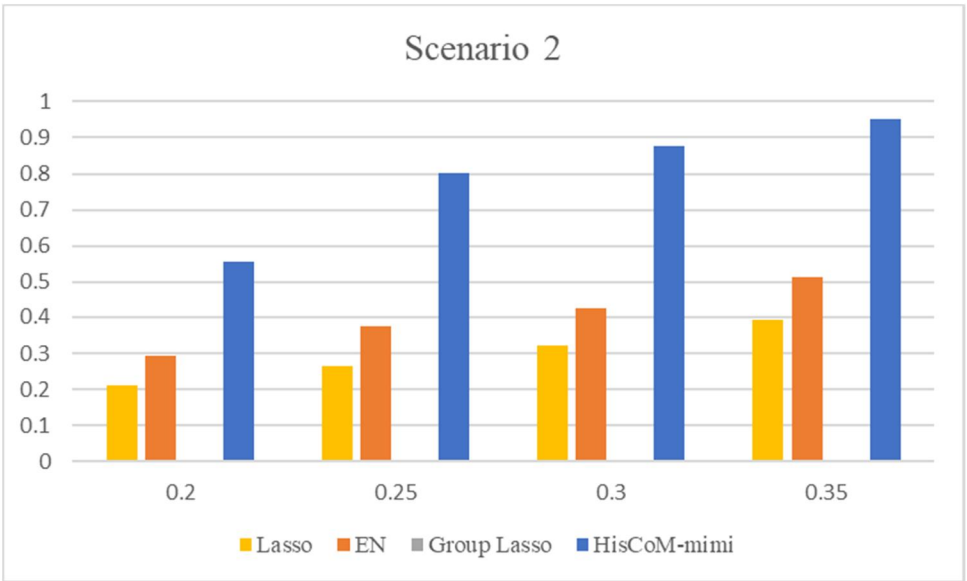

Fig. 4.4 Power comparison for scenario 2

Table 4.3 False positive rate when varying the number of selected mRNAs for lasso and EN

T	5	4	3	2	1
Lasso	0	0	0.007	0.022	0.053
EN	0	0.002	0.014	0.055	0.204

4.3.2 Constructing miRNA-mRNA subnetwork for SNUH PDAC dataset

To use human mRNA and miRNA probes, we first filtered out non-annotated mRNA probes and non-human miRNA probes. After filtering, there were 22,077 mRNA probes and 3,391 miRNA probes. To construct miRNA-mRNA subnetworks, we checked predicted target mRNAs, for each miRNA, from TargetScan 7.1 (targetscan.org) (48, 52). Among predicted targets, we only selected mRNAs having significant Pearson correlation coefficients with a specific miRNA. After filtering, there were 55 miRNAs, and 2,411 edges connected with mRNAs.

4.3.3 Integration analysis for the SNUH PDAC data

Table 4.3. shows the top significant weights of miRNA-mRNA integrations derived from HisCoM-mimi. To perform multiple comparison, we used FDR q-values summarized in the 7th column (53). We could only find twelve miRNAs having q-values below 0.05. Tables 4.5 and 4.6 show the lists of the selected markers by lasso and EN, respectively. Since lasso and EN select markers without any group information, they selected miRNA and mRNA markers independently. There were no miRNAs selected by lasso or EN directly, with lasso yielding only two significant mRNAs, both related to miR-326. Other mRNAs were independently selected from different miRNAs. Consequently, there were only 12 markers selected by lasso. For EN, 58 mRNAs were selected. Similar to the lasso result, there were no selected

miRNAs, although four miRNAs (miR-206, miR-3064, miR-222, and miR-326) connected to more than three mRNAs. Fig. 4.5 shows a Venn diagram of the number of miRNAs selected by each method. Each number represents the total number of detected miRNAs and one in the parenthesis does the number of detected miRNAs whose relationship with pancreatic cancer were reported. HisCoM-mimi selected larger number of unique miRNAs and the majority of them were already were reported.

Table 4.4 Significant miRNAs produced by HisCoM-mimi.

Order	miRNA	number of inhibited miRNAs	Number of significant miRNAs	β_{mimi}	$P_{HisCoM\text{-}mimi}$	$q_{HisCoM\text{-}mimi}$
1	miR-133b	81	29	0.319	0.0008	0.0126
2	miR-141	105	57	0.638	0.0008	0.0126
3	miR-222	127	70	0.587	0.0010	0.0126
4	miR-532	11	0	0.190	0.0010	0.0126
5	miR-93	80	36	-0.573	0.0014	0.0126
6	miR-219	26	3	0.278	0.0016	0.0126
7	miR-590	24	4	-0.183	0.0016	0.0126
8	miR-326	13	0	0.172	0.0022	0.0151
9	miR-203	65	11	-0.261	0.0026	0.0159
10	miR-132	4	0	-0.204	0.0034	0.0187
11	miR-96	109	42	0.701	0.0038	0.0190
12	miR-708	43	3	-0.181	0.0102	0.0468

Table 4.5 Selected markers by lasso. 12 markers (12 mRNAs) were selected. No miRNAs were selected.

Selected marker	beta	Connected miRNA
NSD1	-0.704	miR-206
EMX2	-0.336	miR-222
BBC3	0.329	miR-222
GSG1	0.005	miR-3064
ZRANB3	-0.414	miR-326
MLEC	0.051	miR-362
PLCE1	0.129	miR-1271
TFCP2	0.112	miR-497
AKAP7	-0.017	miR-1297
MAMDC2	1.044	miR-670
DRGX	0.393	miR-96
FBXL2	-0.187	miR-133b

Table 4.6 Markers selected by EN.

Selected mRNA	Beta	Connected miRNA	Selected mRNA	Beta	Connected miRNA
NSD1	-0.340	miR-206	CTRC	-0.007	miR-326
FRS2	-0.046	miR-206	MLEC	0.034	miR-362
MGAT4A	0.004	miR-206	NOTCH1	0.003	miR-367
SLC8A1	0.022	miR-206	SH3PXD2A	0.014	miR-367
PI4KA	0.027	miR-206	PTDSS1	0.017	miR-372
MATR3	0.034	miR-206	CATSPER4	0.002	miR-378
OSBPL8	0.088	miR-206	TRIM55	0.071	miR-378
EMX2	-0.275	miR-222	SLC35E2B	-0.128	miR-488
KIAA0430	-0.039	miR-222	SALL4	-0.080	miR-1271
AXIN2	0.003	miR-222	MAGI3	0.009	miR-1271
PRUNE	0.013	miR-222	PLCE1	0.198	miR-1271
SHISA9	0.016	miR-222	TFCP2	0.216	miR-497
SHC3	0.031	miR-222	KDM5B	0.040	miR-524
RBL1	0.044	miR-222	RNASEH2C	-0.043	miR-670
SOCS1	0.053	miR-222	MAP3K10	0.163	miR-670
SH3BP4	0.057	miR-222	MAMDC2	0.395	miR-670
BBC3	0.074	miR-222	TCEB3	-0.286	miR-93

SEC23IP	0.077	miR-222	RASL11B	0.036	miR-93
ESR1	0.085	miR-222	KIAA0087	0.182	miR-96
DGKI	-0.003	miR-330-5p	DRGX	0.249	miR-96
NUP214	-0.103	miR-3064	HS3ST2	0.016	miR-100
TCP11	-0.077	miR-3064	SYDE2	0.098	miR-107
BCL2L13	-0.022	miR-3064	AKAP7	-0.207	miR-1297
SLC16A10	-0.016	miR-3064	FBXL2	-0.373	miR-133b
GSG1	0.034	miR-3064	CLIP2	0.005	miR-141
LRRC34	-0.159	miR-326	LYPD3	0.188	miR-152
ZRANB3	-0.127	miR-326	PAQR9	0.308	miR-152
AQP2	-0.037	miR-326	SCN1A	0.017	miR-203
CCPG1	0.070	miR-211	BGN	-0.161	miR-32

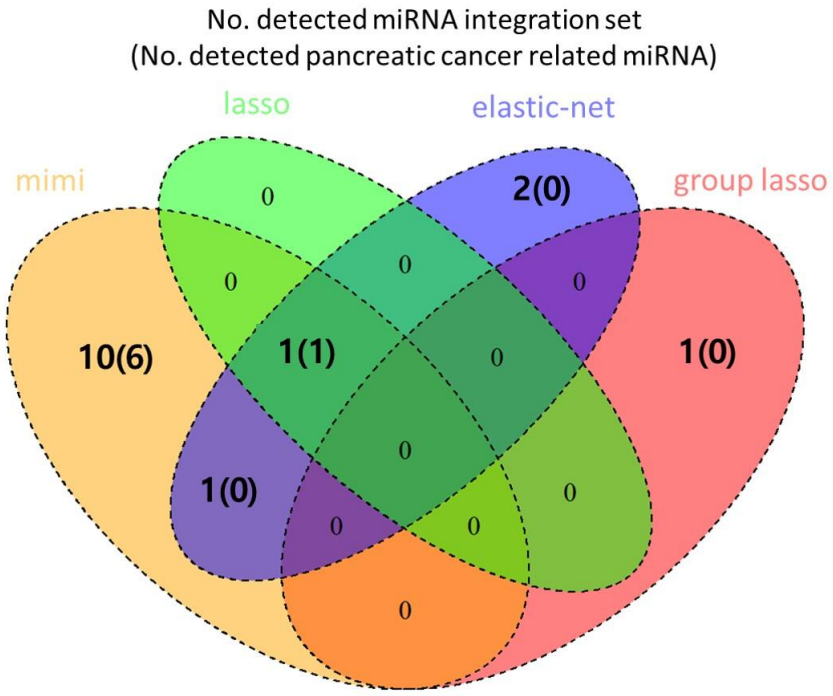

Fig. 4.5 Venn Diagram for number of detected miRNAs for each method

For the lasso group only one miRNA (miR-32) and whose related two mRNA (COL1A2, and BGN) were selected. Although miR-32 is not reported as pancreatic cancer marker, there were some reports that miR-32 is related with other cancers (54, 55).

Table 4.7 summarizes miRNAs detected by HisCoM-mimi, lasso, EN, or GL. Previously, miR-93, miR-219, miR-141, miR-222, miR-203, miR-132, miR-96, and miR-206 were reported to be pancreatic cancer-related markers (39, 56-63). Although other miRNAs detected by HisCoM-mimi, lasso, EN, or GL have not been reported for pancreatic cancer relation, miR-532, miR-590, miR-133b, miR-326, miR-708, miR-3064, and miR-32 were reported to associate with other cancer types (54, 64-70)

Table 4.7 Cancer related miRNAs detected by methods

Method	miRNA	Number of used mRNA	Reported cancer Relationship	Method	miRNA	number of used mRNA	Reported cancer relationship
HisCoM-mimi	miR-93	80	Pancreas	HisCoM-mimi	miR-132	4	Pancreas
HisCoM-mimi	miR-219	26	Pancreas	HisCoM-mimi	miR-96	109	Pancreas
HisCoM-mimi	miR-532	11	Other	HisCoM-mimi	miR-708	43	Other
HisCoM-mimi	miR-590	24	Other	Lasso	miR-222	2	Pancreas
HisCoM-mimi	miR-141	105	Pancreas	EN	miR-206	7	Pancreas
HisCoM-mimi	miR-133b	81	Other	EN	miR-222	12	Pancreas
HisCoM-mimi	miR-222	127	Pancreas	EN	miR-3064	5	Other
HisCoM-mimi	miR-203	65	Pancreas	EN	miR-326	4	Other
HisCoM-mimi	miR-326	13	Other cancer	GL	miR-32	2	Other

Table 4.8 shows the cross-validation (CV) results for comparing prediction performance for marker-sets selected by HisCoM-mimi, Lasso, EN, and Group Lasso. The first column indicates methods used to construct prediction model and the second column does the method to select marker sets. The third column shows the area under the Receiver Operating Characteristic curve (AUC) results performed by leave-one-out cross validation (LOOCV). This setting is from the previous study of Kwon et al (52). The fourth column indicates the average AUC values performed by four-fold CV with a hundred iterations. Here, we used four-fold and eight-fold CV to balance the number of samples in CV datasets. The fifth column indicates the average AUC values performed by eight-fold CV with a hundred iterations. For all selected marker-sets, all prediction models built by HisCoM-mimi showed the best performances yielding AUC values higher than 0.9 except the marker-set selected by Group lasso in which the number of markers is less than five and one path coefficient exists.

Table 4.8 Evaluation of Prediction performance for marker set selected by HisCoM-mimi, Lasso, EN, or Group Lasso in PDAC samples

Marker set	Method	AUC-loocv	AUC-4-fold CV	AUC-8-fold CV
HisCoM-mimi	HisCoM-mimi	**0.997**	**0.996**	**0.997**
	Lasso	0.948	0.947	0.948
	EN	0.975	0.969	0.971
	Group Lasso	0.889	0.888	0.895
Lasso	HisCoM-mimi	0.976	0.975	0.976
	Lasso	0.938	0.928	0.939
	EN	0.970	0.953	0.963
	Group Lasso	0.910	0.910	0.918
EN	HisCoM-mimi	0.976	0.976	0.976
	Lasso	0.939	0.927	0.935
	EN	0.969	0.957	0.965
	Group Lasso	0.911	0.912	0.915

4.4 Discussion

In this chapter, we proposed and developed a novel method, hierarchical structured component analysis of microRNA-mRNA integration ("HisCoM-mimi"), to construct a component model to identifying significantly integrated miRNA-target-mRNA cognate pairs. Since HisCoM-mimi could use subgroup information, it yelded more results, as related to phenotypes (e.g. cancer, metabolic syndrome, and etc.), than those of other existing methods that lack network information.

In simulation studies, we compared the performances of HisCoM-mimi, lasso, EN, and GL. From that comparison, HisCoM-mimi showed better performance than the other three methods. Controlling type I error, by HisCoM-mimi, was easier for controlling FPRs than other methods, because HisCoM-mimi uses permutation based p-values. In particular, HisCoM-mimi could identify miRNA-mRNA integration sets in a much more flexible way, due to better use of a standard multiple testing framework, as compared to the other methods. In real data analysis, HisCoM-mimi succesfully identified more miRNA-mRNA integration sets for pancreatic ductal adenocarcinoma (PDAC) diagnosis, compared to the other methods. Among 12 miRNAs, whose q-values were below 0.05 by HisCoM-mimi, 7 miRNAs were previously reported to associate with a panreatic cancer (39, 56-63). EN found two miRNAs (miR-222, and miR-206) [30,34]. Among two miRNAs selected by lasso, only miR-222 was reported to associate with pancreatic cancer.

Although HisCoM-mimi worked well for the PDAC data sets, further biological verification of those results are needed. In future studies, we will perform additional simulation analyses to evaluate the performance of HisCoM-mimi, under numerous conditions. Furthermore, HisCoM-mimi can be extended in many ways, for other types of phenotypes, such as time to event. Second, it can be easily applied to other cancer studies to identify miRNA-mRNA integration sets for early diagnosis and prognosis. Third, it can be extended to combine other types of omics data such as genomics, epigenomics, and proteomics data. It is now established that dysregulated miRNAs play substantial roles in a myriad of diseases (71). We firmly believe that these methods for miRNA identification and their target transcripts could yield effective biomarkers and therapeutic targets, in addition to providing better understanding of disease mechanisms and etiology.

Chapter 5

Hierarchical structural component miRNA-mRNA integration model for survival phenotype (mimi-surv)

5.1 Introduction

In chapter 5, we proposed a hierarchical structured component analysis of miRNA-mRNA integration (HisCoM-mimi) to interpret how miRNAs affect phenotypes, indirectly, by statistically regulating the expression of distinct mRNAs. Since HisCoM-mimi is based on a generalized linear model (GLM), it can be applied to many phenotypes following an exponential family distribution. Thus, when we previously applied HisCoM-mimi to discriminate miRNA-mRNA expression being from cancerous vs. normal tissues, and this method showed more biological and intuitive interpretations than other

methods. However, HisCoM-mimi cannot be directly applied to survival time analysis, because it is based on a GLM.

In this chapter, we propose a miRNA-mRNA integration model for survival time (″mimi-surv″), which is an extension of HisCoM-mimi to the survival phenotype based on a Cox model (17, 72). mimi-surv is also a component-based analysis, such as, generalized structure component analysis GSCA, and PHARAOH (73).

We applied our new approach, mimi-surv, to both microarray-based data from PDAC patients, at SNUH, and high-throughput sequencing data, obtained from data deposited in The Cancer Genome Atlas (TCGA). From those data, we performed survival analysis on integrated miRNA-mRNA sets, using mimi-surv to find prognostic factors for survival after surgery of PDAC.

Many studies have identified cell-free miRNAs as prognostic markers of PDAC (46). However, although some prognostic miRNAs have been identified, few studies have interpreted role their precise role(s) in the progression of PDAC. Finally, we performed simulation to compare type I error and power with other survival data analysis methods.

5.2 Methods

5.2.1 mimi-surv model

Our mimi-surv model has three structures: miRNA-mRNA structure, miRNA integration latent structure, and phenotype-latent structure. For

miRNA-mRNA structure, we considered linear combination structures for miRNA and target mRNAs, according to the Eq. (14):

$$\hat{X}_{ijk} = x_{ijk} - \gamma_{jk} z_j,\ j = 1, \dots, G_j, \tag{14}$$

where x_{ijk} is the mRNA expression of the kth gene, z_j is the jth miRNA expression, γ_{jk}, the inhibition coefficient for the jth miRNA for the kth gene, and G_j is the number of inhibited mRNAs by the jth miRNA. By estimating the miRNA inhibition coefficients γ_{jk}, mRNA expression, after removing the inhibition effect of miRNA, can be obtained.

$$f_{ij} = \gamma_{j0} z_j + \sum_{k=1}^{G_j} \hat{X}_{ijk} w_{jk} \tag{15}$$

The miRNA latent variable, f_{ij} defined in Eq. (15), is built by linearly combining miRNA expression values, while γ_{j0} denotes the direct effect of the miRNA on the phenotype. Then, the latent variable, f_{ij}, represents the global effect of the miRNA′s activity, as measured by its inhibition of its target mRNA(s) expression.

To construct the phenotype-latent structure, we assume that the survival time be associated with the risk factors under a Cox proportional hazards model(72). The Cox proportional hazards (Cox-PH) model is specified as Eq. (16):

$$h(y_i|F_i) = h_0(y_i)\, exp\left(\sum_{j=1}^{J}\left[\sum_{k=1}^{G_j}\gamma_{j0}z_j + \hat{X}_{ijk}w_{jk}\right]\beta_j\right)$$
$$= h_0(Y)\, exp\left(\sum_{j=1}^{J} f_{ij}\beta_j\right) \quad (16)$$

Here the phenotype variable y_i denotes the survival time and $h(y_i|F)$ denotes the hazard function of the ith sample. In addition, $h_0(Y)$ is the baseline hazard function, and β_j represents the effect of f_{ij} on the hazard rate, as a risk factor.

5.2.2 Fitting the mimi-surv model

To estimate the parameters for mimi-surv, we adopted our previously developed HisCoM-mimi algorithm, which is based on the alternating least squares algorithm for the penalized log-likelihood function, with penalty parameters. Then, the objective function to be maximized is given as follows:

$$\phi = \sum_{i:C_i=1}\left(\sum_{j=1}^{J} f_{ij}\beta_j - log\sum_{l:Y_l\geq Y_i} exp\left(\sum_{j=1}^{J} f_{lj}\beta_j\right)\right)$$
$$-\frac{1}{2}\lambda_m\sum_{j=1}^{J}\sum_{k=1}^{G_j}P_{\lambda_{mm}}(w_{jk}) - \frac{1}{2}\lambda_{mm}\sum_{j=0}^{J}P_{\lambda_m}(\beta_j) \quad (17)$$

where the first summation part represents partial likelihoods of observations, and λ_m and λ_{mm} are ridge parameters for the miRNA-mRNA pairs of interest, representing the integrated latent components (72).

To maximize the objective function, ϕ, an alternating least squares algorithm is used as following steps.

Let $\eta = XWB$, $u = \frac{\partial l}{\partial \eta}$, $A = -\frac{\partial^2 l}{\partial \eta \eta^T}$, $z = \eta + A^{-1}u$, then

$$l(W, B) \approx (z - \eta)^T A (z - \eta) \tag{18}$$

1) Fix s_W, s_B and initialize $\hat{B}, \widehat{W}$=0.
2) Compute η, u, A and z based on the current value of $\hat{B}, \widehat{W}$.
3) Minimize $(z - \eta)^T A (z - \eta)$ with fixed $\widehat{W}$ subject to $\sum |W_{ij}| \leq s_W$.
4) Repeat steps 2,3 until $\hat{B}$ does not changes
5) Compute η, u, A and z based on the current value of $\hat{B}, \widehat{W}$
6) Minimize $(z - \eta)^T A (z - \eta)$ with updated $\hat{B}$ subject to $\sum |B_i| \leq s_B$.
7) Repeat steps 5,6 until $\widehat{W}$ does not changes.
8) Repeat steps 2-7 until l(W,B) does not changes.

To perform statistical inference, a permutation test is used to obtain p-values. and q-values for multiple testing adjustment (53). Since the number of miRNA integration set is around a few decades, the number of permutations is set 10,000.

5.2.3 Comparative model for survival phenotype

To compare the results of mimi-surv with those of other methods, we considered several types of Cox proportional hazards model (Cox-PH). Firstly,

we considered a single miRNA Cox-PH model (single), as the following equation for each miRNA, z_{ij}.

$$h(y_i|z_{ij}) = h_0(y_i)\, exp(\delta_j z_{ij})\,,\ j = 1, \dots, J \tag{19}$$

Secondly, we considered penalized Cox-PH regression models such as ridge, lasso, elastic-net (EN), and group lasso (grplasso) (49, 50, 74). Eq. (20). shows the objective function for ridge, lasso and EN Cox-PH models.

$$\varphi_1 = \sum_{i:C_i=1}\left(\sum_{j=1}^{J} \delta_j z_{ij} - log \sum_{l:Y_l \geq Y_i} exp\left(\sum_{j=1}^{J} \delta_j z_{ij}\right)\right) - P_\theta(\delta_j), \tag{20}$$

where $P_\theta(\delta_j) = \theta \sum_{j=1}^{J} \delta_j^2$ for ridge, $P_\theta(\delta_j) = \theta \sum_{j=1}^{J} |\delta|$ for lasso, and $P_\theta(\delta_j) = \theta \left(\frac{1}{2}\sum_{j=1}^{J} \delta_j^2 + \sum_{j=1}^{J} |\delta_j\right)$ for EN. Here θ is the tuning parameter to adjust the strength of the penalty function.

For a grplasso Cox-PH model (51), using the group information from the miRNAs and mRNAs, the following regression model is given

$$\begin{gathered} h(Y|F) = h_0(Y)\, exp\left(\sum_{j=1}^{J} \delta_j z_{ij} + \sum_{j=1}^{J} \sum_{k=1}^{G_j} \lambda_{jk}\, \hat{x}_{ijk}\right), \\ \text{(subject to } \left(|\delta_j| + \sum_{k=1}^{G_j} |\lambda_{jk}|\right) \leq t) \end{gathered} \tag{21}$$

To find the optimal tuning parameter θ, we performed 10-fold cross-validation and then determined the value of θ, which minimizes the value of the objected function for the validation set.

5.2.4 Simulation Study

To compare which method had a better power to discover the true signal miRNA-mRNA pair, we performed simulation studies to compute type I

errors and power of mimi-surv and grplasso, using the miRNA expression values of the SNUH PDAC dataset. In this simulation, we did not consider lasso, EN, or ridge methods because they cannot find significant miRNA-mRNA integration set together by using group information. To generate a simulation dataset, we used the same simulation settings as we did for our previous HisCoM-mimi analysis (17). We generated two kinds of simulation data sets, considering two scenarios to compare the statistical power of each method. We next assumed a true model as given in eq. (9) (below). We considered that only one true signal miRNA, having an effect size of β_{miRNA}, and its regulated target mRNAs, having the same effect size, β_{mRNA}. We then considered two scenarios: (1) a true miRNA containing two regulated mRNAs; and (2) an assumption of five mRNAs. Table 5.1 lists the miRNAs and their regulated mRNAs.

$$h(Y|X,Z) = h_0(Y)exp(\beta_{miRNA}z_1 + \sum_{i=1}^{p} \beta_{mRNA}\hat{x}_i) \quad (22)$$

For each scenario, we used thirteen miRNAs and twenty-seven mRNAs. The miRNAs having less than six target mRNAs, in our identification of mRNA-miRNA pair analysis, and their target mRNAs were included in simulation analysis. For scenario one, we chose miR-212, miR-219, or miR-32 as causal for regulating two mRNAs, in each simulation. For each causal miRNA, we generated 1,000 datasets, and calculated the average type I error and power for this first scenario. For scenario two, we chose miR-204 as a

causal miRNA, and generated 1,000 datasets to calculate the power and type I error.

For type I error evaluation, we assumed that $\beta_{miRNA}, \beta_{mRNA} = 0$. For power comparison, we generated simulation data sets under the assumption that $\beta_{miRNA}, \beta_{mRNA} = 0.1, 0.2$ or 0.3. For the given 95 patients′ expression levels, we simulated 1,000 datasets. We used a significance level of $\alpha = 0.05$.

Table 5.1 miRNA-mRNA lists used in simulation

miRNA name	Regulated mRNA lists
miR-204	GRIN2B, HMGA2, ARNTL2, ACADL, TDRD6
miR-212	PAX5, SHISA9
miR-219	HMGA2, EGR3
miR-32	PRKAB2, SNX2
miR-25	EPHA3
miR-362	PLAT, SMAD2, CHRD1
miR-373	VHL
miR-1297	MCL1, RLF, C20orf24, EDEM3
miR-485	STRBP
miR-508	NR5A2
miR-132	SHANK2
miR-133b	SLC7A2
miR-200b	SLIT2, BNC2, CDH11

5.3 Results

5.3.1 Simulation results

In our simulation study, we first determined the type I error of each method, for several simulation settings (Fig. 5.1). The first x-axis shows the censoring fraction of each simulation setting, and the second x-axis shows the number of target mRNA of the miRNAs. These results show that mimi-surv, and grplasso, effectively yield a type I error of 0.05. However, when the censoring fraction was 0.15, and the number of mRNAs was 5, mimi-surv with ridge penalty had a slightly inflated type I error. On the contrary, mimi-surv with lasso penalty had deflation in type I error for the less power than the other methods.

Fig. 5.2A shows a bar plot comparing the power of each method when $\beta_{mRNA} = 0.2$, the censoring fraction was 0.15, and the number of target mRNAs connected to the causal miRNAs was two. In this case, mimi-surv with ridge penalty showed the largest power between all the compared methods. Ridge had second largest power in the compared methods. Lasso, mimi-EN and mimi-lasso had smaller power than the other methods. These tendencies were constantly maintained when changing β_{mRNA}, β_{miRNA}, or the number of connected mRNAs.

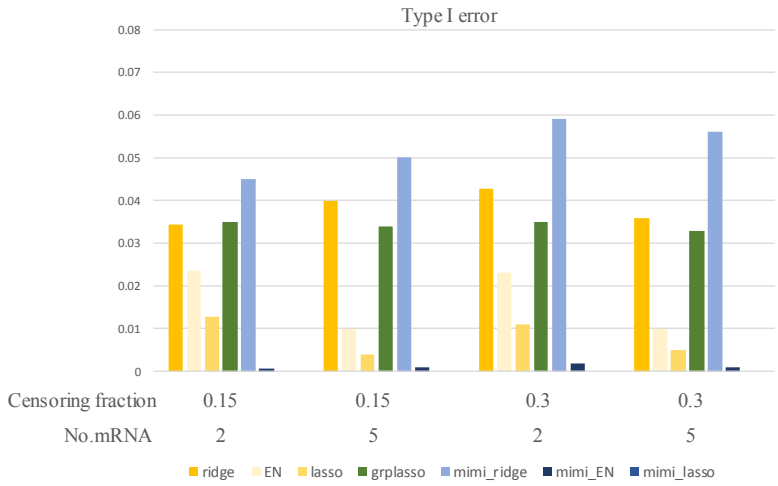

Fig. 5.1. Bar plots of method for comparing type I error. First six bars show the type I error of each method when the number of mRNAs targeted by each miRNA is two.

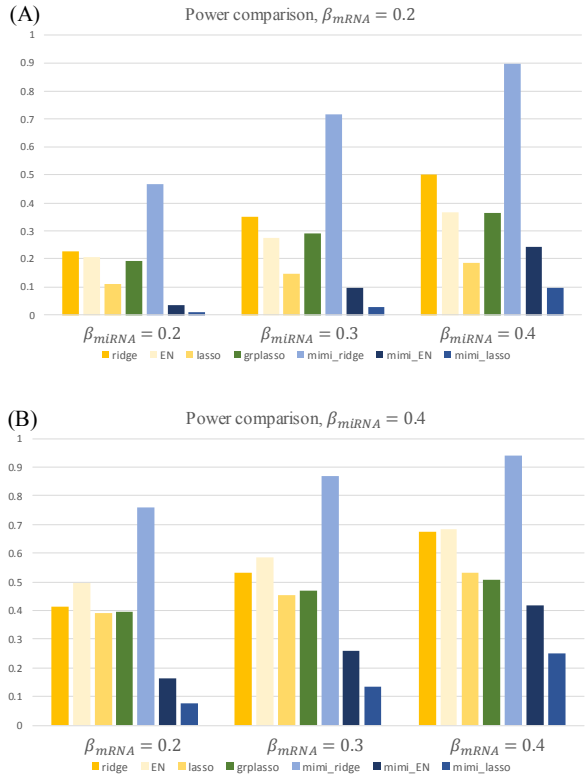

Fig. 5.2. Bar plots of the seven methods for comparing power when $\beta_{mRNA} = 0.2$ and censoring fraction is 0.15. (A) shows the results when, and the number of target mRNAs is two, and (B) shows the results when the number of target mRNAs is five.

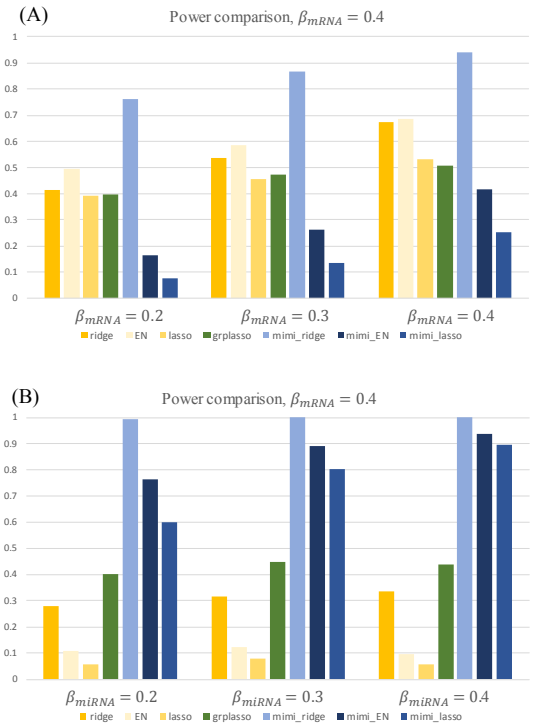

Fig. 5.3. Bar plots of the seven methods for comparing power when $\beta_{mRNA} = 0.4$ and censoring fraction is 0.15. (A) shows the results when, and the number of target mRNAs is two, and (B) shows the results when the number of target mRNAs is five.

Fig. 5.2B shows the bar plot which compared the power of each method when the number of mRNAs connected to target miRNA was five. As Fig. 5.2A results, mimi-surv with ridge penalty showed the largest power. Unlike the results from Fig 5.2A, mimi-EN had larger power than ridge, EN, lasso, and grplasso. The same tendency was observed for various values of β_{miRNA} and β_{mRNA}.

Fig 5.3 shows the power results when $\beta_{mRNA} = 0.4$ and the censoring fraction was 0.15. Fig 5.3A and 5.3B show the cases when the number of target mRNAs connected to the causal miRNAs was two and five, respectively. The patterns are similar to those in Fig 5.2 showing that mimi-surv with ridge penalty showed the largest power among all the compared methods. In this case, mimi-EN and mimi-lasso performed better than ridge, EN, lasso, and grplasso when the number of target mRNAs is five.

5.3.2 SNUH dataset analysis results

In constructing a procedure to identify miRNA-mRNA integration sets, 54 miRNA-mRNA integrations were constructed. For the set of 54 integrations, we applied mimi-surv, and compared other methods for their ability to find significant miRNA-mRNA integration. In real data analysis, we focused on finding significant miRNAs with different methods. Thus, we compared significant miRNA lists from single, ridge, lasso, EN, grplasso, and mimi-surv. Fig. 5.4 shows a Venn diagram comparing the number of miRNAs detected by each method. The number without brackets shows the number of miRNAs reported in other studies, and those within brackets show the total

number of miRNAs that were found significant by each method. As shown, single marker analysis rejected more miRNAs than the other methods. In total, the mimi-surv methods detected six miRNAs. Among them, four miRNAs were reported in other PDAC analyses (46, 75-78), while the penalized Cox-PH method rejected less miRNAs than the other methods, and ridge had the largest rejection rate. Among the rejected miRNAs, miR-204 was commonly detected by all methods. It has also been reported that the expression patterns of miR-204 for PDAC stage 1 samples were different from those of stage II-IV samples (46) and additionally, miR-204 was used to distinguish solid pseudopapillary tumors from pancreatic malignancies (75).

Table 5.2 shows information for the miRNAs detected by mimi-surv. To find significant markers, we used nominal p-values. Unfortunately, when we applied a family-wise error rate or false discovery rate, no markers were rejected by penalized regression-based methods or mimi-surv. Although miR-93 had more than 901 mRNAs contained in same set, only 7 miRNAs were found by elastic net, and 9 miRNAs were found significant by lasso. These results show that we could reduce the number of candidate miRNA-mRNA combinations when using mimi-surv. Although the simulation study showed that performance of mimi-surv with ridge penalty had better power than other penalties, mimi-surv with EN or lasso penalty detected more miRNAs in real PDAC data analysis. Although simulation studies were performed only for small miRNA-mRNA networks, numerous miRNA-mRNA networks contained more than one hundred mRNAs. Thus, these differences between

simulation studies and real data analyses were due to the number of mRNAs connected with significant miRNAs.

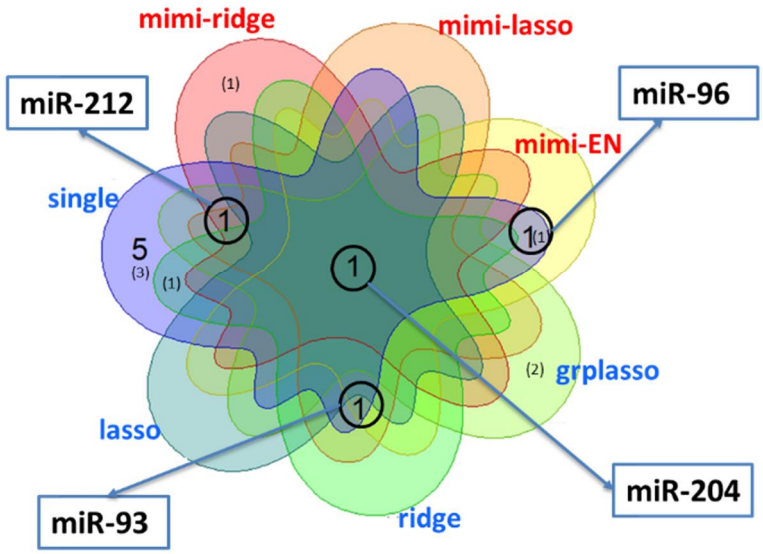

Fig. 5.4. Venn diagram for the number of miRNAs detected by each method in analysis of PDAC data from SNUH. The numbers without brackets show the numbers of miRNAs found in other PDAC analyses, while those within brackets show the number of miRNAs not previously identified.

Table 5.2. Information about the miRNAs detected by mimi-surv.

No	miRNA	number of inhibited mRNAs	number of significant mRNAs	β_{mimi}	P_{mimi}	*Penalty*
1	miR-204	5	0	-0.018	**0.015**	ridge
			1 (GRIN2B)	-0.179	**0.004**	lasso
			1 (GRIN2B)	-0.142	**0.031**	EN
2	miR-93	901	9	-0.406	**0.012**	lasso
			7	-0.544	**0.003**	EN
3	miR-212	2	0	0.015	**0.045**	ridge
			1 (PAX5)	0.008	**0.033**	lasso
4	miR-96	34	2 (GPM6B, EPHA3)	0.209	**0.017**	EN
5	miR-497	189	2 (LRRC14, PHF13)	-0.252	**0.036**	EN
6	miR-339	46	0	0.024	**0.045**	ridge

5.3.3 TCGA dataset analysis results

We downloaded a TCGA PDAC dataset from the Genomic Data Commons (GDC) data portal of the U.S. National Cancer Institute (https://portal.gdc.cancer.gov/) (47). For mRNA and miRNA expression profiling, Illumina Hiseq was used. To normalize mRNA-seq and miRNA-seq datasets, Fragments Per Kilobase Million (FPKM) was measured for each read count. We first collected 185 TCGA PDAC data samples for analysis. In sample quality control procedure, we excluded 25 non-PDAC samples, and 47 PDAC samples whose follow-up time was less than 3 months, because the cause of death was hard to determine. The PDAC patients′ average age was 63.9 years (standard deviation (SD): 11.1 years), 48 patients were male, and 64 were female. The median survival time was 585 days

In our procedure for constructing miRNA-mRNA integration sets, 52 miRNA-mRNA integrations were constructed, to which we applied mimi-surv, and other compared methods, for our TCGA PDAC data analysis. Fig. 5.5 shows a Venn diagram containing the number of miRNAs detected by each method. For the TCGA data analysis, the penalized regression methods could not find any miRNAs. However, miR-93 was detected by mimi-surv, with both lasso and EN. miR-93 was detected not only in the TCGA data analysis, but also in SNUH data analysis. Among our detected miRNAs, miR-129, miR-200a, miR-200c, miR-25, and miR-93 were previously reported in other PDAC studies (75, 76, 79-81).

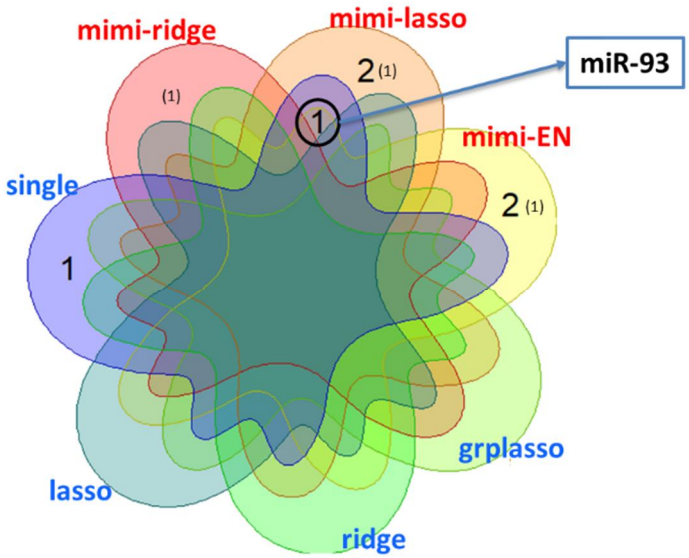

Fig. 5.5. Venn-diagram for the number of miRNAs detected by each method in analysis of TCGA data

5.4 Discussion

In this chapter, we proposed a new method, mimi-surv, to find significant miRNA-mRNA pairs associated with survival time. In simulation studies, we compared the performances of mimi-surv, ridge, lasso, EN, and group lasso. To find more suitable penalties, we compared the performances of mimi-surv with ridge, lasso, and EN. From that comparison, mimi-surv with ridge penalty performed better than the other methods. In analysis of real pancreatic ductal adenocarcinoma (PDAC) data, from both Seoul National University Hospital (SNUH) and The Cancer Genome Atlas (TCGA), mimi-surv performed well, in identifying miRNA-mRNA integration sets for survival time (Cancer Genome Atlas Research, et al., 2013).

Although mimi-surv worked well for the PDAC data sets, it could not find significant miRNA sets using multiple testing criteria. Since our mimi-surv method is more complicated than the others we assessed, it is difficult to find significant miRNAs with a small number of patients. Also, our results from permutation-based p-values indicate it takes too many iterations to check p-values less than 10^{-7}

Chapter 6

Summary and Conclusions

In this study, we propose a structured component-based analysis, for integrating omics data for identifying multiple accurate biomarkers. It is well known that miRNAs affect phenotypes indirectly, by regulating mRNA expression or protein translation (13). Herein, we propose hierarchical structured component analysis of miRNA-mRNA integration (HisCoM-mimi) analysis, which models biological relationships as structured components, to efficiently yield integrated markers.

In chapter 3, we proposed and developed a novel method, hierarchical structured component analysis of microRNA-mRNA integration ("HisCoM-

mimi″), to construct a component model to identifying significantly integrated miRNA-target-mRNA cognate pairs. Since HisCoM-mimi could use subgroup information, it yelded more results, as related to phenotypes (e.g. cancer, metabolic syndrome, and etc.), than those of other existing methods that lack network information.

In simulation studies, we compared the performances of HisCoM-mimi, lasso, EN, and GL. From that comparison, HisCoM-mimi showed better performance than the other three methods. In particular, HisCoM-mimi could identify miRNA-mRNA integration sets in a much more flexible way, due to better use of a standard multiple testing framework, as compared to the other methods. In real data analysis, HisCoM-mimi succesfully identified more miRNA-mRNA integration sets for PDAC diagnosis, compared to the other methods. Among 12 miRNAs, whose q-values were below 0.05 by HisCoM-mimi, 7 miRNAs were previously reported to associate with a panreatic cancer (39, 56-63). EN found two miRNAs (miR-222, and miR-206) [30,34]. Among two miRNAs selected by lasso, only miR-222 was reported to associate with pancreatic cancer.

In chapter 4, we proposed a different type of HisCoM-mimi, mimi-surv, to find significant miRNA-mRNA pairs associated with survival time. In simulation studies, we compared the performances of mimi-surv, ridge, lasso, EN, and group lasso. To find more suitable penalties, we compared the performances of mimi-surv with ridge, lasso, and EN. From that comparison,

mimi-surv with ridge penalty performed better than the other methods. In analysis of real PDAC data, from SNUH and TCGA, mimi-surv performed well, in identifying miRNA-mRNA integration sets for survival time (Cancer Genome Atlas Research, et al., 2013).

In summary, miRNA-mRNA integration analysis can increase the power to detect association signals in cancer. Using miRNA without mRNA expression cannot give information how miRNA works at cancer. On the contrary, using mRNA without miRNA expression could miss the right signaling pathway for cancer relationship. In this thesis, we suggest appropriate miRNA-mRNA integration method which find the right signaling pathway for cancer relationship of miRNA-mRNA interaction. Thus, our proposed method, HisCoM-mimi could find more cancer related miRNA and their target mRNAs efficiently than other methods.

Although our methods could find biologically meaningful marker combinations efficiently, it cannot be found markers which have non-linear relations between phenotype and other omics markers. Also, we need to find deeper omics integrations such as integration between SNP, mRNA, miRNA, and proteins. In further study, we will apply HisCoM-mimi to integrate more than two levels of omics dataset and provide deeper biological interpretation with multi-omics analysis.

Bibliography

1. Reis-Filho JS. Next-generation sequencing. Breast cancer research : BCR. 2009;11 Suppl 3:S12.
2. Bock C, Reither S, Mikeska T, Paulsen M, Walter J, Lengauer T. BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. Bioinformatics. 2005;21(21):4067-8.
3. Anderson L, Hunter CL. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Mol Cell Proteomics. 2006;5(4):573-88.
4. Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Bio. 2006;7(3):198-210.
5. Gilbert-Diamond D, Moore JH. Analysis of gene-gene interactions. Curr Protoc Hum Genet. 2011;Chapter 1:Unit1 14.
6. Kim Y, Lee S, Choi S, Jang JY, Park T. Hierarchical structural component modeling of microRNA-mRNA integration analysis. BMC Bioinformatics. 2018;19(Suppl 4):75.
7. Crick F. Central Dogma of Molecular Biology. Nature. 1970;227(5258):561-&.
8. Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. 2014;15(8):509-24.
9. Enerly E, Steinfeld I, Kleivi K, Leivonen SK, Aure MR, Russnes HG, et al. miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors. Plos One. 2011;6(2):e16915.
10. Farazi TA, Hoell JI, Morozov P, Tuschl T. MicroRNAs in Human Cancer. Adv Exp Med Biol. 2013;774:1-20.
11. Kang SM, Lee HJ. MicroRNAs in human lung cancer. Exp Biol Med. 2014;239(11):1505-13.
12. Navarro A, Monzo M. MicroRNAs in Human Embryonic and Cancer Stem Cells. Yonsei Med J. 2010;51(5):622-32.
13. Negrini M, Ferracin M, Sabbioni S, Croce CM. MicroRNAs in human cancer: from research to therapy. J Cell Sci. 2007;120(11):1833-40.
14. Nam S, Li M, Choi K, Balch C, Kim S, Nephew KP. MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. Nucleic acids research. 2009;37(Web Server issue):W356-62.
15. Buffa FM, Camps C, Winchester L, Snell CE, Gee HE, Sheldon H, et al. microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. Cancer Res. 2011;71(17):5635-45.
16. Cho JH, Gelinas R, Wang K, Etheridge A, Piper MG, Batte K, et al. Systems biology of interstitial lung diseases: integration of mRNA and microRNA expression changes. BMC medical genomics. 2011;4:8.

17. Kim. Y, Lee. S, Choi. S, Jang. J-Y, Park T. Hierarchical structural component modeling of microRNA-mRNA integration analysis. BMC Bioinformatics. 2018;In press.
18. Freeman WM, Walker SJ, Vrana KE. Quantitative RT-PCR: pitfalls and potential. Biotechniques. 1999;26(1):112-22, 24-5.
19. Chang TW. Binding of cells to matrixes of distinct antibodies coated on solid surface. J Immunol Methods. 1983;65(1-2):217-23.
20. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270(5235):467-70.
21. Wang D, Carroll GT, Turro NJ, Koberstein JT, Kovac P, Saksena R, et al. Photogenerated glycan arrays identify immunogenic sugar moieties of Bacillus anthracis exosporium. Proteomics. 2007;7(2):180-4.
22. Yetisen AK, Akram MS, Lowe CR. Paper-based microfluidic point-of-care diagnostic devices. Lab Chip. 2013;13(12):2210-51.
23. Banta-Wright SA, Steiner RD. Tandem mass spectrometry in newborn screening: a primer for neonatal and perinatal nurses. J Perinat Neonatal Nurs. 2004;18(1):41-58; quiz 9-60.
24. Lange V, Picotti P, Domon B, Aebersold R. Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol. 2008;4:222.
25. Alam SI, Uppal A, Gupta P, Kamboj DV. Multiple-reaction monitoring for multiplex detection of three bacterial toxins using liquid chromatography-tandem mass spectrometry. Lett Appl Microbiol. 2017;64(3):217-24.
26. Luo P, Dai W, Yin P, Zeng Z, Kong H, Zhou L, et al. Multiple reaction monitoring-ion pair finder: a systematic approach to transform nontargeted mode to pseudotargeted mode for metabolomics study based on liquid chromatography-mass spectrometry. Anal Chem. 2015;87(10):5050-5.
27. Consortium M, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature biotechnology. 2006;24(9):1151-61.
28. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nature biotechnology. 2010;28(8):827-38.
29. Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu TM, Bao W, et al. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nature biotechnology. 2006;24(9):1140-50.
30. Arikawa E, Sun Y, Wang J, Zhou Q, Ning B, Dial SL, et al. Cross-platform comparison of SYBR Green real-time PCR with TaqMan PCR, microarrays and other gene expression measurement technologies evaluated in the MicroArray Quality Control (MAQC) study. BMC genomics. 2008;9:328.

31. Chen JJ, Hsueh HM, Delongchamp RR, Lin CJ, Tsai CA. Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. BMC bioinformatics. 2007;8:412.
32. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28(16):2184-5.
33. Wilson CL, Miller CJ. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. Bioinformatics. 2005;21(18):3683-5.
34. Lee EK, Yi SG, Park T. arrayQCplot: software for checking the quality of microarray data. Bioinformatics. 2006;22(18):2305-7.
35. Kim G-t, Kim Y, Kwon M-S, Park T, editors. Quality control plot for high dimensional omics data. Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on; 2016: IEEE.
36. Hwang HS, Takane Y. Generalized structured component analysis. Psychometrika. 2004;69(1):81-99.
37. Vandenwollenberg AL. Redundancy Analysis an Alternative for Canonical Correlation Analysis. Psychometrika. 1977;42(2):207-19.
38. Lee S, Choi S, Kim YJ, Kim BJ, Consortium Td-G, Hwang H, et al. Pathway-based approach using hierarchical components of collapsed rare variants. Bioinformatics. 2016;32(17):i586-i94.
39. Greither T, Grochola LF, Udelnow A, Lautenschlager C, Wurl P, Taubert H. Elevated expression of microRNAs 155, 203, 210 and 222 in pancreatic tumors is associated with poorer survival. International journal of cancer. 2010;126(1):73-80.
40. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. Ca-Cancer J Clin. 2013;63(1):11-30.
41. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res. 2014;74(11):2913-21.
42. Oh CM, Won YJ, Jung KW, Kong HJ, Cho H, Lee JK, et al. Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2013. Cancer Res Treat. 2016;48(2):436-50.
43. Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. Nature. 2013;501(7467):328-37.
44. Namkung J, Kwon W, Choi Y, Yi SG, Han S, Kang MJ, et al. Molecular subtypes of pancreatic cancer based on miRNA expression profiles have independent prognostic value. Journal of gastroenterology and hepatology. 2016;31(6):1160-7.
45. Frampton AE, Krell J, Jamieson NB, Gall TM, Giovannetti E, Funel N, et al. microRNAs with prognostic significance in pancreatic ductal adenocarcinoma: A meta-analysis. Eur J Cancer. 2015;51(11):1389-404.
46. Debernardi S, Massat NJ, Radon TP, Sangaralingam A, Banissi A, Ennis DP, et al. Noninvasive urinary miRNA biomarkers for early detection of pancreatic adenocarcinoma. Am J Cancer Res. 2015;5(11):3455-66.

47. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20.
48. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. eLife. 2015;4.
49. Tibshirani R. Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met. 1996;58(1):267-88.
50. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Roy Stat Soc B. 2005;67:301-20.
51. Meier L, van de Geer SA, Buhlmann P. The group lasso for logistic regression. J R Stat Soc B. 2008;70:53-71.
52. Kwon MS, Kim Y, Lee S, Namkung J, Yun T, Yi SG, et al. Integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer. BMC genomics. 2015;16 Suppl 9:S4.
53. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res. 2001;125(1-2):279-84.
54. Yang D, Ma M, Zhou W, Yang B, Xiao C. Inhibition of miR-32 activity promoted EMT induced by PM2.5 exposure through the modulation of the Smad1-mediated signaling pathways in lung cancer cells. Chemosphere. 2017;184:289-98.
55. Chen R, Liao JY, Huang J, Chen WL, Ma XJ, Luo XD. Downregulation of SRC Kinase Signaling Inhibitor 1 (SRCIN1) Expression By MicroRNA-32 Promotes Proliferation and Epithelial-Mesenchymal Transition in Human Liver Cancer Cells. Oncology research. 2017.
56. Chen S, Chen X, Sun KX, Xiu YL, Liu BL, Feng MX, et al. MicroRNA-93 Promotes Epithelial-Mesenchymal Transition of Endometrial Carcinoma Cells. PloS one. 2016;11(11):e0165776.
57. Lahdaoui F, Delpu Y, Vincent A, Renaud F, Messager M, Duchene B, et al. miR-219-1-3p is a negative regulator of the mucin MUC4 expression and is a tumor suppressor in pancreatic cancer. Oncogene. 2015;34(6):780-8.
58. Xu L, Li Q, Xu D, Wang Q, An Y, Du Q, et al. hsa-miR-141 downregulates TM4SF1 to inhibit pancreatic cancer cell invasion and migration. International journal of oncology. 2014;44(2):459-66.
59. Lee CL, He H, Jiang YJ, Di Y, Yang F, Li J, et al. Elevated expression of tumor miR-222 in pancreatic cancer is associated with Ki67 and poor prognosis. Med Oncol. 2013;30(4).
60. Park JK, Henry JC, Jiang J, Esau C, Gusev Y, Lerner MR, et al. miR-132 and miR-212 are increased in pancreatic cancer and target the retinoblastoma tumor suppressor. Biochem Biophys Res Commun. 2011;406(4):518-23.
61. Zhang S, Hao J, Xie F, Hu X, Liu C, Tong J, et al. Downregulation of miR-132 by promoter methylation contributes to pancreatic cancer development. Carcinogenesis. 2011;32(8):1183-9.

62. Feng J, Yu J, Pan X, Li Z, Chen Z, Zhang W, et al. HERG1 functions as an oncogene in pancreatic cancer and is downregulated by miR-96. Oncotarget. 2014;5(14):5832-44.
63. Keklikoglou I, Hosaka K, Bender C, Bott A, Koerner C, Mitra D, et al. MicroRNA-206 functions as a pleiotropic modulator of cell proliferation, invasion and lymphangiogenesis in pancreatic adenocarcinoma by targeting ANXA2 and KRAS genes. Oncogene. 2015;34(37):4867-78.
64. Hu S, Zheng Q, Wu H, Wang C, Liu T, Zhou W. miR-532 promoted gastric cancer migration and invasion by targeting NKD1. Life sciences. 2017;177:15-9.
65. Bai L, Wang H, Wang AH, Zhang LY, Bai J. MicroRNA-532 and microRNA-3064 inhibit cell proliferation and invasion by acting as direct regulators of human telomerase reverse transcriptase in ovarian cancer. PloS one. 2017;12(3):e0173912.
66. Sheikholeslami A, Nabiuni M, Arefian E. Suppressing the molecular signaling pathways involved in inflammation and cancer in breast cancer cell lines MDA-MB-231 and MCF-7 by miR-590. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine. 2017;39(4):1010428317697570.
67. Yang D, Zhao D, Chen X. MiR-133b inhibits proliferation and invasion of gastric cancer cells by up-regulating FBN1 expression. Cancer biomarkers : section A of Disease markers. 2017.
68. Li D, Xia L, Chen M, Lin C, Wu H, Zhang Y, et al. miR-133b, a particular member of myomiRs, coming into playing its unique pathological role in human cancer. Oncotarget. 2017.
69. Wu H, Wang Y, Wu C, Yang P, Li H, Li Z. Resveratrol Induces Cancer Cell Apoptosis through MiR-326/PKM2-Mediated ER Stress and Mitochondrial Fission. Journal of agricultural and food chemistry. 2016;64(49):9356-67.
70. Ji S, Zhang B, Kong Y, Ma F, Hua Y. MiR-326 inhibits gastric cancer cell growth through down regulating NOB1. Oncology research. 2016.
71. Esteller M. Non-coding RNAs in human disease. Nature reviews Genetics. 2011;12(12):861-74.
72. Cox DR. Regression Models and Life-Tables. J Roy Stat Soc B. 1972;34(2):187-+.
73. Lee S, Choi S, Kim YJ, Kim BJ, Hwang H, Park T, et al. Pathway-based approach using hierarchical components of collapsed rare variants. Bioinformatics. 2016;32(17):586-94.
74. Simon N, Friedman J, Hastie T, Tibshirani R. A Sparse-Group Lasso. J Comput Graph Stat. 2013;22(2):231-45.
75. Li P, Hu Y, Yi J, Li J, Yang J, Wang J. Identification of potential biomarkers to differentially diagnose solid pseudopapillary tumors and pancreatic malignancies via a gene regulatory network. J Transl Med. 2015;13:361.

76. Cheng Y, Yang H, Sun Y, Zhang H, Yu S, Lu Z, et al. RUNX1 promote invasiveness in pancreatic ductal adenocarcinoma through regulating miR-93. Oncotarget. 2017;8(59):99567-79.
77. Ma C, Nong K, Wu B, Dong B, Bai Y, Zhu H, et al. miR-212 promotes pancreatic cancer cell growth and invasion by targeting the hedgehog signaling pathway receptor patched-1. J Exp Clin Cancer Res. 2014;33:54.
78. Tanaka M, Suzuki HI, Shibahara J, Kunita A, Isagawa T, Yoshimi A, et al. EVI1 oncogene promotes KRAS pathway through suppression of microRNA-96 in pancreatic carcinogenesis. Oncogene. 2014;33(19):2454-63.
79. Wu X, Wu G, Wu Z, Yao X, Li G. MiR-200a Suppresses the Proliferation and Metastasis in Pancreatic Ductal Adenocarcinoma through Downregulation of DEK Gene. Transl Oncol. 2016;9(1):25-31.
80. Bryant JL, Britson J, Balko JM, Willian M, Timmons R, Frolov A, et al. A microRNA gene expression signature predicts response to erlotinib in epithelial cancer cell lines and targets EMT. Br J Cancer. 2012;106(1):148-56.
81. Yu Q, Xu C, Yuan W, Wang C, Zhao P, Chen L, et al. Evaluation of Plasma MicroRNAs as Diagnostic and Prognostic Biomarkers in Pancreatic Adenocarcinoma: miR-196a and miR-210 Could Be Negative and Positive Prognostic Markers, Respectively. Biomed Res Int. 2017;2017:6495867.

초 록

효과적인 다중 마커의 발굴은 개인별 맞춤의학의 시대를 열기 위한 주요한 논제의다. 단일 오믹스 (Omics) 자료를 이용하여 단일 마커를 발굴하는 방법론들이 많이 개발된 것에 비하여 다중 오믹스 자료에 대한 다중마커 발굴 방법론은 아직 많이 개발되지 않았다는 단점이 있다. 길이가 매우 짧은 전사체인 miRNA 는 형질변화에 직접 영향을 끼치지 않지만 다른 단백질로 번역될 수 있는 전사체들의 발현에 제약을 줌으로써 간접적으로 형질변화에 영향을 끼치는 것으로 알려져 있다. 이 학위논문에서는 이러한 간접적인 영향을 생물학적으로 해석할 수 있고 관련된 다중마커를 효과적으로 같이 발굴할 수 있는 구조 방법론에 기반한 miRNA-mRNA 통합 분석 방법론인 (“HisCoM-mimi”) 모형을 제안하고자 한다.

이 학위논문에서는 두가지 종류의 HisCoM-mimi 를 제안한다. 첫번째 HisCoM-mimi 는 판별분석에 사용할 수 있는 방법론이다. 시뮬레이션 결과 HisCoM-mimi 는 다른 비교 방법론보다 더 효율적으로 miRNA-mRNA 다중마커를 발굴하고 판별분석을 진행할 수 있는 것을 확인할 수 있었다. 또한 실제 자료 분석에 있어서 HisCoM-mimi 는 췌장암 환자에 대한 진단분석에서 다른 방법론들에 비해 더 설명력 있는 miRNA-mRNA 다중마커를 발굴하였다.

두번째 종류의 HisCoM-mimi 는 생존분석을 위해 개발된 방법론이다 (mimi-surv). 판별분석을 위한 HisCoM-mimi 의 시뮬레이션 결과와 마찬가지로 다른 방법론에 비해 통계적 검정력이 좋은 것을 확인할 수 있었다. 췌장암 환자에 대한 실제 자료 분석에 있어서도 mimi-surv 는 췌장암 환자의 생존시간과

관련된 miRNA-mRNA 다중마커들을 잘 발굴하는 것을 확인할 수 있었다. 흥미롭게도 췌장암과 관련된 miRNA 로 많이 언급되는 miR-93 의 경우 서울대병원에서 수집한 자료를 이용한 연구결과와 The Cancer Genome Atlas 라는 암환자에 대한 오믹스 자료 통합 데이터베이스의 자료를 이용한 연구결과에서 공통적으로 발굴되는 것을 확인할 수 있었다. 또한 다른 여러 췌장암 관련 miRNA 들을 miRNA-mRNA 관련 구조를 이용한 방법론들에서 더 잘 발굴하는 것을 확인할 수 있었다.

시뮬레이션 결과와 실제 췌장암 자료 분석 결과에서 확인할 수 있듯이 본 학위 논문에서 제안한 HisCoM-mimi 방법론은 여러 오믹스 자료를 이용한 생물학적인 의미가 있는 다중마커 조합을 찾기에 매우 효과적이다.

주요어: 전사체, 소형 전사체, 오믹스 자료 통합, 생존분석, 판별분석

학 번: 2013-30080