Title
Application of Statistical Methods to Integrative Analysis of Genomic Data

Author
Kim, Kyung Pil

Publication Date
2013

Peer reviewed|Thesis/dissertation

Application of Statistical Methods to Integrative Analysis of Genomic Data

by

Kyung Pil Kim

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Haiyan Huang, Chair
Professor Peter J. Bickel
Professor Lewis J. Feldman

Spring 2013

Application of Statistical Methods to Integrative Analysis of Genomic Data

Abstract

Application of Statistical Methods to Integrative Analysis of Genomic Data

by

Kyung Pil Kim

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Haiyan Huang, Chair

The *genomic revolution* has resulted in both the development of techniques for obtaining large quantities of genomic data rapidly and a striking increase in our knowledge on genomics. At the same time, the genomic revolution also created numerous open questions and challenges in analyzing the enormous amount of data required to gain insights on the underlying biological mechanisms. This dissertation addresses these challenges by answering fundamental questions arising from two closely related fields, *functional genomics* and *pharmacogenomics*, utilizing the nature and biology of microarray datasets.

In the *functional genomic* study, we try to identify pathway genes which are a group of genes that work cooperatively in the same pathway constituting a fundamental functional grouping in a biological process. Identifying pathway genes has been one of the major tasks in understanding biological processes. However, due to the difficulty in characterizing/inferring different types of biological gene relationships, as well as several computational issues arising from dealing with high-dimensional biological data, deducing genes in pathways remains challenging. In this study, we elucidate higher level gene-gene interactions by evaluating the conditional dependencies between genes, *i.e.* the relationships between genes after removing the influences of a set of previously known pathway genes. These previously known pathway genes serve as *seed genes* in our model and guide the detection of other genes involved in the same pathway. The detailed statistical techniques involve the estimation of a precision matrix whose elements are known to be proportional to partial correlations (*i.e.* conditional dependencies) between genes under appropriate normality assumptions. Likelihood ratio tests on two forms of precision matrices are further performed to see if a candidate pathway gene is conditionally independent of all the previously known pathway genes. When used effectively, this is shown to be a promising technique to recover gene relationships that would have otherwise gone undetected by conventional methods. The advantage of the proposed method is demonstrated using both simulation studies and real datasets. We also demonstrate the importance of taking into account experimental dependencies in the simulation and real data studies.

In the *pharmacogenomic* study, genetic variants causing inter-individual variation in drug response are investigated. Specifically, signature genes which contribute to the high and low responder variation in statin efficacy are discovered. Using Nonnegative Matrix Factorization (NMF) method, we first identify two distinct molecular patterns between the high and low responder groups. Based on this separation, the modified Significance Analysis Microarrays (SAM) method further searches for signature genes which had gone undetected by the original SAM method. In the biological validation studies, our gene signatures are shown to be significantly enriched with HMGCR-correlated genes. Furthermore, a notable difference is observed in the amount of HMGCR enzymatic activity change between the high and low responder groups - the high responder group shows a bigger activity decrease, implying that statin inhibits the HMGCR enzymatic activity more efficiently in the high responder groups. This helps us understand why the high responder group shows a greater decrease in low density lipoprotein cholesterol (LDLC) level and higher statin efficacy than the low responder group. Overall, the discovered gene signatures are shown to have high biological relevance to the cholesterol biosynthesis pathway, which HMGCR mainly acts on.

To my family, Jiho, Jihoon and husband

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 High throughput genomic data

The advent of high-throughput technologies in genomics and proteomics promotes the generation of enormous amounts of data that are being produced on a daily bases in laboratories around the world. The size, type and structure of these data have also been growing at an unprecedented rate. Gene expression, single nucleotide polymorphism (SNP), copy number variation (CNV) and protein-protein/gene-gene interactions are some examples of genomic and proteomic data produced using high throughput technologies such as microarrays [76], array comparative hybridization, aCGH [63] and mass spectrometry [1]. The amount and type of biological data keeps increasing even further in various fields such as methylation, alternative splicing, transcriptomic and metabolomic.

Each of these distinct data types provides a different, partly independent and partly complementary, view of the whole genome. However, understanding the underlying biological functions of genes, proteins and other aspects of the genome requires more information than provided by each of the datasets. Thus, integrating data from different sources becomes an indispensable part of current research to gain broad interdisciplinary views of proliferating genomic and proteomic field [35, 68, 88]. In functional genomics, for example, defining functions and interactions of all the genes in the genome of an organism is a daunting task and achieving this goal requires integrating information from different experiments [45]. Also, there are efforts to combine similar types of data across different studies, which can be done through meta-analytic approaches. For instance, with the accumulated number of publicly available independent microarray datasets, several applications have proven the utility of data integration by combining studies which address similar hypothesis [3, 14, 69].

The overarching goals of data integration are to obtain more precision, better accuracy and greater statistical power than any individual dataset would provide. Moreover, integration can be useful in comparing, validating and assessing results from different studies and datasets. It is likely that whenever information from multiple independent sources agree, it is more likely for the findings to be valid and reliable than information from a single source

[45].

Nevertheless, many challenges exist in the integrative analysis of these vast amount of diverse types of genomic data. The challenges may be of conceptual, methodological or practical nature and may relate to issues that arise due to computational or statistical complexities. For example, genomic data are often subject to varying degrees of noise, the curse of high dimensionality and small sample size. Furthermore, they are generated from various sources and provided in different format such as vectors, graphs or sequences. Hence, they need to be converted into a common format and dimension before they can be combined to extract the most information out of each dataset. However, data from different sources might have different quality and informativity. Even for similar data types, data from different sources might have different quality depending on the experimental conditions and design.

To fully utilize the benefit of data integration and to best answer to the sophisticated, higher level biological questions arising from the various genomic data, there has been an ever-increasing need for developing novel statistical methodologies and models to handle the enormous amounts of complex and noisy data frequently confronted in genomics.

1.2 Statistical applications in functional genomics

The Human Genome Project was completed in 2003 and it is regarded as the first step in understanding human beings at the molecular level. Though the project is complete, many questions still remain unanswered, including the function of most of the estimated 30,000 human genes.

This naturally brings our attention to *functional genomics* which aims at understanding the functions of genes and their interplay with proteins and the environment to create complex and dynamic living systems. High throughput functional genomic technology is not restricted to only fundamental transcriptome studies, however, and is expanding its potential into biomedical applications such as pharmacogenomics, prognostic biomarker determination and disease subtype discovery. In pharmacogenomics, for instance, gene expression profiles are used to characterize the influence of genomic variation on drug response in patients and classify therapeutic targets to ensure maximum drug efficacy with minimal adverse effects. Such approaches promise the advent of *personalized medicine*, in which drugs and drug combinations are optimized for each individual's unique genetic makeup [82, 40].

Assessing the function of genes can be tackled by different approaches. One avenue of research focuses on gene expression level acquired from repeated measurements of its RNA transcripts. For this purpose, microarray technologies are popularly adopted as the standard tool to obtain a panoramic view of the activity of the genome of entire organisms. Microarray technologies enable us to simultaneously observe thousands of genes in action and to dissect the functions, regulatory mechanisms and interaction pathways of the entire genome [22, 53, 54].

A fundamental component in integrating and understanding the data generated by the microarray technologies is the development of statistical methods. Due to the complicated and high-dimensional nature of the microarray data, statistical modeling of gene expression data often becomes challenging. Despite the intrinsic obstacles of microarrays, statistical methods have shown great promises in microarray analysis. Existing statistical methods could be classified into four categories - *class comparison*, *class discovery*, *class prediction* and *pathway analysis*.

- *Class comparison* The issue in the class comparison is selecting *differentially expressed genes*, *i.e.* genes whose expression is significantly different between conditions. Statisticians have been involved in differential expression analysis since the introduction of microarray technologies. As a result, many methods have been developed. Some are based on parametric models (e.g. *MAANOVA* [95], *limma* [79])[1] while others rely on non-parametric approaches to tackle the difficulties associated with distributional assumptions (e.g. *SAM* [89]). A comparative review of all methods are given in [65, 15]

- *Class discovery* Clustering or unsupervised classification is used in gene expression analysis to identify similarly expressed groups of genes/samples. The results are then verified based on biological rationale. The underlying idea is that the functionally related genes are believed to be either co-expressed or inversely expressed, thus they are likely to be grouped together [23]. Clustering can be applied gene-wise or sample-wise; the former helps identify groups of co-regulated genes while the latter helps identify new biological classes such as new tumor subtypes. Alternatively, clustering can also be done over genes and samples simultaneously. Some of the well known algorithms include *hierarchical clustering* [23], *k-means* [46], *partition around medoids* (*PAM*) and *self organizing maps* (*SOM*) [19].

- *Class prediction* The goal of class prediction is to develop a multivariate function for predicting the status of patients (phenotype) using gene expression profiles. Given expression profiles and phenotypes for patients, a classification rule is built by learning this training dataset. Then the objective is to predict the status of new undiagonized patients according to their expression profiles. As such, class prediction methods play critical role in the biomedical area by assigning tumors to one of predefined subtypes [33] or building a prognostic model for various tumors [24, 90]. There are a huge number of available classification methods, ranging from relatively simple methods *logistic regression*, *discriminant analysis* [20, 33, 21], *k-nearest neighbour* to complex multi-categorical methods *support vector machines* (*SVM*) borrowed from the machine learning field [81, 2].

[1] *MAANOVA* uses analysis of the variance (ANOVA) to capture the main sources of variability in the experiment. *limma* adopts a parametric approach using linear models and empirical bayes.

- *Pathway analysis* *Class comparison* generates long lists of genes which have been selected based on statistical significance. With these gene lists (query), researchers try to find biological interpretation by relating the query with functional annotation databases such as the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). To test the significant enrichment of specific functional categories in a given query, *Gene Enrichment Analysis* (*GE*) and *Gene Set Enrichment Analysis* (*GSEA*) [83] are most popularly adopted.[2] These methods enable us to discover the common biological functions or properties represented within the query which finally provide great biological insight, such as pathway discovery or association studies with disease.

Although massive amount of data has been generated with microarrays, there is no consensus on what is the best quantitative methods to analyze them. Many methods lack appropriate measures of uncertainty, and make dubious distributional assumptions. Furthermore, little is known on how to design informative experiments, how to assess whether the experiment protocols were properly observed, or how to evaluate the reliability of the results obtained. To further exacerbate the situation, the unique character of gene expression data keeps generating many challenging statistical questions. Statisticians and scientists in many disciplines are thus diving in to tackle the urgent need to develop suitable statistical methods in recent days.

1.3 Scope and contributions

The major challenge in using gene expression data is to develop or choose the appropriate statistical methods which can provide immediate but profound biological insights on the variety of questions we have. This dissertation addresses this challenge by answering fundamental questions arising from two closely related fields, *functional genomics* and *pharmacogenomics*, utilizing the nature and biology of microarray data.

Project I (chapter 2) is categorized as a *functional genomic* study. The main purpose of this project is to identify *pathway genes* which are a group of genes that work cooperatively in the same pathway constituting a fundamental functional grouping in a biological process. Identifying pathway genes has been one of the major tasks in understanding biological processes. However, due to the difficulty in characterizing/inferring different types of biological gene relationships, as well as several computational issues arising from dealing with high-dimensional biological data, deducing genes in pathways remains challenging.

This project contributes to the field of pathway gene discovery in the following manner:

- We elucidate higher level gene-gene interactions by evaluating the conditional dependencies between genes, *i.e.* the relationships between genes after removing the influ-

[2] More methods are available in [47, 72]

ences of a set of previously known pathway genes. These previously known pathway genes serve as *seed genes* in our model and will guide the detection of other genes involved in the same pathway.

- Based on our distance measure, we formulate a novel statistical model by estimating a precision matrix whose elements are known to be proportional to partial correlations (*i.e.* conditional dependencies) between genes under appropriate normality assumptions.

- We also take into account the presence of experiment dependencies in the gene expression data when estimating the precision matrix.

- To test the conditional independence of a candidate pathway gene with the seed genes, likelihood ratio tests are performed on two forms of precision matrices.

Both simulation studies and real data analysis confirm that our method outperforms other existing methods. By taking into account the experiment dependencies in the gene expression data, we could better infer pathway gene relationship than the methods which ignore the experiment dependencies. In the flavonoid biosynthesis (FB) pathway studies, our method also identifies genes from neighbouring pathways by considering the indirect relationships between genes. This finding will find its utility in future studies which are targeted to discovering the cooperative nature of genes in the pathways. Overall, we have shown that our method is a promising approach to recover gene relationships that would have otherwise been missed by conventional methods.

Project II (chapter 3) is categorized as a *pharmacogenomic* study. This project aims to study the genetic variants affecting patients' different drug response. Specifically, we try to identify gene signatures which contribute to the high and low responder variation in statin efficacy. When a gene variant is associated with a particular drug response in a patient, one could potentially make clinical decisions based on genetics by adjusting the dosage or choosing a different drug.

This project contributes to the discovery of genetic features in pharmacogenomic study in the following manner:

- To demonstrate the existence of distinct molecular patterns between the high and the low responder groups in gene expression data, Nonnegative Matrix Factorization (NMF) method is adopted which is popularly used in molecular pattern recognition. The result assures the existence of clear-cut separation between the two responder groups.

- To identify signature genes based on the separation defined by the NMF analysis, a new algorithm from Significance Analysis Microarrays (SAM) method is proposed. By stabilizing the variance of the test statistic and estimating the gene-specific null

distribution of the test statistic, a set of signature genes is identified which had gone undetected by the original SAM method.

- In biological validation studies, (*i*) our signature genes are shown to be significantly enriched with HMGCR-correlated genes, and (*ii*) a notable difference is observed in the amount of HMGCR enzymatic activity change between the high and low responders. Specifically, the high responder group shows a bigger activity decrease, implying that statin inhibits the HMGCR enzymatic activity more efficiently in the high responder group. This helps us understand why the high responder group exhibits a greater LDLC decrease and higher statin efficacy than the low responder groups. Overall, these results imply that our gene signature is biologically relevant to the cholesterol biosynthesis pathway, which HMGCR mainly acts on.

With the aid of novel statistical methods such as NMF and SAM, we successfully identified signature genes of our interest. Together with the biological validations we performed, our results will shed a light on understanding the inter-individual variation in statin efficacy.

Chapter 2

Project I: Using Biologically Interrelated Experiments to Identify Pathway Genes in Arabidopsis

2.1 Introduction

2.1.1 Biological pathway and pathway genes

A biological pathway is a series of chemical reactions that form an integral and critical part of every biological process. Pathway genes, or genes involved in the same biological pathway, constitute a fundamental functional grouping in a biological process. A major task in understanding biological processes is to identify a set of genes in the same biological pathways and elucidating the relationships between them.

2.1.2 Overview of existing methods

Using gene expression data, there have been two popular computational approaches for finding pathway genes: *clustering analysis* and *network models*. *Clustering analysis* uses a co-expression measure to quantify similarities in gene expressions and then assigns similar genes into clusters [23]. Genes in each cluster are considered to be functionally related, and thus likely to be in the same pathway. This approach works when the pathway genes exhibit strong co-expressions with one another. *Network models* generally model a pathway as a network, with the genes represented as nodes and the gene relationships represented as edges linking the nodes, e.g. the work in [27]. Starting with a full network, a typical pathway can be identified as a connected (sub)network after all the weak or insignificant edges are removed by a backward edge exclusion technique. Or alternatively, starting with an empty network, strong or significant edges can be added gradually to form a (sub)network using the method of forward inclusion of edges. Both have been widely used in literature to construct biological networks [6, 9, 29, 17, 42, 34, 57, 64, 75, 74, 93].

The property of a network mainly relies on how to evaluate the edges between genes. There are two ways to assign edge weights. One is based on gene covariance matrix, which measures marginal similarity/correlation between any two genes. The other is based on inverse covariance matrix of genes, leading to a graph concerning conditional independence relationships. The latter is equivalent to using partial correlations as similarities.

Despite their appealing features, the approaches described above have limitations. One limitation comes from the high dimensionality of microarray data. The well-known large p, small n problem can result in an unreliable co-expression measure and hence a very high rate of false discoveries in clustering analysis. In the network models, this raises concern about the stability and accuracy of the model inference; it is almost impossible to employ the network models on a genomic scale as the estimation of covariance or its inverse matrices becomes problematic. Although there has been recent work such as regularized network models to overcome this problem [75, 74], the accuracy of the results remains unclear. As such, in practice, these approaches are usually applied to a rather small number of genes or among a small number of clusters of genes preselected based on some prior knowledge, which as a consequence, makes it difficult for us to explore the whole genomic scale of information.

Another concern is related to the limited biological inference of these approaches. Cluster methods are based on a marginal co-expression measure between two genes independent of other genes. Similarly in a network model using covariance matrix, an edge only connects genes with strong marginal correlations. Such approaches potentially expose us to the risk of missing higher level interactions such as group interactions, i.e. gene A interacts with a group of genes but it does not possess any strong relationship with the individual ones. This group interaction is frequently observed in real biological pathways when a group of genes cooperatively regulate one gene. Using the inverse covariance matrix of genes in a network model has a better hope for detecting such kinds of higher level interactions. The inverse covariance matrix is also known as the precision matrix, whose elements have an interpretation in terms of partial correlations (i.e. the correlation between any two genes conditioned on one or several other genes). However, in the current literature, partial correlation is mostly calculated conditioned on either all the available genes or a more-or-less arbitrary subset of them that likely contain noisy (i.e. non-pathway or biologically unrelated) genes. It is reported that conditioning on all genes simultaneously can introduce spurious dependencies which are not from a direct causal effect or common ancestors [29]. This problem may be circumvented to some extent by considering lower order partial correlations, e.g. calculating a partial correlation of two genes conditioned on every other individual variables (first-order partial correlation), and on every other two variables (second-order partial correlation) [29, 55, 93, 92]. However, one concern on lower order partial correlation is its insensitivity for inferring higher level gene associations such as group interactions. More importantly, if the conditioned genes are biologically unrelated, the corresponding conditional dependence properties would be difficult to interpret and verifying the biological relevance of the recovered networks becomes challenging. Further discussions on the adverse effects of conditioning on noisy genes are given in Sections 2.2 and 2.3.

2.1.3 A new pathway gene search algorithm based on partial correlations

We introduce a new pathway gene search algorithm, designed based on evaluating partial correlations between genes, for a particular biological pathway of interest. The motivation of using partial correlation is based on its ability to detect complex gene relationships under appropriate normality assumptions of the data: (i) a strong partial correlation between two genes suggests a direct interaction despite a weak marginal correlation; (ii) a negligible partial correlation suggests no direct relationship after removing influences from other genes and the two genes are conditionally independent.

To overcome the concerns and limitations of current methods for using partial correlations, we require a few (e.g. 3 - 5) preselected biologically related pathway genes, upon which the partial correlation is conditioned on, to guide the search. Specifically, we perform the likelihood ratio tests to see if a candidate gene is conditionally independent of all the preselected known pathway genes.

The requirement of pre-known pathway genes seems a limitation of our approach. However, by incorporating this small amount of biological knowledge, huge advantages on biological inference can be gained and false positive discoveries have been reduced dramatically (Section 2.3). Furthermore, by conditioning on preselected pathway genes, the resulting partial correlation coefficients can be directly interpreted as a similarity measure to the considered pathway. In addition to suggesting satisfying mathematical and biological properties, the proposed approach is also advantageous computationally since we only need to estimate a moderate dimensional precision matrix once for each candidate gene.

Moreover, we also take into account the presence of experiment dependencies in the gene expression data when estimating a precision matrix elements in a precision matrix are proportional to partial correlations between genes [75]. In current studies of gene relationships, the presence of expression dependencies attributable to the biologically interrelated experiments has been widely ignored. When unaccounted for these (experiment) dependencies can result in inaccurate inferences of functional gene relationships, and hence incorrect biological conclusions [86]. Our simulation and real data study supports this conclusion and confirms that considering those dependencies indeed plays a critical role in correctly inferring pathway gene relationships.

2.2 Methods

As we discussed earlier, in contrast to marginal calculation (e.g. the Pearson correlation), partial correlation can work as a more effective tool for inferring complex gene interactions in pathways when it is properly computed.

Below, we first provide a brief review on the concept of partial correlation, followed by a detailed description of a new search strategy, designed based on a likelihood ratio test on partial correlations, for finding pathway genes.

2.2.1 Partial correlation

When an expression matrix (with genes in rows and experimental conditions in columns) is multivariate normally distributed, standard graphical model theory [42] shows that the partial correlation between genes can be equivalently represented by the corresponding elements in the precision matrix $(\Sigma^G)^{-1}$, where Σ^G is the covariance matrix. That is, for a set of genes W , the partial correlation between gene i and gene j can be expressed as

$$\rho_{ij} = cor\left(i, j \,|W \backslash \{i, j\}\right) = \begin{cases} -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}, & i \neq j \\ 1, & i = j \end{cases} \tag{2.1}$$

where ω_{ij} are elements in the inverse gene covariance (or inverse gene correlation) matrix [42, 75]. With the normality assumption on expression measurements, when ρ_{ij} vanishes, two genes i and j are conditionally independent given the remaining genes.

2.2.2 Method motivation

A negligible element in the precision matrix suggests conditional independence between two genes. This motivates us to use precision matrix as a key component in our method for detecting higher level gene interactions, e.g. group gene interactions in a pathway.

However, the successful use of partial correlation highly relies on two issues. *One issue* is about the selection of a proper set of genes upon which the correlation is conditioned on, i.e. $W \setminus \{i, j\}$ in Equation 2.1. When this set of genes contains noisy (i.e. non-pathway) genes, the derived partial correlation would be unreliable for detecting gene relationships. We can see this explicitly through a linear regression interpretation of partial correlation. In terms of linear regression, the partial correlation ρ_{ij} between gene i and j conditioned on a set of genes Z is simply the correlation $cor(\varepsilon_1, \varepsilon_2)$ of the residuals ε_1 and ε_2 resulting from linearly regressing gene i and gene j against the genes in Z, respectively. Assume we have known pre-pathway genes x and y, and a non-pathway gene h that is independent of genes x and y in Z. Now we consider two candidate genes $u = x + y$ and $v = \delta(x + y) + h$ (note that these two equations only represent the expression relationship between the genes), where δ is small and close to 0. Clearly, u is more likely to be a pathway gene due to its direct and strong relationship with two pre-known pathway genes x and y, while v is more likely to be a non-pathway gene since it is almost a replicate of the non-pathway gene h. However, the partial correlations $cor(u, x|Z\backslash\{x\}) = cor(v, x|Z\backslash\{x\}) = cor(u, y|Z\backslash\{y\}) = cor(v, y|Z\backslash\{y\}) = 1$, showing no advantages of gene u over gene v for their partial correlations with the pathways genes x and y. This undesired performance is due to the inclusion of *noisy* genes in the gene set Z upon which the partial correlation was computed. Recognizing this, we decide to build up our approach by conditioning only on a small set of preknown pathway genes to reduce noise in partial correlation estimation. We call this set of pre-known pathway genes as *seed genes*. Though the requirement of seed genes seems a limitation, only 3 - 5 seed genes are really needed for our method to run and generate reliable results. In brief, by incorporating

a small amount of prior biological information, we can gain huge advantages in detecting genes involved in a particular pathway (Section 2.3). Furthermore, in Section 2.3, by using both simulation and real data, we additionally demonstrate the adverse effects of having noisy genes in the set of seed genes in detecting pathways genes.

The other issue critical to the proper use of partial correlation is on the estimation of gene precision matrix (see Equation 2.1). Given a gene expression matrix with genes in rows and experiments in columns, an effective estimation of gene precision matrix is challenging especially when there are experiment dependencies (or when the row-wise and column-wise dependencies co-exist) in the original gene expression. Experiment dependencies can be defined as the dependencies in gene expression between experiments due to the similar or related cellular states induced by the experiments [86]. Such dependencies cause dependent elements in a gene expression vector. When unaccounted for, they can result in inaccurate inferences of gene relationships, and hence incorrect biological conclusions. To take into account the experiment dependencies in partial correlation estimation, we adapt a model and an estimation procedure, named *Knorm* from Teng and Huang (2009), for inferring gene correlation matrix when there are both the gene-wise and experiment-wise dependencies in the gene expression matrix. The main aspect of the framework is the use of a Kronecker product covariance matrix to model the gene-experiment interactions. The *Knorm* estimation is mainly achieved by an iterative estimation of the two covariance matrices: one covariance matrix is estimated through a weighted correlation formula assuming the other covariance matrix is known. In addition, a row subsampling technique (to enable a comparable number of rows and columns in estimation) and a covariance shrinkage technique (to stabilize the estimated covariance matrices) are employed to ensure a robust estimation. Compared with the Pearson coefficient, the *Knorm* correlation has a smaller estimation variance when experiment dependencies exist. More details of incorporating *Knorm* in our estimation procedure are presented in next section.

2.2.3 Likelihood ratio tests for pathway gene search

Let $S = \{g_1, ..., g_k\}$ denote the set of seed genes for a pathway of interest and G denote the set of all genes whose expression measurements in T experiments (each experiment may have > 1 replicates) are available. Usually $|G| \gg |S| = k$ and $T > |S|$. Motivated by the arguments in the above section, we formulate a searching strategy, based on performing likelihood ratio tests, for pathway genes as follows.

(i) We first estimate the experiment correlation matrix Σ^E using the *Knorm* R package provided by Teng and Huang (2009). The input data are the expression measurements of the $|G|$ genes in T experiments, and there are > 1 replicated samples for each experiment. To generate expression matrices, we randomly choose one replicate from each experiment to compose a sample matrix $\mathbf{X}_b$ of dimension $|G| \times T$ and by repeating this process, we generate B sample matrices $\mathbf{X}_1, ..., \mathbf{X}_B$ with B large enough. By the model in Teng and Huang (2009), $\mathbf{X}_b$ is considered to be generated from a multivariate

normal distribution with mean $\mathbf{M}$ (a matrix of dimension $|G| \times T$) and a covariance matrix $\Sigma^G \otimes \Sigma^E$, where Σ^G represents the gene covariance matrix and Σ^E is the experiment correlation matrix. The output of the *Knorm* R package is the estimated $\mathbf{M}$ and Σ^E, denoted as $\hat{\mathbf{M}}$ and $\hat{\Sigma}^E$, by an iterative estimation procedure. More details on the *Knorm* estimation procedure can be found in [86].

(ii) For a candidate gene $g_c \in G \backslash S$ (S is the set of k seed genes), we estimate the gene covariance matrix for genes in $S \cup g_c$ by

$$\hat{\Sigma}_c = \frac{1}{B} \sum_{b=1}^{B} \hat{\Sigma}_{c,b} \tag{2.2}$$

where

$$\hat{\Sigma}_{c,b} = \frac{\left(\mathbf{X}_{c,b} - \hat{\mathbf{M}}\right) \left(\hat{\Sigma}^E\right)^{-1} \left(\mathbf{X}_{c,b} - \hat{\mathbf{M}}\right)'}{T}. \tag{2.3}$$

$\mathbf{X}_{c,b}$ represents one of the sample matrices [of dimension $(k+1) \times T$] constructed by bootstrapping the replicates of each experiment for the expression measurements of the genes in $S \cup g_c$ (the first k rows correspond to the k seed genes and the row $k+1$ corresponds to the candidate gene g_c).

(iii) Obtain the precision matrix, $\hat{\Sigma}^*_{c,1} = (\hat{\Sigma}_c)^{-1}$. Note that as we usually require $T > 1.5k$ so that $\hat{\Sigma}_c$ is usually invertible. When $\hat{\Sigma}_c$ is not invertible, we use its pseudo inverse. $\hat{\Sigma}^*_{c,1}$ will be used as an approximate Maximum Likelihood Estimate (MLE) of the precision matrix under the alternative model in Equation 2.7. We further write

$$\hat{\Sigma}^*_{c,1} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,k} & a_{1,k+1} \\ \vdots & \ddots & \vdots & \vdots \\ a_{k,1} & \cdots & a_{k,k} & a_{k,k+1} \\ a_{k+1,1} & \cdots & a_{k+1,k} & a_{k+1,k+1} \end{bmatrix} \tag{2.4}$$

where $a_{i,j} = a_{j,i}$ for $i = 1, ..., k+1$ and $j = 1, ..., k+1$.

(iv) Obtain matrix $\hat{\Sigma}^*_{c,0}$ from $\hat{\Sigma}^*_{c,1}$ by replacing the offdiagonal elements in the bottom row and rightmost column of $\hat{\Sigma}^*_{c,1}$ by zeros. That is,

$$\hat{\Sigma}^*_{c,0} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,k} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ a_{k,1} & \cdots & a_{k,k} & 0 \\ 0 & \cdots & 0 & a_{k+1,k+1} \end{bmatrix}. \tag{2.5}$$

$\hat{\Sigma}^*_{c,0}$ will be used as an approximate MLE of the precision matrix under the null hypothesis in Equation 2.7.

(v) Perform the following hypothesis test

$$H_0 : \Sigma^* \in \Omega_0 \quad \text{vs} \quad H_1 : \Sigma^* \in \Omega \backslash \Omega_0, \tag{2.6}$$

where Ω_0 is the collection of precision matrices (for the genes in $S \cup g_c$) with zero offdiagonal elements in the bottom row and rightmost column, and Ω is the collection of all possible precision matrices. The null hypothesis assumes conditional independence between the candidate gene and each of the seed genes given all other seed genes. Then the test statistic is

$$\begin{aligned} & -2 \log LR^* \\ = \; & -2 \log \frac{\sup_{\Sigma^* \in \Omega_0} L\left(\Sigma^*; X_1 \cdots, X_B\right)}{\sup_{\Sigma^* \in \Omega} L\left(\Sigma^*; X_1, \cdots, X_B\right)} \\ \approx \; & -2 \left(l\left(\hat{\Sigma}^*_{c,0}; X_1, \cdots, X_B, \hat{M}\right) - l\left(\hat{\Sigma}^*_{c,1}; X_1, \cdots, X_B, \hat{M}\right)\right), \end{aligned} \tag{2.7}$$

where $L(\cdot)$ and $l(\cdot)$ denote the likelihood and the log-likelihood function, respectively. When a candidate gene has no relationship with the pathway seed genes, the corresponding elements in a precision matrix will be close to zeros (i.e. null is true and the test statistic will be small). In contrast, if a candidate gene has a significant association with the pathway genes, those values will be far from zero and naturally the test statistic will be large and declared as significant under the test.

(vi) Repeat steps (ii)-(v) for all candidate genes in $G \backslash S$. Given the test statistic values for all the candidate genes, we rank them in decreasing order. It is a natural interpretation that the higher a candidate gene is ranked in the list, the more likely that gene is associated with the seed genes. Based on the list, we can decide how many of them should be declared as pathway genes using statistical thresholds (see Section 2.3.2.4 for p-value calculations) and/or biological cutoff. We call this method as *pwsrc.knorm*.

If there are questions on which known pathway genes to include as seed genes or on which expression datasets to use, an optional method is to repeatedly run *pwsrc.knorm* to derive a set of *frequently identified* pathway genes under different datasets with different possible sets of seed genes tried. The candidate pathway genes identified this way would be robust against the change of data and the choice of seed genes.

2.2.4 Other methods for comparison

For performance comparison, four additional methods, *pwsrc.null*, *pearson.mean*, *pearson.max* and *GLM* are considered. The first *pwsrc.null*, is designed by replacing $\hat{\Sigma}^E$ in Equation 2.3 with the identity matrix to represent the case ignoring experiment dependencies. *pearson.mean* and *pearson.max* adopt Pearson correlation as a distance measure. Specifically, they calculate pair-wise correlation coefficients between each candidate gene and

the seed genes and take either mean (*pearson.mean*) or maximum (*pearson.max*) of them. *GLM* adopts the regression model of the candidate gene g_c on the seed genes in S as follows:

$$g_c = \alpha_0 + \sum_{j=1}^{|S|} \alpha_j g_{k_j} + \varepsilon, \tag{2.8}$$

where ε is assumed to be normally distributed with zero mean. Since a negligible residue implies a possible interaction between the candidate gene and the seed genes, naturally we can use the residuals as our test statistics (all the genes are scaled to have unit norm before doing regression analysis). For a fair comparison with our method, the experiment dependencies in the gene expression are removed by projecting the data matrix onto the eigenspace of $\hat{\Sigma}^E$ by $\mathbf{X}^* = (\mathbf{X} - \hat{\mathbf{M}}) \cdot (\hat{\Sigma}^E)^{-1/2}$.

2.3 Results

We evaluate the performance of the proposed method in identifying pathway genes using simulation data and genomic scale *Arabidopsis thaliana* datasets obtained from four different types of environmental stresses (oxidation, wounding, UV-B light and drought). We examine the effects of these stresses by focusing on the genes associated with the glucosinolate (GSL) and flavonoid biosynthesis (FB) pathways.

In both studies, we calculated $precision = TP/(TP+FP)$ and $recall = TP/(TP+FN)$ to assess the results from our approach and several other methods mentioned in Section 2.2. Here TP is the number of true positive findings of pathway genes, FP is the number of false positives and FN is the number of false negatives. Note that *precision* and *recall* are popular measures for evaluation of classification performance. In the context of this study, they can be regarded as a measure of *exactness* and *completeness* of our pathway gene searching results, respectively.

In our study, the pathway genes are defined as composed of structural genes that encode an enzyme, whereas regulator genes are defined as genes controlling the expression of the structural genes.

2.3.1 Simulation study

We simulate a microarray dataset consisting of 500 genes and 30 experiments, with 5 replicates for each experiment. To make the approach more realistic we introduce experiment dependencies, multiple distinct pathways and some random noise into the simulated data. The simulation parameters are as follows:

(i) Experiment correlation matrix, Σ^E. This matrix characterizes the experiment dependencies. For illustrative purposes, we set the experiment correlation matrix to have various dependencies such as 10, 33, 50 and 67 %. In the case of a 33 % dependency, for example, $\sim$ 33 % of the experiments have high dependencies while the remaining

experiments are uncorrelated with one another, i.e. the first 10×10 elements in Σ^E lie between 0.5 and 0.6, with the rest being zeros. Diagonals on Σ^E are set to 1. Figure 2.1 shows the heatmaps for three of the four experiment correlation matrices mentioned above.

(ii) Gene covariance matrix, Σ^G. This matrix characterizes the gene dependencies among one another. As an illustrative example, we introduce two distinct pathways with 15 genes in each pathway; genes in the same pathway have high correlation while genes not in the same pathway are uncorrelated. Specifically, in each pathway the first four genes designated to be seed genes show high correlation (correlation coefficient changes between 0.5 and 0.6) between each other. The remaining 11 genes are separated into three subgroups and are designed to have high correlation with 1, 2 or 4 of the seed genes, respectively, and low correlation with the others (correlation coefficient changes between 0.1 and 0.2).

The simulated data is generated as follows. First, we generate a 500×30 gene expression matrix $\mathbf{X}$, with $vec(\mathbf{X}^T)$, from a multivariate normal distribution with mean $\mathbf{X}$ (zero matrix) and a covariance matrix $\Sigma^G \otimes \Sigma^E$. To make the pathway genes more realistic, for each pathway two randomly chosen genes in each subgroup are linearly combined to make a new pathway gene. The same procedure generated all the final 11 pathway genes for each pathway (replacing the original 11 pathway genes generated above). Using the final 500×30 gene expression matrix, we add random noise with a small SD (e.g. 0.01) to each column (i.e. experiment) to generate the 5 replicates for each experiment. Repeating this process, we generate 1000 simulation datasets.

In this analysis, we compare our approach to that of others and evaluate the performance using *precision* and *recall* measures. All the approaches are implemented as follows: given seed genes, run the pathway search algorithms as described in Section 2.2 and rank the genes by their measured relationships to the seed genes. Calculate *precision* and *recall* for the top n (i.e, $n = 1, ..., 15$) genes.

As this is a simulation study and we know the true experiment correlation matrix, we add one more method *pwsrc.true* into the comparison. The only difference between *pwsrc.true* and *pwsrc.knorm* is that *pwsrc.true* uses the true experiment correlation matrix (Σ^E_{true}) instead of the estimated one in Equation 2.3. We denote the estimated correlation matrix used by our method as $\hat{\Sigma}^E_{knorm}$ for clarity.

The results are summarized in Figure 2.2. When the dependencies among experiments are low, *pwsrc.knorm* performs worse than *pwsrc.null*. However, this performance discrepancy becomes smaller as the experiment dependency increases and finally *pwsrc.knorm* outperforms *pwsrc.null* when the experiment dependency exceeds 33 %. This situation can be easily understood in Figure 2.1. When the dependencies among experiments are low, the noisy signals in the off diagonal elements in $\hat{\Sigma}^E_{knorm}$ become non-negligible and so Σ^E_{null} becomes a better estimate for Σ^E_{true} even though it totally misses capturing the experiment

dependencies. However, when the experiment dependency increases up to 33 %, $\hat{\Sigma}^E_{knorm}$ estimates Σ^E_{true} better than Σ^E_{null} as it is critical to capture the dependent structure now.

These results emphasize the importance of considering experiment dependencies when they exist at a non-negligible level in data, which is actually the case in real applications. Our approach overall achieves higher *precision* and *recall* than *pearson.mean* and *pearson.max* [Figure 2.2], whereas the *GLM* method provides about the same result as our method due to the way we simulated the data (results are not shown here).

To determine the importance of the seed gene quality, we added two randomly chosen, non-pathway genes into the seed-gene-set which is originally composed of four pathway genes. The results are summarized in Figure 2.3. Regardless of the experiment dependencies, the performance of *pwsrc.knorm* becomes worse when the seed-gene-set contains noisy genes.

2.3.2 Application to real datasets

We next test the validity of our approach by applying it to biological pathways composed of genes that are known to operate in tandem. For this test set, we selected two secondary metabolic pathways from the model plant *A.thaliana*: the pathway leading to GSLs, sulfur-rich amino acid-containing compounds which become active in response to tissue damage, and believed to offer a protective function [36, 80, 91, 97], and the pathway leading to flavonoids, compounds of diverse biological activities such as anti-oxidants, functioning in UV protection, in defense, in auxin transport inhibition, and in flower colouring [30, 62, 85, 94].

In *Arabidopsis*, the regulators and structural genes in glucosinolate (GSL) and flavonoid biosynthesis (FB) pathways have been extensively characterized. A considerable number of genes in both pathways are induced by broad environmental stresses, and regulated at the transcriptional level. Furthermore, several research groups have applied transcriptome co-expression to analyze the two pathways [30, 37, 98], thus providing us with a rich source of data for validating our results.

Known genes in each pathway were selected and their conditional dependencies examined using the approach outlined in Section 2.2. For this effort, we used public ATH1 microarray datasets from the AtGenExpress consortium.[1] Among stress serial microarray experiments, we selected four datasets for analysis. A summary of the experimental sets used is listed in Table 3.1, whereas a detailed description of their experimental parameters is provided in Table 3.2.

We then asked, under these varied conditions whether we could recover these known pathway genes by our approach. Finally, having investigated the validity of this approach, and demonstrating that our approach is much more effective than any previous approaches for detecting the known pathway genes, we asked whether we could identify other possible candidate pathway (new) genes.

[1] www.arabidopsis.org/portals/expression/microarray/ATGenExpress.jsp

	Oxidation	Wounding	UV-B light	Drought
Data counts (biosamples/replicate sets)	52 (26/26)	60 (30/30)	60 (30/30)	60 (30/30)
Number of genes	22810	22810	22810	22810
Experimental variables	Methyl viologen time shoot, root	wounding time shoot, root	UV-B light[a] time shoot, root	drought time shoot, root
Submission number	ME00340	ME00330	ME00329	ME00338

[a]UV-B light: Ultraviolet radiation with a range of 280-320 nanometers

Table 2.1: Description of the *A. thaliana* microarray datasets with four different types of stress.

Initially, we investigated the two pathways gene sets in shoot tissue only, but then later expanded the study to include root tissue.

2.3.2.1 Data preprocessing using the RMA normalization

We pre-processed the array data from different experiments using the RMA (Robust Multiarray Average) normalization method [43, 7, 44] which is available from the Bioconductor website. This normalization consists of three steps: a background adjustment, quantile normalization, and summarization of the probe sets. Using RMA normalization, most of the gene replicates are summarized into unique measures. For example, the microarray data with oxidative stress, 99.2 % (20832 out of 21009 genes) of genes represent no replicates after RMA normalization. None of our seed genes (*seed-gene-set I*, *II*, *III* and *IV*) is represented more than once.

2.3.2.2 Studies on the GSL pathway

Based on an extensive literature search, we determined 64 genes that can be associated with the GSL pathway [Table 3.3]. These 64 genes include, in addition to core genes involved in GSL biosynthesis, regulators of this biosynthesis, early steps of side chain elongation/modification and late steps of catabolism [Figure 2.4 in detail].

For our study, two seed-gene-sets are proposed: (*i*) *seed-gene-set I*: AT5G60890 (*ATR1*), AT4G39950 (*CYP79B2*), AT2G20610 (*SUR1*), AT4G31500 (*CYP83B1*) and (*ii*) *seed-gene-set II*: AT5G60890 (*ATR1*), AT5G07690 (*MYB29*), AT5G61420 (*MYB28*), AT4G39940 (*AKN2*). In *seed-gene-set I*, only *ATR1* encodes a transcription factor (TF), whereas three other genes encode enzymes. In contrast to *seed-gene-set I* (comprised of the core pathway genes), *seed-gene-set II* is composed of four regulatory genes [Figure 2.4].

Using two seed-gene-sets, we first analyzed only the shoot tissue dataset from tissues subjected to oxidative stress. This dataset is composed of 22810 genes and 13 experiments with two biological replicates for each experiment. The number of identified GSL pathway genes is summarized in Table 3.4(a)-(b) for the top 10, 20, 30 and 50 genes from the list

(a)

Experiment number (shoot)	Treatment	Time points	Experiment number (root)
1	Control	0 h	14
2	Control	0.5 h	15
3	Control	1h	16
4	Control	3 h	17
5	Control	6 h	18
6	Control	12 h	19
7	Control	24 h	20
8	MV, $10\mu M$	0.5 h	21
9	MV, $10\mu M$	1 h	22
10	MV, $10\mu M$	3 h	23
11	MV, $10\mu M$	6 h	24
12	MV, $10\mu M$	12 h	25
13	MV, $10\mu M$	24 h	26

(b)

Experiment number (shoot)	Treatment	Time points	Experiment number (root)
1	Control	0 h	16
2	Control	0.25 h	17
3	Control	0.5 h	18
4	Control	1h	19
5	Control	3 h	20
6	Control	6 h	21
7	Control	12 h	22
8	Control	24 h	23
9	Stress	0.25 h	24
10	Stress	0.5 h	25
11	Stress	1 h	26
12	Stress	3 h	27
13	Stress	6 h	28
14	Stress	12 h	29
15	Stress	24 h	30

Table 2.2: Description of the experiment conditions used to generate the *A. thaliana* microarray dataset with (a) oxidative stress and (b) wounding, UV-B light and drought stresses.

obtained by *pwsrc.knorm*, *pwsrc.null*, *pearson.mean*, *pearson.max* and *GLM*. With *seed-gene-set I* in Table 3.4(a), *pwsrc.knorm* works best, finding 4, 6, 7 and 8 pathway genes out of the top 10, 20, 30 and 50 genes, respectively. With *seed-gene-set II* in Table 3.4(b), a significant increase is observed in the number of identified pathway genes, especially for *pwsrc.knorm*. For example, among the top 30 genes in the list, *pwsrc.knorm* finds 7 more pathway genes, while *pwsrc.null*, *pearson.mean*, *pearson.max* and *GLM* find 2, 2, 1 and 4 more genes compared to Table 3.4(a), respectively. This increase demonstrates that *seed-gene-set II* indeed carries more influential information than *seed-gene-set I*, which enables us to examine the GSL pathway more thoroughly. Furthermore, our method pushes the pathway genes to rank higher positions in the list so that the final *precision* becomes 60, 55 and 47 %, respectively, for the top 10, 20 and 30 genes.

Next, the dataset is expanded to include the root tissue as well, so now the dataset consists of 26 experiments with two replicates each. Again, the combined dataset is analyzed with the two seed sets as above and the results are summarized in Table 3.4(c)-(d). For *pwsrc.knorm*, a dramatic increase is observed with the *seed-gene-set I* [compare Table 3.4(a) and (c)], in contrast to the *seed-gene-set II* [compare Table 3.4(b) and (d)]. This finding emphasizes the importance of designing the seed-gene-set. When the seed set is appropriately designed for the pathway of our interest, i.e. *seed-gene-set II*, pathway searches could proceed more efficiently with a smaller set of data, but if not, more information (a larger dataset) would be needed to achieve the same performance. *pwsrc.null* finds no pathway genes in this data, which demonstrates the importance of considering experiment dependency, especially as the dataset dimension expands. Different to *pwsrc.null* and the Pearson correlation-based measures, *GLM* shows a prominent increase, and we believe that the extra information added by the root tissue helps *GLM* perform better. The graphical summary of Table 3.4 is given in Figure 2.5. For each method, *precision* and *recall* are calculated for the top 10, 20, 30, 50 and 100 gene lists and plotted accordingly.

In contrast to and different from the oxidative stress, wounding stress is known to induce the expression of *MYB28* and *MYB29* [32], which are the two of four seed genes in *seed-gene-set II* and which regulate Met-derived GSL biosynthesis. Based on our success in finding additional GSL pathway genes using *seed-gene-set II* and oxidative stress as the environmental input, we predicted that we would have similar success using wounding as the environmental input. We expected under wounding stress conditions, that structural genes in the GSL pathway would have stronger association with *seed-gene-set II* than under oxidative stress condition.

Data from the shoot only subjected to wounding are first analyzed by considering 22810 genes, and 15 experiments, each with two biological replicates. The results are summarized in Table 2.5(a)-(b). Again, a significant increase in the number of identified pathway genes is observed from *seed-gene-set I* [Table 2.5(a)] to *seed-gene-set II* [Table 2.5(b)]. Next, the dataset from the root portion is also included, now comprising 30 experiments in total, with two biological replicates for each experiment. No matter what seed-gene-set we use, *pwsrc.knorm* works best [Table 2.5(c)-(d)]. The *precisions* for the top 10, 20 and 30 ranked genes are 100, 70, 50 % with the *seed-gene-set I*, and 90, 65, 60 % with the *seed-gene-set II*.

It is also noteworthy that with *seed-gene-set I* and *II*, *pwsrc.knorm* finds 10 and 9 genes to be in the same biological pathway from the top 10 genes, respectively. A graphical summary of Table 2.5 is given in Figure 2.6.

The performances in the analysis of the last two datasets subjected to UV-B light and drought stresses are similar to the previous results and the results are summarized in Tables 2.6 and 2.7, and Figure 2.7 and 2.8.

2.3.2.3 Studies on the FB pathway

The flavonoid pathway is derived from the upstream phenylpropanoid pathway, beginning at coumaroyl-CoA [Figure 2.9]. Based on an extensive literature search, we found that at least 26 genes can be associated with the FB pathway [Table 2.8]. Genes encoding enzymes in this pathway are regulated by at least 12 TFs belonging to different families, including bZIP WD40, WRKY, MADSbox, R2R3-MYB, and the basic helixloophelix (bHLH) family [98].

It is also worth noting that the genes we considered for the two pathways (GSL and FB) are exclusive to each other, and thus there is no overlap in the genes of the pathways we consider.

It is reported that structural genes (encoding enzymes) in the FB pathway are regulated at the transcriptional level, suggesting that the regulation genes would be good candidates as seed genes, as indicated by the result of GSL pathway study in Section 2.3.2.2. Then we selected two different seed-gene-sets from four different types of TFs [AT4G09820 (*TT8*), AT5G23260 (*TT16*), AT5G24520 (*TTG1*), AT2G37260 (*TTG2*)] and one structural gene [AT5G08640 (*FLS*)]: (*i*) *seed-gene-set III*: AT4G09820 (*TT8*), AT5G23260 (*TT16*), AT5G24520 (*TTG1*), AT5G08640 (*FLS*) and (*ii*) *seed-gene-set IV*: AT4G09820 (*TT8*), AT5G23260 (*TT16*), AT2G37260 (*TTG2*), AT5G08640 (*FLS*).

In this FB pathway study, we present the results using both shoot and root tissues. The number of genes identified by *seed-gene-set III* and *IV* using four different datasets is summarized in Table 2.9. Overall, *pwsrc.knorm* outperforms other methods regardless of seed-gene-sets and stress types. It is worth noting that both seed-gene-sets detected several genes from the upstream phenylpropanoid pathway by *pwsrc.knorm* and *GLM*. To elucidate the cooperative nature of these pathways, we designate the number of identified genes from the upstream pathways (phenylpropanoid pathway) in the parenthesis adjacent to the total number of identified genes [Table 2.9]. For example, the dataset with drought stress 33 % (by *pwsrc.knorm*), and 20 % (by *GLM*) of the identified genes from top 20 are derived from the upstream pathways. Table 2.10 lists all the identified drought stress pathway genes from the top 20 list, and designates the original pathway to which each gene belongs.

In Figure 2.9, all the identified top 20 genes in Table 2.9(d) by *pwsrc.knorm* are visualized. It is noteworthy that *pwsrc.knorm* not only detects six core genes - AT3G51240 (*F3H*), AT3G55120 (*CHI*), AT5G13930 (*CHS*), AT5G07990 (*F3′H*), AT5G17050 (*UGT78D2*), AT1G78570 (*RHM1*) - in the FB pathway, but additionally finds three more genes, AT1G65050 (*4CL3*), AT2G23910 (*CCR6*) and AT2G37040 (*PAL1*), located at the branch points of phenyl-

Group	AGI code	Gene name
Regulator genes	AT1G18570	MYB51
	AT1G66340	ETR1
	AT3G54640	TRP3/TSA1
	AT4G12030	BAT5
	AT5G03280	EIN2
	AT5G07690	MYB29
	AT5G07700	MYB76
	AT5G46330	FLS2
	AT5G60890	ATR1/Myb34
	AT5G61420	MYB28
	AT1G07640	OBP2
	AT3G09710	IQD1
GSL biosynthesis pathway (verified by experiment)	AT1G12140	GS-OX5
	AT1G16400	CYP79F2
	AT1G16410	CYP79F1
	AT1G18590	SOT17
	AT1G24100	UGT74B1
	AT1G62540	GS-OX2
	AT1G62560	GS-OX3
	AT1G62570	GS-OX4
	AT1G65860	GS-OX1
	AT1G74090	SOT18
	AT1G74100	SOT16
	AT2G20610	SUR1
	AT2G22330	CYP79B3
	AT2G25450	GS-OH
	AT2G43100	IPMI2
	AT3G19710	BCAT4
	AT3G49680	BCAT3
	AT3G58990	IPMI1/AtLeuD2
	AT4G03050	AOP3
	AT4G03060	AOP2
	AT4G13430	AtLeuC1
	AT4G13770	CYP83A1
	AT4G31500	CYP83B1/SUR2/ATR4
	AT4G39950	CYP79B2
	AT5G05260	CYP79A2
	AT5G14200	AtIMD1
	AT5G23010	MAM1
	AT5G23020	MAM3
	AT5G57220	CYP81F2
	AT1G31180	IPMDH1
	AT4G30530	GGP1
GSL biosynthesis pathway (predicted)	AT1G78370	GSTF20
	AT5G07460	PMSR2
	AT5G36160	
	AT2G30860	GSTF9
	AT2G30870	GSTF10
	AT2G31790	UGT74C1
	AT3G03190	GSTF11
GSL biosynthesis pathway (co-substrate pathway)	AT2G14750	AKN1
	AT4G39940	AKN2
	AT4G23100	PAD2
	AT1G65880	BZO1
	AT5G65940	CHY1
	AT1G04580	AAO4
	AT5G63980	FIERY1/SAL1
GSL-catabolic pathway	AT5G44070	PCS1
	AT1G47600	TGG4
	AT1G54030	MVP1
	AT1G59870	PEN3
	AT2G44490	PEN2
	AT1G54040	ESP
	AT3G14210	ESM1

Table 2.3: List of 64 GSL metabolism pathway genes.

	Top	*pwsrc.knorm*	*pwsrc.null*	*pearson.mean*	*pearson.max*	*GLM*
(a)	10	4	0	2	1	0
	20	6	1	3	4	0
	30	7	1	3	5	0
	50	8	1	4	7	0
(b)	10	6	0	4	2	2
	20	11	3	4	4	3
	30	14	3	5	6	4
	50	14	5	8	8	5
(c)	10	9	0	3	0	3
	20	11	0	3	0	5
	30	12	0	3	1	7
	50	12	0	3	2	10
(d)	10	6	0	5	1	4
	20	10	0	7	1	5
	30	13	0	9	1	9
	50	19	0	9	1	12

Table 2.4: The number of identified GSL pathway genes in the *A. thaliana* microarray dataset from tissues subjected to oxidative stress using (a) shoot tissue only, *seed-gene-set I*; (b) shoot tissue only, *seed-gene-set II*; (c) shoot and root tissues, *seed-gene-set I*; (d) shoot and root tissues, *seed-gene-set II*.

	Top	*pwsrc.knorm*	*pwsrc.null*	*pearson.mean*	*pearson.max*	*GLM*
(a)	10	3	1	2	3	0
	20	4	1	2	3	0
	30	4	1	3	3	0
	50	4	1	3	6	0
(b)	10	4	0	6	4	0
	20	8	0	9	5	1
	30	11	2	13	5	1
	50	12	5	15	8	1
(c)	10	10	3	3	0	6
	20	14	4	3	1	10
	30	15	5	3	1	12
	50	16	5	3	2	14
(d)	10	9	0	6	0	7
	20	13	0	7	0	8
	30	18	0	8	0	10
	50	22	2	10	1	16

Table 2.5: The number of identified GSL pathway genes in the *A. thaliana* microarray dataset from tissues subjected to wounding stress using (a) shoot tissue only, *seed-gene-set I*; (b) shoot tissue only, *seed-gene-set II*; (c) shoot and root tissues, *seed-gene-set I*; (d) shoot and root tissues, *seed-gene-set II*.

	Top	*pwsrc.knorm*	*pwsrc.null*	*pearson.mean*	*pearson.max*	*GLM*
(a)	10	3	2	1	0	1
	20	5	2	1	1	2
	30	5	2	2	1	3
	50	5	2	3	1	3
(b)	10	6	1	3	2	3
	20	9	2	3	4	4
	30	9	2	3	5	5
	50	10	2	4	6	6
(c)	10	7	1	1	0	6
	20	10	1	2	0	7
	30	10	1	3	0	7
	50	10	1	3	0	7
(d)	10	8	0	1	1	6
	20	11	0	1	2	9
	30	12	0	1	2	10
	50	16	0	1	2	11

Table 2.6: The number of identified GSL pathway genes in the *A. thaliana* microarray dataset from tissues subjected to UV-B light stress using (a) shoot tissue only, *seed-gene-set I*, (b) shoot tissue only, *seed-gene-set II*, (c) shoot and root tissues, *seed-gene-set I*, (d) shoot and root tissues, *seed-gene-set II*.

	Top	*pwsrc.knorm*	*pwsrc.null*	*pearson.mean*	*pearson.max*	*GLM*
(a)	10	3	2	1	3	0
	20	4	2	2	3	1
	30	5	2	3	6	1
	50	6	3	6	6	2
(b)	10	5	2	4	5	3
	20	8	3	4	8	3
	30	8	3	6	9	4
	50	8	4	8	12	6
(c)	10	9	0	3	2	5
	20	10	0	3	3	5
	30	12	0	3	3	9
	50	15	0	3	3	12
(d)	10	5	0	4	1	5
	20	8	1	4	1	5
	30	9	1	4	1	6
	50	10	2	6	1	7

Table 2.7: The number of identified GSL pathway genes in the *A. thaliana* microarray dataset from tissues subjected to drought stress using (a) shoot tissue only, *seed-gene-set I*, (b) shoot tissue only, *seed-gene-set II*, (c) shoot and root tissues, *seed-gene-set I*, (d) shoot and root tissues, *seed-gene-set II*.

Pathways	FB	Phenylpropanoid biosynthesis
AGI code (gene name)	AT4G09820 (*TT8*) [a,b]	AT2G37040 (*PAL1*)
	AT5G23260 (*TT16*)[a,b]	AT3G53260 (*PAL2*)
	AT5G24520 (*TTG1*)[a]	AT5G04230 (*PAL3*)
	AT2G37260 (*TTG2*)[b]	AT3G10340 (*PAL4*)
	AT5G08640 (*FLS*)[a,b]	AT2G30490 (*C4H*)
	AT5G13930 (*CHS*)	AT1G51680 (*4CL1*)
	AT3G55120 (*CHI*)	AT3G21240 (*4CL2*)
	AT3G51240 (*F3H*)	AT1G65060 (*4CL3*)
	AT5G07990 (*F3′H*)	AT3G21230 (*4CL5*)
	AT5G42800 (*DFR*)	AT1G15950 (*CCR1*)
	AT1G61720 (*BAN*)	AT2G23910 (*CCR6*)
	AT5G17220 (*GST*)	
	AT3G59030 (*TT12*)	
	AT5G35550 (*TT2*)	
	AT1G06000 (*UGT89C1*)	
	AT5G17050 (*UGT78D2*)	
	AT1G78570 (*RHM1*)	
	AT4G14090 (*UGT75C1*)	
	AT1G30530 (*UGT78D1*)	
	AT3G29590 (*A5G6999MaT*)	
	AT5G54160 (*OMT1*)	
	AT3G62610 (*MYB11*)	
	AT2G47460 (*MYB12*)	
	AT5G49330 (*MYB111*)	
	AT1G56650 (*PAP1*)	
	AT1G66390 (*PAP2*)	

[a] *seed-gene-set III*
[b] *seed-gene-set IV*

Table 2.8: List of FB and phenylpropanoid biosynthesis pathway genes related to our study.

propanoid pathway to the FB pathway or to the lignin biosynthesis pathway at coumaroyl-CoA [Figure 2.9]. Among those additionally found genes, *CCR6* and *PAL1* are uniquely detected by our method.

Thus, in contrast to, and differing from the other methods, *pwsrc.knorm* enables us to find additional genes from closely related pathways by considering the indirect relationships between genes. This finding can be useful for future studies targeted to discovering the cooperative nature of genes in the FB pathways.

We also compared our results with the seed-gene-sets containing some noisy genes. For example, we applied our method to the dataset with drought stress with the seed-gene-set composed of *seed-genes-set III* and two of GSL pathway genes. As summarized in Table 2.11, the number of pathway genes identified decreased which implies that the pathway gene search becomes less efficient when the seed-gene-set contains noisy genes not biologically related to the pathway of our interest.

	seed-gene-set	Top rank	*pwsrc.knorm*	*pwsrc.null*	*pearson.mean*	*pearson.max*	*GLM*
(a)	*III*	10	7 (2)	1	0	4	5 (1)
		20	8 (3)	2	0	4	6 (1)
		30	10 (4)	2	0	4	7 (2)
		50	11 (4)	2	0	5 (1)	8 (3)
	IV	10	7 (2)	2	3	5 (1)	4 (1)
		20	9 (3)	3	3	6 (1)	5 (1)
		30	10 (4)	3	3	6 (1)	6 (1)
		50	11 (4)	4	4	6 (1)	7 (2)
(b)	*III*	10	6 (1)	3	0	4	5 (1)
		20	9 (3)	3	0	4	5 (1)
		30	9 (3)	4	0	4	5 (1)
		50	11 (4)	4	0	4	7 (2)
	IV	10	6 (1)	4	0	4	4 (1)
		20	9 (2)	4	1 (1)	4	5 (1)
		30	10 (3)	4	2 (2)	4	6 (1)
		50	11 (4)	4	2 (2)	5 (1)	7 (2)
(c)	*III*	10	6 (1)	1	0	2	4 (1)
		20	8 (3)	1	0	3	4 (1)
		30	11 (4)	1	0	3	4 (1)
		50	11 (4)	1	0	4 (1)	5 (2)
	IV	10	7 (2)	1	0	5 (1)	4 (1)
		20	8 (3)	1	0	6 (1)	4 (1)
		30	10 (3)	1	0	6 (1)	4 (1)
		50	11 (4)	1	1 (1)	6 (1)	4 (1)
(d)	*III*	10	6 (2)	0	0	2	4 (1)
		20	9 (3)	1	0	3	5 (1)
		30	10 (3)	1	0	4	6 (2)
		50	11 (3)	1	0	4	6 (2)
	IV	10	6 (2)	2	1	4	5 (1)
		20	9 (3)	2	2	4	5 (1)
		30	10 (3)	2	2	4	6 (2)
		50	13 (4)	2	2	4	6 (2)

Table 2.9: The number of pathway genes identified from FB and phenylpropanoid biosynthesis pathways in the *A. thaliana* microarray dataset from shoot and root tissues subjected to (a) oxidation, (b) wounding, (c) UV-B light and (d) drought stresses. The number of identified genes from phenylpropanoid pathways is designated in the parenthesis adjacent to the total number of identified genes.

Method	*seed-gene-set III*	*seed-gene-set IV*
pwsrc.knorm	AT1G65060 (*4CL3*)[b]	AT1G65060 (*4CL3*)[b]
	AT1G78570 (*RHM1*)[a]	AT3G51240 (*F3H*)[a]
	AT3G51240 (*F3H*)[a]	AT1G78570 (*RHM1*)[a]
	AT3G55120 (*CHI*)[a]	AT3G55120 (*CHI*)[a]
	AT5G13930 (*CHS*)[a]	AT5G13930 (*CHS*)[a]
	AT2G23910 (*CCR6*)[b]	AT2G23910 (*CCR6*)[b]
	AT5G17050 (*UGT78D2*)[a]	AT5G07990 (*F3′H*)[a]
	AT5G07990 (*F3′H*)[a]	AT5G17050 (*UGT78D2*)[a]
	AT2G37040 (*PAL1*)[b]	AT2G37040 (*PAL1*)[b]
pwsrc.null	AT1G78570 (*RHM1*)[a]	AT1G78570 (*RHM1*)[a]
		AT5G13930 (*CHS*)[a]
pearson.mean		AT4G14090 (*UGT75C1*)[a]
		AT5G42800 (*DFR*)[a]
pearson.max	AT1G78570 (*RHM1*)[a]	AT1G78570 (*RHM1*)[a]
	AT5G13930 (*CHS*)[a]	AT5G13930 (*CHS*)[a]
	AT3G55120 (*CHI*)[a]	AT3G55120 (*CHI*)[a]
		AT3G51240 (*F3H*)[a]
GLM	AT1G65060 (*4CL3*)[b]	AT1G65060 (*4CL3*)[b]
	AT1G78570 (*RHM1*)[a]	AT3G51240 (*F3H*)[a]
	AT3G51240 (*F3H*)[a]	AT3G55120 (*CHI*)[a]
	AT3G55120 (*CHI*)[a]	AT1G78570 (*RHM1*)[a]
	AT5G13930 (*CHS*)[a]	AT5G13930 (*CHS*)[a]

[a] FB pathway genes
[b] Phenylpropanoid biosynthesis pathway genes

Table 2.10: List of identified genes from top 20 list in Table 2.9(d). Each gene is designated by the original pathway to which it belongs.

Top	(a)	(b)	(c)	(d)
10	6	3	3	3
20	9	5	5	4
30	10	6	7	5

Table 2.11: The number of pathway genes identified from FB and phenylpropanoid biosynthesis pathways in the *A. thaliana* microarray dataset from shoot and root tissues subjected to drought stresses. For comparison, different seed genes sets are used: (a) *seed-gene-set III*, (b) *seed-gene-set III*, *ATR1*, *AKN2*, (c) *seed-gene-set III*, *MYB28*, *AKN2*, (d) *seed-gene-set III*, *ATR1*, *MYB28*, *AKN2*.

	GSL pathway (shoot)		GSL pathway (shoot & root)		FB pathway (shoot & root)	
	seed-gene-set I	*seed-gene-set II*	*seed-gene-set I*	*seed-gene-set II*	*seed-gene-set III*	*seed-gene-set IV*
Oxidation	98	152	88	123	30	28
Wounding	50	104	45	82	25	20
UV-B light	63	262	43	62	32	29
Drought	203	336	73	75	40	25

Table 2.12: The number of top candidate genes whose p-values are less than or equal to the threshold $\alpha = 0.05$ for the GSL and FB pathways.

2.3.2.4 Calculating p-values for the test statistics using chi-square approximation

To calculate p-values for our test statistics, we assumed that Equation 2.7 has an approximate asymptotic chi-square distribution with k degrees of freedom, where k represents the number of seed genes.

$$-2 \log LR^* \approx \chi^2_{|k|} \tag{2.9}$$

Calculated p-values are adjusted by the Bonferroni correction by multiplying the number of outcomes being tested, *i.e.* the number of candidate genes which is 22806. Setting $\alpha = 0.05$ as our threshold, we counted the number of candidate genes whose p-values are less than or equal to the threshold and the results are summarized in Table 2.12. For example, the dataset from shoot tissues subjected to oxidative stress with GSL pathway, we found top 98 and 152 genes are significant with *seed-gene-set I* and *II*, respectively. Even though it is still questionable regarding the chi-square approximation, we do not see any better alternative for the moment.

2.3.2.5 Robustness of our method

To check the sensitivity of our method with respect to the number of known pathway genes, we studied the dataset subjected to oxidative stress from Section 2.3.2.2 with *seed-gene-set II*. For simplicity, we used only the shoot tissue data. Given *seed-gene-set II* which is composed of 4 regulatory genes (*ATR1*, *MYB29*, *MYB28*, *AKN2*: these are denoted by *a*, *b*, *c* and *d* in Table 2.13 respectively), we worked with all the possible subsets of size 2 and 3 and summarized the results in Table 2.13.

We first want to point out the importance of the *AKN2* gene. It is obvious that the relationship between the pathway genes is not equivalent, which is why the seed-gene selection would play a critical role in our method. To illustrate this point here, we examined correlations between the seed genes and the pathway genes [Table 2.14] from the public databases, and found that *AKN2* is the gene most strongly correlated with many other pathway genes, including the genes in indole GSL branch and aliphatic GSL branch [Figure 2.4 and Table 2.14]. As an additional filter for selecting seed genes we considered the biological evidence linking the candidate seed genes with the pathways of interest. As such, in the presence

Top	*a,b*	*a,c*	*a,d*	*b,c*	*b,d*	*c,d*	*a,b,c*	*a,b,d*	*a,c,d*	*b,c,d*	*a,b,c,d*
10	3	0	6	3	7	8	1	7	5	7	6
20	4	0	11	3	13	14	2	10	10	12	11
30	4	0	14	3	15	14	3	13	12	13	14
50	6	0	15	5	17	17	4	15	14	15	14

Table 2.13: The number of pathway genes identified using the subsets of *seed-gene-set II*.

of the *AKN2* gene, the performance of our method looks pretty robust no matter which subsets we used. For example, from top 30 in Table 2.14, six subsets which include *AKN2* gene found 14, 15, 14, 13, 12, 13 genes respectively, which is close to what we found with *seed-gene-set II*. However in the absence of *AKN2* gene, the performance got worse and fluctuated dramatically. This again emphasizes the importance of seed-gene-set designing. Unless the seed-gene-set misses significant genes playing a critical role, our method performs in a robust way with respect to the size of seed gene set.

2.3.2.6 Comparison with other studies

Finally, we compared our results with other literatures on the discovery of GSL and FB pathway genes [30, 37, 71, 98].

Using pearson correlation coefficients, Hirai *et al.* [37] constructed co-expression relationships (correlation coefficient > 0.65) of *Myb28* and *Myb29* with other genes. In our work these two genes were used in the *seed-gene-set II* to study the GSL pathway. Compared to Hirai's results, we found 11 overlapped pathway genes. Furthermore, we discovered more pathway genes (e.g. *PMSR2*, *GSTF11* and *GS-OX1-5*, Figure 2.4) which are not listed in their results.

In another study, Saito *et al.* [71] constructed the co-expression networks of a general phenylpropanoid pathway using 54 pre-selected 'guide genes' (13 transcription factors and 41 enzymes involved in flavonoid and phenylpropanoid pathways). These are composed of 4 modules such as flavonoid, anthocyanidin, proanthocyanidin and lignin. Especially, the flavonoid module consists of 16 genes. Comparing our finding from top 20 genes in Table 2.9 with theirs in flavonoid module, 7 pathway genes (including *FLS*, which is one of the seed genes in *seed-gene-set IV*) are commonly detected. The 3 seed genes encoding TFs (*TT8*, *TT16* and *TTG2*) in our study are not present in their flavonoid module. Instead, they are located in their proanthocyanidin module, which is downstream of flavonoid pathway. However, in the existence of external stress, their results suggested that these 3 TFs could network with the flavonoid network module, which validates our findings.

Seed-gene	Correlated gene	Pearson's correlation coefficient
AKN2	*AKN1*	0.923
	UGT74B1	0.913
	SOT17	0.912
	CYP79B3	0.808
	SUR1	0.804
	GGP1	0.771
	CYP83B1/SUR2/ATR4	0.766
	CYP79B2	0.757
	FIERY1/SAL1	0.733
	ATR1/Myb34	0.724
	SOT16	0.718
	PMSR2	0.707
MYB28	*CYP83A1*	0.821
	UGT74C1	0.808
	IPMI1/AtLeuD2	0.792
	SOT18	0.775
	BAT5	0.771
	CYP79F1	0.769
	MAM1	0.750
	GSTF11	0.735
	IPMI2	0.724
ATR1	*UGT74B1*	0.755
	AKN2	0.724
	SOT17	0.712
MYB29	*GS-OX1*	0.707

Table 2.14: The relationships between *seed-gene-set II* and other pathway genes. The seed gene was queried on *http://prime.psc.riken.jp/?action=coexpression_index* with following criteria: Matrix - stress treatments v.1 (298 data); Method - union of sets; Threshold value - 0.5; Display limit - 1000. The correlated genes were filtered by applying pearson's correlation coefficient cutoff ≥ 0.7.

2.4 Discussion

We have proposed a novel approach to identify genes associated with a pathway specified by a set of seed genes. This approach considers the space of pathway genes as a span generated by its pathway genes and uses partial correlation as a distance measure to determine genes interacting with the previously identified pathway genes. This approach differs from many existing approaches in the following aspects: (*i*) it uses the partial correlation conditioned upon identified pathway genes, not on all genes; (*ii*) it enables us to identify genes having higher level interaction (i.e. group interaction) although their pair-wise marginal correlations are weak; (*iii*) it considers experiment dependencies when inferring gene relationships; (*iv*) its computational workload is less demanding.

The first aspect above implies our method is pathway specific. It focuses its search only on

a particular pathway among all existing multiple (unknown) pathways in a dataset. This is a limitation of our approach. But we note that this specific search has shown huge advantages in reducing false positive discoveries in both our simulation and real data studies, and has also led to a deeper and insightful biological interpretation of the results. This approach can potentially be extended to the situation that the seed genes or target pathways are not available, if the seed genes for different pathways can be originated from analysis of other sources of biological data.

Although our approach has yielded encouraging biological results in a real dataset application, there is still room for further improvement, including exploration of properties of this approach to answer questions like, 'What are the biological properties of the identified genes?' and, 'How reliable is the set of identified genes?'. Further biological understanding of the identified pathway genes would give us deeper insights into the biological process under consideration.

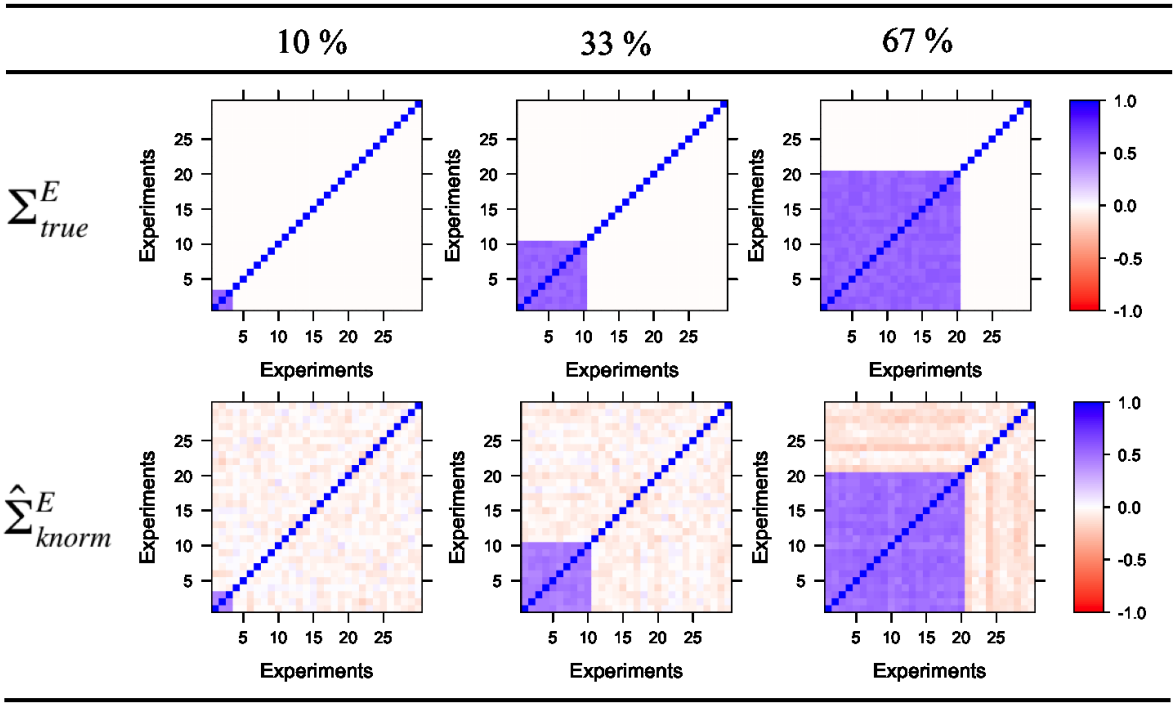

Figure 2.1: Heatmaps of (top) true and (bottom) estimated experiment correlation matrices of the simulation datasets having different experiment dependencies (10, 33 and 67 %).

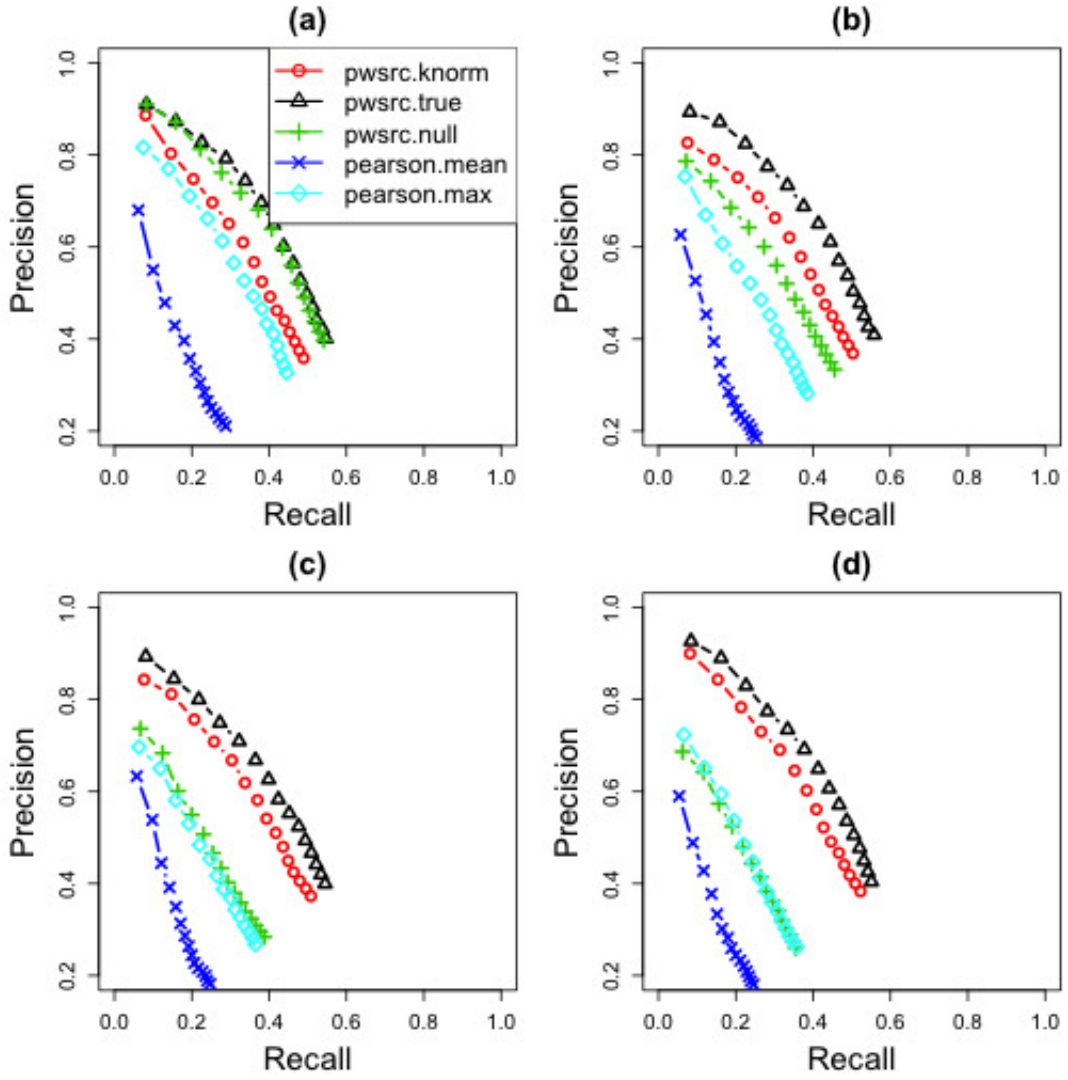

Figure 2.2: Graphical summary of the simulation study. Simulation datasets are generated with different experiment dependencies (a) 10 %, (b) 33 %, (c) 50 % and (d) 67 %. For each plot, *precision* and *recall* are calculated from the top $n(n = 1, ..., 15)$ genes in the list obtained by five different methods.

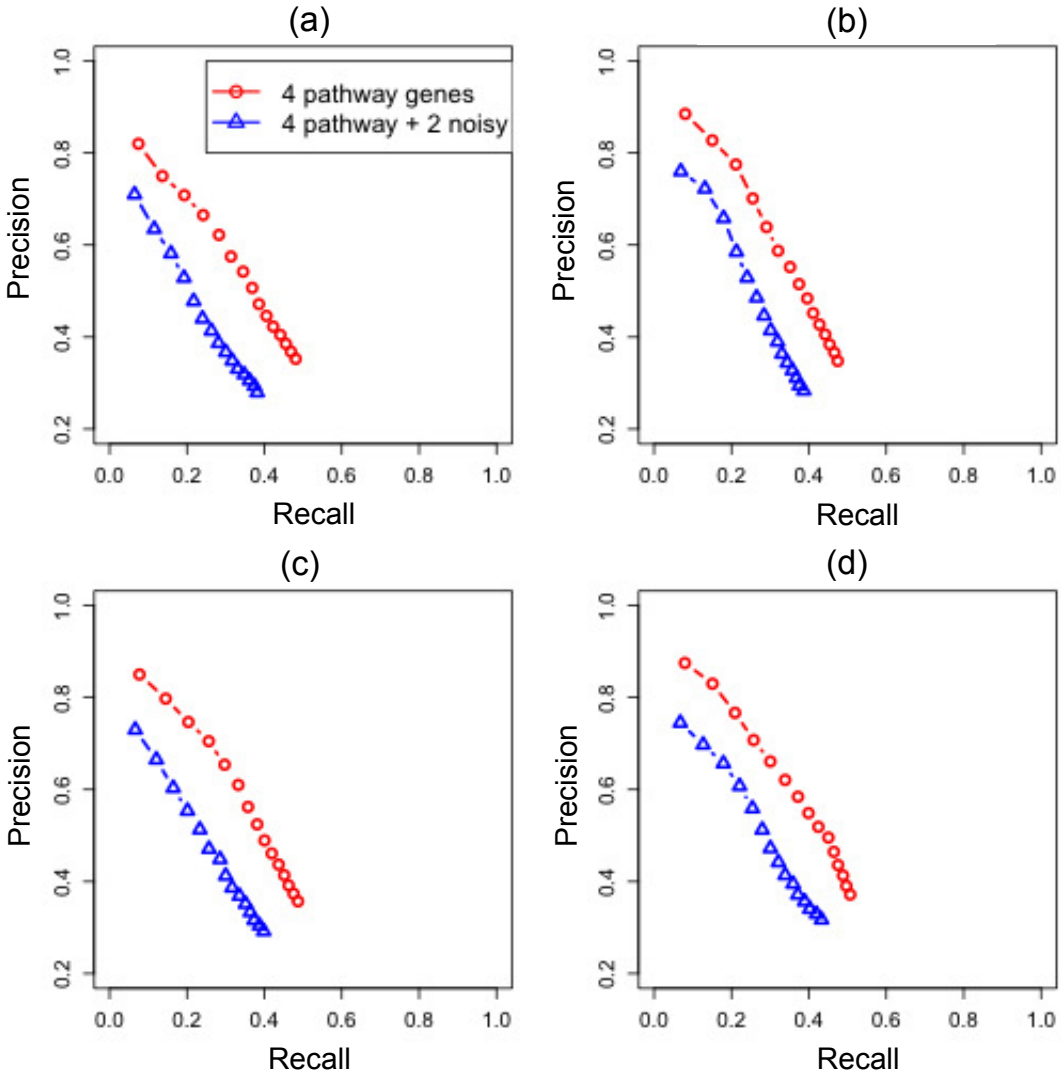

Figure 2.3: Graphical summary of the simulation study. Simulation datasets are generated with different experiment dependencies (a) 10 %, (b) 33 %, (c) 50 % and (d) 67 %. For each plot, *precision* and *recall* are calculated from the top n ($n = 1, , 15$) genes in the list obtained by *pwsrc.knorm* with seed gene set composed of four pathway genes (red dots) or four pathway genes and 2 non-pathway genes (blue dots).

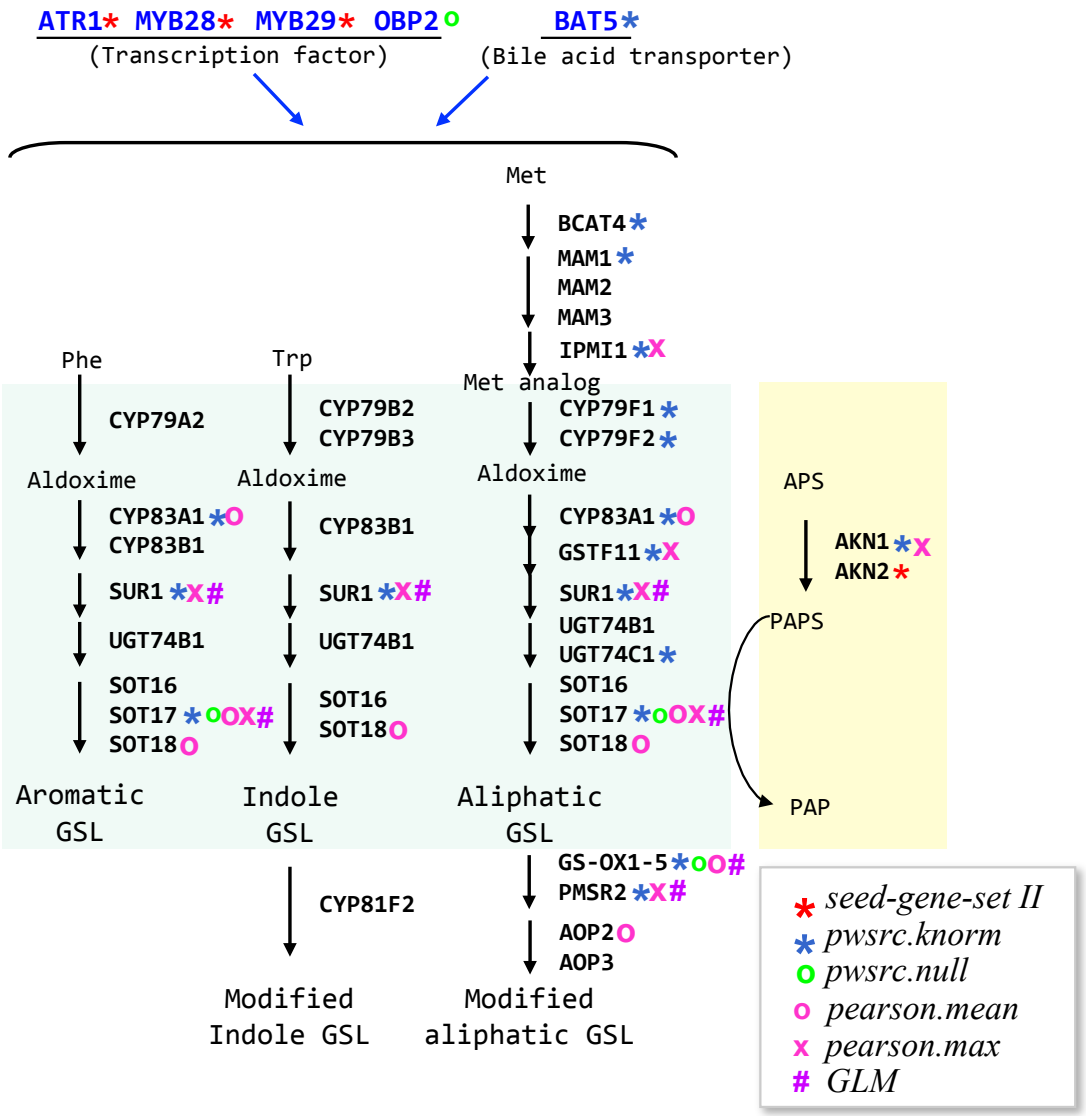

Figure 2.4: Simplified schematic representation of GSL metabolic pathway. Enzymes and regulators are indicated by bold, capital letters. The GSL pathway genes from the top 30 lists identified by different methods are designated by different markers. *A. thaliana* dataset from shoot tissues subjected to oxidative stress and *seed-gene-set II* are used. Compared to other methods, our method uniquely finds six genes, *BAT5*, *BCAT4*, *MAM1*, *CYP79F1*, *CYP79F2*, *UGT74C1* and misses three genes, *OBP2*, *SOT18*, *AOP2*.

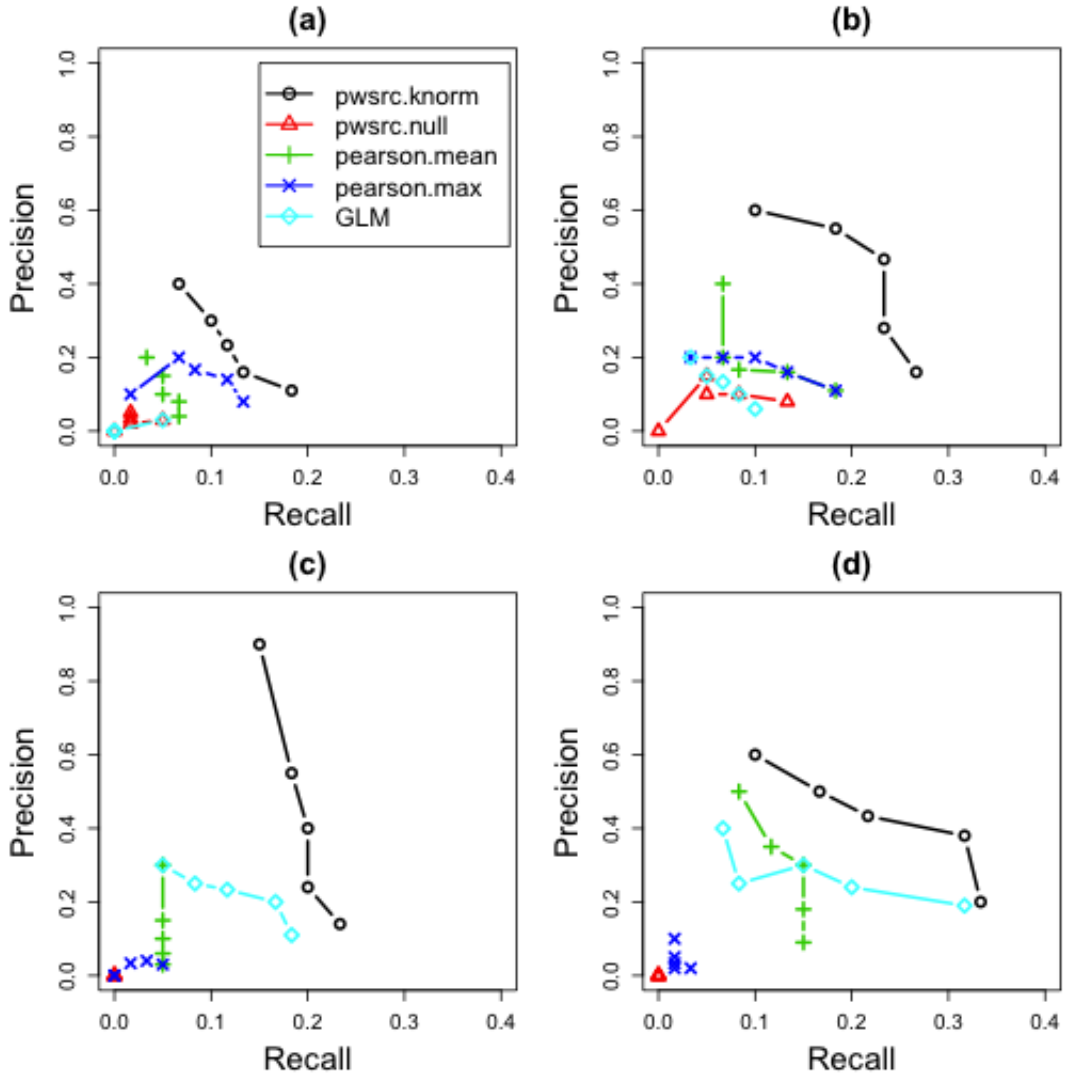

Figure 2.5: Graphical summary of the *A. Thaliana* microarray dataset subjected to oxidative stress; (a) shoot tissue, *seed-gene-set I*, (b) shoot tissue, *seed-gene-set II*, (c) shoot and root tissues, *seed-gene-set I*, (d) shoot and root tissues, *seed-gene-set II*. *Precision* and *recall* are calculated from the top 10, 20, 30, 50 and 100 genes in the list obtained by different methods.

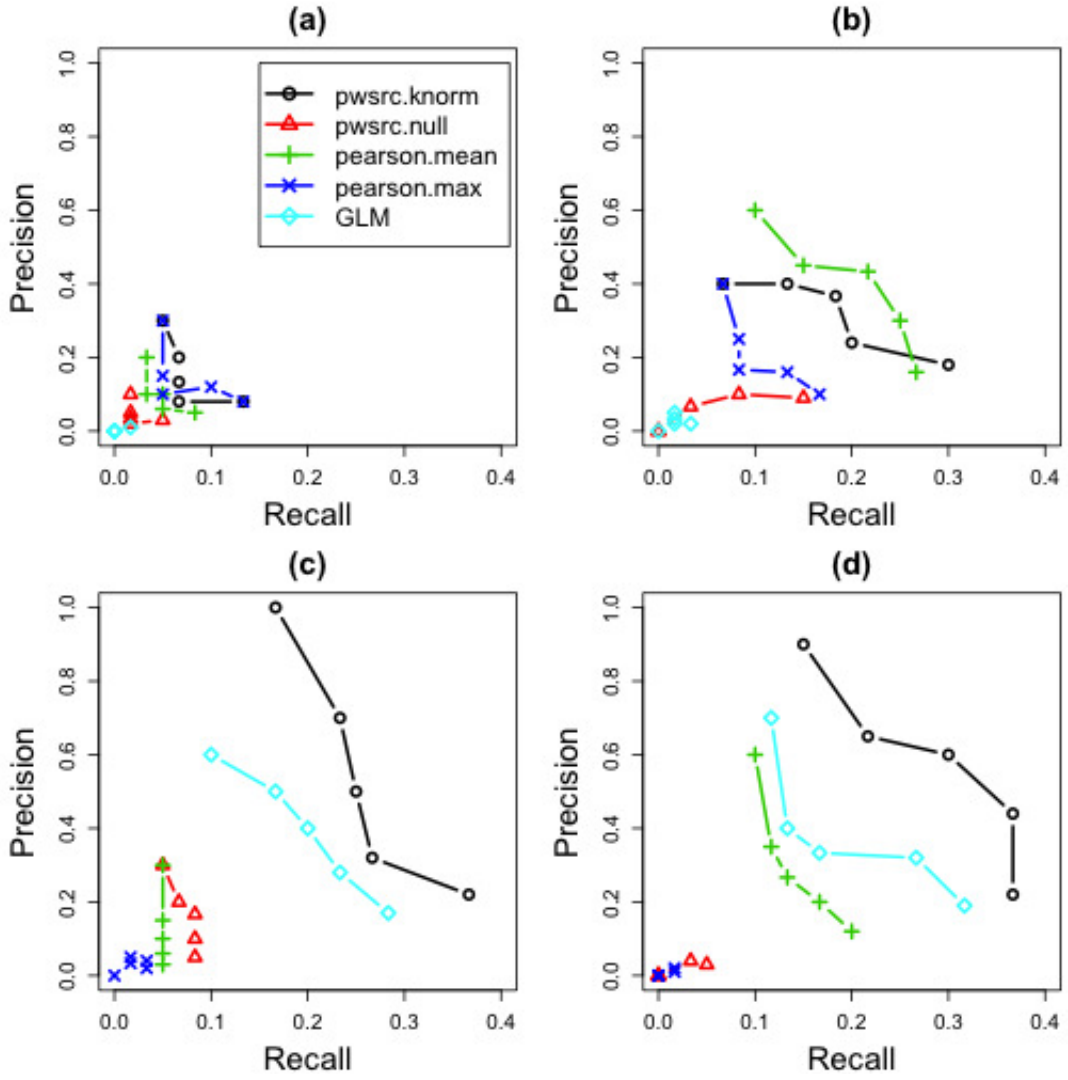

Figure 2.6: Graphical summary of the *A. Thaliana* microarray dataset subjected wounding stress; (a) shoot tissue, *it seed-gene-set I*, (b) shoot tissue, *seed-gene-set II*, (c) shoot and root tissues, *seed-gene-set I*, (d) shoot and root tissues, *seed-gene-set II*. *Precision* and *recall* are calculated from the top 10, 20, 30, 50 and 100 genes in the list obtained by different methods.

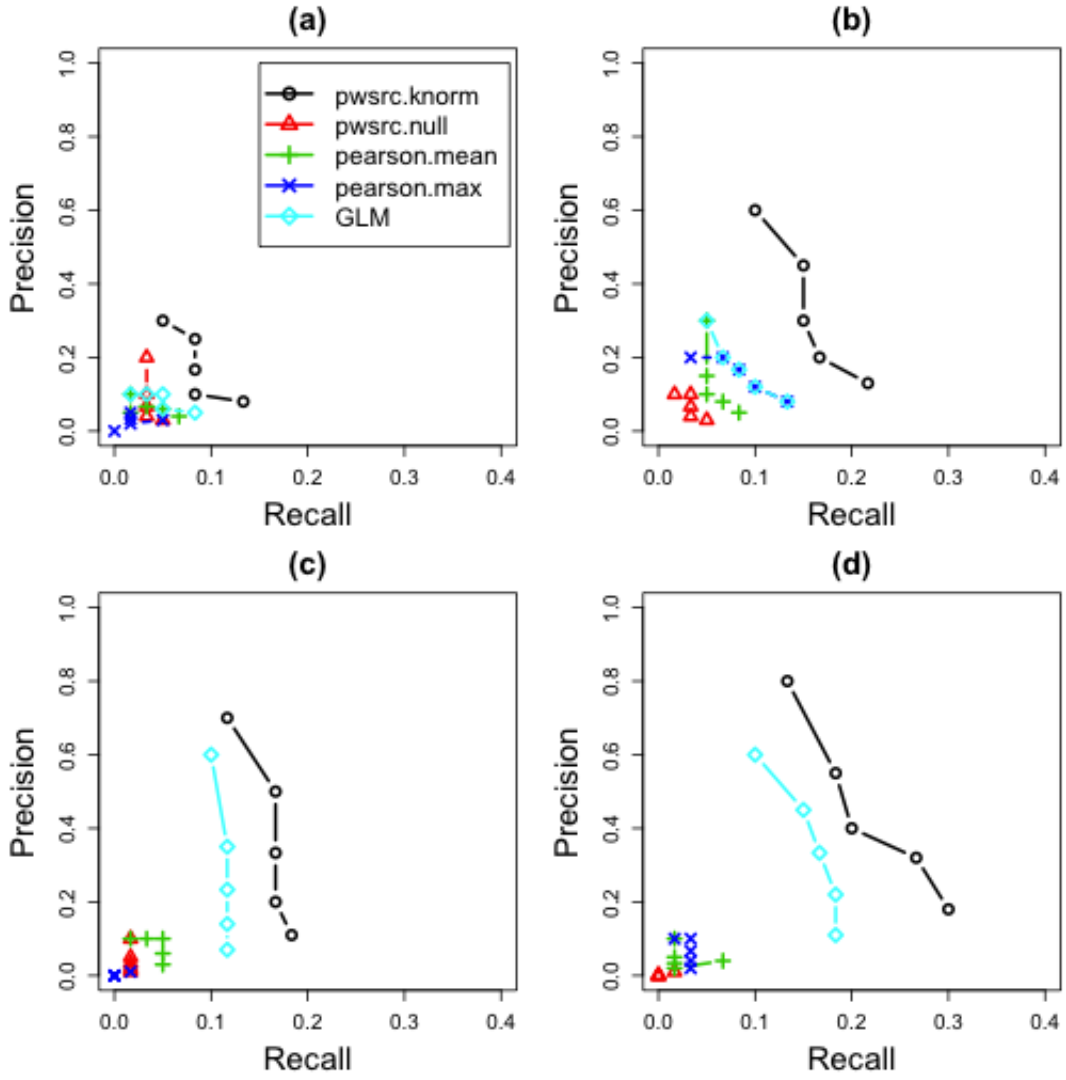

Figure 2.7: Graphical summary of the *A. Thaliana* microarray dataset subjected to UV-B light stress; (a) shoot tissue, *it seed-gene-set I*, (b) shoot tissue, *seed-gene-set II*, (c) shoot and root tissues, *seed-gene-set I*, (d) shoot and root tissues, *seed-gene-set II*. *Precision* and *recall* are calculated from the top 10, 20, 30, 50 and 100 genes in the list obtained by different methods.

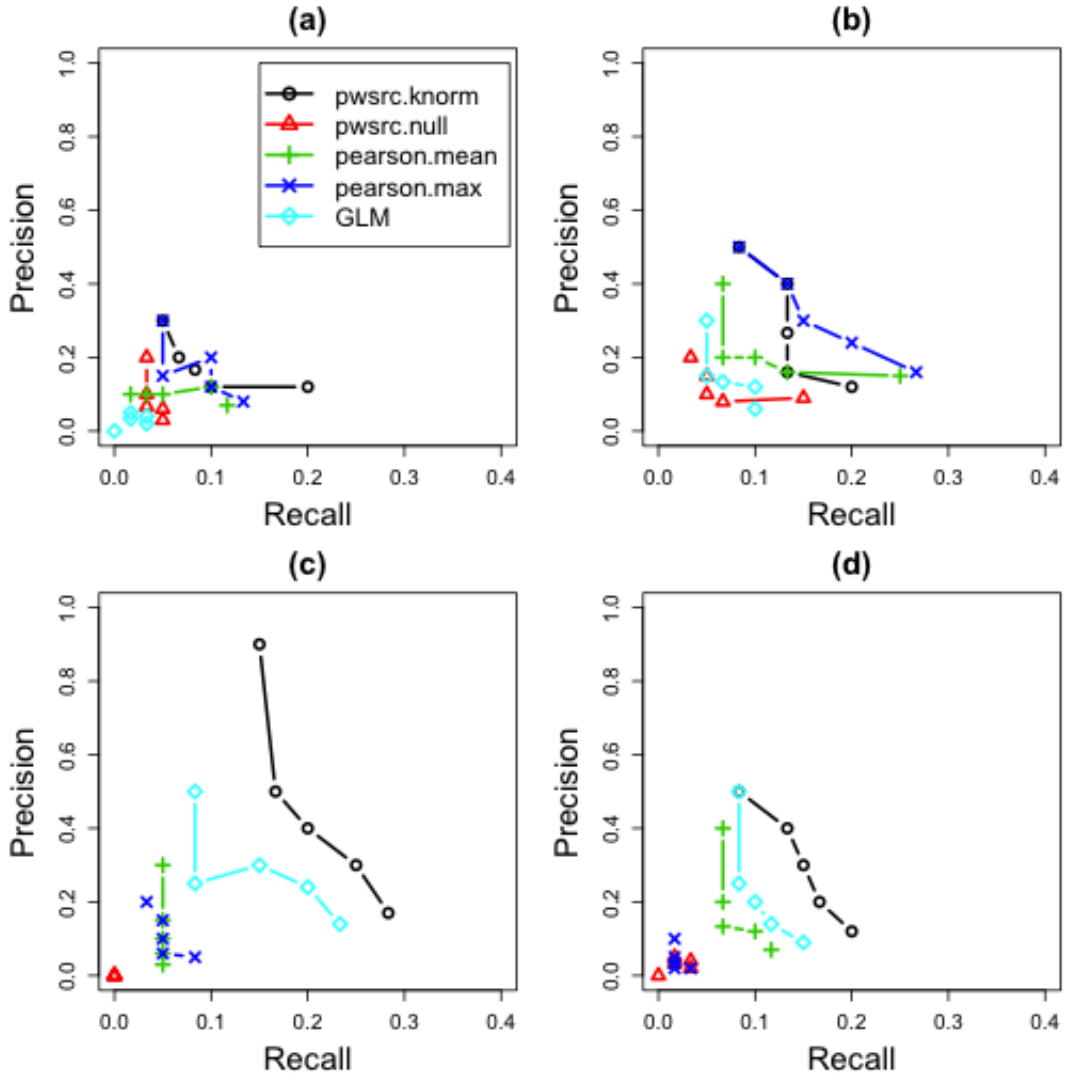

Figure 2.8: Graphical summary of the *A. Thaliana* microarray dataset subjected to drought stress; (a) shoot tissue, *it seed-gene-set I*, (b) shoot tissue, *seed-gene-set II*, (c) shoot and root tissues, *seed-gene-set I*, (d) shoot and root tissues, *seed-gene-set II*. *Precision* and *recall* are calculated from the top 10, 20, 30, 50 and 100 genes in the list obtained by different methods.

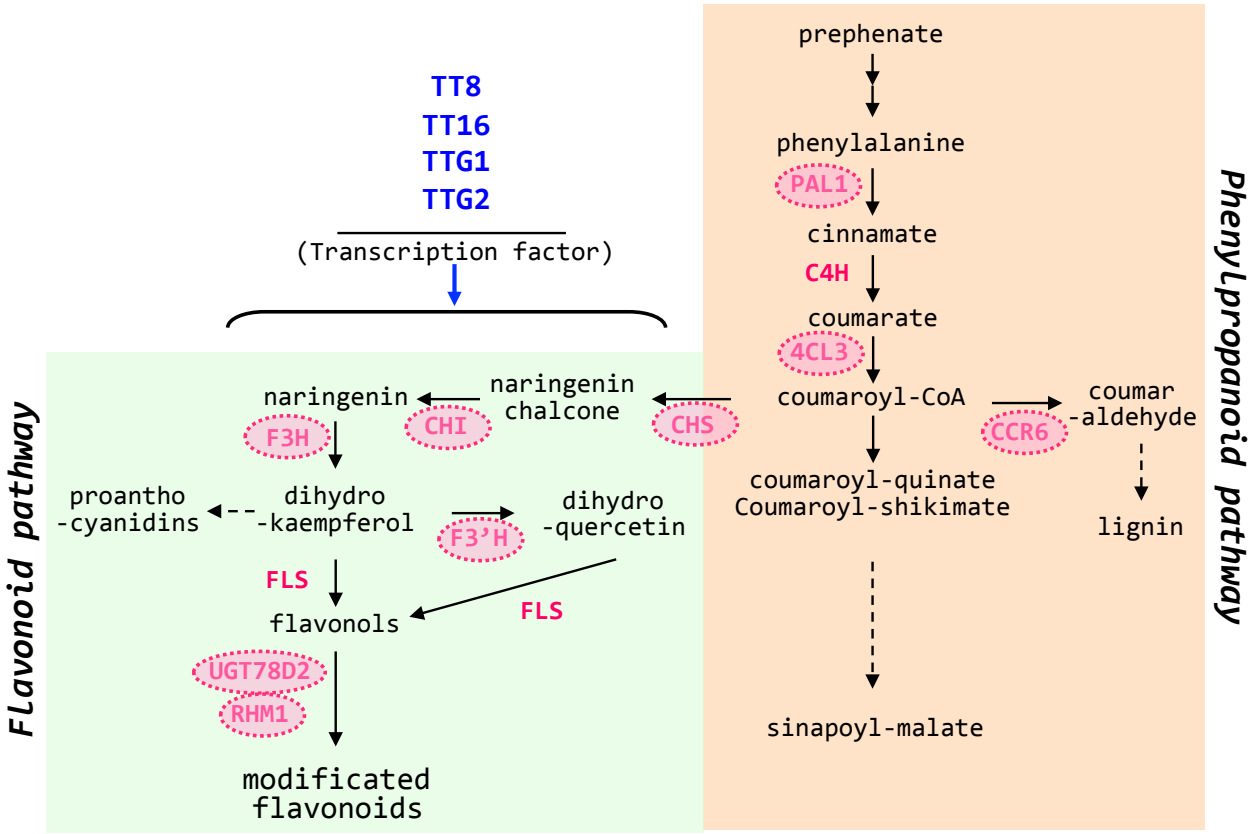

Figure 2.9: Simplified schematic representation of FB and phenylpropanoid biosynthesis pathways. Enzymes and regulator are indicated by bold, capital letters. Pathway genes identified by *pwsrc.knorm* from top 20 list in Table 2.9(d) are marked by dotted circles.

Chapter 3

Project II: Identifying Gene Signatures Contributing to Inter-Individual Variation in Statin Efficacy

3.1 Introduction

Statins are widely prescribed drugs that lower the risk of cardiovascular disease (CVD) by reducing low density lipoprotein cholesterol (LDLC) levels. They act by inhibiting 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGCR), the enzyme which catalyzes the rate limiting step of cholesterol biosynthesis in liver [Figure 3.1 and 3.2]. Although numerous trials have demonstrated statin efficacy in the reduction of CVD risk, there is substantial variation between individuals in the magnitude of plasma LDLC reduction [78, 70, 58]. For example, in the Cholesterol and Pharmacogenetics (CAP) study, the LDLC changes for the 372 participants after simvastatin therapy were collected [Figure 3.9].[1] While the result appears normally distributed, it is clear that there are some extreme responders who are either sensitively responding (red colored), or poorly or not responding (blue colored) to the treatment.

Variation in the LDLC response to statin treatment has been attributed to both genetic and nongenetic factors. These include single nucleotide polymorphisms (SNPs), haplotypes in genes encoding key regulators of cholesterol metabolism including HMGCR, APOE, PCSK9, ACE, LDLR and ABC B1 [12, 18, 50, 87], as well as phenotypic predictors such as race, age and smoking status [78]. However, the extent of variation in statin efficacy explained by these polymorphisms is limited [87, 59]. Recently, whole genome approaches have identified one SNP (*rs8014194*) that might be associated with variation in the magni-

[1] CAP is one of two pharmacogenetics studies conducted by the Pharmacogenetics and Risk of Cardiovascular disease (PARC) study group.

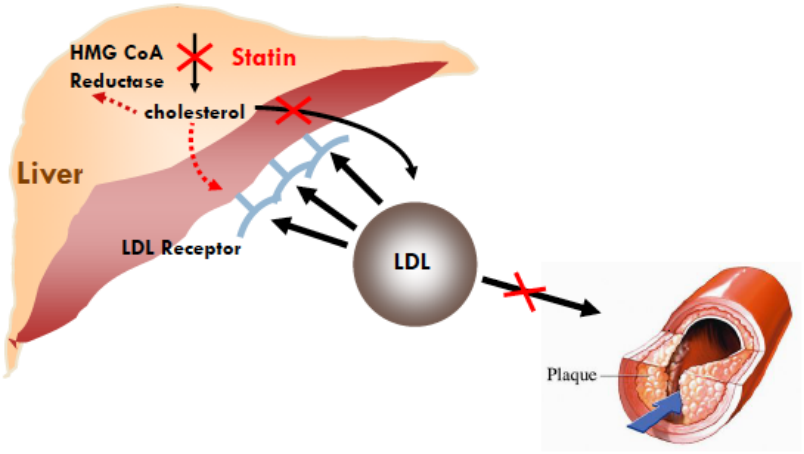

Figure 3.1: Statins reduce cardiovascular disease by lowering LDLC.

tude of statin-mediated reduction in total cholesterol and LDLC [5]. However, this result is somewhat preliminary and only explains a small portion of variation in statin efficacy.

This project aims to perform more comprehensive studies to identify gene signatures which cause inter-individual variation in drug treatment. Specifically, we try to discover gene signatures responsible for the the high and low responder variation in statin efficacy. For this, NMF analysis is used to demonstrate the existence of distinct patterns between the high and low responder groups from gene expression data. Based on the separated molecular patterns, pattern specific genetic markers are identified using a modified Significance Analysis of Microarrays (SAM) method. Biological validation processes demonstrate the relevance of the identified gene signatures with the biological mechanisms contributing to inter-individual variation in statin efficacy.

3.2 Method

In this section, brief overview of the NMF and its algorithm, as well as the model selection rule is described. Then, a novel strategy on searching for signature genes is explained.

3.2.1 NMF and its algorithm

Nonnegative Matrix Factorization (NMF) is an unsupervised, parts-based learning algorithm in which a nonnegative matrix $\mathbf{A}$ is decomposed into two nonnegative matrices $\mathbf{A} \approx \mathbf{WH}$ through a series of multiplicative updates. Introduced by Lee and Seung in 1999 in the context of text mining and facial pattern recognition [52], NMF has been widely applied in areas such as image processing, natural language processing [77, 11], sparse coding [38, 39], speech recognition, video summarization and internet research [56, 13].

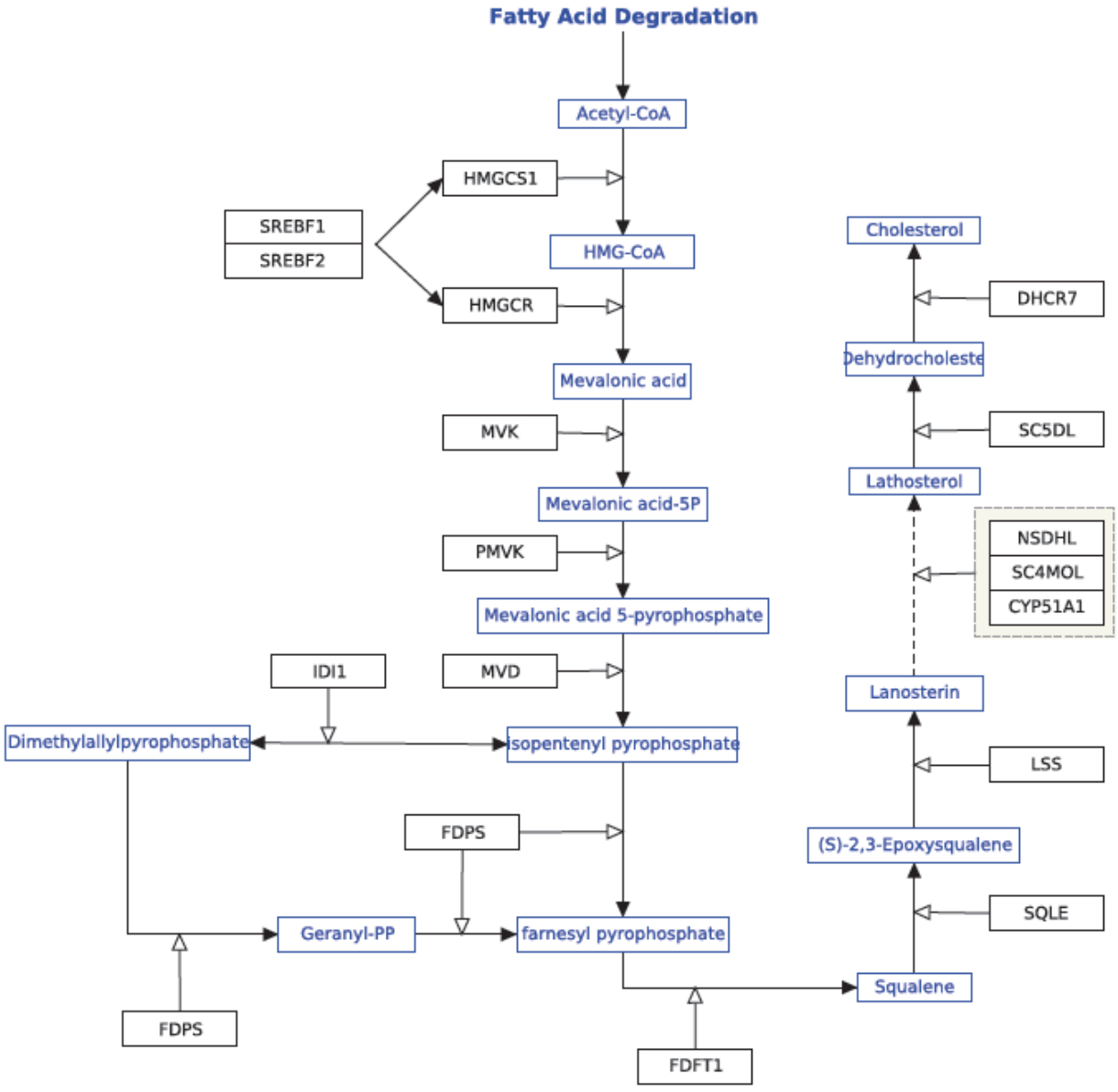

Figure 3.2: Cholesterol biosynthesis pathway in homo sapiens.

Three features distinguishing NMF as a unique and popular choice over traditional decomposition techniques such as principal component analysis (PCA), independent component analysis (ICA) or singular value decomposition (SVD) are as follows. First, the matrix factors are constrained to have nonnegative entries which allows direct interpretation as real underlying components within the context defined by the original data. For example, Brunet *et al.*[8] interpreted the resulting basis factors as metagenes capturing gene expression patterns specific to different groups of samples. Second, NMF often generates sparse basis vectors with few non-zero entries which allows us to discover parts-based basis vectors [52]. This feature is especially useful in identifying clusters of samples and their characterization. Due to the sparseness of the results, each set of metagene-specific genes is small which eases the interpretation of the results. Finally, NMF does not require the basis components to be orthogonal or independent, but allows them to overlap. This feature is useful in searching for multiple pathways or processes in the context of gene expression microarrays, since overlapping metagenes could suggest collaborative pathways implicated in a complex network structure [8, 31].

In light of this, NMF recently has been popularly adopted for analyzing large scale gene expression data to obtain new insights into cancer type discovery [8, 31, 48, 14, 4, 28, 84], functional characterization of genes [10, 67, 48, 49, 25], predicting cis-regulating elements from positional word count matrices [41] and phenotype prediction using cross-platform microarray data [96].[2] In one notable example, Collisson *et al.* [14] used NMF with combined transcriptional profiles of primary pancreatic ductal adenocarcinoma (PDA) samples to successfully define three PDA subtypes and presented evidence for clinical outcome and therapeutic response differences among them. They further defined subtype specific gene signatures that may have utility in stratifying patients for treatment, namely subtype specific therapies. As such, NMF is a promising tool for the purpose of clustering gene expression data to identify distinct functional molecular patterns.

There are several modified versions of the NMF algorithm [51, 31, 48, 66, 4] but the original version suggested by Brunet *et al.* [8] is used in our study. NMF aims to extract a small number of features (ranks), each defined as a positive linear combination of the n genes, and express the gene expression level of the samples as a positive linear combination of these pre-defined features.

Given a nonnegative gene expression data matrix $\mathbf{A}$, which has n genes and m samples, NMF factorizes $\mathbf{A}$ into two matrices with positive entries $\mathbf{W}$ and $\mathbf{H}$ with rank k.[3]

$$\mathbf{A}_{n\times m} \approx \mathbf{W}_{n\times k}\mathbf{H}_{k\times m} \quad \text{s.t.} \min_{\mathbf{W},\mathbf{H}\geq 0}[D(\mathbf{A},\mathbf{WH}) + R(\mathbf{W},\mathbf{H})], \quad k \ll min(n,m) \tag{3.1}$$

$$\mathbf{H}_{au} \leftarrow \mathbf{H}_{au}\frac{\sum_i \mathbf{W}_{ia}\mathbf{A}_{iu}/(\mathbf{WH})_{iu}}{\sum_k \mathbf{W}_{ka}} \tag{3.2}$$

$$\mathbf{W}_{ia} \leftarrow \mathbf{W}_{ia}\frac{\sum_u \mathbf{H}_{au}\mathbf{A}_{iu}/(\mathbf{WH})_{iu}}{\sum_v \mathbf{H}_{av}} \tag{3.3}$$

[2]See [16] for a comprehensive review of applications of NMF in bioinformatics.

[3]k is the number of factors or features which is much smaller than n and m.

After randomly initializing the matrices, $\mathbf{W}$ and $\mathbf{H}$ are iteratively updated using the coupled divergence equations [52] in Equation (3.2) and (3.3) to minimize the objective function $D(\mathbf{A}, \mathbf{WH}) + R(\mathbf{W}, \mathbf{H})$ in Equation (3.1). Notice that D in Equation (3.1) is defined as *Kullback-Leibler* loss function,

$$\mathbf{D}_{Kullback-Leibler} : \mathbf{A}, \mathbf{B} \mapsto KL(\mathbf{A} \,||\, \mathbf{B}) = \sum_{ij} a_{ij} log \frac{a_{ij}}{b_{ij}} - a_{ij} + b_{ij}, \tag{3.4}$$

and R is an optional regularization function defined to enforce smoothness or sparsity on the matrices $\mathbf{W}$ and $\mathbf{H}$ which is ignored in our study. To illustrate, Figure 3.3 provides a graphical summary of NMF analysis with rank $k = 2$. Matrix $\mathbf{W}$ has size $n \times k$, with each of the k columns defining a rank; entry w_{ij} is the coefficient of gene i in rank j. If gene i has a higher coefficient for the first rank than the second rank, this gene is regarded as contributing more to the first class than the second one. Matrix $\mathbf{H}$ has size $k \times m$, with each of the m columns representing the rank expression pattern of the corresponding sample; entry h_{ij} represents the expression level of rank i in sample j. As such, the profile of $\mathbf{H}$ enables us to assign m samples to k classes. For example, in the continuous profile in Figure 3.3, the first few samples have higher coefficient for the first rank while the last few samples have higher coefficient for the second rank. Thus we can easily assign the first group of samples into the first class and the second group into the second class.

3.2.2 Model selection, choice of k

For any rank k, NMF groups the samples into k clusters. But the key issue is to assess whether a given rank k provides a biologically meaningful decomposition of the data.

Brunet *et al.* [8] developed a model selection rule based on consensus clustering [61] for evaluating the quality of the resulting clusters. The procedures is as follows.

(i) For each run, calculate a connectivity matrix $\mathbf{C}$ of size $m \times m$ with $c_{ij} = 1$ if sample i and j are assigned to the same class and $c_{ij} = 0$ otherwise. Due to the random initialization of $\mathbf{W}$ and $\mathbf{H}$, NMF will not necessarily converge to the same solution on each run. However, if clustering into k classes is strong, the sample assignment rarely changes from run to run.

(ii) Average the connectivity matrices from many runs to obtain a consensus matrix $\bar{\mathbf{C}}$. The entries of $\bar{\mathbf{C}}$ range from 0 and 1 and reflect the frequency with which the two samples cluster together. If the clustering is stable, the connectivity matrices from different runs will vary little, resulting in a consensus matrix with entries close to 0 or 1.

(iii) Using the off diagonal entries as a measure of similarity among samples, the consensus matrix is reordered using average linkage hierarchical clustering to obtain $\bar{\mathbf{C}}_{HC}$.

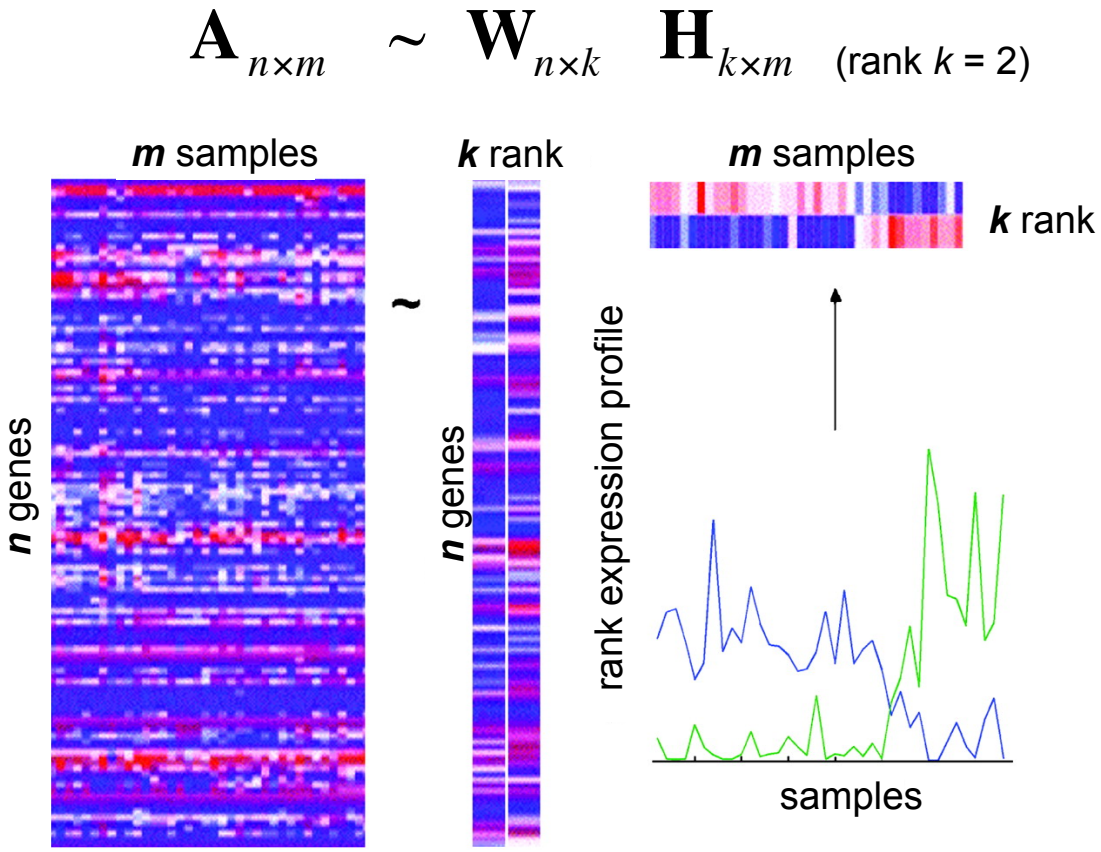

Figure 3.3: Graphical explanation of NMF analysis. A rank 2 reduction of a microarray dataset of n genes and m samples is obtained by NMF to give $\mathbf{W}$ and $\mathbf{H}$ of size $n \times 2$ and $2 \times m$, respectively. All matrix expression levels are color coded by using a heatmap from blue (minimum) to red (maximum).

(iv) To evaluate the stability of the clustering associated with a given rank k, the cophenetic correlation coefficients $\rho_k(\bar{\mathbf{C}})$ is calculated as $\rho_k(\bar{\mathbf{C}}) = cor(I - \bar{\mathbf{C}}, \bar{\mathbf{C}}_{HC})$. If the clustering does not vary at all, $\rho_k(\bar{\mathbf{C}}) = 1$; otherwise, the value of $\rho_k(\bar{\mathbf{C}})$ falls between 0 and 1.

By tracking how the cophenetic correlation coefficient changes as k changes, we can select the rank k where the cophenetic correlation coefficient begins to decrease.

3.2.3 Signature gene selection

Once NMF analysis identifies the molecular patterns that distinguish the high and low responder groups, we select signature genes that are differentially expressed between the two groups.

For this purpose, we referred to Significance Analysis of Microarrays (SAM) method, which has been popularly used to identify differentially expressed genes among groups based on statistical significance [89].

This method starts with assigning a score *relative difference*, $d(i)$, to each gene on the basis of change in the gene expression relative to the *gene specific scatter*, $s(i)$, which is the standard deviation of repeated measurements. In particular, the relative difference is defined as

$$d(i) = \frac{\bar{x}_H(i) - \bar{x}_L(i)}{s(i) + s_0} \tag{3.5}$$

$$s(i) = \sqrt{a\left(\sum_m [x_m(i) - \bar{x}_H(i)]^2 + \sum_n [x_n(i) - \bar{x}_L(i)]^2\right)} \tag{3.6}$$

where $\bar{x}_H(i)$ and $\bar{x}_L(i)$ are defined as the average level of expression for gene (i) in the high and low responder groups, respectively; a is a multiplicative factor of $a = (1/n_H + 1/n_L)/(n_H + n_L - 2)$, where n_H and n_L are the sample sizes of the high and low responder groups, respectively. Since the variance of $d(i)$ is sensitive to the general expression level of each gene, a small positive constant s_0 is introduced in the denominator in Equation (3.5) to stabilize the variance of $d(i)$.[4]

For genes with scores greater than an adjustable threshold, SAM uses permutations of the repeated measurements to estimate the percentage of genes identified by chance. Specifically, genes are first ranked by their $d(i)$ values and the expected relative difference $d_E(i)$ under the null distribution is computed via random permutation of samples, *i.e.* for each permutation p, relative differences $d_p(i)$ were calculated and ranked by magnitude and the expected relative difference is defined as the average over all the permutations $d_E(i) = \overline{d_p(i)}$. Then, gene i is deemed differentially expressed if $|d(i) - d_E(i)| > \Delta$. A more detailed discussion can be found in [89].

[4]The coefficient of variation $d(i)$ was computed as a function of $s(i)$ in moving windows across the data and the value for s_0 was chosen to minimize the coefficient of variation.

In the original paper [89], Tusher *et al.* found 34 differentially expressed genes from ionizing radiation response data using the SAM method. Those genes are represented by points displaced from the $d(i) = d_E(i)$ line by a distance greater than a threshold Δ in Figure 3.4(a).

However, the SAM method was not able to discover any significantly differentially expressed genes in our data [see Figure 3.4(b)]. This can be attributed to the following reasons. First, the variance in gene expression levels in our dataset is smaller than that of the ionizing data, a result of the gene-wise quantile normalization performed in our data pre-processing step. Second, the variance in $d(i)$ has not been completely stabilized, a more pronounced problem when $s(i)$ is small. Finally, the null distribution calculated for the test statistic in the original SAM method is not gene specific. Thus, it does not reflect the genes' true null distributions in an appropriate way.

To address these issues, we propose a new algorithm for extracting signature genes as follows.

(i) Calculate the relative difference $d(i)$ as in Equation (3.5) with varying positive s_0 values instead of a fixed one. Specifically, begin with a small constant value, decreasing it toward 0 as $s(i)$ increases. This enables us to stabilize $d(i)$ more efficiently especially when $s(i)$ is small.

(ii) To estimate the significance of the observed gene expression level, we calculate the null distribution of the test statistic, $d(i)$, based on random permutations of the samples within each gene. In contrast to SAM's general null distribution, using samples within each gene permits estimation of a gene-specific null distribution.

(iii) Obtain the corresponding p-value of the observed relative difference $d(i)$ based on the null distribution calculated in (ii).

(iv) Select the signature genes with p-values lower than a pre-selected cutoff, set to 0.01 in our study.

3.3 Datasets

3.3.1 Data descriptions

We used data generated by the Cholesterol and Pharmacogenetics (CAP) study, which enrolled 944 participants in a 6-week simvastatin trial (40mg/day) with the objective of examining the genetic factors affecting simvastatin-mediated changes in lipids and lipoproteins [78].

Participants were healthy adult volunteers who self-reported either recent African ($N = 335$) or recent Caucasian ($N = 609$) ancestry. Baseline health and demographic information were obtained at the time of enrollment. Fasting plasma was collected at two pre-treatment

time points (screen visit and enrollment visit) and at two post-treatment time points (four and six weeks of treatment). Lipids and lipoproteins were measured at all four time points. Total cholesterol (TC), triglyceride, apolipoprotein B (apoB), and high density lipoprotein cholesterol (HDLC) levels were measured using an Express 550 Plus analyzer (Ciba Corning, Oberlin, Ohio) in a laboratory that was monitored for consistency by the Centers for Disease Control - National Heart, Lung, and Blood Institute standardization program as described in [78]. LDLC was calculated by the Friedewald equation [26]. Because TC, HDLC and triglycerides were not significantly different between screen and enrollment, the average of these two measurements was used as the pretreatment value to minimize technical variation. For the same reason, the average of four- and six-week measurements was used as the post-treatment value.

RNA quality and quantity was assessed using a Nanodrop ND-1000 spectrophotometer and Agilent bioanalyzer. Paired RNA samples from 568 lymphoblastoid cell lines (LCLs),[5] selected based on RNA quality and quantity, were amplified and biotin labeled using the Illumina TotalPrep-96 RNA amplification kit, hybridized to Illumina HumanRef-8v3 beadarrays (Illumina), and scanned using an Illumina BeadXpress reader. Data were read into GenomeStudio and samples were selected for inclusion based on quality control criteria: (i) signal to noise ratio ($95^{th} : 5^{th}$ percentiles), (ii) matched gender between sample and data, and (iii) average correlation of expression profiles within three standard deviations of the within-group mean ($r = 0.99 \pm 0.0093$ for control-treated and $r = 0.98 \pm 0.0071$ for simvastatin-treated beadarrays). In total, viable expression data was obtained from 1040 beadarrays including 480 sets of paired samples for 10195 genes. Genes were annotated through biomaRt from ensMBL Build 5438.

Our study is limited to the 480 paired samples, all of Caucasian ancestry. Of those 480 samples, we excluded 64 smoker samples to eliminate a potential confounding factor. We also checked the sample identity using Eric Shadt's Bayesian method [73]. This is based on a Baysian approach to predict SNP genotypes using RNA expression data. The genotype predicted from our RNA expression data accurately and uniquely identified individuals in our population, excluding 44 mismatched samples which were finally removed from the data.

Figure 3.5 provides a graphical summary of the final 372 samples. Our final sample is composed of 171 female and 201 male participants, with ages ranging from 30 to 90. Their BMI distribution suggests that the participants in this study are somewhat overweight since the median BMI is above the normal BMI range (18.5 - 25) which is delineated using red dotted lines in Figure 3.5.

HMGCR enzymatic activity was measured for the 204 CAP participants (72 African and 132 Caucasian ancestry)[6] in the presence or absence of simvastatin (25, 50 nmol/L) [60]. This experiment was performed in 18 consecutive batches so that each batch had a balanced number of samples of Caucasian and African ancestry. Activity was expressed as picomole

[5]LCLs, immortalized by Epstein-Barr virus transformation of lymphocytes isolated from whole blood, were derived from European-American participants in the CAP trial.

[6]Among 132 Caucasian participants, 90 are overlapped with our 372 participants.

of mevalonate formed per minute per 500000 homogenized live cells.[7,8]

3.3.2 Data preprocessing

3.3.2.1 LDLC phenotype

The mean LDLC levels before and after simvastatin treatment are denoted as α and β, respectively, for ease of notation. To measure LDLC change, a multiplicative measure $log(\beta/\alpha)$, which we will refer to *deltaLnLDLC*, is preferred over the arithmetic difference $\beta - \alpha$ because it is comparatively less correlated with baseline values [Figure 3.6]. Since we are interested in statin response, not in the baseline, it is necessary to remove as much of the baseline component as possible. The logarithmic transformation removes outlier effects to some extent as well.

Next, clinical covariate effects on *deltaLnLDLC* are adjusted with linear regression. Among the three candidate covariates (sex, age and BMI) we considered, only age effects were significant [Figure 3.7]. So we adjusted for the age covariate to get *adj.deltaLnLDLC* which is used as a measure of LDLC change in our study. Figure 3.8 graphically summarizes the effect of covariate adjustment. Before adjustment [Figure 3.8(a)], the younger age group showed less LDLC change than the older groups, but this discrepancy is eliminated after adjustment [Figure 3.8(b)].

The distribution of LDLC change, *adj.deltaLnLDLC*, is shown in Figure 3.9 where the extremely high and low responders to the simvastatin treatment are color coded as red and blue, respectively.

3.3.2.2 Gene expression dataset

The treated and untreated gene expression data sets were quantile transformed to the overall average empirical distribution across all arrays, and pooled together for further processing. This gives a matrix with 10195 genes and 744 samples. To avoid problems caused by outliers or other deviations from normality in later association tests for eQTL analysis, the expression values for each gene were also quantile normalized.[9]

To find the clinical covariate effects, boxplots for seven known covariates (date, cell growth rate, RNA labeling batch, beadarray hybridization batch, gender, age and BMI) were generated as in Figure 3.10(a). Each boxplot summarizes the correlation coefficients of the expression levels before and after adjusting for each covariate. It turns out that four covariates - date, RNA labeling batch, beadarray hybridization batch and gender -

[7]The traditional units of per 1 mg protein may be inappropriate because protein may be quantified from both living and dead cells, whereas only living homogenized cells contribute to activity.

[8]Cell count was determined on the Guava Personal Cell Analysis fluorescent-activated cell sorting (Guava Technologies Inc, Hayward, Calif).

[9]This gene expression data was originally prepared for eQTL analysis purpose and we followed the protocol they used to preprocess the array data.

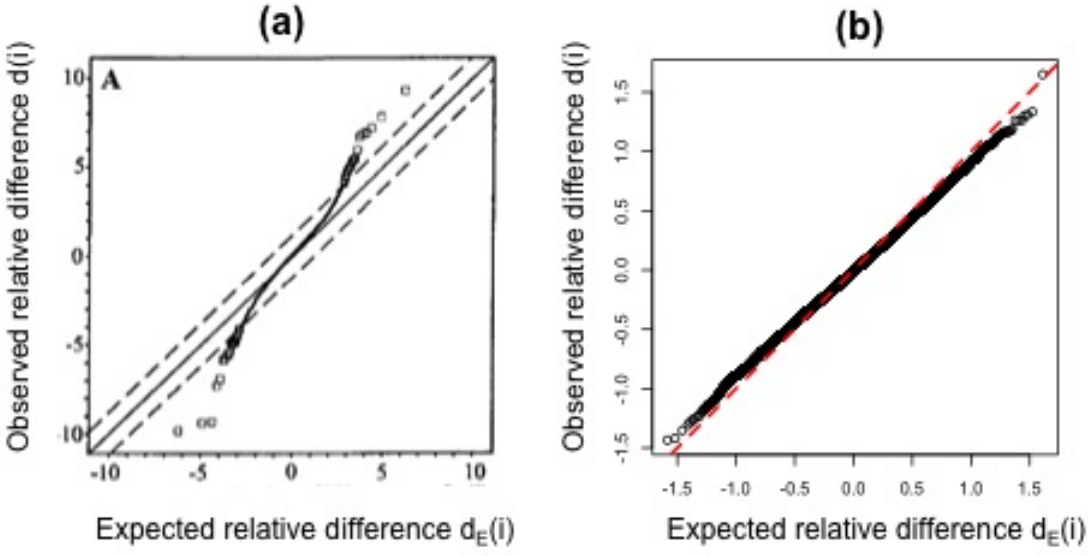

Figure 3.4: Scatter plots of the observed relative difference $d(i)$ versus the expected relative difference $d_E(i)$ demonstrating the performance of SAM method with (a) ionizing radiation response data in [89] and (b) our statin dataset.

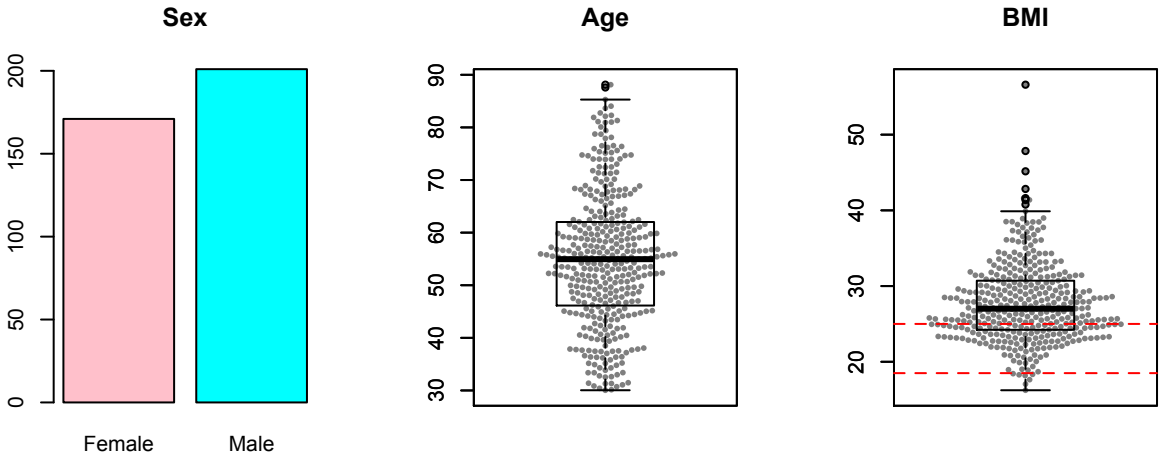

Figure 3.5: Graphical summary of 372 Caucasian participants by sex, age and BMI.

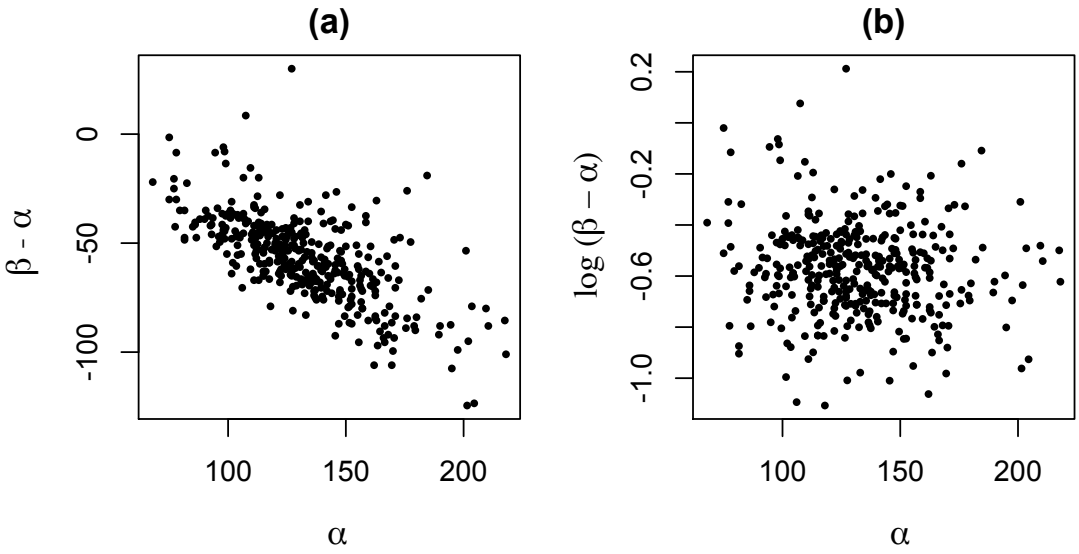

Figure 3.6: Comparison of (a) $\beta - \alpha$ versus (b) $log(\beta - \alpha)$.

```
Call:
lm(formula = deltaLnLDLC ~ Sex + Age + BMI)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52544 -0.11259  0.00384  0.09823  0.75471

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.455e-01  6.771e-02  -6.579 1.63e-10 ***
SexMale      3.051e-02  1.917e-02   1.592  0.11226
Age         -2.597e-03  7.560e-04  -3.435  0.00066 ***
BMI         -7.409e-05  1.810e-03  -0.041  0.96737
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.7: Summary of the linear regression result of the clinical covariates on *deltaLnLDLC*.

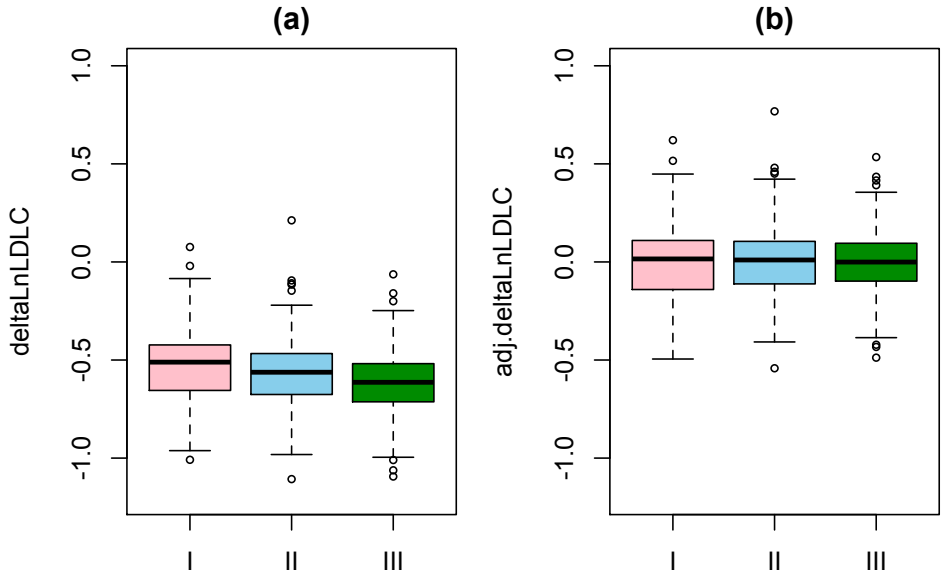

Figure 3.8: Comparison of *deltaLnLDLC* versus *adj.deltaLnLDLC* levels across different age groups. Age is arbitrarily split into three groups: (I) younger than or equal to 45, (II) older than 45 and younger than 60, (III) older than or equal to 60.

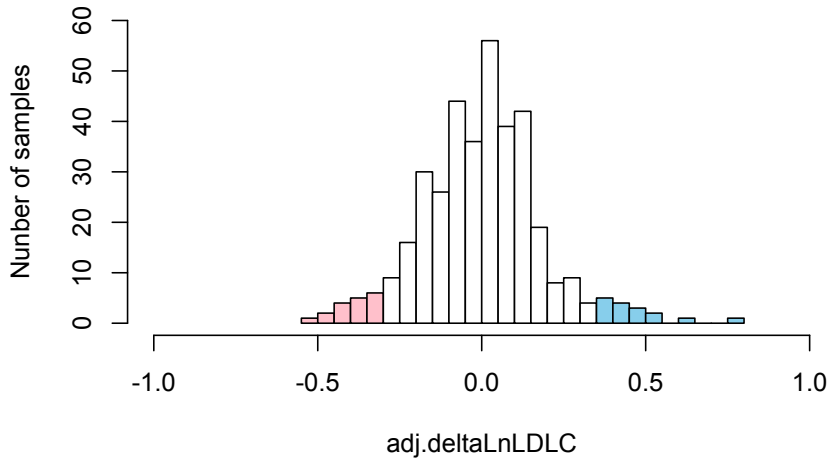

Figure 3.9: Distribution of LDLC change (*adj.deltaLnLDLC*), for 372 samples. 13 each of the highest and lowest responders for the simvastatin treatment are color coded with red and blue, respectively.

significantly affected the expression. So the data was adjusted to control these four covariates by quantile normalizing the residuals from a linear model with each gene.

Then the data was split into the treated ($T372$) and untreated ($U372$) groups to obtain the delta data, which is $D372 = T372 - U372$. We again quantile normalized each gene in $D372$.

Finally, to meet the nonnegative requirement of NMF, each entry in $D372$ was replaced with the value of its p-value subtracted from one, where the p-values were calculated under the assumption of normality.

3.3.2.3 HMGCR enzymatic activity data

The treated (T) and untreated (U) activity data were both adjusted for the total protein values. Then, samples with increased enzymatic activity were excluded, leaving 193 samples.

Due to the original experiment design, the data exhibits huge batch effects as well as dose effects [see Figure 3.11]. Consequently, we used linear regression to adjust for these covariates in the treated (T') and untreated (U') data separately.

Finally, the HMGCR enzymatic activity change was computed as $(T' - U')/U'$.

3.4 Results

In NMF analysis, we first investigate how many samples to use for the best separation between the high and low responders.

To evaluate the cluster quality for each number of samples, *purity* and *entropy* [100, 99] measures were used. Purity and entropy are two widely used measures to evaluate the performance of unsupervised learning algorithms. Purity evaluates the coherence of a cluster, *i.e.* the extent to which a cluster contains samples from a single class. Given a particular cluster of C_i of size n_i, the purity of C_i is defined as

$$P(C_i) = \frac{1}{n_i} \max_h (n_i^h) \tag{3.7}$$

where $\max_h(n_i^h)$ is the number of samples from the dominant class in cluster C_i and n_i^h is the number of samples from cluster C_i assigned to class h. The purity of clustering result is the weighted sum of the purity of k individual clusters given by

$$purity = \sum_{h=1}^{k} \frac{n_h}{n} P(C_h). \tag{3.8}$$

Entropy measures how the various classes of samples are distributed in a given cluster, which is defined as

$$E(C_i) = -\frac{1}{\log k} \sum_{h=1}^{k} \frac{n_i^h}{n_i} \log \frac{n_i^h}{n_i} \tag{3.9}$$

where k is the total number of classes in the dataset. The entire entropy is defined to be the sum of the individual entropies weighted according to the cluster size:

$$Entropy = \sum_{h=1}^{k} \frac{n_h}{n} E(C_h). \tag{3.10}$$

The entropy measure is more comprehensive than purity because it considers the overall distribution of all the categories in a given cluster rather than just the number of objects in the dominant category. When the clustering is perfect, the purity and the entropy are 1 and 0, respectively. As such, higher purity and lower entropy correspond to better clustering results.

Given the best clustering results, we identified the signature genes and performed further biological validation studies.

3.4.1 Implementation of NMF analysis

3.4.1.1 Selecting the best number of samples

Given the LDLC change profile from our samples [Figure 3.9], one question which naturally arises is '*How many samples to choose from both groups for best separation?*'. With too few samples, key features of the two groups may be omitted. However, with too many samples, the NMF algorithm might have difficulty distinguishing the two groups.

To obtain a suitable set of samples that reflect the biological distinctions between the two groups, we adopted the following random sampling strategy.

(i) Choose N each of the highest and lowest responding samples from both tails, $2N$ in total. N is increased from 10 to 20.[10]

(ii) Select the G genes with the greatest variation between the two groups as follows. Rank genes within each sample. For each gene, find the mean rank both in the high ($\bar{R}_H$) and low ($\bar{R}_L$) responder groups and calculate the absolute rank difference $d = |\bar{R}_H - \bar{R}_L|$. Select G genes with the highest d values. In our study, G was selected to be 2000.

(iii) Perform NMF analysis and evaluate the cluster quality using the purity and entropy measures.

(iv) Repeat (i) through (iii) for another randomly selected set of $2N$ samples.

By comparing the NMF clustering quality between two sampling processes, we can choose a reasonable number of samples.

The resulting purity and entropy plots are given in Figure 3.12. Overall, the samples from two extreme responder groups achieved higher purity and lower entropy than the randomly

[10] N is originally tested from 5 to 30, but the results for 10 to 20 are shown for simplicity.

selected samples. NMF analysis resulted in a correct categorization when up to 26 (13 high, 13 low responders) samples were used. As more samples were included, the purity and entropy worsened. For the analysis described in Section 3.4.2 and 3.4.3, 26 samples were used. For comparison, we also used 32 samples which was the largest number of samples that still yielded a reasonable purity value (0.9).

The NMF results with 26 samples are summarized in Figure 3.13. A plot of the cophenetic correlation coefficients in Figure 3.13(b) shows that the best rank is two. Further, the reordered consensus matrix plots in Figure 3.13(a) show that the clustering result is most stable and robust at rank two. This confirms the existence of a clear-cut separation between two responder groups.

The analysis with 32 samples had similar results, but the overall cophenetic correlation coefficients were smaller [Figure 3.14]. This underscores the trade-off relationship between clustering quality and sample numbers. In other words, the addition of more samples to the NMF analysis lowers the clustering quality and vice versa.

3.4.1.2 The effects of the number of genes in NMF analysis

In our NMF analysis, the number of genes was fixed at 2000, a number widely used in other applications as well. To see how the number of genes affect the NMF analysis, we evaluate the NMF clustering quality using different numbers of genes.

Figure 3.15 shows the purity and entropy computed using the NMF clustering based on the 1000, 2000, 3000, 4000 and 5000 most variable genes. For each setting, we again increased the number of samples from 20 to 40. For smaller number of genes like 1000 or 2000, the overall clustering quality stayed perfect up to 28 and 26 samples, respectively. However, with more genes, the overall clustering quality worsens; the purity is reduced even in the 20 sample case.

More information could be extracted if more genes are used. However, this observation shows that not all genes carry useful information for separating the high and low responders, only adding noise to the analysis. This again confirms that 2000 is a reasonable number of genes to use for clustering.

3.4.1.3 NMF analysis with the high, low and mild responder groups

Other questions arise when analyzing data from the high and low responder groups, such as *'What if we add the mild responders in NMF analysis? Would that be clustered separately or together with others?'*. An answer to this question would provide a more comprehensive understanding of the inter-individual variation; subsequently, we performed the NMF analysis with three groups - high, low, and mild responder groups. The mild responder samples were taken from the middle area of the distribution in Figure 3.9 and the NMF analysis was performed together with the pre-selected high and low responders.

Figure 3.16 shows the cophenetic correlation coefficients and reordered consensus matrices for a total of 30, 39 and 48 samples, with a third of the samples taken from each of the three

Number of high, mild, low responders	Predicted	High responders	Mild responders	Low responders
30	class 1	10	9	2
	class 2	0	1	8
39	class 1	13	8	0
	class 2	0	5	13
48	class 1	13	10	2
	class 2	3	6	14

Table 3.1: Summary of the NMF clustering results with the high, mild and low responder groups.

responder groups. For example, 10 samples were taken from each of the high, mild and low responders for a total of 30 samples. In all of the cases, two remains the most desired rank value.

Table 3.1 summarizes the detailed clustering results from Figure 3.16. When we used 39 or 48 samples, roughly 62 % of the mild responders were clustered with the high responders and the rest were assigned to the low responders. However, if we take 30 most extreme samples, most of the mild responders (9 out of 10) were clustered with the high responders. This implies that the expression level from the mild group was not distinct enough to be clustered separately from the high and low responders. Instead, they behave more like the high responder group as the high and low responder groups move toward the extreme tails.

This is also supported by Figure 3.13(c) and 3.14(c). These heatmaps compare the expression levels among the high, mild and low responder groups using the signature genes identified in Section 3.4.2. While the expression levels from the high and low responders are distinct from each other, the ones from the mild group are somewhat random and intermediate. However, we can still see the rough similarity of the expression levels between the high and mild groups.

3.4.2 Signature gene selection

The original SAM method found no minimum of the coefficient of variation of $d(i)$ within the range of $s(i)$ in our dataset [Figure 3.17(b) right]. In the scatter plot of the gene specific scatter $s(i)$ versus the relative difference $d(i)$ [Figure 3.17(b) left], the variance of $d(i)$ is shown to be unstable especially when $s(i)$ is small. For comparison, the scatter plot with $s_0 = 0$ is given in Figure 3.17(a).

When the varying s_0 values described in Section 3.2.3 were used instead, a prominent minimum of the coefficient of variation of $d(i)$ was found, which is 0.13 [Figure 3.17(c)]. The corresponding scatter plot in Figure 3.17(c) looks less correlated than the one in Figure 3.17(b). This indicates that the dependency of $d(i)$ on $s(i)$ has been successfully removed.

The identified signature genes are denoted by red dots in the scatter plots of $d(i)$ versus $s(i)$ [Figure 3.18(d)]. For comparison, three more signature gene sets were obtained as

follows. Figure 3.18(a)(c) and Figure 3.18(b)(d) were generated by holding $s_0 = 0$ fixed at 0 and allowing s_0 to vary, respectively. The null distribution of $d(i)$ is empirically estimated through random permutations, within either the 26 high and low responder group [Figure 3.18(a)(b)] or all 372 population samples [Figure 3.18(c)(d)].

Again, when s_0 is allowed to vary, the variance of $d(i)$ becomes more stable [compare Figure 3.18(a)(c) and Figure 3.18(b)(d)]. Further, using all of the population samples in random permutation yields a more uniform signature gene distribution, irrespective of $s(i)$ values [compare Figure 3.18(b) and (d)]. For ease of notation, the signature genes identified in Figure 3.18(a)-(d) are named as *SG1*, *SG2*, *SG3* and *SG4*.

3.4.3 Biological validation

SG4 achieves greater statistical significance than *SG1-SG3*. Toward biological validation, two studies were performed: first, a correlation study with the cholesterol biosynthesis pathway genes; second, a comparison study with HMGCR enzymatic activity data. The genes in *SG4* are listed in Table 3.2.

3.4.3.1 A correlation study with the cholesterol biosynthesis pathway genes

Since statin directly works on the cholesterol biosynthesis pathway by inhibiting HMGCR enzymatic activity [Figure 3.2], genes in this pathway are highly affected by statin treatment, though the high and low responder groups may be affected differently. To demonstrate the relevance of our signature genes with the cholesterol biosynthesis pathway, we examined the extent to which the signature gene sets were enriched with cholesterol biosynthesis pathway genes.

For this purpose, thirteen representative genes were selected from the cholesterol biosynthesis pathway [Figure 3.2].[11] For each pathway gene, (*i*) compute the pairwise correlations with the remaining 10194 genes, (*ii*) select the significantly correlated genes at a 0.05 FDR adjusted p-value cutoff, and (*iii*) count how many of the correlated genes are overlapped with each of *SG1-SG4*. Both the untreated and treated gene expression datasets were used in this study. The results are summarized in Table 3.3 and 3.4, respectively.

Table 3.3(a) and 3.4(a) show the proportions of overlapped genes in *SG1-SG4*, while Table 3.3(b) and 3.4(b) give their corresponding p-values, calculated from the hypergeometric distribution. With the untreated data, for example, 47 % of *SG4* genes are overlapped with HMGCR-correlated genes (p-value $= 1.26 \times 10^{-4}$). It is worth noting that only *SG4* has a statistically significant overlap with the HMGCR-correlated genes in both datasets, which implies that *SG4* is highly enriched with HMGCR-correlated genes. The same correlation study was also performed with the delta data ($D372$), but no genes achieved statistical significance.

[11] For comparison, we added LDLR gene from the cholesterol uptake pathway.

Gene symbol	Ensemble ID	sign($d(i)$)	Gene symbol	Ensemble ID	sign($d(i)$)
C1orf218	ENSG00000213047	-1	ACVR1B	ENSG00000135503	1
FBXO16	ENSG00000214050	-1	ALPPL2	ENSG00000163286	1
MYLK2	ENSG00000101306	-1	FSD1CL	ENSG00000106701	1
SETBP1	ENSG00000152217	-1	HIST1H2BK	ENSG00000197903	1
B4GALT4	ENSG00000121578	-1	HIST1H4H	ENSG00000182217	1
GOPC	ENSG00000047932	-1	NTRK2	ENSG00000148053	1
EPOR	ENSG00000187266	-1	SDC4	ENSG00000124145	1
KIAA1333	ENSG00000092140	-1	SYPL1	ENSG00000008282	1
PDPR	ENSG00000090857	-1	B3GALNT1	ENSG00000169255	1
DNAJC7	ENSG00000168259	-1	CNKSR3	ENSG00000153721	1
SLC44A1	ENSG00000070214	-1	EOMES	ENSG00000163508	1
ARHGDIB	ENSG00000111348	-1	KLF11	ENSG00000172059	1
CEACAM21	ENSG00000007129	-1	LRRC26	ENSG00000184709	1
LARGE	ENSG00000133424	-1	PEX7	ENSG00000112357	1
TMEM1	ENSG00000160218	-1	SDPR	ENSG00000168497	1
UBE2CBP	ENSG00000118420	-1	BRWD2	ENSG00000120008	1
ADAT1	ENSG00000065457	-1	CRIM1	ENSG00000150938	1
CYP51A1	ENSG00000001630	-1	D4S234E	ENSG00000168824	1
NUDT4P1	ENSG00000177144	-1	LTB	ENSG00000204487	1
PITPNC1	ENSG00000154217	-1	LYPLAL1	ENSG00000143353	1
CYorf15A	ENSG00000099749	-1	NCF4	ENSG00000100365	1
KIAA1407	ENSG00000163617	-1	TBX15	ENSG00000092607	1
SEC16A	ENSG00000148396	-1	UCHL1	ENSG00000154277	1
TTC3	ENSG00000182670	-1	CYB5A	ENSG00000166347	1
GRPEL2	ENSG00000164284	-1	EIF1AY	ENSG00000198692	1
LMTK2	ENSG00000164715	-1	HIGD2A	ENSG00000146066	1
NPAT	ENSG00000149308	-1	LOC389816	ENSG00000184709	1
ADAM19	ENSG00000135074	-1	MYO6	ENSG00000196586	1
ARMC6	ENSG00000105676	-1	TSPAN1	ENSG00000117472	1
CDRT4	ENSG00000175106	-1	ADORA2B	ENSG00000170425	1
GALM	ENSG00000143891	-1	EIF4E2	ENSG00000135930	1
ABCA9	ENSG00000154258	-1	EPDR1	ENSG00000086289	1
LRBA	ENSG00000198589	-1	HIST1H2BG	ENSG00000168242	1
SUZ12	ENSG00000178691	-1	LCK	ENSG00000182866	1
ZNF398	ENSG00000197024	-1	MBP	ENSG00000197971	1
			PON2	ENSG00000105854	1
			ACACB	ENSG00000076555	1
			HIST1H1C	ENSG00000187837	1
			KREMEN1	ENSG00000183762	1
			RNF125	ENSG00000101695	1
			TSPAN4	ENSG00000214063	1
			TTC9C	ENSG00000162222	1
			LOC391356	ENSG00000184924	1
			MRLC2	ENSG00000118680	1
			NMNAT3	ENSG00000163864	1
			ANKRD29	ENSG00000154065	1
			HIBCH	ENSG00000198130	1
			LY86	ENSG00000112799	1
			MGC42105	ENSG00000177453	1
			RGS1	ENSG00000090104	1
			RNASEH2A	ENSG00000104889	1
			SYNGR1	ENSG00000100321	1
			TGFB1I1	ENSG00000140682	1
			TMEM60	ENSG00000135211	1
			C9orf21	ENSG00000158122	1
			LIN7A	ENSG00000111052	1
			ADCK2	ENSG00000133597	1
			C14orf100	ENSG00000050130	1
			C5orf13	ENSG00000134986	1
			CYP2W1	ENSG00000073067	1
			GABRE	ENSG00000102287	1
			MS4A7	ENSG00000166927	1
			SH2D2A	ENSG00000027869	1
			TRAPPC2	ENSG00000196459	1

Table 3.2: List of 99 signature genes in *SG4*.

(a)

Gene symbol	SG1	SG2	SG3	SG4
DHCR7	0.3	0.27	0.3	0.27
SC5DL	0.45	0.41	0.46	0.53
NSDHL	0.31	0.32	0.35	0.29
SC4MOL	0.34	0.34	0.32	0.39
CYP51A1	0.17	0.13	0.17	0.16
LSS	0.27	0.24	0.26	0.24
SQLE	0.21	0.21	0.26	0.23
FDFT1	0.17	0.17	0.16	0.17
MVD	0.27	0.29	0.27	0.3
PMVK	0.23	0.22	0.21	0.25
MVK	0.2	0.21	0.21	0.2
HMGCR	0.39	0.37	0.37	0.47
HMGCS1	0.19	0.2	0.21	0.26
LDLR	0.45	0.46	0.45	0.48

(b)

Gene symbol	SG1	SG2	SG3	SG4
DHCR7	0.942	0.985	0.935	0.966
SC5DL	0.922	0.976	0.864	0.324
NSDHL	1	0.999	0.997	1
SC4MOL	1	0.999	1	0.948
CYP51A1	1	1	1	0.999
LSS	0.998	1	0.999	0.999
SQLE	1	1	0.998	0.998
FDFT1	0.968	0.958	0.99	0.932
MVD	1	0.998	1	0.991
PMVK	0.978	0.985	0.997	0.877
MVK	0.99	0.962	0.983	0.963
HMGCR	0.013	0.041	0.038	1.26E-04
HMGCS1	1	0.999	0.997	0.868
LDLR	0.728	0.632	0.736	0.379

Table 3.3: (a) The proportions of the overlapped genes in *SG1-SG4* with the list of genes correlated with each of cholesterol biosynthesis pathway genes. Correlation is calculated using the untreated gene expression data. 3713, 5173, 4728, 4895, 3022, 3963, 3830, 2440, 4324, 3171, 2914, 3111, 3258, 4838 genes are identified to be significantly correlated with each of the pathway genes, respectively. *SG1-SG4* are composed of 150, 123, 146 and 99 genes, respectively. The corresponding p-values are shown in (b).

(a)

Gene symbol	SG1	SG2	SG3	SG4
DHCR7	0.31	0.3	0.31	0.33
SC5DL	0.4	0.36	0.37	0.39
NSDHL	0.35	0.37	0.38	0.35
SC4MOL	0.25	0.26	0.25	0.24
CYP51A1	0.17	0.17	0.18	0.18
LSS	0.36	0.33	0.37	0.23
SQLE	0.19	0.21	0.23	0.19
FDFT1	0.07	0.06	0.07	0.11
MVD	0.39	0.4	0.38	0.26
PMVK	0.21	0.21	0.22	0.26
MVK	0.18	0.2	0.23	0.15
HMGCR	0.16	0.19	0.18	0.28
HMGCS1	0.07	0.09	0.09	0.13
LDLR	0.29	0.33	0.32	0.27

(b)

Gene symbol	SG1	SG2	SG3	SG4
DHCR7	0.744	0.808	0.781	0.526
SC5DL	0.939	0.99	0.988	0.91
NSDHL	1	0.999	0.999	0.999
SC4MOL	1	1	1	1
CYP51A1	0.998	0.997	0.995	0.983
LSS	0.693	0.855	0.598	0.999
SQLE	1	0.995	0.993	0.997
FDFT1	0.993	0.995	0.991	0.689
MVD	0.976	0.952	0.986	1
PMVK	1	0.999	0.999	0.953
MVK	1	0.997	0.991	1
HMGCR	0.864	0.575	0.616	0.016
HMGCS1	0.985	0.908	0.932	0.447
LDLR	1	0.995	0.999	1

Table 3.4: (a) The proportions of the overlapped genes in *SG1-SG4* with the list of genes correlated with each of cholesterol biosynthesis pathway genes. Correlation is calculated using the treated gene expression data. 3487, 4747, 5246, 4372, 2848, 3909, 3269, 1365, 4859, 3530, 3241, 2025, 1356, 4591 genes are identified to be significantly correlated with each of the pathway genes, respectively. *SG1-SG4* are composed of 150, 123, 146 and 99 genes, respectively. The corresponding p-values are shown in (b).

3.4.3.2 Comparison of HMGCR enzymatic activity change

Observing that $SG4$ is significantly enriched with HMGCR-correlated genes naturally lead us to hypothesize that there might be a notable difference in the amount of HMGCR enzymatic activity change between the high and low responders.

To verify this hypothesis, the amount of HMGCR enzymatic activity change between the two groups was compared. Since HMGCR activity data is not available for all the 372 samples, we only compared the available activity data which were ranked within the top 13, 20 and 50 high or low responders.

Boxplots comparing the activity change as well as the baseline activity are shown in Figure 3.19. The number of high and low responders used to generate each boxplot is noted in the parentheses. A t-test for comparing two groups in each boxplot is performed and the resulting p-values are given at the top of each plot. Comparing the top 13 high and low responders (seven and four samples of activity data are available, respectively), the low responder group showed less activity change than the high responder group, though the difference was not significant (p-value = 0.194). However, the difference became significant (p-value = 0.049) when we compared the top 50 high and low responders (26 and 24 samples, respectively). This demonstrates that there is a significant difference in the amount of HMGCR enzymatic activity change between the high and low responder groups.

Due to the limited data, our hypothesis has not been fully verified. As more HMGCR enzymatic activity data becomes available in the future, a more precise comparison between the two groups could be made to better verify our hypothesis.

3.5 Discussion

Using the NMF analysis, we identified two distinct molecular patterns between the high and low responder groups. The number of samples that produced the best separation between these groups was found to be 13 each of the highest and lowest responders. Using this optimal number of samples, our modification of the SAM method identified 99 signature genes that had gone undetected by the original SAM method.

Two independent studies were performed to validate our findings biologically. The correlation study showed that our signature genes were significantly enriched with HMGCR-correlated genes; 47 % (p-value = 1.26×10^{-4}) and 28 % (p-value = 1.60×10^{-2}) of our signature genes are overlapped with the HMGCR-correlated genes in the untreated and treated gene expression datasets, respectively. In the comparison study, we observed that there is a notable difference in the amount of HMGCR enzymatic activity change between the high and low responder groups. The high responder group exhibits a bigger activity decrease than the low responder group, implying that statin inhibits the HMGCR enzymatic activity more efficiently in the high responder groups. These results will help us understand why the high responder group shows a greater LDLC decrease and higher statin efficacy than the low responder group.

Our studies suggest the plausible relevance of our signature genes to the cholesterol biosynthesis pathway, which HMGCR mainly acts on. However, the results shown here are preliminary and there is still room for further validation. Further research is being done to ascertain if there is more concrete biological evidence of our findings.

- In the correlation study in Section 3.4.3.1, we noticed that only 47 % or 28 % of our signature genes are enriched with the HMGCR-correlated genes from the untreated and treated data, respectively. One naturally arising question is *'What about other genes? What biological pathways or processes are they involved in?'* To answer this question, we need to take into account other biologically related pathways such as cholesterol uptake and export pathways.

- We can look for more references on each signature gene to better understand its biological relevance to our study. For example, our data showed that the 'CYP51A1' gene was highly activated for the low responder group. As it turns out, this gene is known to be involved in lipid and lipoprotein metabolism, as well as cholesterol biosynthesis.

- We are also planning to perform genotypic association studies between the two groups using expression quantitative trait loci (eQTL) data. Using comprehensive genetic association studies with our signature genes, we expect to identify single nucleotide polymorphisms (SNPs) that are associated with the variation in statin response.

In this study, we only considered the LDLC phenotype which is known to have high physiological relevance to CVD risk. There are also other phenotypes that are relevant to our study, such as total cholesterol, high density lipoprotein (HDL) cholesterol, apolipoprotein B (apoB) as well as the rate of LDL cholesterol uptake. Since phenotypes each provide different intrinsic information, using these phenotypes within our analytical framework might shed more light on the inter-individual variation in statin efficacy.

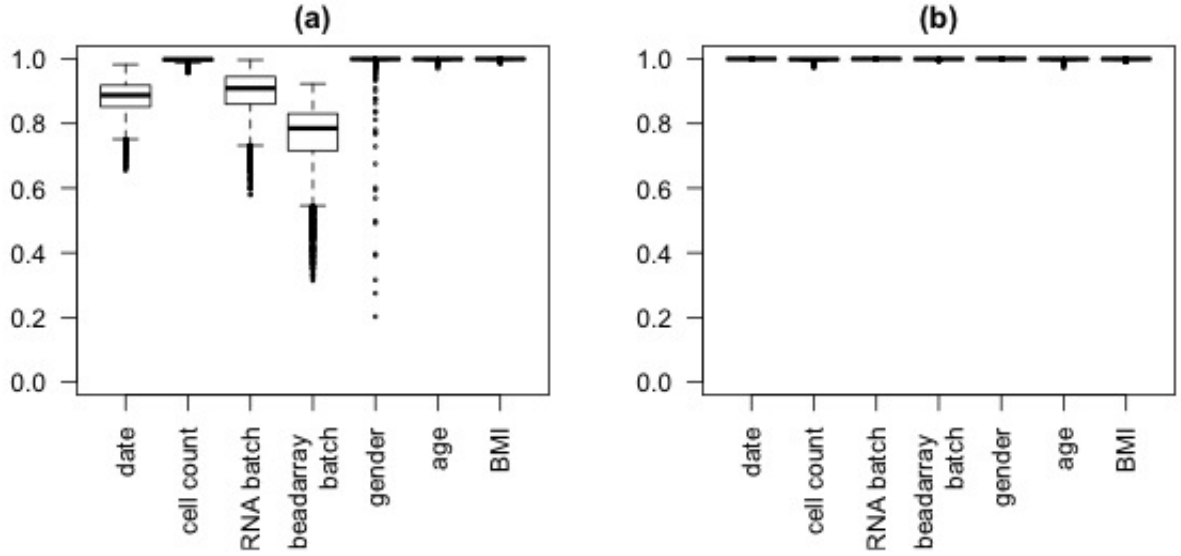

Figure 3.10: Boxplots showing the covariates effect on the expression data (a) before and (b) after adjustment. Each boxplot is generated by the correlation coefficients of the expression levels before and after adjusting each covariate.

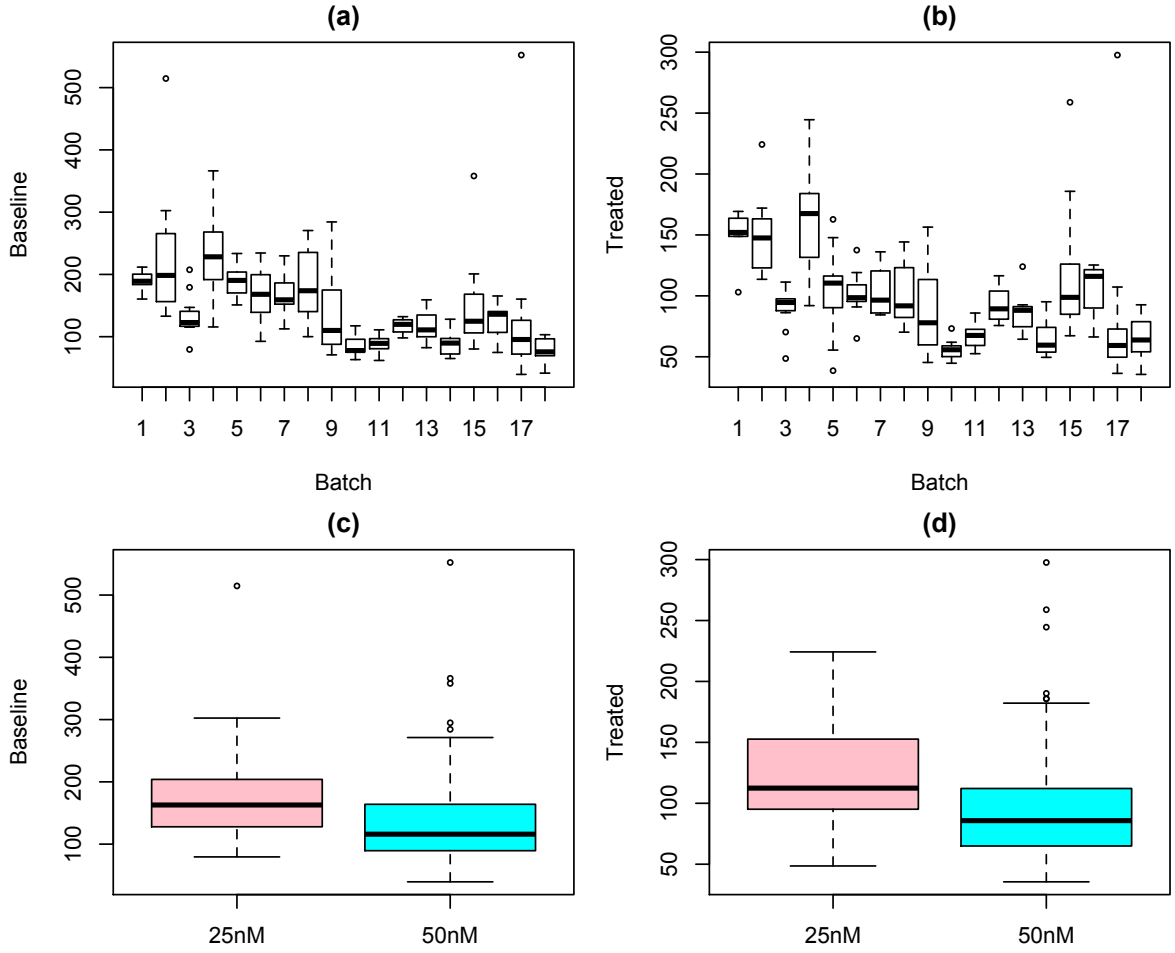

Figure 3.11: Plots showing (a)(b) the batch and (c)(d) dose effects on HMGCR enzymatic activity data. Panels (a) and (c) corresponds to pre-treatment, (b) and (d) to post-treatment of simvastatin.

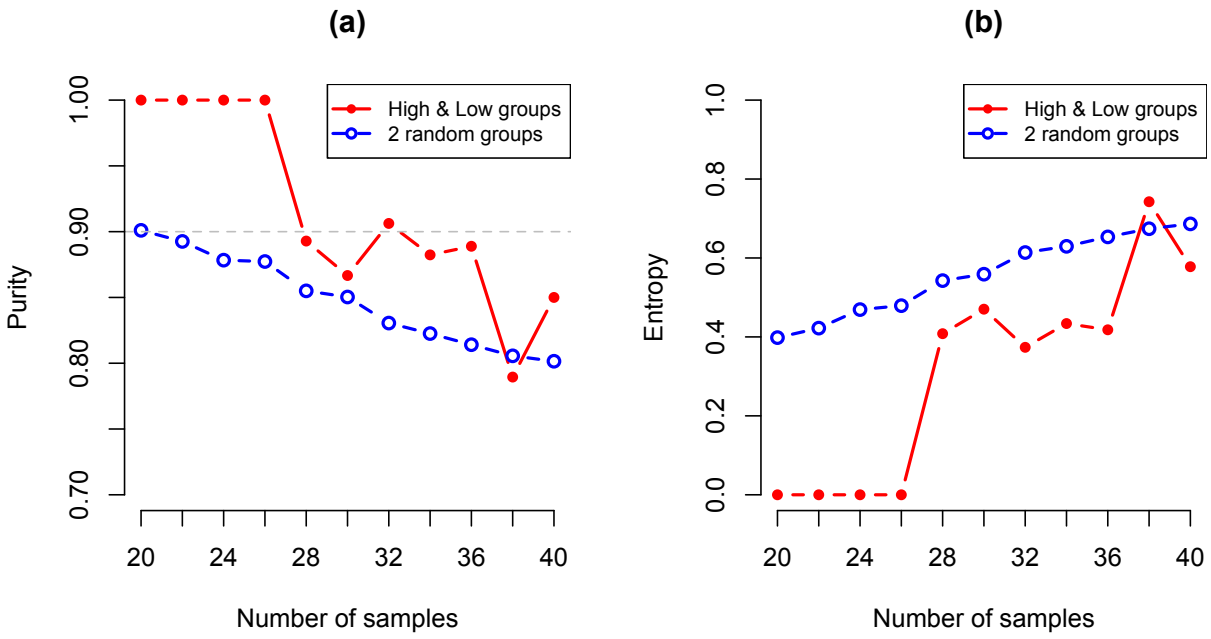

Figure 3.12: (a) Purity and (b) entropy plots from the NMF analysis with changing number of samples. The most varying 2000 genes are used in this analysis.

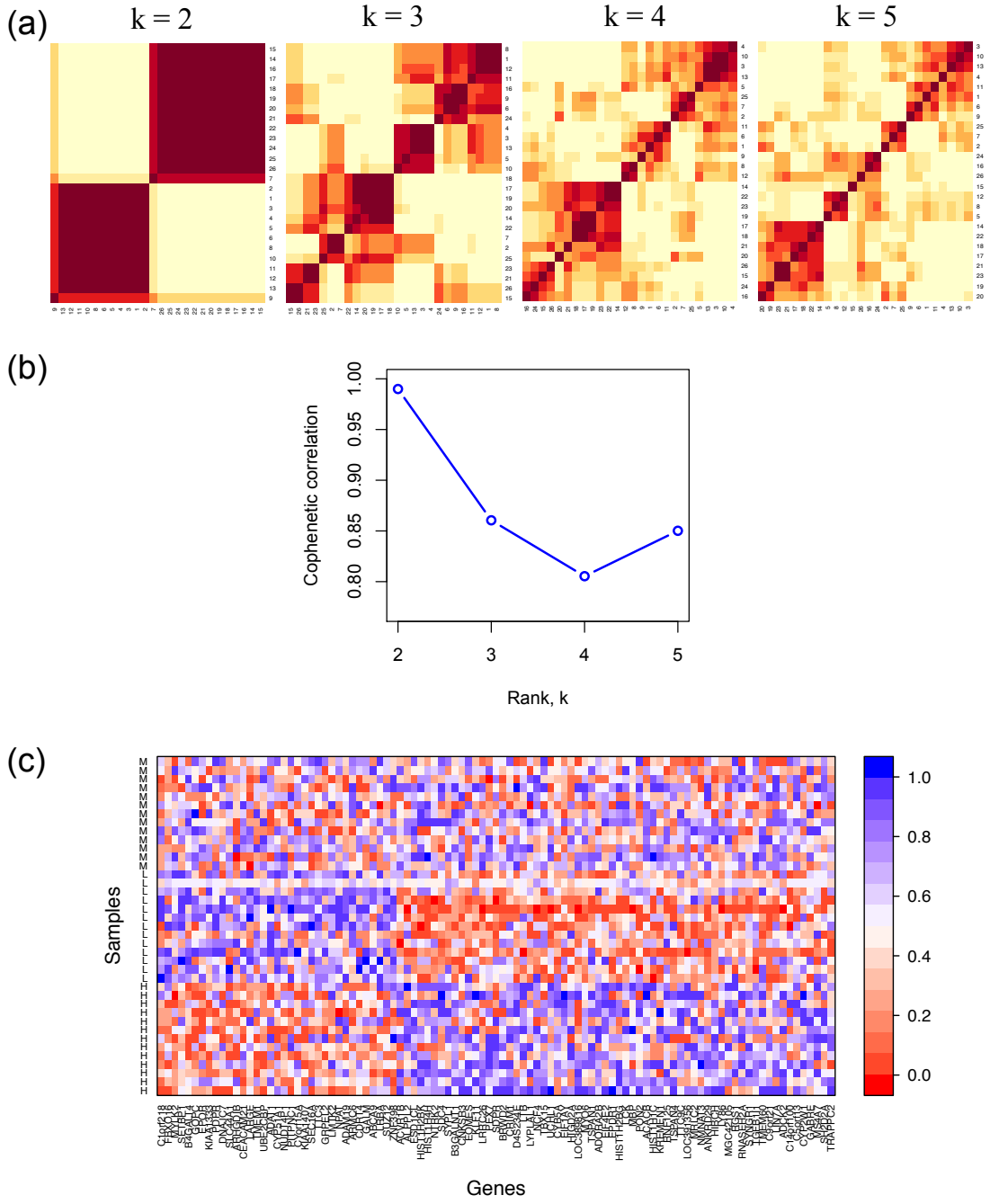

Figure 3.13: NMF analysis results with 13 each of the highest and lowest responder samples. (a) Reordered consensus matrices, $\bar{\mathbf{C}}_{HC}$, averaging 1000 connectivity matrices computed at $k = 2 - 5$. (b) Cophenetic correlation coefficients computed at $k = 2 - 5$. (c) A heatmap showing the gene expression levels of the 99 signature genes between the high (H) and low (L) responders to the statin treatment. Samples from the mild (M) responders are added on top for comparison.

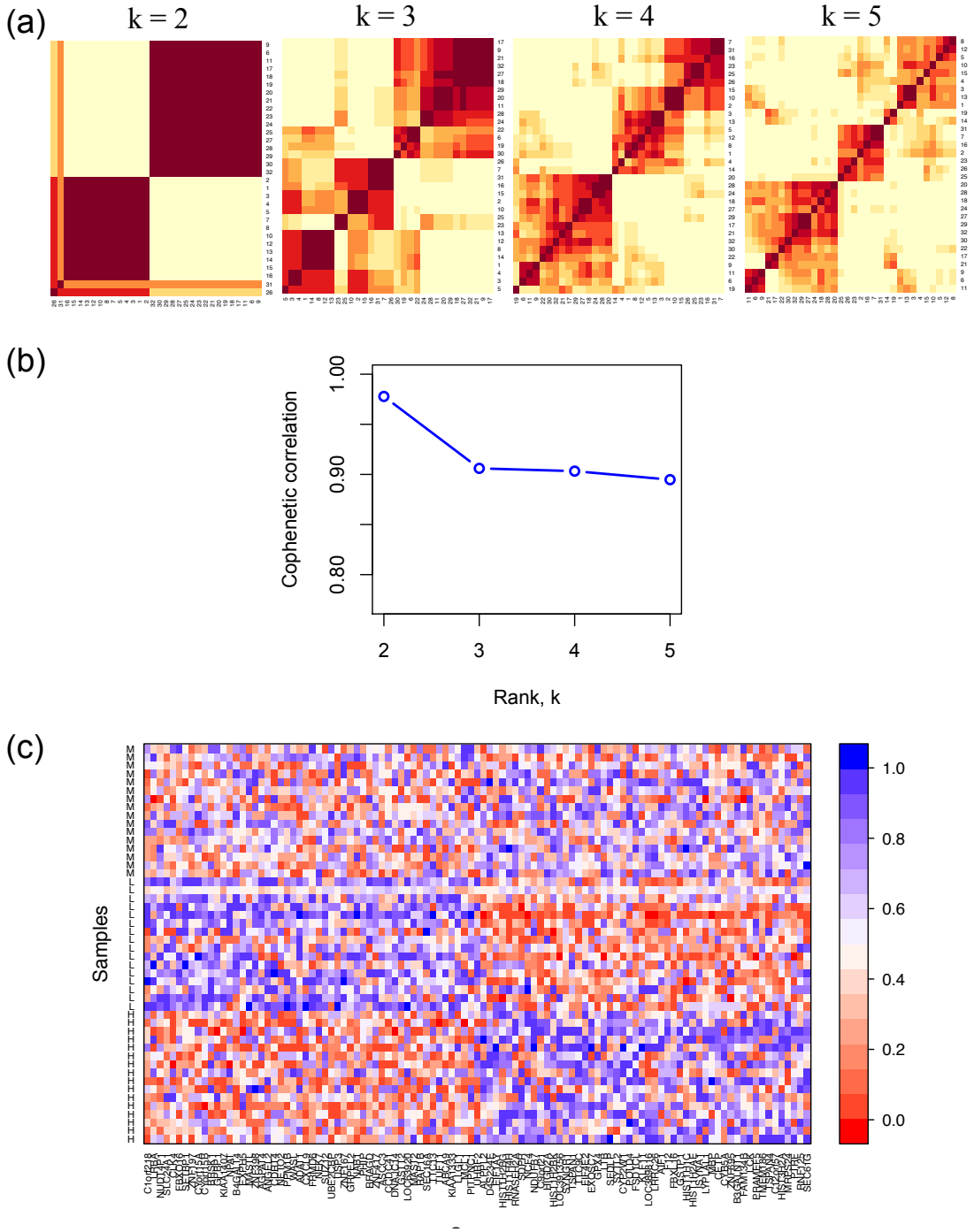

Figure 3.14: NMF analysis results with 16 each of the highest and lowest responder samples. (a) Reordered consensus matrices, $\bar{\mathbf{C}}_{HC}$, averaging 1000 connectivity matrices computed at $k = 2 - 5$. (b) Cophenetic correlation coefficients computed at $k = 2 - 5$. (c) A heatmap showing the gene expression levels of the 105 signature genes between the high (H) and low (L) responders to the statin treatment. Samples from the mild (M) responders are added on top for comparison.

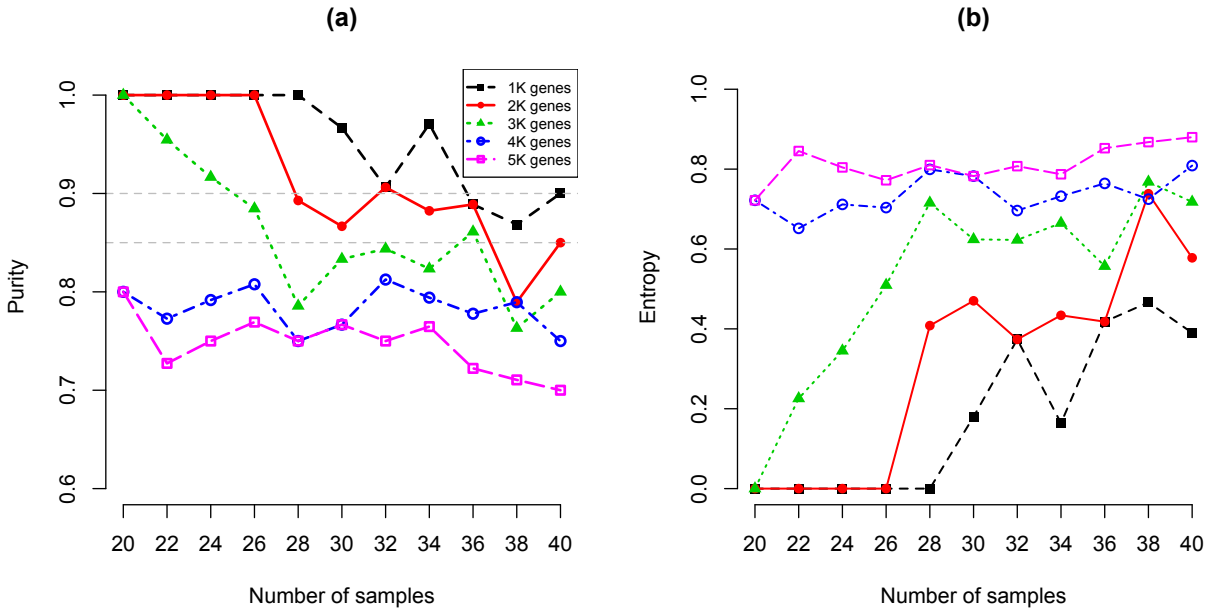

Figure 3.15: (a) Purity and (b) entropy plots from the NMF analysis with changing number of genes from 1000 to 5000.

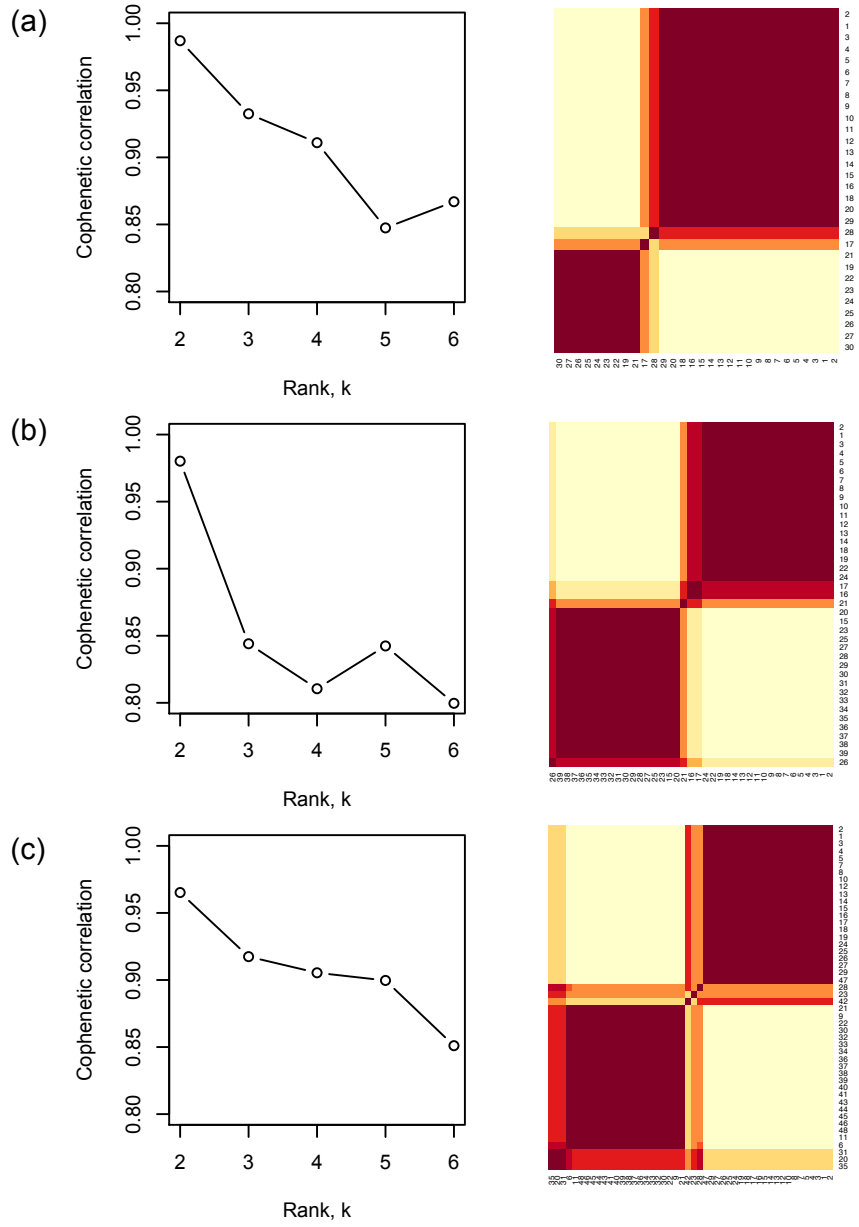

Figure 3.16: NMF analysis with the three (high, mild, low) responder groups. Cophenetic correlation coefficients and reordered consensus matrices are shown for (a) 30, (b) 39 and (c) 48 total number of samples, with a third of the samples taken from each of the three responder groups. Detailed clustering results are summarized in Table 3.1

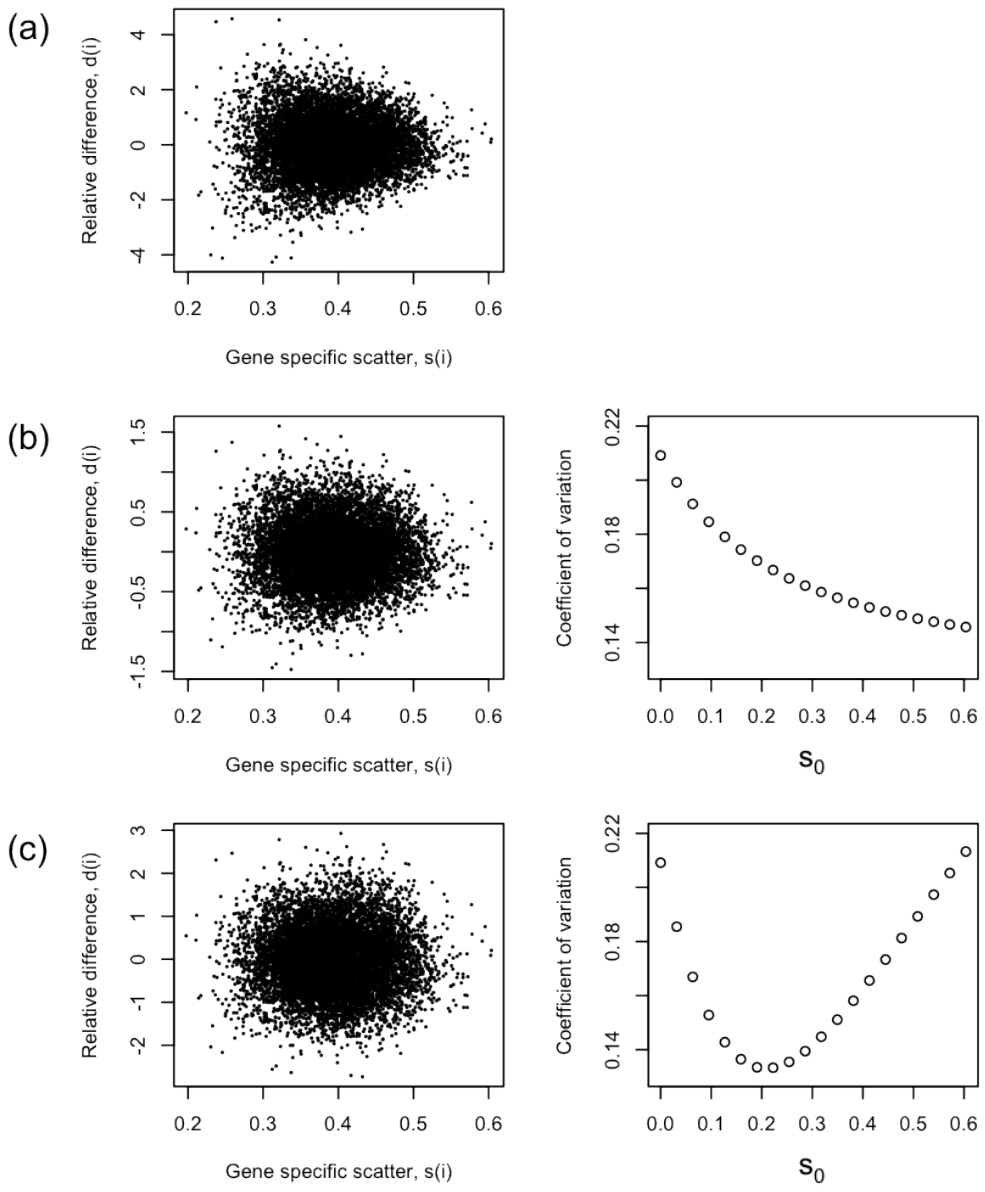

Figure 3.17: Scatter plots of the gene specific scatter $s(i)$ versus the relative difference $d(i)$ with (a) $s_0 = 0$, (b) s_0 from SAM, and (c) the varying s_0 values. The corresponding plots of the coefficients of variation of $d(i)$ are shown next to (b) and (c).

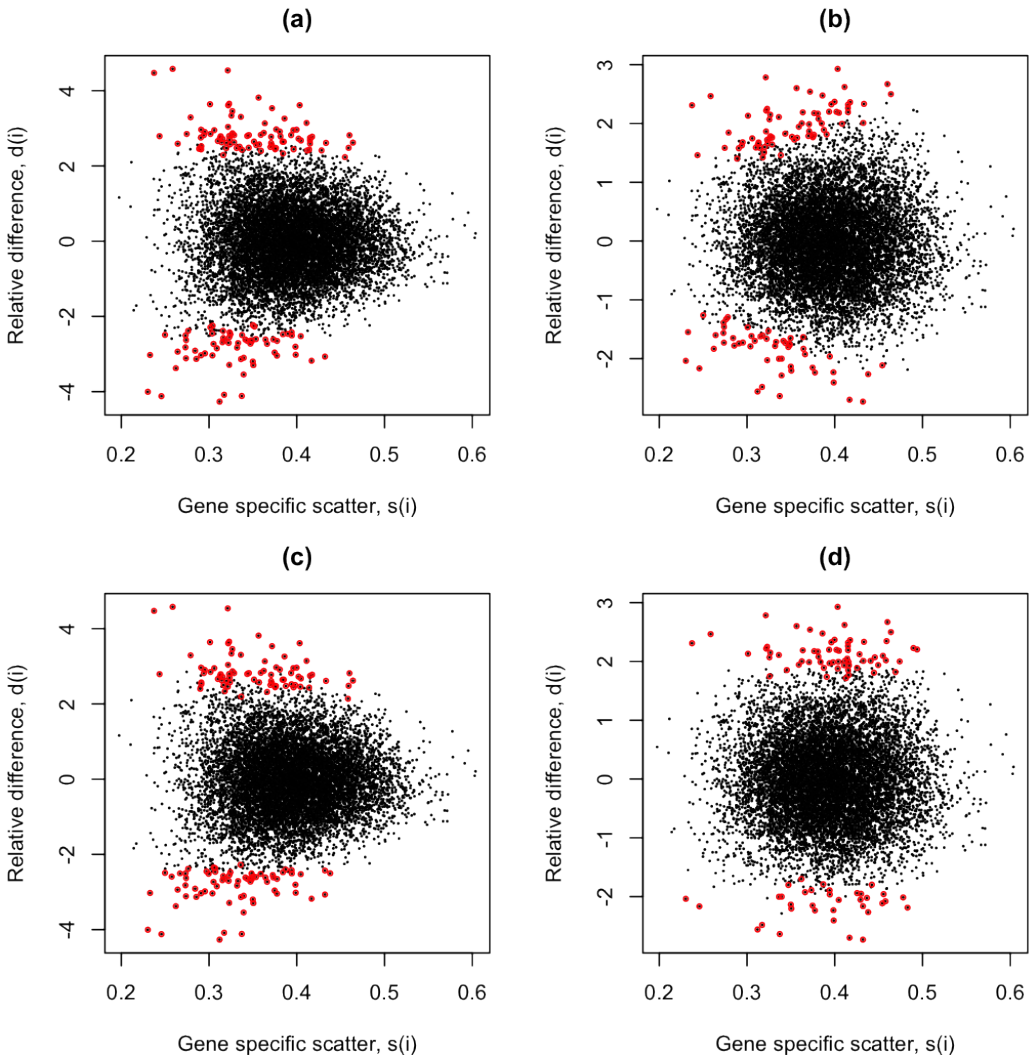

Figure 3.18: Identification of signature genes differentially expressed between 13 each of the highest and lowest responders. The signature genes are denoted with red dots. Scatter plots of the gene specific scatter $s(i)$ versus the relative difference $d(i)$ are generated with (a)(c) $s_0 = 0$ and (b)(d) the varying s_0 values. The null distribution of $d(i)$ is empirically estimated through random permutations, within either (a)(b) the 26 high and low responders or (c)(d) all 372 population samples.

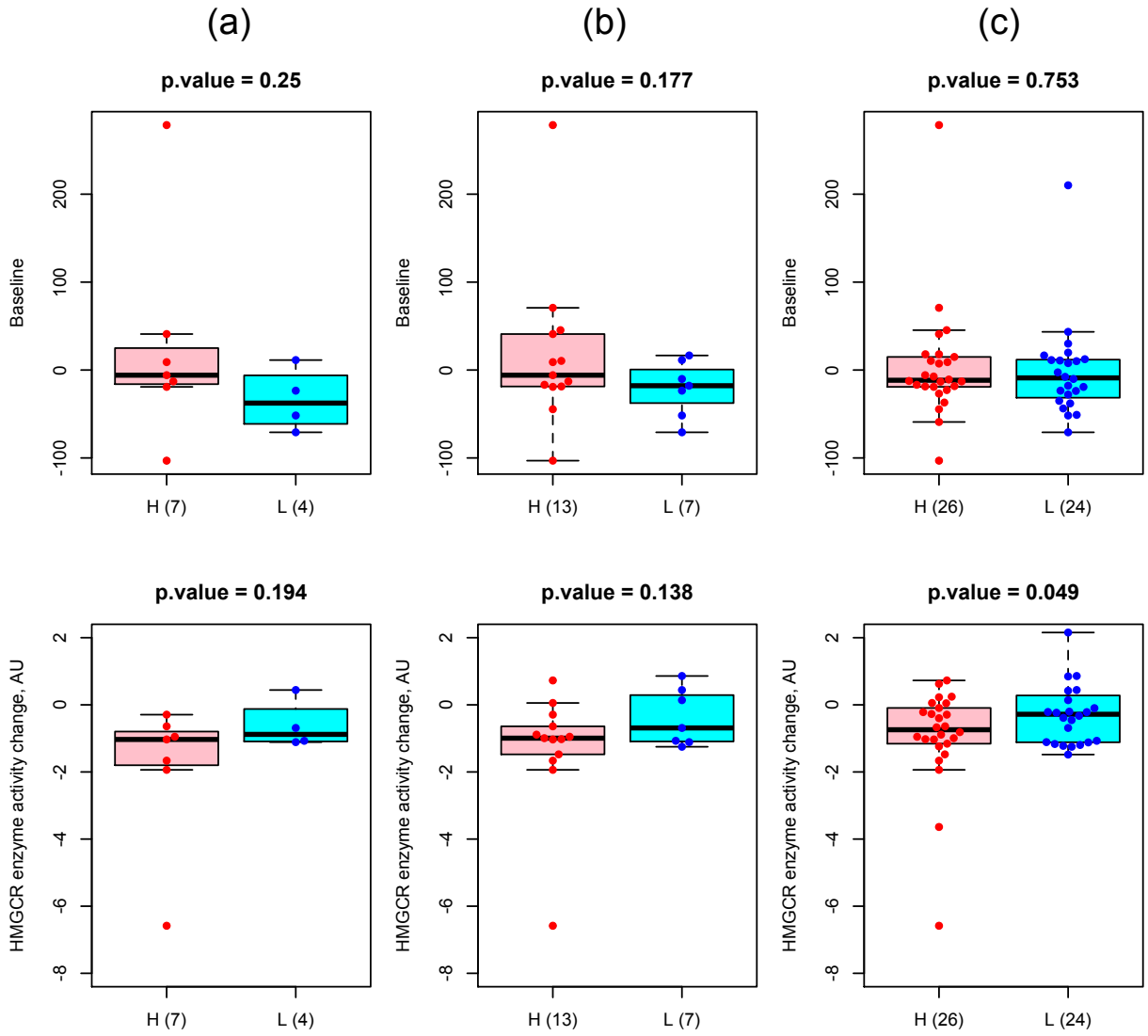

Figure 3.19: Boxplots comparing the amount of HMGCR enzymatic activity change between the (a) top 13, (b) top 20 and (c) top 50 high and low responder groups. The number of the high (H) and low (L) samples used to generate each boxplot is designated in the parentheses right next to H and L letters. For a better comparison, a t-test is performed and the resulting p-value is displayed at the top of each plot. HMGCR baseline enzymatic activity is also given at the top.

Chapter 4

Summary and Concluding Remarks

The *genomic revolution* has resulted in both the development of techniques for obtaining large quantities of genomic data rapidly and a striking increase in our knowledge on genomics. At the same time, the genomic revolution also created numerous open questions and challenges in analyzing the enormous amount of data required to gain insights on the underlying biological mechanisms.

This dissertation addresses these challenges by answering fundamental questions arising from two closely related fields, *functional genomics* and *pharmacogenomics*, utilizing the nature and biology of microarray datasets.

In the study on *functional genomics* (chapter 2), we tried to understand the underlying biology within pathways, which form the backbone of any biological process, by identifying functionally related pathway genes. To infer the complicated and higher level pathway gene interactions, partial correlation method was adopted. Based on this distance measure, a novel statistical model was built by estimating a precision matrix under appropriate normality assumptions. In the process of estimation, we also considered the existence of experiment dependencies in the gene expression data. Finally, likelihood ratio tests were performed to test the conditional dependencies of a candidate pathway gene. This gives us a list genes which is believed to be in the same pathway with the given seed genes. We demonstrated that our approach outperforms other existing methods in uncovering true gene relationships using both simulation and real gene expression data from the glucosinolate (GSL) and flavonoid biosynthesis (FB) pathways. In the FB pathway studies especially, we found that our method can identify genes from neighbouring pathways by considering the indirect relationships between genes. This finding will find its utility in future studies targeted on discovering cooperative nature of genes in the pathways.

In the study on *pharmacogenomics* (chapter 3), genetic variants causing inter-individual variation in drug response was investigated. Specifically, signature genes which contribute to the high and low responder variation in statin efficacy were discovered. Using Nonnegative Matrix Factorization (NMF) analysis, we (*i*) identified two distinct molecular patterns between the high and low responder groups and (*ii*) found the number of samples that produce the best separation between these groups was 13 each of the highest and lowest responder

samples. The modified Significance Analysis Microarrays (SAM) method identified 99 signature genes that had gone undetected by the original SAM method. In the correlation study, we showed that our signature genes were significantly enriched with HMGCR-correlated genes; 47 % (p-value $= 1.26 \times 10^{-4}$) and 28 % (p-value $= 1.60 \times 10^{-2}$) of our signature genes are overlapped with the HMGCR-correlated genes in the untreated and treated gene expression datasets, respectively. Furthermore, we observed that there is a notable difference in the amount of HMGCR enzymatic activity change between the high and low responder groups. The high responder group exhibited a bigger activity decrease than the low responder group, implying that statin inhibits the HMGCR enzymatic activity more efficiently in the high responder groups. This helps us understand why the high responder group shows a greater LDLC decrease and higher statin efficacy than the low responder groups. Overall, the discovered gene signatures are shown to have high biological relevance to the cholesterol biosynthesis pathway, which HMGCR mainly acts on.

This dissertation contributes to the understanding of biological processes with the aid of statistical frameworks for handling and analyzing high throughput genomic data. The proposed statistical methods were shown to perform well under practical applications. We believe that the proposed methods should be applicable to various other fields with suitable modifications.

Bibliography

[1] Ruedi Aebersold and Matthias Mann. "Mass spectrometry-based proteomics". In: *Nature* 422.6928 (2003), pp. 198–207.

[2] David B. Allison. *DNA microarrays and related genomics techniques: Designs, analysis, and interpretation of experiments*. CRC Press, 2006.

[3] Dhivya Arasappan et al. "Meta-analysis of microarray data using a pathway-based approach identifies a 37-gene expression signature for systemic lupus erythematosus in human peripheral blood mononuclear cells". In: *BMC Medicine* 9.1 (2011), p. 65.

[4] Liviu Badea. "Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization". In: *Pacific Symposium on Biocomputing* 13 (2008), pp. 279–290.

[5] Mathew J. Barber et al. "Genome-wide association of lipid-lowering response to statins in combined study populations". In: *PLoS ONE* 5.3 (2010), e9763.

[6] Hamid Bolouri and Eric H. Davidson. "Modeling transcriptional regulatory networks". In: *BioEssays* 24 (2002), pp. 1118–1129.

[7] Ben M. Bolstad et al. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". In: *Bioinformatics* 19.2 (2003), pp. 185–193.

[8] Jean-Philippe Brunet et al. "Metagenes and molecular pattern discovery using matrix factorization". In: *Proceedings of the National Academy of Sciences* 101.12 (2004), pp. 4164–4169.

[9] Atul J. Butte and Isaac S. Kohane. "Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements". In: *Pacific Symposium on Biocomputing* 24 (2000), pp. 418–429.

[10] Pedro Carmona-Saez et al. "Biclustering of gene expression data by non-smooth nonnegative matrix factorization". In: *BMC Bioinformatics* 7 (2006), p. 78.

[11] Monica Chagoyen et al. "Discovering semantic features in the literature: A foundation for building functional associations". In: *BMC Bioinformatics* 7 (2006), p. 41.

[12] Daniel I. Chasman et al. “Pharmacogenetic study of statin therapy and cholesterol reduction”. In: *The Journal of the American Medical Association* 291.23 (2004), pp. 2821–2827.

[13] Yang Chen et al. “Phoenix: A weight-based network coordinate system using matrix factorization”. In: *IEEE Transactions on Network and Service Management* 8.4 (2011), pp. 334–347.

[14] Eric A. Collisson et al. “Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy”. In: *Nature Medicine* 17.4 (2011), pp. 500–503.

[15] Xinping Cui et al. “Improved statistical tests for differential gene expression by shrinking variance components estimates”. In: *Biostatistics* 6.1 (2005), pp. 59–75.

[16] Karthik Devarajan. “Nonnegative matrix factorization: An analytical and interpretive tool in computational biology”. In: *PLoS Computational Biology* 4.7 (2008), e1000029.

[17] Adrian Dobra et al. “Sparse graphical models for exploring gene expression data”. In: *Journal of Multivariate Analysis* 90 (2004), pp. 196–212.

[18] Louise A. Donnelly et al. “A paucimorphic variant in the HMG-CoA reductase gene is associated with lipid-lowering response to statin treatment in diabetes: A GoDARTS study”. In: *Pharmacogenet Genomics* 18.12 (2008), pp. 1021–1026.

[19] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern recognition*. John Wiley and Sons, 2001.

[20] Sandrine Dudoit, Jane Fridlyand, and Terence P. Speed. “Comparison of discrimination methods for the classification of tumors using gene expression data”. In: *Journal of the American Statistical Association* 97.457 (2002), pp. 77–87.

[21] Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C. Boldrick. “Multiple hypothesis testing in microarray experiments”. In: *Statistical Science* 18 (2003), pp. 71–103.

[22] David J. Duggan et al. “Expression profiling using cDNA microarrays”. In: *Nature Genetics* 21 (1999), pp. 10–14.

[23] Michael B. Eisen et al. “Cluster analysis and display of genome-wide expression patterns”. In: *Proceedings of the National Academy of Sciences* 95 (1998), pp. 14863–14868.

[24] Greg Finak et al. “Stromal gene expression predicts clinical outcome in breast cancer”. In: *Nature Medicine* 14.5 (2008), pp. 518–527.

[25] Paul Fogel et al. “Inferential, robust non-negative matrix factorization analysis of microarray data”. In: *Bioinformatics* 23.1 (2007), pp. 44–49.

[26] William T. Friedewald, Robert I. Levy, and Donald S. Fredrickson. “Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge”. In: *Clinical Chemistry* 18.6 (1972), pp. 499–502.

[27] Nir Friedman et al. "Using Bayesian networks to analyze expression data". In: *Journal of Computational Biology* 7.3-4 (2000), pp. 601–620.

[28] Attila Frigyesi and Mattias Höglund. "Non-negative matrix factorization for the analysis of complex gene expression data: Identification of clinically relevant tumor subtypes". In: *Cancer Informatics* 6 (2008), pp. 275–292.

[29] Alberto de la Fuente et al. "Discovery of meaningful associations in genomic data using partial correlation coefficients". In: *Bioinformatics* 20.18 (2004), pp. 3565–3574.

[30] Claire M.M. Gachon et al. "Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: Functional and evolutionary implications". In: *Plant Molecular Biology* 58.2 (2005), pp. 229–245.

[31] Yuan Gao and George Church. "Improving molecular cancer class discovery through sparse non-negative matrix factorization". In: *Bioinformatics* 21.21 (2005), pp. 3970–3975.

[32] Tamara Gigolashvili et al. "The plastidic bile acid transporter 5 is required for the biosynthesis of methionine-derived glucosinolates in Arabidopsis thaliana". In: *The Plant Cell* 21 (2009), pp. 1813–1829.

[33] Todd R. Golub et al. "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring". In: *Science* 286.5439 (1999), pp. 531–537.

[34] *Graphical models*. Clarendon Press, Oxford., 1996.

[35] Jemila S. Hamid et al. "Data integration in genetics and genomics: Methods and challenges". In: *Human Genomics and Proteomics* 1.1 (2009), p. 869093.

[36] Kim E. Hammond-Kosack and Jonathan D.G. Jones. *Responses to plant pathogens*. Ed. by Jones RL Buchanan BB Gruissem W. Biochemistry and Molecular Biology of Plants. American Society of Plant Physiologists, Rockville, MD, 2001.

[37] Masami Y. Hirai et al. "Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis". In: *Proceedings of the National Academy of Sciences* 104.15 (2007), pp. 6478–6483.

[38] Patrik O. Hoyer. "Modeling receptive fields with non-negative sparse coding". In: *Neurocomputing* 52-54 (2003), pp. 547–552.

[39] Patrik O. Hoyer. "Nonnegative matrix factorization with sparseness constraints". In: *Journal of Machine Learning Research* 5 (2004), pp. 1457–1469.

[40] Jean-Sébastien Hulot. "Pharmacogenomics and personalized medicine: Lost in translation?" In: *Genome Medicine* 2 (2010), p. 13.

[41] Lucie N. Hutchins et al. "Position-dependent motif characterization using non-negative matrix factorization". In: *Bioinformatics* 24.23 (2008), pp. 2684–2690.

[42] *Introduction to Graphical Modeling*. Springer-Verlag New York, Inc., 1995.

[43] Rafael A. Irizarry et al. “Exploration, normalization, and summaries of high density oligonucleotide array probe level data”. In: *Biostatistics* 4.2 (2003), pp. 249–264.

[44] Rafael A. Irizarry et al. “Summaries of Affymetrix GeneChip probe level data”. In: *Nucleic Acids Research* 31.4 (2003), e15.

[45] Ronald Jansen et al. “Integration of genomic datasets to predict protein complexes in yeast”. In: *Journal of Structural and Functional Genomics* 2.2 (2002), pp. 71–81.

[46] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data. An introduction to cluster analysis*. John Wiley and Sons, 1990.

[47] Purvesh Khatri and Sorin Draghici. “Ontological analysis of gene expression data: Current tools, limitations, and open problems”. In: *Bioinformatics* 21.18 (2005), pp. 3587–3595.

[48] Hyunsoo Kim and Haesun Park. “Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis”. In: *Bioinformatics* 23.12 (2007), pp. 1495–1502.

[49] Philip M. Kim and Bruce Tidor. “Subsystem identification through dimensionality reduction of large-scale gene expression data”. In: *Genome Research* 13 (2003), pp. 1706–1718.

[50] Ronald M. Krauss et al. “Variation in the 3-hydroxyl-3-methylglutaryl coenzyme A reductase gene is associated with racial differences in low-density lipoprotein cholesterol response to simvastatin treatment”. In: *Circulation* 117 (2008), pp. 1537–1544.

[51] Daniel D. Lee and H. Sebastian Seung. “Algorithms for non-negative matrix factorization”. In: *Advances in Neural Information Processing Systems* 13 (2001), pp. 556–562.

[52] Daniel D. Lee and H. Sebastian Seung. “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.21 (1999), pp. 788–791.

[53] Robert J. Lipshutz et al. “High density synthetic oligonucleotide arrays”. In: *Nature Genetics* 21 (1999), pp. 20–24.

[54] David J. Lockhart and Elizabeth A. Winzeler. “Genomics, gene expression and DNA arrays”. In: *Nature* 405 (2000), pp. 827–836.

[55] Paul M. Magwene and Junhyong Kim. “Estimating genomic coexpression networks using first-order conditional independence”. In: *Genome Biology* 5 (2004), R100.

[56] Yun Mao, Lawrence K. Saul, and Jonathan M. Smith. “IDES: An internet distance estimation service for large networks”. In: *IEEE Journal on Selected Areas in Communications* 24.12 (2006), pp. 2273 –2284.

[57] Tetsuya Matsuno et al. “Graphical gaussian modeling for gene association structures based on expression deviation patterns induced by various chemical stimuli”. In: *IEICE Transactions on Information and Systems* E89D.4 (2006), pp. 1563–1573.

[58] Marisa W. Medina. “The relationship between HMGCR genetic variation, alternative splicing, and statin efficacy”. In: *Discovery Medicine* 9.49 (2010), pp. 495–499.

[59] Marisa W. Medina and Ronald M. Krauss. “The role of HMGCR alternative splicing in statin efficacy”. In: *Trends in Cardiovascular Medicine* 19.5 (2009), pp. 173–177.

[60] Marisa W. Medina et al. “Alternative splicing of 3-hydroxy-3-methylglutaryl coenzyme A reductase is associated with plasma low-density lipoprotein cholesterol response to simvastatin”. In: *Circulation* 118.4 (2008), pp. 355–362.

[61] Stefano Monti et al. “Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data”. In: *Machine Learning* 52 (2003), pp. 91–118.

[62] Marina A. Naoumkina et al. “Genome-wide analysis of phenylpropanoid defence pathways.” In: *Molecular Plant Pathology* 11.6 (2010), pp. 829–846.

[63] Angela E. Oostlander, Gerrit A. Meijeroos, and Bauke Ylstra. “Microarray-based comparative genomic hybridization and its applications in human genetics”. In: *Clinical Genetics* 66.6 (2004), pp. 488–495.

[64] Rainer Opgen-Rhein and Korbinian Strimmer. “From correlation to causation networks: A simple approximate learning algorithm and its application to high-dimensional plant gene expression data”. In: *BMC Systems Biology* (2007), p. 37.

[65] Wei Pan. “A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments”. In: *Bioinformatics* 18.4 (2002), pp. 546–554.

[66] Alberto Pascual-Montano et al. “Nonsmooth nonnegative matrix factorization (nsNMF)”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.3 (2006), pp. 403–415.

[67] Petri Pehkonen, Garry Wong, and Petri Törönen. “Theme discovery from gene lists for identification and viewing of multiple functional groups”. In: *BMC Bioinformatics* 6 (2005), p. 162.

[68] David M. Reif, Bill C. White, and Jason H. Moore. “Integrated analysis of genetic, genomic and proteomic data”. In: *Expert Review of Proteomics* 1.1 (2004), pp. 67–75.

[69] Daniel R. Rhodes et al. “Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression”. In: *Proceedings of the National Academy of Sciences* 101.25 (2004), pp. 9309–9314.

[70] Paul M. Ridker et al. “Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein”. In: *The New England Journal of Medicine* 359.21 (2008), pp. 2195–2207.

[71] Kazuki Saito, Masami Y. Hirai, and Keiko Yonekura-Sakakibara. “Decoding genes with coexpression networks and metabolomics - majority report by precogs”. In: *Trends in Plant Science* 13.1 (2008), pp. 36–43.

[72] A. Sánchez-Pla and J.L. Mosquera. *The quest for biological signicance*. Ed. by G. Platero L.L. Bonilla M. Moscoso and J.M. Vega. Progress in Industrial Mathematics at ECMI 2006. Springer, New York, 2007.

[73] Eric E. Schadt, Sangsoon Woo, and Ke Hao. "Bayesian method to predict individual SNP genotypes from gene expression data". In: *Nature Genetics* 44.5 (2012), pp. 603–608.

[74] Juliane Schäfer and Korbinian Strimmer. "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics". In: *Statistical Applications in Genetics and Molecular Biology* 4.1 (2005), p. 32.

[75] Juliane Schäfer and Korbinian Strimmer. "An empirical Bayes approach to inferring large-scale gene association networks". In: *Bioinformatics* 21.6 (2005), pp. 754–764.

[76] Mark Schena et al. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". In: *Science* 270.5235 (1995), pp. 467–470.

[77] Farial Shahnaz et al. "Document clustering using nonnegative matrix factorization". In: *Information Processing and Management: An International Journal* 42.2 (2006), pp. 373–386.

[78] Joel A. Simon et al. "Phenotypic predictors of response to simvastatin therapy among african-americans and caucasians: The Cholesterol and Pharmacogenetics (CAP) Study". In: *The American Journal of Cardiology* 97.6 (2006), pp. 843–850.

[79] Gordon K. Smyth, Joëlle Michaud, and Hamish S. Scott. "Use of within-array replicate spots for assessing differential expression in microarray experiments". In: *Bioinformatics* 21.9 (2005), pp. 2067–2075.

[80] Ida E. Sønderby, Fernando Geu-Flores, and Barbara A. Halkier. "Biosynthesis of glucosinolates - gene discovery and beyond." In: *Trends in Plant Science* 15.5 (2010), pp. 283–290.

[81] Terence P. Speed. *Statistical analysis of gene expression microarray data*. Boca Raton, FL: Chapman & Hall/CRC Press, 2003.

[82] Alessio Squassina et al. "Realities and expectations of pharmacogenomics and personalized medicine: Impact of translating genetic knowledge into clinical practice". In: *Pharmacogenomics* 11.8 (2010), pp. 1149–1167.

[83] Aravind Subramanian et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.

[84] Pablo Tamayo et al. "Metagene projection for cross-platform, cross-species characterization of global transcriptional states". In: *Proceedings of the National Academy of Sciences* 104.14 (2007), pp. 5959–5964.

[85] Loverine P. Taylor and Erich Grotewold. "Flavonoids as developmental regulators". In: *Current Opinion in Plant Biology* 8.3 (2005), pp. 317–323.

[86] Siew Leng Teng and Haiyan Huang. "A statistical framework to Infer functional gene relationships from biologically interrelated microarray experiments". In: *Journal of the American Statistical Association* 104.486 (2009), pp. 465–473.

[87] John F. Thompson et al. "Comprehensive whole-genome and candidate gene analysis for response to statin therapy in the treating to new targets (TNT) cohort". In: *Circulation Cardiovascular Genetics* 2.2 (2009), pp. 173–181.

[88] Olga G. Troyanskaya et al. "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)". In: *Proceedings of the National Academy of Sciences* 100.14 (2003), pp. 8348–8353.

[89] Virginia G. Tusher, Robert Tibshirani, and Gilbert Chu. "Significance analysis of microarrays applied to the ionizing radiation response". In: *Proceedings of the National Academy of Sciences* 98.9 (2001), pp. 5116–5121.

[90] Laura J. van't Veer et al. "Gene expression profiling predicts clinical outcome of breast cancer". In: *Nature* 415.6871 (2002), pp. 530–536.

[91] Ruud Verkerk et al. "Glucosinolates in Brassica vegetables: The influence of the food supply chain on intake, bioavailability and human health". In: *Molecular Nutrition Food Research* 53.Suppl. 2 (2009), S219.

[92] Anja Wille and Peter Bühlmann. "Low-order conditional independence graphs for Inferring genetic networks". In: *Statistical Applications in Genetics and Molecular Biology* 5.1 (2006), p. 1.

[93] Anja Wille et al. "Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana". In: *Genome Biology* 5 (2004), R92.

[94] Ho-Hyung Woo, Byeong Ryong Jeong, and Martha C. Hawes. "Flavonoids: From cell cycle regulation to biotechnology". In: *Biotechnology Letters* 27.6 (2005), pp. 365–374.

[95] Hao Wu et al. "MAANOVA: A software package for the analysis of spotted cDNA microarray experiments". In: *The Analysis of Gene Expression Data: Methods and Software* (2003), pp. 313–341.

[96] Min Xu et al. "Automated multidimensional phenotypic profiling using large public microarray repositories". In: *Proceedings of the National Academy of Sciences* 106.30 (2009), pp. 12323–12328.

[97] Xiufeng Yan and Sixue Chen. "Regulation of plant glucosinolate metabolism". In: *Planta* 226.6 (2007), pp. 1343–1352.

[98] Keiko Yonekura-Sakakibaraa et al. "Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in Arabidopsis". In: *The Plant Cell* 20 (2008), pp. 2160–2176.

[99] Ying Zhao and George Karypis. "Empirical and theoretical comparisons of selected criterion functions for document clustering". In: *Machine Learning* 55 (2004), pp. 311–331.

[100] Ying Zhao and George Karypis. “Evaluation of hierarchical clustering algorithms for document datasets”. In: *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (2002), pp. 515–524.