Statistical Methods for Biological Pathway Analysis

A Dissertation Presented

by

Xiao Wu

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

August 2011

Stony Brook University

The Graduate School

Xiao Wu

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Wei Zhu – Dissertation Advisor
Professor, Department of Applied Mathematics and Statistics

Haipeng Xing - Chairperson of Defense
Assistant Professor, Department of Applied Mathematics and Statistics

Xiangmin Jiao – Defense committee member
Assistant Professor, Department of Applied Mathematics and Statistics

Daniel van der Lelie – Outside member
Professor, Biological department, Brookhaven national laboratory

Safiyh Taghavi – Outside member
Professor, Biological department, Brookhaven national laboratory

Ellen Li – Outside member
Professor, Department of Medicine, Stony Brook University

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Statistical Methods for Biological Pathway Analysis

by

Xiao Wu

Doctor of Philosophy

in

Department of Applied Mathematics and Statistics

Stony Brook University

2011

This thesis features a novel theoretical development, as well as a novel application of the structural equation modeling (SEM) framework for biological pathway and biological measurement platform comparisons respectively. For the SEM methodology development, we have extended the covariate structural equation modeling (cSEM) method (Sharpe, 2010) for pathway comparisons that was limited to continuous variables on the pathway nodes and categorical variables as pathway covariates only, to allow both continuous and categorical variables as pathway nodes as well as pathway covariates. This novel mixed variable cSEM method will permit researchers to implement a pathway with both continuous variables such as gene expression levels, and categorical variables such as genotypes on the pathway nodes, and compare the pathway between different groups (diseased, normal etc.) as well as evaluate the impact of continuous variables such as age on the pathway links (i.e. connecting patterns and strengths).

Culture-independent phylogenetic analysis of 16S ribosomal RNA gene sequences has emerged as an incisive method of identifying bacteria present in a specimen. However multiple competing measurement platforms are often available to enumerate the abundances of the bacteria, including Sanger sequencing, pyrosequencing, and quantitative PCR. Here we present a novel application of the latent variable SEM to estimate the reliabilities of, and the similarities between different measurement platforms, and subsequently, weigh these measures optimally for a unified analysis of the true latent microbiome composition. The latent variable SEM contains the usual repeated measures ANCOVA as special cases and, as a more general, realistic and optimal model, features superior model goodness-of-fit as well as more reliable analysis results.

The third and final contribution of this thesis is the establishment of two bioinformatics pipelines in a systems biology framework to integrate incremental biological knowledge obtained through the analysis of newly available experimental data, to existing biological knowledge database, and subsequently evolve such knowledgebase to the next level. Two examples, one from the molecular study of the human inflammatory bowel diseases, and one from the study of endophytic bacteria known to impact the growth rate of certain plant, are provided to illustrate these novel pipelines.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor Professor Wei Zhu from the bottom of my heart. She is the usher leading me into the statistical world. Prof. Zhu has provided me intellectual freedom, deep insights, and great pictures in my research. Without her guidance, support and encouragement, this dissertation would never have come true. Her endless effort in pursuing truth, passion to life, and positive attitude has always been a source of inspiration. I have been fortunate to have her as my advisor.

Special thanks go to Dr. Niels van der Lelie and Dr. Safiyh Taghavi, for enjoyable and fruitful collaborations that contributed to part of this thesis. They are supportive to me all the time financially and mentally. I am also very grateful to Dr. Ellen Li for introducing us the exciting and challenging project. My gratitude also extends to Dr. Haipeng Xing and Dr. Xiangmin Jiao for serving on my thesis committee.

I want to thank Dr. Sean McCorkle for his help on preprocessing RNA-seq data; thank Dr. Sergei Maslov for the suggestive assistance on the regulatory network analysis; and thank Dr. Yian-biao Zhang for performing biological experiments of *Enterobacter sp.* 638 project; also thank Dr. Sebastien Monchy for his guidance on comparative genomic analysis.

Thanks also go to my friends and colleagues, to Dr. Tianyi Zhang for helping me with all sorts of problems; to Dr. Kathryn Sharpe for her previous work and kind help; to Hongyan Chen, Shirley Leong, Shaonan Zhang, Xiao Xu, Jun Huang, Han Hao, Lin Chen, Yueting Zheng and Ding Wang for the friendship and support.

Last but not least, I thank my parents for their care and support, my dear husband for his love. This thesis is dedicated to them.

Vita, Publications and Fields of Study

EDUCATION

STONY BROOK UNIVERSITY (Sep 2007-present) Cumulative GPA: 3.9/4.0

- Ph. D. in Statistics (Expected in August 2011)
- M. S. in Statistics (May 2010)
- Related Coursework: Categorical Data Analysis | Regression Analysis | Mathematical Statistics | Exploratory Data Analysis | Stochastic Models | Experimental Design | Multivariate Analysis

ZHEJIANG UNIVERSITY (Sep 2003-Jun 2007) Major GPA: 4.0 /4.0 | Cumulative GPA: 3.9/4.0

- B. S. in Biotechnology | Chu Kochen Honors College
- Related Coursework: Biostatistics and Experiment Design | Bioinformatics and Data Processing | Probability and Statistics | Mathematical Analysis | Linear Algebra | Fundamentals of Programming in C-Language & Its Lab. | Algorithms and Data Structure | Molecular Biology

RESEARCH AND TEACHING EXPERIENCE

Research Assistant,
Dept. of Applied Math & Statistics, SUNY Stony Brook (Jan 2008-Present)

- Multimodal microbiome analysis using latent variable structural equation modeling and pathway Analysis in Inflammatory Bowel Disease studies (*Department of Medicine, Washington University at St. Louis).*
- Whole transcriptome analysis on mRNA sequencing data (*Brookhaven National Laboratory).*
- Comparative genomic and functional analysis on *Pseudomonas putida* (*Brookhaven National Laboratory).*
- Compound Hierarchical Cluster Analysis on neuronal microarray data (*Cold Spring Harbor Laboratory*).
- Differential gene expression analysis to identify biomarkers and discriminant analysis to distinguish among etiologies of thrombocytosis by microarray data (*Stony Brook University Medical Center*).

Instructor of "Introduction to Plant Physiology"

- For 70 high-school students preparing for National Olympic Biology Competition (May 2005)

PUBLICATIONS

- **Wu X**, Monchy S, Taghavi S, Zhu W, Ramos J and van der Lelie D. "Comparative genomics and functional analysis of niche-specific adaptation in *Pseudomonas putida*". *FEMS Microbiology Reviews*. 35: 299–323 (2011).
- Matilla MA, Pizarro-Tobias P, Roca A, Fernández M, Duque E, Molina L, **Wu X**, van der Lelie D, Gómez MJ, Segura A, Ramos JL. "Complete genome of the plant growth-promoting rhizobacterium *Pseudomonas putida* BIRD-1". *J Bacteriol*. 193(5):1290 (2011).
- van der Lelie D, Taghavi S, Monchy S, Schwender J, Miller L, Ferrieri R, Rogers A, **Wu X**, Zhu W, Weyens N, Vangronsveld J and Newman L. "Poplar and its bacterial endophytes: Coexistence and harmony". *Critical Reviews in Plant Sciences*. 28(5):346–358, (2009).

- **Wu X**, Sharpe K, Frank D, Li E, Zhu W. "Multimodal Microbiome Analysis with Latent Variable Structural Equation Modeling". Manuscript to be submitted to *Bioinformatics*.

PRESENTATION

- “Comparative genetic pathway analysis using structural equation modeling”, 2011 IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS ’11), Orlando, Feb. 2011.

PROGRAMMING SKILLS

- **Software:** SAS Certificated (Score 96% SAS Base Programming Exam), Proficient in Excel and R

Chapter 1 Introduction and overview

In recent years, various statistical pathway analysis methods have gained enormous popularity in biomedical research because of the increasing focus on systems biological studies. As noted by the nation's leading scientists, "The (traditional) reductionist approach has successfully identified most of the components and many of the interactions but, unfortunately, offers no convincing concepts or methods to understand how system properties emerge...the pluralism of causes and effects in biological networks is better addressed by observing, through quantitative measures, multiple components simultaneously and by rigorous data integration with mathematical models" (Sauer, Heinemann et al. 2007). The biostatistics and bioinformatics community have quickly taken up the initiatives developing statistical methods for biological pathway analysis.

In the first part of this thesis, we will introduce several novel methodological development and application of structural equation modeling (SEM) for biological pathway analysis that include, especially, (1) *an efficient modeling and computation scheme for mixed variables SEM* – that is, SEM with both categorical and continuous variables as pathway nodes, (2) the development of the new *mixed variable **covariate** SEM framework* for comparative analysis of pathways with mixed categorical variables (e.g. genotypes, phenotypes) and continuous variables (e.g. gene expression levels, age, etc.) as both pathway nodes and pathway covariates, (3) the proposition of *a joint statistical and biological pathway analysis pipeline consisting of SAM, IPA and covariate SEM for comparative genetic network analysis* utilizing both known biological database and newly available experimental data, and (4) the *novel application of the latent variable SEM for microbiome measurement platform comparison and combination*. These original methods and applications are illustrated through an on-going study of the human inflammatory bowel diseases led by our collaborator, Dr. Ellen Li, at the Stony Brook University Medical School.

In contrast to the first part, the novel biological pathway analysis by structural equation modeling with application to a human study, the second part of this thesis focuses on developing *a novel bioinformatics pipeline for a systems biology study of the endophytic bacteria*, a plant colonizing and growth promoting bacterium, *at the level of the genome, transcriptome and metabolome*. The aim of this study is to identify mechanisms by which the endophytic bacteria can regulate the growth of poplar, to quantify the degree to which major regulatory pathways of endophytic bacteria are involved, and to ultimately, model these pathways to optimize the production of poplar biomass on marginal soils as a feedstock for bio-refineries. The endophytic bacteria project is led by our collaborators Dr. Daniel van der Lelie and Dr. Safiyh Taghavi from the Biological Department of the Brookhaven National Laboratory.

Part I

Novel Methodological Development and Application of Structural Equation Modeling (SEM)

1.1 An overview of SEM

Biological pathway analysis almost invariably boils down to the discovery, confirmation, and comparison of networks consisting of nodes and links where nodes stand for biological components such as genes and proteins, and links represent the relations among these components. In this dissertation we will focus on the analysis of biological pathways based on structural equation modeling (SEM) – a modern statistical technique for hypothesis-driven confirmatory and comparative network analyses (Bollen 1989). As shown in Figure 1 (A), SEM evaluates the strength of relations among genes A, B and C simultaneously by a multi-equation system, where some variables can be both the dependent as well as the independent variables such as gene C.

SEM is not a data-driven network discovery tool, instead it is a hypothesis driven confirmatory pathway analysis method designed to integrate existing biological network knowledgebase with newly available experimental data to evolve our understanding of the underlying pathways. In addition, SEM can incorporate latent variables that are not measured directly or accurately, but rather inferred from several measured indicators. This allows researchers to explicitly capture the unreliability of measurements from biological experiments.

SEM is a very general framework that includes a vast array of major statistical analysis methods such as factor analysis, regression analysis and time series analysis as special cases.

To further adapt SEM to modern systems biological studies, in this thesis, we develop the novel mixed variable covariate SEM analysis methods where both the network nodes and pathway covariates (i.e. variables that could potentially influence the strength of pathway links) can be a mixture of categorical and continuous variables (Figure 1 D). Other new extensions and applications to classical SEM include efficient modeling and estimation of mixed variable SEM, a joint statistical and biological covariate SEM pathway analysis pipeline, and latent variable SEM for measurement platforms comparison and integration.

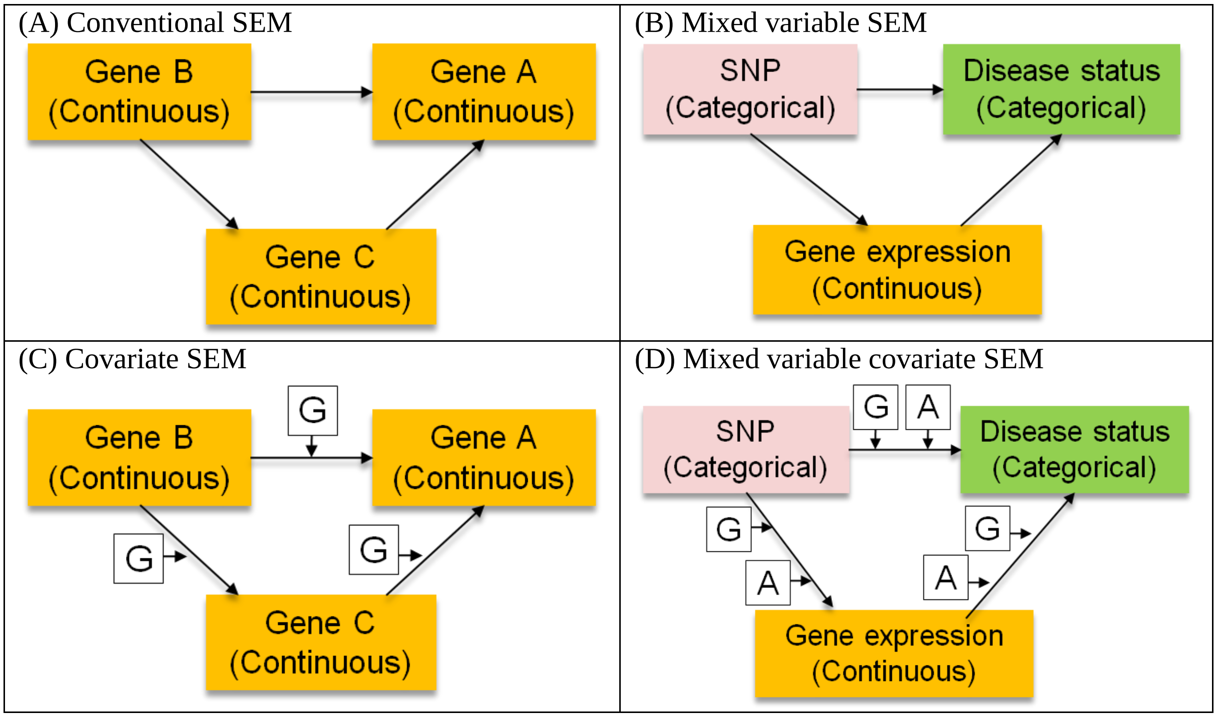
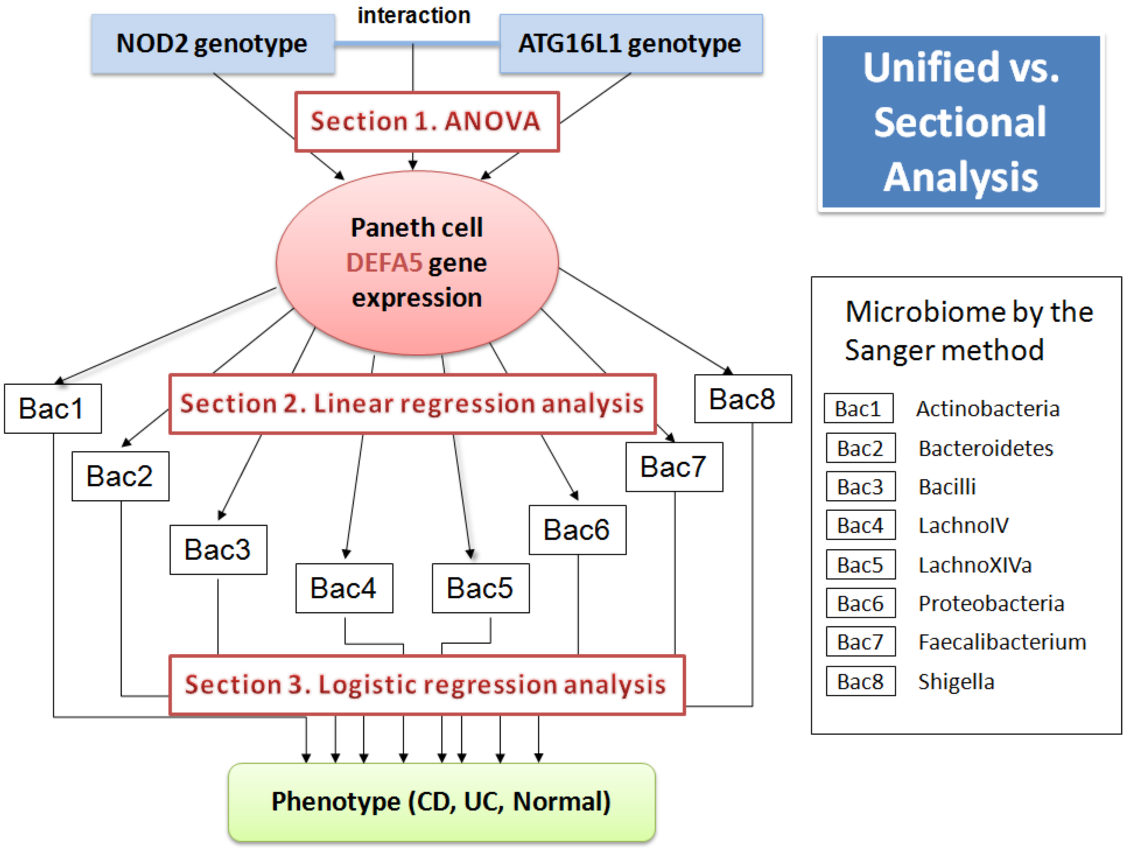


Figure 1. Illustration of structural equation models for different types of biological pathways. (A) Conventional SEM for the pathway with all continuous variables (gene expression values) as nodes. (B) Mixed variable SEM for the pathway with both continuous (gene expression values) and categorical (genotype and phenotype) nodes. (C) Covariate SEM with continuous nodes and categorical covariates (G, such as gender) (Sharpe, 2010). (D) Mixed variable covariate SEM where both the pathway nodes and covariates can be either categorical or continuous variables (G, categorical covariate such as gender; A, continuous covariate such as age).

1.2 Mixed variable SEM

Traditional SEM features exclusively continuous pathway nodes. However, biological pathways often call for both continuous and categorical pathway nodes (Figure 1 B). Figure 2 presents a more detailed mixed variable pathway for the study of inflammatory bowel diseases where the network contains both categorical and continuous variables as nodes. Furthermore, this pathway features categorical variables as both independent (genotypes) and dependent (phenotypes) variables. Before the advent of mixed variable SEM, this pathway can only be analyzed in two or three sections, separately as illustrated in Figure 2. However, the sectional approach ignores relations among the links and is thus myopic and less powerful.

The first general-purpose mixed variable SEM framework, the generalized linear latent and mixed models (GLLAMM), was proposed by Dr. Rabe-Hesketh and colleagues (Skrondal and Rabe-Hesketh 2004). Their multi-level latent variable modeling approach, and their estimation algorithm based on the Gauss-Hermite quadrature, however, are cumbersome and slow. In this thesis, we simplified the modeling framework by combining traditional continuous variable SEM and the generalized linear model (GLM), and at the same time, developed a much efficient computational algorithm based on likelihood factorization.


interaction
NOD2 genotype
ATG16L1 genotype
Section 1. ANOVA
Unified vs. Sectional Analysis
Paneth cell DEFA5 gene expression
Section 2. Linear regression analysis
Bac1
Bac2
Bac3
Bac4
Bac5
Bac6
Bac7
Bac8
Section 3. Logistic regression analysis
Phenotype (CD, UC, Normal)
Microbiome by the Sanger method

Bac1	Actinobacteria
Bac2	Bacteroidetes
Bac3	Bacilli
Bac4	LachnoIV
Bac5	LachnoXIVa
Bac6	Proteobacteria
Bac7	Faecalibacterium
Bac8	Shigella

Figure 2. *Mixed variables SEM*. The entire biological pathway diagram shown here includes both categorical variable nodes (genotypes and phenotypes) and continuous variable nodes (gene expression and bacteria expression). Prior to the development of the mixed variables SEM, the pathway can only be analyzed in sections rather than simultaneously as a whole. For

example, the analysis of variance (ANOVA) can be performed in Section 1 to study the causal relationship between genotypes and Paneth cell alpha defensin 5 (DEFA5) gene expression. Section 2 can be featured as a multivariate linear regression analysis model, to check the relation between the levels of DEFA5 and 8 bacteria expressed in unaffected ileum tissues. Finally, Section 3 can be analyzed via a logistic regression model to examine the links between Microbiome and disease phenotypes.

1.3 Covariate SEM

A fundamental quest in biological network studies is to compare the pathway connection/link strengths between different groups (genotypes, phenotypes, gender etc.). Our group has successfully developed the covariate SEM analysis for network comparison when the underlying nodes are continuous as illustrated by Figure 1 (C) (Sharpe, 2010). In this thesis, we generalize the existing covariate SEM framework featuring continuous variables as pathway nodes and categorical variables as the pathway covariates (i.e. variables modulating the values/strengths of the pathway links), to *mixed variable covariate SEM* by allowing both the nodes and the covariates to be either continuous or categorical.

To incorporate the pathway "covariate" in biological networks is critical, especially for complex diseases where the eventual phenotype is not determined by a single genetic factor but rather, a set of biological and environmental variables and their interactions. Furthermore, the interaction term can often be modeled as a covariate pointing to the pathway between the interacting variables as shown in Figure 3. This example originates from a recent work on the joint analysis of genotypes and gene expression data (Parts, Stegle et al. 2011) where they proposed that the gene expression levels are influenced by genotypes (SNPs), other physiological and/or environmental factors as well as interactions between them. The relation between the expression $y_{g,j}$ of gene g in observation j and its potential covariates is modeled in an additive model as follows:

$$y_{g,j} = \mu_g + \sum_{n=1}^{N} \theta_{g,n} s_{n,j} + \sum_{k=1}^{K} w_{g,k} x_{k,j} + \sum_{k=1}^{K} \sum_{n=1}^{N} \phi_{g,k,n} (s_{n,j} x_{k,j}) + \psi_{g,j} .$$

Here, μ_g is the mean expression level, $\psi_{g,j}$ is noise. $\theta_{g,n}$ denotes the genotype effect of SNP $s_{n,j}$, $w_{g,k}$ is the effect of intermediate factor effect of $x_{k,j}$, and the strength of interaction effects between them is regulated by weight $\phi_{g,k,n}$. We illustrate their model in terms of an SEM path diagram in Figure 3 (A). By this construction, they have found an abundance of statistical interactions and shown how many of them help to interpret yeast gene expression regulation. Other recent studies also searched for genotype-environment effect, and found many gene expression levels affected by epistatic interactions (Costanzo, Baryshnikova et al. 2010) or interactions between the genotype and environmental factors, such as growth conditions of yeast (Smith and Kruglyak 2008), smoking and alcohol on coronary heart disease (van den Donk, van Engeland et al. 2007), and life stressful experiences on depression (Caspi, Sugden et al. 2003). Alternatively, the interaction term can be incorporated as a node pointing to the link of the pathway, that is, as a pathway covariate (Figure 3 B). This further motivated our work on covariate SEM. Along this direction, we extended the mixed variable SEM by incorporating pathway covariates, both categorical and continuous, to better model complex biological pathways. In addition, the green node "other factors" in Figure 3 can be formulated as either known factors such as age, smoking etc., or hidden factors such as latent cell status.

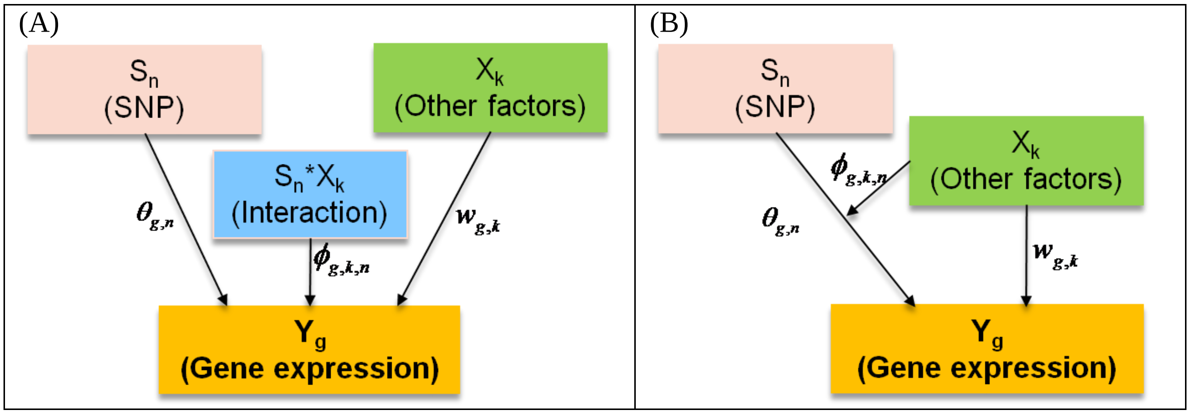


Figure 3. Relation between the interaction term and pathway covariate – illustrated through a gene expression variation model in Parts, et al., 2010. The full model combines genetic factor S (red), other factors X (green) and their interaction S*X (blue) to explain the observed gene expression levels Y (yellow). (A) Interaction as one node pointing to the response gene expression level Y. (B) An alternative way to visualize the interaction term by considering factors X as pathway covariates. Except its path to the gene expression Y, its interaction with SNP was incorporated by adding arrows pointing to the link from SNP to the gene expression Y.

1.4 Application to a study on inflammatory bowel diseases

Inflammatory bowel disease (IBD) is a group of inflammatory conditions of human colon and small intestine. The major types of IBD are Crohn's disease (CD) and ulcerative colitis (UC) (Baumgart and Carding 2007). UC is a specific disease of the large intestine or colon; while CD can affect any part of the gastrointestinal tract, but most commonly involves the terminal ileum. What is the worst, CD is often recursive following removal of affected part. IBD is a complex disease and is found associated with multiple factors, including multiple genetic, microbial, and environmental factors (Podolsky 2002; Renz, von Mutius et al. 2011), however the cause of this disease is not clear yet.

Current inflammatory bowel disease project features a variety of mixed data for each subject include: gene expression data (continuous), genotype data (categorical), microbiome (compositional) and other covariates such as age, BMI, gender, etc. With our newly developed mixed variable covariate SEM method, the question we tried to answer is: "How strong are the relations among these diverse factors on the disease pathway, and how the strengths differentiate under different covariate conditions?" As a result, we found some significant gene expression pathways (e. g. pregnane X receptor pathway), as well as that in the disease model NOD2 risk alleles and Paneth cell related gene expressions increase risk of Crohn's disease while the abundance of bacteria *C. Coccoides* has a negative effect.

1.5 Summary of Part I

Part I of this thesis is presented as follows. In Chapter 2, we provide a thorough literature review on existing biological pathway analyses methods dealing with continuous and categorical variables separately or together. In Chapter 3, background of the structural equation modeling -- model specification and estimation -- is introduced.

In Chapter 4, we propose a bioinformatics pipeline for comparative gene expression pathway studies by integrating the data-driven significance analysis of microarray (Tusher, Tibshirani et al. 2001), the knowledge-driven ingenuity pathway database, and finally the covariate SEM analysis framework previously proposed by our group (Sharpe 2010).

In Chapter 5, we presented a novel application of the latent variable SEM for microbiome data analysis encompassing multiple measurement platforms including Sanger sequencing, pyrosequencing and quantitative PCR. The latent variable SEM is applied to estimate the reliabilities of, and the similarities between different measurement platforms, and subsequently, weigh these measures optimally for a unified analysis of the true latent microbiome composition integrating potential covariates. The latent variable SEM contains the usual repeated measures ANCOVA as special cases and, as a more general, realistic and optimal model, features superior model goodness-of-fit as well as more reliable analysis results as shown in Figure 4 below.

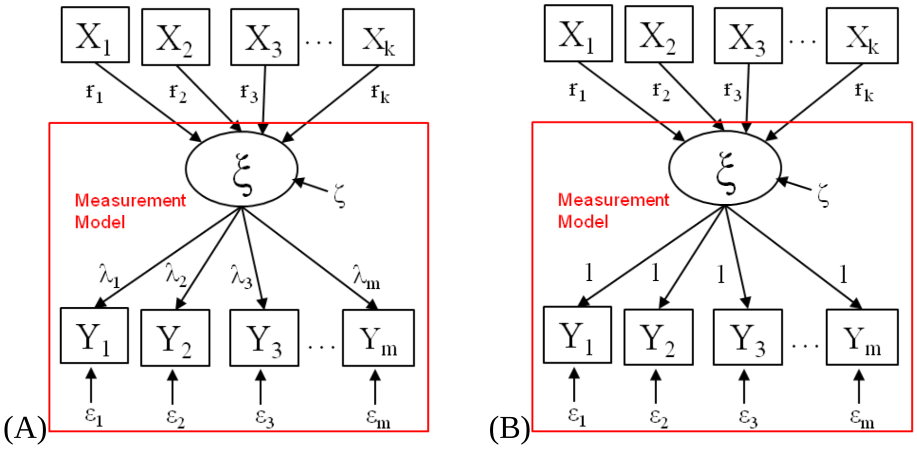


Figure 4. Path diagrams for (A) latent variable SEM, and (B) repeated measures ANOVA. Their difference lies in the measurement model where the repeated measures ANOVA assumes equal path coefficients, and for its univariate approach the measurement error variances are assumed to be equal as well.

In the ensuing Chapter 6-8, we have extended the covariate structural equation modeling (cSEM) method (Sharpe, 2010) for pathway comparisons that was limited to continuous variables on the pathway nodes and categorical variables as pathway covariates only, to allow both continuous and categorical variables as pathway nodes as well as pathway covariates. Chapter 6 presented the derivation of mixed variable SEM containing categorical endogenous (response) variable, while Chapter 7 illustrated the mixed variable SEM in two examples with disease phenotypes as endogenous variable on the pathway. The non-parametric method bootstrapping was also adopted when variables are non-normal. In Chapter 8 mixed variable SEM was further generalized to allow categorical or continuous pathway covariates, and illustrated through the analysis of a study on inflammatory bowel disease.

Part II

Novel Bioinformatics Pipeline and Application

1.6 Bioinformatics work flow

Biological systems such as cells, regulatory gene networks and protein interaction complexes cannot be understood based on individual components (genes, mRNA, proteins etc) alone, but only through an analysis involving multiple components. In recent years, systems biology studies aiming at unraveling the complex interactions in biological systems based on large scale genomic, transcriptomic, metabolic and proteomic data and technologies is becoming increasingly common (Ideker 2004). Although the concept of systems biology has been around for over fifty years, it has only been truly feasible since the 1990’s with the birth of functional genomics and the inventions of high-throughput quantitative sequencing technologies (Zhu and Snyder 2002). The study of the complex biological systems in turn calls for the development of more sophisticated bioinformatics pipelines that will be able to process experimental data, integrate existing biological knowledge-bases, as summarized in the following link: http://www.biochemweb.org/systems.shtml, and also integrate and develop necessary computational tools. The development of novel bioinformatics pipelines, in a sense, has become a critical component in systems biology studies.

In this part of thesis we present a novel integrative bioinformatics pipeline for a systems biology study of the endophytic bacteria at the level of the genome, transcriptome and metabolome. Endophytic bacteria are bacteria that reside within the living tissue of their host plants without substantively harming it (Misaghi and Donndelinger 1990). Endophytic bacteria have beneficial effect on plant growth, which is of great importance for the use of plants as feedstocks for biofuels and for carbon sequestration through biomass production. Moreover, this is vital when considering the aim of improving biomass production of marginal soils, thus avoiding competition for agricultural resources, which is one of the critical socioeconomic issues of the increased use of biofuels (Taghavi, Garafola et al. 2009).

After isolated the endophytic bacteria from their host poplar, whole genome of bacteria were sequenced. Putative **coding sequences (CDS) annotation and function prediction** were performed via *Magnifying Genome (MaGe) annotation platform* (Vallenet, Engelen et al. 2009). Furthermore, **comparative genomics and functional analysis** to explore plant-associated niche specific adaption of bacteria was completed by a combination of the following tools: *PhyloProfile Synteny* and *Genomic Islands* in MaGe (Vallenet, Engelen et al. 2009), as well as *Prophinder* (Lima-Mendez, Van Helden et al. 2008) and *IS Finder* (http://www-is.biotoul.fr/). Analysis of the genome sequences pointed to a remarkable interaction between one of endophytic bacterium, *Enterobacter sp.* 638, and its poplar host (Taghavi, van der Lelie et al. 2010). Particularly it showed the adjacency of two functional operons: sucrose utilization operon (*scrKYAB*) and acetoin / 2,3-butanediol synthesis operon (*budABC*) on the *Enterobacter sp.* 638 genome (Taghavi et al. 2010, Figure 5). It is possible that these two operons interact and play an important role in the crosstalk between the *Enterobacter* sp. 638 and its plant host. The presence of sucrose -- the major photosynthate -- is a signal of proximity with plants to bacteria, which was hypothesized to trigger the transcription of the *budABC* operon in *Enterobacter* sp. 638, resulting in the synthesis of the phytohormones acetoin and 2,3-butanediol. It is a convincing mechanism proved from the genomic, transcriptional and metabolic analyses (Taghavi, van der Lelie et al. 2010).

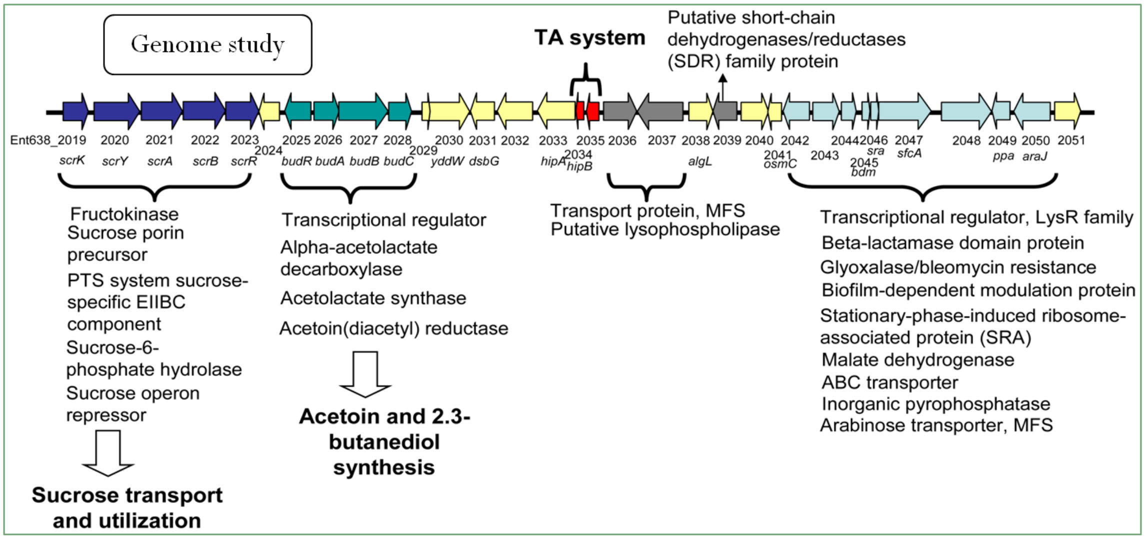

Genome study
TA system
Putative short-chain dehydrogenases/reductases (SDR) family protein
Ent638_2019 2020 2021 2022 2023 2024 2025 2026 2027 2028 2029 2030 2031 2032 2033 2034 2035 2036 2037 2038 2039 2040 2041 2042 2043 2044 2045 2046 2047 2048 2049 2050 2051
scrK scrY scrA scrB scrR budR budA budB budC yddW dsbG hipA hipB algL osmC bdm sra sfcA ppa araJ
Fructokinase
Sucrose porin precursor
PTS system sucrose-specific EIIBC component
Sucrose-6-phosphate hydrolase
Sucrose operon repressor
Sucrose transport and utilization
Transcriptional regulator
Alpha-acetolactate decarboxylase
Acetolactate synthase
Acetoin(diacetyl) reductase
Acetoin and 2.3-butanediol synthesis
Transport protein, MFS
Putative lysophospholipase
Transcriptional regulator, LysR family
Beta-lactamase domain protein
Glyoxalase/bleomycin resistance
Biofilm-dependent modulation protein
Stationary-phase-induced ribosome-associated protein (SRA)
Malate dehydrogenase
ABC transporter
Inorganic pyrophosphatase
Arabinose transporter, MFS

Figure 5. Schematic representation of one genomic region found on the chromosome of *Enterobacter sp.* 638. Putative open reading frames are indicated by arrows, below which the *Enterobacter sp.* 638 gene number and gene annotation are shown. The genes involved in

sucrose transport and utilization, acetoin and 2,3-butanediol synthesis, the toxin-antitoxin (TA system), as well as other putative functions are also indicated.

However it might be a simplified scheme given the distinct phenomenon of this bacterium in sucrose or lactate medium in terms of the growth curve, pH, the extracellular structures and phytohormone productions, etc. There are necessarily more gene transcriptions and regulators involved to respond to presence of sucrose, and finally coordinate a chain of reactions that are as the basis for the strain's adaptation to its endophytic lifestyle. Therefore, we performed whole transcriptome analysis by **RNA-seq** on *Enterobacter sp.* 638 grown either in sucrose or lactate after 6 and 12 hours, in order to gain insights into the differential gene expression profiles under these distinct conditions as shown in Figure 6.

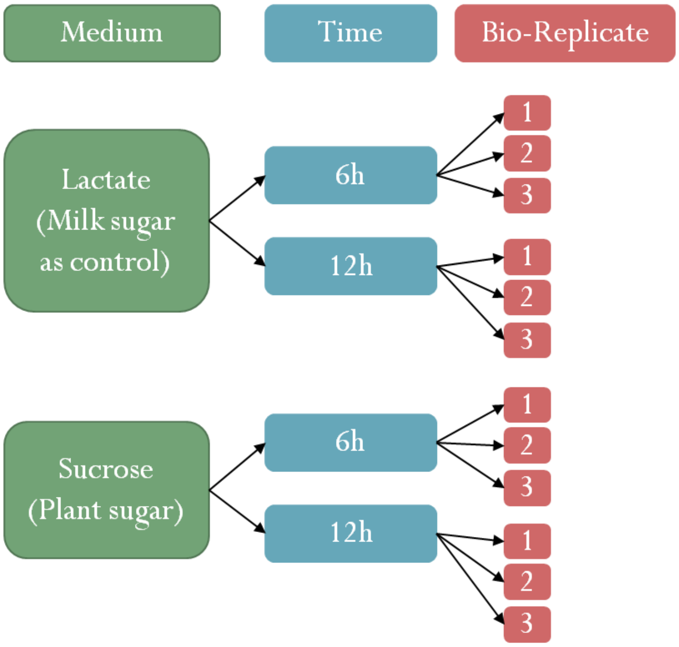


Figure 6. Experimental design of RNA-seq of *Enterobacter sp.* 638 strain. Four distinct conditions are compared with two growth media and two time points: growth media contain lactate or sucrose as sole carbon source, respectively, where sucrose is a plant sugar mimicking the presence of plant while lactate is a milk sugar as a control; two time points are 6 hours or 12 hours after growth. Triplicates of each condition are considered in this experiment.

Raw data from the RNA-seq experiment (Figure 6) are featured by millions of short reads (~36nt) that are aligned to the reference genome of *Enterobacter sp.* 638. Here we adopted an efficient *look-up algorithm Suffix-array* to achieve this goal and transform data to counts of gene expression level. The count data were in turn normalized within samples by *RPKM* (reads per kilobase of exon model per million mapped reads) (Mortazavi, Williams et al. 2008) and between samples by *quantile normalization* that is widely used in microarray study (Bolstad, Irizarry et al. 2003). In the next step, normalized data were compared across different experimental conditions to **identify differentially expressed genes**. Many methods have been developed for the analysis of differential expressions for continuous data generated by microarray, such as SAM (Tusher, Tibshirani et al. 2001). However, RNA-seq provides a discrete measurement for each gene. Even log-transformed, measures are not well approximated by continuous distributions, especially in the lower count and for small samples. By introducing one additional parameter for dispersion, the negative-binomial-based analysis is shown well performed for RNA-seq data, especially for small sample size (Robinson and Smyth 2008). Here we implemented this method to identify differentially expressed genes using *R package edgeR* (Robinson, McCarthy et al. 2010), where the exact test having strong parallels to Fisher's exact test, is used to test for differential expressions and to compute the exact *p* values. The studywise significance level is controlled by keeping the false discovery rate at 0.05 (Benjamini and Hochberg 1995).

Obtaining a list of differential expressed genes is not the final step of the analysis. Next, we grouped genes in terms of similar expression patterns via **clustering analysis** and surveyed representative biological functions in each group in the ensuing **functional categories analysis**. Cluster analysis was performed using the *hclust function in R* based on a distance metric of one minus the Spearman correlation. For each cluster with distinct expression pattern, functional categories analysis was performed by *R package GO-seq* (Young, Wakefield et al. 2010). Instead of GO terms that are only well specified for model organisms, the manually curated functional categories in MaGe, bioprocess and biological roles, are used for our newly isolated bacteria.

There is a wide scope for integrating the results of RNA-seq data with other sources of biological data to establish a more complete picture of gene regulation (Hawkins, Hon et al. 2010). For example, integration of expression data with genotype, transcription factor binding, RNA interference, histone modification and DNA methylation information has the potential for greater understanding of a variety of regulatory mechanisms (Montgomery, Sammeth et al. 2010). A few reports of these 'integrative' analyses have emerged recently (Ouyang, Zhou et al. 2009). Although our current experiment did not generate these additional types of biological data, some efforts were made to better understand the **regulatory networks** based on the observed transcriptional changes.

The genome of *Enterobacter sp.* 638 is very close related to *Escherichia coli* K12. *E. coli* K12 is the best known annotated model organism for bacteria. Therefore, we proposed to first map the orthologs from *E. coli* K12 to *Enterobacter sp.* 638 via the *KEGG ortholog database* in MATLAB (http://www.genome.jp/kegg/soap/doc/keggapi_manual .html), and then infer regulatory relationships in *Enterobacter sp.* 638. The database *RegulonDB* (http://regulondb.ccg.unam.mx/html/Database_summary.jsp) records the most comprehensive and updated transcriptional network for *E. coli* K12. We thereby resorted to the regulatory networks of *E. coli* orthologs for insights on their counterparts in *Enterobacter sp.* 638. The resulting regulatory networks are customized and visualized in *Cytoscape* – a visual analysis tool (Smoot, Ono et al. 2011). The entire work flow to integrate these relevant statistical and bioinformatics tools, as well as the biological databases for large-scale genomics and transcriptomics analysis described above is summarized in Figure 7.

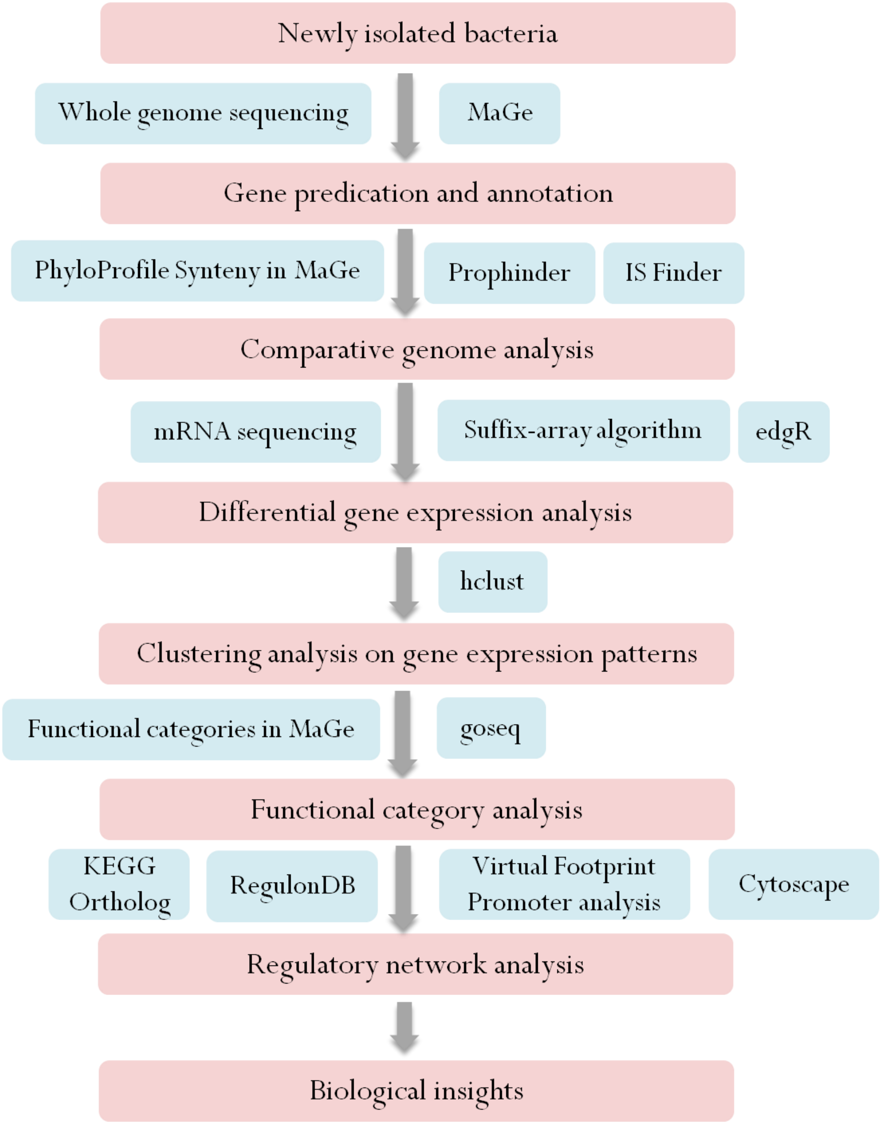

Newly isolated bacteria
Whole genome sequencing
MaGe
Gene predication and annotation
PhyloProfile Synteny in MaGe
Prophinder
IS Finder
Comparative genome analysis
mRNA sequencing
Suffix-array algorithm
edgR
Differential gene expression analysis
hclust
Clustering analysis on gene expression patterns
Functional categories in MaGe
goseq
Functional category analysis
KEGG Ortholog
RegulonDB
Virtual Footprint Promoter analysis
Cytoscape
Regulatory network analysis
Biological insights

Figure 7. Overview of the bioinformatics analysis work flow from newly isolated bacteria to biological discoveries. Major milestones of the pipeline are represented by the red boxes, while methodologies and software used to reach the next milestone are shown in the blue boxes.

1.7 Summary of Part II

In the current study, our objective was achieved by experiments including genome sequencing and mRNA sequencing, and the bioinformatics pipeline shown in Figure 7. We identified an extended set of genes in endophytic bacterium *Enterobacter sp.* 638 involved in plant niche adaptation and beneficial effect to plants: genes that code for putative proteins involved in survival in the rhizosphere (to cope with oxidative stress or uptake of nutrients released by plant roots), root adhesion, colonization/establishment inside the plant (chemiotaxis and flagella), plant protection against fungal and bacterial infections (siderophore production), and improved poplar growth and development through the production of the phytohormones acetoin, 2,3-butanediol and indole acetic acid. We also found that many genes involved in the plant niche adaption appear to under regulation of the RcsAB dual regulator.

In Chapter 9, we elaborated the workflow step by step, starting from genome annotations, comparative genomics and functional analyses of the newly screened organism, and then followed by differential gene expression analysis of mRNA sequencing data, and further biological insight can be gained by exploring patterns of expression changes within clusters and associated functions, and ultimately integrating results to regulatory networks. Results from the endophytic bacteria study were discussed and demonstrated in Chapter 10.

Part I Novel Development and Application of Structural Equation Modeling

Chapter 2 Literature review

2.1 Biological pathway analysis with only continuous variables

Present high throughput biological experiments, such as microarray technique, provide expression levels of tens of thousands of genes simultaneously. The prevalent applications of such experiments have elevated the trend of biological analyses up to a level of pathway perspective. Since outputs from microarray experiments are continuous variables representing corresponding gene expression values, the pathway analyses with continuous variables has been the focus of the biostatistician community. Thus many approaches and tools have been developed in this field. This section will try to provide a through literature review about pathway analyses with only continuous variables, where majority of them were developed in a context of microarray experiments.

2.1.1 Gene set analysis

After determining a list of differentially expressed genes, subsequent pathway analyses are proposed in either of two directions: to connect with existing biological pathways by using public resources; to study the inter-relations among genes suggested by the data. For the first direction, many efforts have been made by biologists to record and construct predefined knowledgebases, such as Gene Ontology (http://www.geneontology.org/) and KEGG (http://www.genome.jp/kegg/). Over-representative test is among the first try to identify the underlying functional profile of a list of genes based on knowledgebase. Such tests often involve the hypergeometric distribution (Cho, Huang et al. 2001), binomial distribution, chi-square test(Fisher and van Belle 1993) and Fisher's exact test (Man, Wang et al. 2000). Alternative approaches include a chi-square test for equality of proportions and Fisher's exact test. In most cases, the differences between these models will not be dramatic. This approach has been adopted with minor variation by many different tools (Hosack, Dennis et al. 2003; Zeeberg, Feng et al. 2003; Al-Shahrour, Diaz-Uriarte et al. 2004; Beissbarth and Speed 2004; Boyle, Weng et al. 2004; Zhang, Schmoyer et al. 2004; Lee, Braynen et al. 2005; Pehkonen, Wong et al. 2005; Yi, Horton et al. 2006). Khatri and Draghici (2005) provided an overview and comparison of such methods. However, a drawback to such gene-list methods is that they rely on the initial gene list in a fundamental way and are sensitive to the choice of both significance criteria and error-control procedure. Moreover, these methods do not consider a gene's relative position in the ranked list.

To overcome the disadvantage of the over-representation approach in its strict cut-off for differential expression of individual genes, methods using the whole vector of *p*-values have been widely used. Gene Set Enrichment Analysis (GSEA) (Mootha, Lindgren et al. 2003; Subramanian, Tamayo et al. 2005) tests whether the ranks of *p*-values of the genes in certain gene set differ from a uniform distribution, using a weighted Kolmogorov-Smirvov test. The idea is similar to the Al-Shahrour (2005) method. In GSEA, genes are ranked based on the correlation between their expression and the phenotype class by using any suitable metric. The priori gene sets S can be defined by GO category, location in the same cytogenetic band, etc. Then the score is calculated by scanning the whole ordered list of gene, increasing when encounter a gene in S and decreasing when meet one is not in S. The enrichment score (ES) is the maximum deviation from zero encountered in the random walk; it corresponds to a weighted Kolmogorov-Smirnov-like statistic. The nominal *p* value of the ES is obtained by performing an empirical phenotype-based permutation test procedure that preserves the gene-gene correlation. They adjust the estimated significance level to account from multiple tests by controlling the proportion of false positives, that is, the false discovery rate (FDR).

GSEA has been improved and extended in many ways. The most notable one is perhaps the Gene Set Analysis (GSA) method proposed by Efron and Tibshirani (2007). This method has been adopted by the Significance Analysis of Microarray (SAM) platform (http://www-stat.stanford.edu/~tibs/SAM/). It uses the max-mean statistic to summarize gene-sets, which is the mean of the positive or the negative part of gene scores in the gene set, whichever is larger in absolute value. The genes are re-standardized before the permutation, and the resulting test statistic is shown to be more powerful than the weighted Kolmogorov-Smirnov-like statistic used in GSEA. The GSA has been extended to account for a versatile array of data including multi-class, survival, and quantitative outcomes (Efron and Tibshirani 2007).

Al-Shahrour (2005) tested genes simultaneously in groups related by common functional properties. First, a list of genes is ordered according to their differential expressions in the experimental conditions by means of a statistical test, then one moves on to establish different partitions across the list by a heuristic method and to check whether partitions of genes with common functional properties are uniformly distributed, or conversely cumulated in one of the tails, where FatiGO (Al-Shahrour, Diaz-Uriarte et al. 2004) has been used to define function categories. Finding of significant asymmetrical distributions of functional terms across the list will suggest groups of genes having in common functional labels are significantly over- or under-expressed as a block. Compared to other approaches based on direct comparison of distributions (e.g. using a Kolmogorov-Smirnov test, or the GSEA), this method is able to find asymmetrical distributions of genes with common biological label across a ranking provided this asymmetry is not too extreme, which means it is relatively sensitive to detect modest asymmetries. However, the disadvantage is apparent due to its necessity to perform a strong adjustment for *P*-values, where for Kolmogorov-Smirnov-based tests; only one test per term is required.

The significance analysis of function and expression (SAFE) (Barry, Nobel et al. 2005) was also an extension of GSEA. It shares very similar idea with GSEA in terms of the two-stage and subject-permutation method. However, it is claimed that SAFE calculates permutation-based *p*-values using a separate null distribution, while GSEA used pooling to compute a FWER-adjusted *p*-value for the largest Kolmogorov-Smirnov statistic, after scaling the statistic based on the different category sizes, where one might ignore the unknown correlation genes. Besides, this paper adopted biological annotation source other than GO: SWISS-PROT, providing a group of keywords for each gene, based on a taxonomy that includes pathways, diseases and general biological process (http://ca.expasy.org/sprot/); The InterPro (http://www.ebi.ac.uk/interpro/) and Protein Family (Pfam) database (http://pfam.sanger.ac.uk/), classifying genes using homology-based domains in the protein sequence.

There are other similar methods existing, for example, researchers proposed "functional class scoring" (FCS), the geometric mean of the *p*-values of the genes in the gene set (Pavlidis, Qin et al. 2004). This paper also showed that the FCS method outperformed the over-representation method in terms of considering all available genomic information rather than predetermined threshold of significance.

2.1.2 Tests based on expression data

Goeman and colleagues tested whether subjects with similar gene expression profiles have similar class labels, based on a logistic regression model (Goeman, van de Geer et al. 2004). Specifically, they proposed a global score test that is based on the random-effects modeling of parameters corresponding to the coefficients of the individual genes in the gene set. Their method addressed the binary and continuous phenotypes as follows. Denoting the phenotype by *Y* and the gene set expressions by ($x_1, \ldots, x_m$), the proposed test statistic is

$$Q = \frac{(\mathbf{Y}-\boldsymbol{\mu})^T \mathbf{R}(\mathbf{Y}-\boldsymbol{\mu})}{\sigma^2}$$

where $\mathbf{R} = (1/m)\mathbf{X}\mathbf{X}^T, \mathbf{X} = (x_1, x_2, \ldots, x_m)$ is a matrix with columns of gene expression vectors, $\mathbf{Y}$ is the vector of outcomes, $\boldsymbol{\mu} = E_{null}(\mathbf{Y})$ is the mean outcome under the null hypothesis of no association, and $\sigma^2 = Var_{null}(\mathbf{Y})$ is the variance of the outcome under the null hypothesis of no association. Later, Goeman et al. extended the method to the censored-survival phenotype with use of a modeling framework incorporating random effects and Cox proportional hazards (Goeman, Oosting et al. 2005).

From different aspect angle, Mansmann and Meister (2005) proposed an ANCOVA global test on main phenotypic effect and gene-phenotype interaction in a two-way layout linear model. This method is equivalent to Goeman et al.'s global test in a setting of independent genes, while in their simulation of correlated genes, this method showed better performance regarding to the power. In particular, among cases where the asymptotic distribution cannot be used, the stratified use of the ANCONA global test outperformed Goeman et al.'s test.

Tomfohr et al. (2005) presented the method that quantifies the level of activity of each pathway in different samples. The predefined gene sets (pathways) are obtained from KEGG (Kyoto Encyclopedia of Genes and Genomes) and Biocarta websites. The activity level is defined in terms of the first eigenvector in the singular value decomposition (SVD) of the matrix of expression levels for its capture of the main component of variation in the full expression matrix. Then the usual significance methods can be applied, where the usual two-sample *t*-test are often performed to evaluate which pathways have activity levels that are significantly different between groups. In addition, the comparison to the GSEA method was demonstrated in an example. In that specific example, their results seem biologically reasonable and are more consistent with previous experimental evidence than those obtained from GSEA.

Dinu et al. (2007) proposed a test called SAM-GS for assessing differential expression of pathways between two phenotypic groups. It addressed the issue of low-variability characteristics of microarray data by adjusting the popular individual-gene analysis, significance analysis of microarray (SAM) (Tusher, Tibshirani et al. 2001; Tibshirani 2006; Efron and Tibshirani 2007). The SAM-GS statistic for pathway analysis with binary phenotype is

$$SAMGS = \sum_{p=1}^{m} d_p^2$$

where *m* is the number of genes in the pathway of interest, and

$$d_p = \frac{\overline{x_p}(1) - \overline{x_p}(2)}{s_p + s_0}.$$

$\overline{x_p}(k)$ is the average expression for the *p*th gene in the pathway for the *k*th class of the binary phenotype (k =1, 2), where s_p is the pooled standard deviation. The constant s_0 was added to adjust for the small variability characteristics of microarray data (Tusher, Tibshirani et al. 2001).

Further, Adewale et al. (2008), based on the SAM-GS and Goeman et al.'s global test, extended pathway analysis to diverse phenotypes, including multi-class, continuous, and censored-survival phenotypes, while allowing covariate adjustments and correlated data. The generality of the proposed method is achieved by using the regression methods. The proposed pathway statistic is defined as:

$$W = \sum_{p=1}^{m} \left(\frac{r_p}{s_p} \right)^2$$

where r_p is any appropriate measure of association between the phenotype Y and the expression x_p for the *p*th gene in the pathway, and s_p is the standard error of r_p. They took r_p as the regression coefficient from modeling the *p*th gene as a predictor of the phenotype Y in an appropriate regression framework. The form of the test statistic is a sum of squares of the Wald statistics for individual genes constituting the pathway.

To sum up, approaches above have one fundamental drawback: it does not render directed or non-directed relationships among the genes. Therefore, it is a method that will only identify potential key players on a biological pathway. Other methods, biological and/data oriented, must be utilized in order to elucidate the entire pathway structure – including key players and their relations as introduced in the following sections.

2.1.3 Inter-relations between genes on biological pathway analysis

In contrast to gene set analysis, current analyses consider not only nodes but also the links between them, which represent the complete pathway property. This introduces the problem that how does one determine the association between genes.

Researchers employed many metrics to represent the gene-gene relationships, including Pearson correlation, partial correlation and conditional probability, etc.

For continuous data, the Pearson correlation coefficient can be used with its classic definition given below

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}$$

It has been assumed that similar patterns in genetic profiles would suggest relationships between genes in that genes with strong correlation in mRNA expression profiles tend to be regulated by the same transcriptional factor and moreover have similar cellular functions (Yu, Luscombe et al. 2003; Allocco, Kohane et al. 2004).

Partial correlation coefficients describe the correlation between two nodes while controlling for the effects of all other nodes in the system. Following the theory of normal distribution, the partial correlation coefficients $\pi_{i,j}$ can be computed from the inverse of the covariance matrix ($\Omega = \Sigma^{-1}$) by:

$\pi_{i,j} = \frac{-\omega_{i,j}}{\sqrt{\omega_{i,i}\omega_{j,j}}}$, where $\omega_{i,j}$ are elements of matrix Ω.

The disadvantage of this procedure is that the empirical covariance matrix can only be inverted if the number of observations exceeds the number of nodes in the network, that is, if the matrix is nonsingular. However, it is usually the case for the gene expression data to have many more genes than observations. Methods based on *covariance selection* provide one way to solve this problem by utilizing the sparse property of the partial correlation matrix. For example, Schafer and Strimmer (2007) have proposed a shrinkage covariance estimator Σ, which is guaranteed to be nonsingular, in a *Gaussian Graphical Model*. This method is based on the assumption that the data follows a multivariate normal distribution N (μ, Σ). The key idea is as follows. It is known that the (unconstrained) maximum likelihood estimator Σ_{ML} has a high variance if the number of nodes exceeds the number of observations, while there are several potentially constrained estimators that have a certain bias but a much lower variance. The shrinkage approach combines the MLE with one constrained estimators Σ_C in a weighted average:

$\Sigma = (1-\lambda)\Sigma_{ML} + \lambda\Sigma_C$, where $\lambda \in [0,1]$ denotes the shrinkage intensity.

Peng and colleagues (2009), inspired by the Gaussian Graphical Model, proposed a method using joint sparse regression techniques for the determination of nonzero partial correlations. They have successfully applied their method to identify hub genes based on partial correlations of microarray data for breast cancer.

2.2 Pathway analysis with only categorical variables

Sole transcriptional changes (gene expressions) are not sufficient to explain all biological phenomena, especially for some complex diseases. Now biological scientists began to explore more, such as SNP (single-nucleotide polymorphism) on the chromosome, for answers. Genomic instability, aberrations of SNP in chromosomes, plays a critical role in the development of many diseases (Klein and Klein 1985). High throughput genotyping experiments have been performed to study genomic instability in diseases. The output of such experiments can be summarized as high-dimensional binary vectors, where each binary variable records genotype status at one marker locus. Therefore, several approaches have been proposed to accommodate this data type in order to understand how SNPs may interact with each other, as it provides insight into the process of the disease development.

2.2.1 Mutual Information

Instead of correlation used for defining the association between genes expressions, the concept of information system has been borrowed to compute the entropy of gene expression patterns and the mutual information between each pair of genes (Butte and Kohane 2000). This method can deal with non-linear associations, whereas only applicable to discrete variables or discretized continuous data.

The entropy of expression pattern is a measure of the information content in that pattern, and is calculated by:

$$H(A) = -\sum_{i=1}^{n} p(x_i) log2(p(x_i))$$

Higher entropy for a gene means its expression levels are more randomly distributed. For continuous expression data, to calculate the discrete probabilities, the researchers use a histogram technique. First calculate the range of values of each gene, and then divide into *n* sub-ranges. *P(*x_i*)* equals to the proportions of measurements in sub-range x_i. The Mutual information is a measure of additional information known about one gene expression pattern when given another, as shown below:

$$MI(A,B) = H(A) - H(A|B) = H(A) + H(B) - H(A,B)$$

Thus the mutual information of zero means the joint distribution of expression values has no more information than genes separately. Higher mutual information between two genes means that one gene is non-randomly associated with the other. In this way, it can be used as a metric between two genes regards to their degree of independence. The hypothesis underlying is that the higher mutual information between two genes, the more likely it is they have a biological relationship.

2.2.2 Graphic model

Alternatively, SNP pathways can be compactly represented by graphs, in which vertices represent SNPs and edges represent interactions between SNPs. Tools developed for graphical models (Lauritzen 1996) can therefore be employed to infer interactions among SNPs. There is a rich literature on fitting graphical models for a small number of variables (Whittaker 1990; Edward 2000; Drton and Perlman 2004). However, in genome-wide SNP profiles, the number of SNPs *p* is typically much larger than the number of samples N. Under such high-dimension-low-sample-size scenarios, Wang et al. recently tackled this challenge by proposing sparse logistic regression with a lasso penalty term and extend it to account for the spatial correlations within chromosomes (Wang, Chao et al. 2011). In their method, they derived a joint probability distribution of the *p* binary variables, which leads to a set of *p* logistic regression models with the combined *p*p* coefficient matrix being enforced symmetric. By assuming symmetric coefficients, they had half-reduced number of parameters. Also they employed psedudolikelihood estimators, a recent work by Hofling and Tibshirani (2009), rather than the exact likelihood estimators. As a result their method was shown high efficiency and fast algorithm by simulation studies.

2.2.3 Boolean Network

The concept of system-level modeling has been extensively studied in engineering and can be utilized towards the modeling of gene regulatory systems (Pomerance, Ott et al. 2009). A Boolean network is a directed graph (network) for discrete state models, whose nodes represent the elements of a system (e.g. genes), characterized by an On (expressing its target protein) or Off state, and directed links between genes indicated that one gene influence the expression of the other either through the expressed protein binding to DNA, or by other signal pathway that modulate transcription of a gene.

In the standard Boolean network model, the system evolves in discrete time steps and at each step the state of every node is simultaneously updated according to some function of its inputs. This function approximates the action of activators (proteins that act to increase expression of a given gene) or inhibitors (proteins that act to reduce expression). Although this method provided a good way to evaluate “switch-type” regulatory pathways, the cutoff expression might seem to be an oversimplification considering the complex mechanism involved in all steps of transcriptional pathways.

2.3 Pathway analysis with mixed variables

Besides pathway analyses dealing with only continuous or categorical variables discussed above, approaches to handle mixed variables (continuous and categorical ones) are in great desire. It is in accordance with the increasing popularity of system biology studies that involve various data types. For instance, data of phenotypes and genotypes are obtained as categorical data, and gene expression as continuous data. Other examples are abundant.

Often although data with mixed variable types are collected in the biological studies, the usual modeling strategy is to consider each outcome separately in a univariate framework. However, the univariate strategy is less efficient in the sense that such an approach ignores the extra information contained in the correlation among the variables. Other advantages of a multivariate setting include avoiding multiple testing and naturally leads to global tests, thus resulting in increased power and better control of Type I error rates Significant efficiency gains over separate univariate analyses have also been reported (Gueorguieva and Sanacora 2006).

Therefore, system biology is in great need of a joint flexible and straightforward analysis to accommodate all mixed variables.

2.3.1 Bivariate mixed variables analysis

The challenge for multivariate methods is the non-existence of obvious multivariate distributions for mixed variables. Several approaches have been proposed, mostly in the context of bivariate mixed response variables (one is categorical and the other is continuous). de Leon, *et al* provided a good review on the bivariate mixed outcome data (de Leon and Chough 2010).

One of the earliest proposals of directly specifying the joint distribution factorizes it into a conditional distribution of one outcome and a marginal distribution of the other. The main idea of the factorization method is to write the likelihood as the product of the marginal and conditional distribution. Cox and Wermuth (1992) discussed two possible factorizations for modeling a continuous and a binary outcome as functions of predictors.

Alternatively, several models using latent variables have been proposed to analyze this problem. Sammel *et al.* (1997) and Arminger and Kusters (1988) discussed models where the outcomes are assumed to be a physical manifestation of a latent variable. A drawback of this model is its non-robustness to misspecification of the covariance because the mean parameters depend heavily on the covariance parameters (Sammel, Lin et al. 1999). Dunson (2000) extended this approach to accommodate non-normal latent variables and non-linear relationships between the observed outcome and the underlying variables. Although very general, Dunson's approach produces a non-identifiable model for the case of a bivariate, binary or continuous outcome. This fact is well known in factor analysis where each latent variable needs three or more indicators in order to for the model to be identifiable; otherwise the parameter space has to be reduced. Often this is achieved by putting constraint on parameters or fixing some parameters to a constant. However, in Dunson's model it is not clear how to constrain the parameters to make the model identifiable without misspecifying the model for the covariance.

In another way, implemented in SAS/STAT software, GLIMMIX procedure performs estimation and statistical inference for generalized linear mixed models (GLMMs) (http://support.sas.com/rnd/app/da/glimmix.html). A GLMM is a statistical model that extends the class of generalized linear models (GLMs) by incorporating normally distributed random effects in the linear predictor and/or by modeling the correlations among the data directly.

Although appealing, all these means in this section are based on bivariate examples, not in context of pathway analysis. Hence they are not directly applicable to the complicated case, for example one can be both independent and dependent variable in different equations on the pathway. This situation can be handled in structural equation modeling (SEM) as its specialty.

2.3.2 Generalized linear latent and mixed model

Dr. Rabe-Hesketh, Pickles and Skrondal have generalized structural equation models (SEM) to accommodate different kinds of responses by a generalized linear latent and mixed models (GLLAMM) (Skrondal and Rabe-Hesketh 2004). SEM specifies relations among variables on the pathway. Chapter 3 will give an introduction on SEM. GLLAMM proposed by Rabe-Hesketh, *et al* is a class of multilevel latent variable models handling various types of responses including continuous, survival data, dichotomous, ordered and unordered categorical data. Typically, they studied the problem of estimating the association between the responses and observed and/or latent explanatory variables. Structural equations are used to specify regressions of latent continuous or discrete variables on explanatory variables as well as relationships among latent variables. The typical model that GLLAMM was designed to analyze is shown in Figure 8.

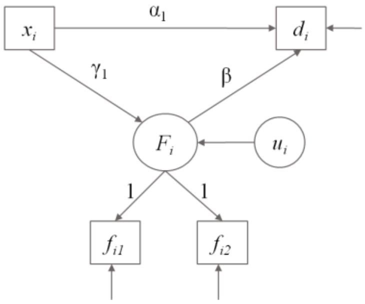


Figure 8. Generalized linear model of the response d_i with the observed variable x_i and the latent variable F_i, where F_i has two observed measurements f_{i1}, f_{i2}, both with measurements errors.

In Figure 8, three sub-models were specified: a latent variable model, a measurement model and an outcome model. The latent variable model of F_i for unit *i* is

$$F_i = \gamma_0 + \gamma_1 x_i + u_i,$$

where x_i is other observed variables, γ_o and γ_1 are regression parameters and u_i is a latent variable representing the deviation of unit *i*'s true value from the mean of x_i.

Next, the classical measurement model assumes that the *r*th covariate measurement for unit *i*, f_{ij}, differs from the latent variable F_i by a normally distributed measurement error ϵ_{ij},

$$f_{ir} = F_i + \varepsilon_{ir} = \gamma_0 + \gamma_1 x_i + u_i + \varepsilon_{ir}$$
$$\varepsilon_{ir} \sim N(0, \sigma_f^2)$$

where ϵ_{ir} and u_i are independent.

The third one, outcome model specifies the relationship between the response and explanatory variables and could also be other forms of generalized linear model. A logistic regression type is

$$\begin{aligned}\text{logit}(P[d_i = 1 | F_i]) &= \alpha_0 + \alpha_1 x_i + \beta F_i \\ &= \delta_0 + \delta_1 x_i + \beta u_i,\end{aligned}$$

where $\delta_0 = \alpha_0 + \beta\gamma_0$ and $\delta_1 = \alpha_1 + \beta\gamma_1$.

Under the normality assumption for u_i, the likelihood is

$$L(\theta_D, \theta_M, \tau) = \prod_i \int P(d_i | u_i; \theta_D) \prod_{r=1}^{n_i} g(f_{ir} | u_i; \theta_M) g(u_i; \tau) \mathrm{d}u_i$$

The likelihood has no closed form but may be integrated numerically using Gauss-Hermite quadrature. They estimate parameters using the Newton-Raphson algorithm and estimate standard errors by inverting the observed information matrix. Although the generality of GLLAMM framework, its time-consuming computation has been criticized by many users, e.g it is reported as “one of the most computationally demanding packages ever” in comparisons among SEM procedures by Dr. Stas Kolenikov (http://repec.org/bost10/kolenikov0712.pdf).

Chapter 3 Structural Equation Modeling (SEM)

Structural equation modeling (SEM) is a methodology for representing, estimating, and testing a network of relationships between variables (measured variables and latent constructs). SEM theory is based on specifying a corresponding model and using data to estimate the values of free parameters. Often the initial hypothesis requires adjustment in light of model evidence. The definition of SEM was articulated by the geneticist Sewall Wright (1921), and the cognitive scientist Herbert Simon (1953) and formally defined by Judea Pearl (2000) using a calculus of counterfactuals (Pearl 2000). SEM can be viewed as a general model of traditional methods like regression, factor analysis and path analysis (Kline 1998).

A suggested process to SEM analysis proceeds from first reviewing the relevant theory and research literature to support model specification; and specification a model (e. g. diagram and equations); determining model identification (e.g. if unique values can be found for parameter estimation and the number of degrees of freedom for model testing is positive); and then collect data, estimate parameters in the model; followed by the evaluation of the model fit and interpret and present results.

3.1 Model specification

Kenneth A. Bollen's textbook gives an excellent introduction theoretical introduction to SEM (1989). The standard path analysis model (SEM with measured variables only) is:

$$Y = \mathrm{B}Y + \Gamma X + \zeta \,.$$

Y is a vector of the endogenous variables studied, and X a vector of the exogenous variables studied. There are some terms used often in SEM. A measured variable refers to a variable that is directly measured whereas in contrast to the latent variable that is a construct not directly or exactly measured. An endogenous—or dependent—variable in the model is one with arrows coming in, i.e. influences of this variable are present in the model. An exogenous—or independent—variable in the model is one with no arrows coming in, i.e. influences of this variable are not present in the current model. In the matrix form, B is a matrix containing path coefficients where the entry $\mathrm{B}_{i,j}$ is the coefficient of the path from endogenous node *j* to endogenous node *i*. Γ is a matrix containing coefficients of paths from exogenous variables to endogenous variables. $\Gamma_{i,j}$ is the coefficient of the path from exogenous node *j* to endogenous node *i*. ζ is a vector containing the error variables in the equations for the path diagram.

Structural equation modeling is often referred to as covariance structure analysis. As this name suggests, interest often focuses on the covariance structure whereas the mean structure is typically eliminated by subtracting the mean from each variable. In our example, both Y and X are centered about their means. The null hypothesis we would like to test in SEM is always $\Sigma = \Sigma(\theta)$, where Σ is the population covariance matrix of the observed variables and $\Sigma(\theta)$ is the covariance matrix written as a function of the free model parameters (the vector θ). The question we want to answer is: "does the covariance matrix predicted by the model is equal to the population covariance matrix?" We can break $\Sigma(\theta)$ into a block matrix as follows.

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix}$$

We will consider each block individually. From the equation $Y = \mathrm{B}Y + \Gamma X + \zeta$, we can obtain the explicit expression of Y:

$$\begin{aligned} &Y - \mathrm{B}Y = \Gamma X + \zeta \\ &(I - \mathrm{B})Y = \Gamma X + \zeta \\ &Y = (I - \mathrm{B})^{-1}(\Gamma X + \zeta) \end{aligned}$$

Thus,

$$\begin{aligned}
\Sigma_{yy}(\theta) &= Cov(Y,Y) \\
&= E[(Y-EY)(Y-EY)'] \qquad (\because EY=0) \\
&= E(YY') \\
&= E[(I-\mathrm{B})^{-1}(\Gamma X+\zeta)((I-\mathrm{B})^{-1}(\Gamma X+\zeta))'] \\
&= E[(I-\mathrm{B})^{-1}(\Gamma X+\zeta)(X'\Gamma'+\zeta')(I-\mathrm{B})^{-1'}] \\
&= (I-\mathrm{B})^{-1}(E(\Gamma XX'\Gamma')+E(\Gamma X\zeta')+E(\zeta X'\Gamma')+E(\zeta\zeta'))(I-\mathrm{B})^{-1'} \\
&= (I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma'+\Psi)(I-\mathrm{B})^{-1'}
\end{aligned}$$

Where Φ is the covariance matrix of X and Ψ is the covariance matrix of error ζ.

Similarly, we can get $\Sigma_{xx}(\theta) = Cov(X,X) = E(XX') = \Phi$ by definition. And the covariance of X and Y:

$$\begin{aligned}
&\Sigma_{xy}(\theta) = Cov(X,Y) \\
&= E[(X-EX)(Y-EY)'] \qquad (\because EX = EY = 0) \\
&= E(XY') \\
&= E(X((I-\mathrm{B})^{-1}(\Gamma X+\zeta))') = E(X(X'\Gamma'+\zeta')(I-\mathrm{B})^{-1'}) \\
&= [E(XX'\Gamma')+E(X\zeta')](I-\mathrm{B})^{-1'} = \Phi\Gamma'(I-B)^{-1'}
\end{aligned}$$

$$\begin{aligned}
&\Sigma_{yx}(\theta) = E(YX') \\
&= [E(XY')]' \\
&= [\Phi\Gamma'(I-B)^{-1'}]' \qquad (\because\ \Phi = \Phi') \\
&= (I-B)^{-1}\Gamma\Phi
\end{aligned}$$

Now we can assemble Σ(θ) as follows:

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix} = \begin{bmatrix} (I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma'+\Psi)(I-\mathrm{B})^{-1'} & (I-\mathrm{B})^{-1}\Gamma\Phi \\ \Phi\Gamma'(I-\mathrm{B})^{-1'} & \Phi \end{bmatrix}$$

Now that we have $\Sigma(\theta)$, we can estimate θ. In estimating θ (a vector of our free paramters—the path coefficients and equation errors), we must choose values of θ in order to minimize the difference between S and $\Sigma(\theta)$. The usual approach of estimating model fit and parameters in SEM is the maximum likelihood (ML) approach.

3.2 Model estimation

The maximum likelihood estimation assumes that the variables in the model are multivariate normal (i.e., the joint distribution of the variables is distributed normally). We will derive the ML of the SEM model based on $Y = \mathrm{B}Y + \Gamma X + \zeta$ as follows.

Assume Y and X are vectors of multivariate normally distributed variables. Assume $Z = \begin{pmatrix} Y \\ X \end{pmatrix}$. Z has length *p*+*q* (*p* is the number of endogenous variables and *q* is the number of exogenous variables in the model). Variables *Z* are centered so that all variables have mean 0. Then because Z is a vector of multivariate normal variables, the distribution of the variables in Z can be written as:

$$f(z;\Sigma) = (2\pi)^{-\frac{p+q}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} z'\Sigma^{-1} z\right)$$

For N independent and identical distributed (iid) observations of the vector Z, the joint density function is

$$\begin{aligned} & f(z_1, z_2, ..., z_N; \Sigma) \\ & = f(z_1;\Sigma) f(z_2;\Sigma) \cdots f(z_N;\Sigma) \qquad (\because \textit{independently distributed}) \\ & = [f(z;\Sigma)]^N \qquad (\because \textit{identically distributed}) \end{aligned}$$

and therefore, the likelihood function is

$$\begin{aligned} & L(\theta; z_1, z_2, ..., z_N) \\ & = f(z_1, z_2, ..., z_N; \Sigma) \qquad . \\ & = (2\pi)^{-\frac{N}{2}(p+q)} |\Sigma(\theta)|^{-\frac{N}{2}} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{N} z_i'\Sigma^{-1}(\theta) z_i\right) \end{aligned}$$

Therefore,

$$\log L(\theta) = -\tfrac{N}{2}(p+q)\log 2\pi - \tfrac{N}{2}\log|\Sigma(\theta)| - \tfrac{1}{2}\sum_{i=1}^{N} z_i'\Sigma^{-1}(\theta) z_i \ .$$

We can simplify $\log L(\theta)$ by dropping and multiplying the constant terms, and finally minimize the resulting fitting function:

$$F_{ML} = \log\left|\Sigma(\theta)\right| + tr\left[S\Sigma^{-1}(\theta)\right] - \log\left|S\right| - (p+q),$$

where S is the sample covariance matrix and p and q are the number of endogenous and exogenous variables, respectively. This fitting function is the basis of ML estimation of SEM. Minimizing the function will yield the appropriate estimates of θ.

Chapter 4 Comparative Genetic Pathway Analysis Using Structural Equation Modeling

In this chapter, we propose a novel genetic pathway discovery and comparison analysis framework integrating newly generated gene expression microarray data and existing biological pathway information. Starting with the significance analysis of microarray (SAM), a list of differentially expressed genes among groups is obtained. This gene list is then imported to the Ingenuity pathway analysis (IPA) to yield potentially relevant biological pathways. Finally, a covariate structural equation modeling method is applied to evaluate pathway connections and group effects. The covariate SEM applied in the final step is based on previous work from Dr. Sharpe in our group (2010). The pathway scheme with all continuous nodes and categorical covariates is considered here. Novel generalizations of this work will be described in later chapters to include both categorical and continuous variables as nodes, as well as covariates on the biological pathways.

Compared to covariate SEM proposed in Sharpe (2010), this comparative genetic pathway analysis is featured by connections to prior knowledge-based pathway construction. We will illustrate this novel pathway analysis pipeline using the whole human genome expression profiling data collected from 99 patients representing three phenotypes: ileal Crohn’s disease (CD), ulcerative colitis (UC) and control subjects without inflammatory bowel diseases (non-IBD).

4.1 Background

In recent years, gene set and pathway analyses have gained increasing popularity over individual gene analyses since the wide availability of the large-scale gene expression data. The analyses of large-scale gene expression profiling dataset have been evolving in two directions:

the data-driven and the knowledge-driven approaches (Viswanathan, Seto et al. 2008). The data-driven analyses are used to generate relationships among gene products (genes or proteins) solely based on experimental data. Various novel statistical methods have been proposed in this direction ranging from the permutation-based *t*-test as implemented in the software suite Significance Analysis of Microarrays (SAM) (Tusher, Tibshirani et al. 2001) to pathway discovery methods, such as the partial correlation network analysis (PCNA) (Peng, Wang et al. 2009). The knowledge-driven analysis, on the other hand, is derived from a detailed pathway knowledgebase for particular domains of interest, such as a cell type, disease or system. The most common statistical method adopted there is a single over-representation test of the gene set based on the hypergeometric model (Cho, Huang et al. 2001) or Fisher's exact test (Man, Wang et al. 2000). This approach has been applied with minor variation by many different tools (Al-Shahrour, Diaz-Uriarte et al. 2004; Beissbarth and Speed 2004; Boyle, Weng et al. 2004; Lee, Braynen et al. 2005). For the knowledge-based analyses, it is critical to adopt a thorough and up-to-date reference database. Many public biological reference databases have been developed including, most notably, Gene Ontology (http://www.geneontology.org/) and KEGG pathway (http://www.genome.jp/kegg/pathway.html). Alternatively, the commercial tool Ingenuity Pathway Analysis (IPA: www.ingenuity.com) has gained tremendous popularity and followings in this field thanks to its up-to-date, integrated and curated knowledge base of canonical pathways and disease-related pathways.

The objective of this study is to develop a pathway analysis pipeline that combines the data- and knowledge-driven approaches. Previous approaches towards combining the two approaches include Gene Set Enrichment Analysis (GSEA) (Mootha, Lindgren et al. 2003; Subramanian, Tamayo et al. 2005), the global test (Goeman, van de Geer et al. 2004) and combination of SAM and IPA (Hever, Roth et al. 2006; Alekseev, Richardson et al. 2009). The SAM/IPA method in particular has gained traction with biological scientists. In order to further evaluate which portions of the biological pathways that have been identified by SAM/IPA are significantly different between groups or conditions, we will adopt the novel covariate structural equation modeling (cSEM) based on a mixed design to this pathway analysis pipeline (Sharpe 2010).

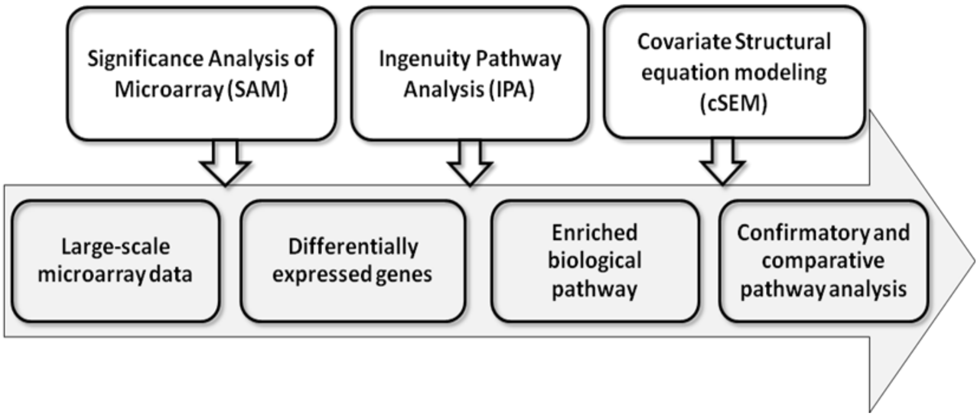

Significance Analysis of Microarray (SAM)
Ingenuity Pathway Analysis (IPA)
Covariate Structural equation modeling (cSEM)
Large-scale microarray data
Differentially expressed genes
Enriched biological pathway
Confirmatory and comparative pathway analysis

Figure 9. Illustration of the proposed comparative pathway analysis pipeline using the SAM, IPA and cSEM.

As illustrated in Figure 9, in the proposed pathway analysis pipeline, differentially expressed genes are first obtained from the SAM gene activation/deactivation analysis, and subsequently imported to the IPA for functional analysis to identify relevant canonical pathways. Genes involved in such canonical pathways are considered for the ensuing cSEM analysis to compare the pathway connection strengths between group / condition etc. Therefore, the SAM analysis will identify pathway nodes/genes that are differentially expressed between groups/conditions, etc., while the cSEM analysis will determine the influence of these covariates (group, condition, etc.) on the pathway connections – i.e. links/paths between the genes.

4.2 Method

4.2.1 Significance Analysis of Microarrays (SAM)

SAM is a statistical analysis technique for identifying significantly activated/deactivated genes using permutation based *t*-test (Tusher, Tibshirani et al. 2001). The multiple hypothesis testing is controlled by false discovery rate (FDR) (Benjamini and Hochberg 1995). One can also choose a fold change threshold. A recent study led by the FDA revealed that SAM is among the top choices to ensure high cross-lab reproducibility in significant findings (Shi, Perkins et al. 2008). We thus chose SAM to identify significantly differentially expressed genes as the first step in the proposed comparative pathway analysis pipeline.

4.2.2 Ingenuity pathway analysis (IPA)

The set of significant genes from SAM is then imported to the IPA (http://www.ingenuity.com/) to identify canonical pathways significantly involved based on existing biological knowledgebase. IPA integrates information available in major public databases and information manually curated by doctoral level researchers from the latest primary literature sources. Thus, in our analysis, we adopted IPA as the reference knowledgebase – the second link on the pipeline.

4.2.3 Covariate Structural Equation Modeling (cSEM)

Structural equation modeling (SEM) is a statistical procedure for confirmatory causal inference proposed in 1921 by the American geneticist Sewall Wright (Wright 1921). However, following its invention, SEM had seen most of its applications in the psychometrics and econometrics fields (Johnston 1972; Bollen 1989) and found little action in the genetic field until the advent of modern microarray studies (Shipley 2000). Xiong et al. (2004) were the first applying SEM to genetic network reconstruction using yeast gene expression data. SEM estimates the hypothesized path model using the covariance (correlation) structure from the data in a maximum likelihood framework. For genetic data, similar gene expression profiles usually suggest relations among genes. It has been shown that genes with strongly correlated mRNA expression patterns tend to be regulated by the same transcriptional factor, and furthermore, have similar cellular functions (Yu, Luscombe et al. 2003; Allocco, Kohane et al. 2004). This justifies the application of SEM to gene expression data.

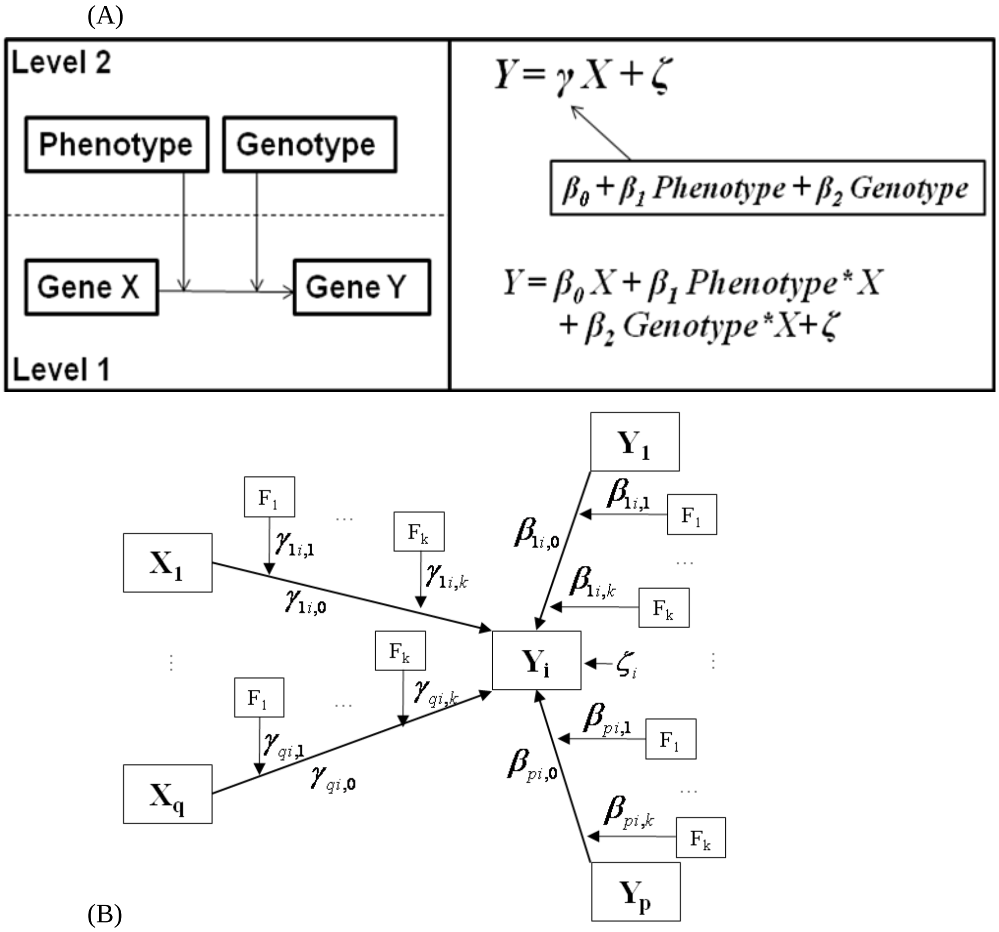


Figure 10. Illustration of the methodology of covariate SEM (cSEM). (A) The two-level parametric model depicts the impact of gene X on gene Y (Level 1) and the impact of two covariates phenotype and genotype in the gene X→Y pathway/interaction (Level 2). The cSEM model was established by re-parameterizing the path coefficient between X and Y to incorporate potential phenotype and genotype influences. (B) A more general path diagram of the response variables $Y_1,...Y_i,...Y_p$ and independent variables $X_1,...,X_q$ incorporating covariates $F_1,...,F_k$. The structural relations and the path coefficients of Y_i are specified according to Equation (1). Similar model structures are defined for other response variables $Y_1,...,Y_p$, but omitted in this path diagram.

Traditional SEM is often inadequate for applications to biological data from certain experimental setting. Recently we have developed a customized SEM procedure and program for mixed designs — a popular paradigm in biomedical studies with multiple groups and repeated measures on each subject (Sharpe 2010). This framework includes the covariate SEM (cSEM) as a special case where our focus is on comparing the pathway connection strengths between several groups (treatment groups, phenotypes, genotypes etc.) where these group and condition factors are referred to as covariates (Figure 10A).

This novel method can analyze multiple independent and/or correlated datasets simultaneously to determine precisely which paths are affected by which factor (group etc.). Correlation of repeated measures is incorporated into the model-implied covariance matrix. It is developed in a maximum likelihood framework through a two-level modeling approach. As illustrated in Figure 10B, the combined two-level SEM model is a series of linear equations:

$$\begin{aligned} y_i &= (\gamma_{1i,0} + \gamma_{1i,1}F_1 + \cdots + \gamma_{1i,k}F_k)x_1 + \cdots + (\gamma_{qi,0} + \gamma_{qi,1}F_1 + \cdots + \gamma_{qi,k}F_k)x_q \\ &\quad + (\beta_{1i,0} + \beta_{1i,1}F_1 + \cdots + \beta_{1i,k}F_k)y_1 + \cdots + (\beta_{pi,0} + \beta_{pi,1}F_1 + \cdots + \beta_{pi,k}F_k)y_p + \zeta_i \end{aligned} \tag{1}$$

for $i = 1 \ldots p$, where p is the number of response variables, F_i is the i^{th} dichotomous factor, x_i's are the independent variables, y_i's are the dependent/response variables, and ζ_i corresponds to the independent normal random error in each equation. The Box-Cox transformation is routinely applied to ensure normality of the data. The maximum likelihood estimators of the model parameters (coefficients) are obtained by minimizing the fitting function,

$$\begin{aligned} F_{ML} = N_0 \log\left|\Sigma_0(\theta)\right| + N_0 tr\left[S_0 \Sigma_0^{-1}(\theta)\right] - N_0 \log\left|S_0\right| - N_0(p+q) + \quad \ldots \quad + \\ N_{G-1} \log\left|\Sigma_{G-1}(\theta)\right| + N_{G-1} tr\left[S_{G-1} \Sigma_{G-1}^{-1}(\theta)\right] - N_{G-1} \log\left|S_{G-1}\right| - N_{G-1}(p+q) \end{aligned} \tag{2}$$

Here N_k is the number of observations taken on the k$^{\text{th}}$ independent group of subjects, S_k is the covariance matrix of $(Y \;\; X)'$ for the k$^{\text{th}}$ independent group of subjects, $\Sigma_k(\theta)$ is the model-implied covariance matrix for the k$^{\text{th}}$ independent group of subjects, and $p+q$ is the number of variables measured for each group (summing the number of variables over all conditions). Once the parameters are estimated, the standard errors can be estimated via the asymptotic covariance matrix (inverse of the Fisher Information matrix). The Fisher Information matrix can be estimated using the matrix of second derivatives of the fitting function, the Hessian matrix of the function. Parameter estimates and corresponding errors are used for significance tests via the asymptotic normal distribution of parameter estimates.

Implemented as the third and final link on the proposed comparative gene expression pathway analysis pipeline, cSEM can evaluate changes in gene-gene interactions (i.e. pathway links or paths) due to groups and/or conditions. Taken together, the proposed analysis pipeline will unravel relevant pathways (SAM+IPA), and identify differences in pathway nodes/gene expression levels (SAM) and in pathway connections/gene-gene interactions (cSEM) across groups/conditions. In the following, we apply the proposed pipeline towards the analysis of a gene microarray study comparing three phenotypes: ulcerative colitis (UC) and Crohn's disease (CD) -- two major inflammatory bowel disease (IBD) phenotypes, and non-IBD controls.

4.3 Application and results

4.3.1 Data set

Ulcerative colitis (UC) and Crohn's disease (CD) are two subtypes of Inflammatory Bowel Disease (IBD). They are chronic inflammatory disorders that are affected by multiple genetic, microbial, and environmental factors (Podolsky 2002). UC is a specific disease of the large intestine or colon; while CD can affect any part of the gastrointestinal tract, but most commonly involves the terminal ileum. This study focused on the dysregulation of genetic factors and their interactions in the ileum of IBD patients. The microarray data were generated from Agilent Human Whole Genome arrays (Agilent No.G4410A). Mucosal mRNA samples were collected from the ileums of 27 UC patients, 47 CD patients and 25 non-IBD controls. See the patients data descriptions in (Zhang, DeSimone et al. 2011). After the data preprocessing, including background filter, normalization and $\log_2$ transformation by limma package in R, and the log ratio was used as expression value of each gene for further analysis.

4.3.2 Differentially expressed genes between UC, CD and non-IBD

The differentially expressed genes between UC, CD and non-IBD patients were identified by two-class unpaired test in SAM for UC vs. non-IBD, CD vs. non-IBD and UC vs. CD (Tusher, Tibshirani et al. 2001). The results contained significant genes, with corresponding FDR < 5%, fold change >1.5 for each of three comparisons. These three gene lists were then combined into one union list (n=2979 genes) for candidate genes to distinguish disease status of UC, CD or non-IBD . These differentially expressed genes would be considered for further analysis. Thus, we obtained information about individual gene expression changes between different IBD phenotypes: up-regulated, down-regulated or no changes.

4.3.3 Associated biological pathways

The called genes, which are potentially related to IBD phenotype, were imported to the IPA for functional analysis of canonical pathways associated with such genes. In terms of the Ingenuity Pathways Knowledge Base, 49 canonical pathways were identified to be over-represented in the gene list at the significant level of 0.05 (Data were not shown). Among the enriched pathway list, the pregnane X receptor (PXR) pathway located at the top significant level and appeared highly significant in ileum samples of UC and non-IBD. Also it is a transcriptional pathway which is suitable for microarray data analysis. Furthermore, it resembled the findings in colon of UC patients from Dr. Langmann's group that dysregulation of PXR target genes in the gut is likely to contribute to the pathophysiology of UC (Langmann, Moehle et al. 2004; Langmann and Schmitz 2006). Thus, we had PXR pathway as one of the most important pathways of our interest to study on.

Based on both canonical pathway in Knowledge Base of Ingenuity and previous experimental evidence, we generated null hypothesis of PXR path diagram for covariate SEM analysis. This pathway includes pregnane X receptor and its target genes involved in phase I, phase II xenobiotics metabolism and transport of xenobiotics, which are critical components in intestinal barrier function against xenobiotics and bacteria (Figure 11).

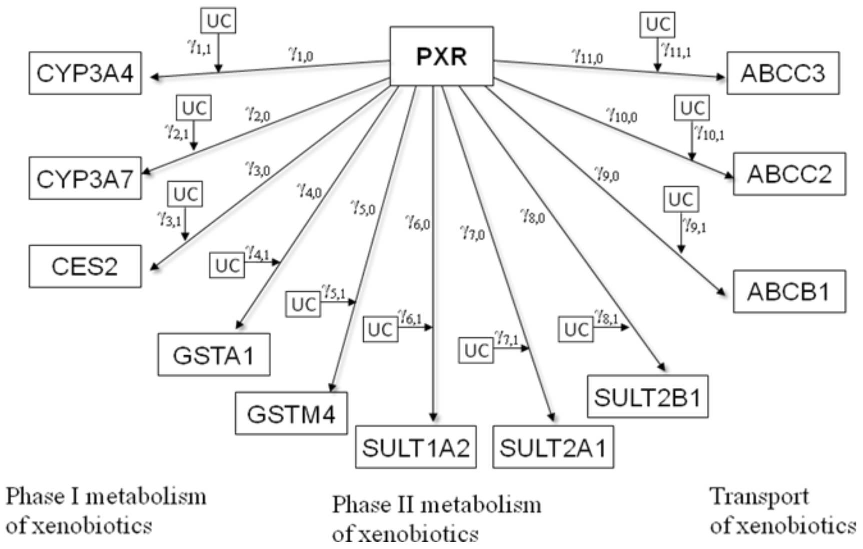


Figure 11. The PXR pathway identified through SAM and IPA analysis, under the cSEM null hypothesis with potential UC effect (versus non-UC) on the pathway connections.

Fold changes of gene expressions in the PXR pathway are shown in Table 1A. The majority of detoxification genes show decreased transcripts in UC compared to non-IBD while no difference between CD and non-IBD. It suggests that, in the ileum, CD and non-IBD patients have similar transcriptional level of PXR pathway and both significantly higher than it is in UC patients as confirmed by the ensuing SAM analysis between UC and non-UC (CD + non-IBD) (Table 1B). Subsequently, we applied cSEM to compare gene-gene interactions on the PXR pathway between UC and non-UC. The work flow of the comparative pathway analysis pipeline on the given IBD study is shown in Figure 12.

Table 1. Fold changes of gene expressions in the PXR pathway

(A) Fold changes in CD, UC ileum samples compared with non-IBD controls

FC against nonUC	**CYP3A4**	**CYP3A7**	**CES2**	**GSTA1**	**GSTM4**	**SULT1A2**
UC	-1.53*	-1.77*	-1.59*	-1.91*	-1.28	-1.70*
CD	-1.14	-1.25	-1.10	-1.35	1.21	-1.09
FC against nonUC	**SULT2A1**	**SULT2B1**	**ABCB1**	**ABCC2**	**ABCC3**	**PXR**
UC	-2.03*	-1.79*	-1.45	-1.75*	1.09	-1.26
CD	1.03	-1.13	-1.25	1.04	-1.17	-1.01

*Significant fold changes have been indicated for the threshold as fold change greater than 1.5.

(B) Fold changes in UC ileum samples compared with non-UC controls

FC against nonUC	**CYP3A4**	**CYP3A7**	**CES2**	**GSTA1**	**GSTM4**	**SULT1A2**
UC	-1.40	-1.53	-1.49	-1.57	-1.46	-1.61
FC against nonUC	**SULT2A1**	**SULT2B1**	**ABCB1**	**ABCC2**	**ABCC3**	**PXR**
UC	-2.07	-1.65	-1.25	-1.80	1.21	-1.26

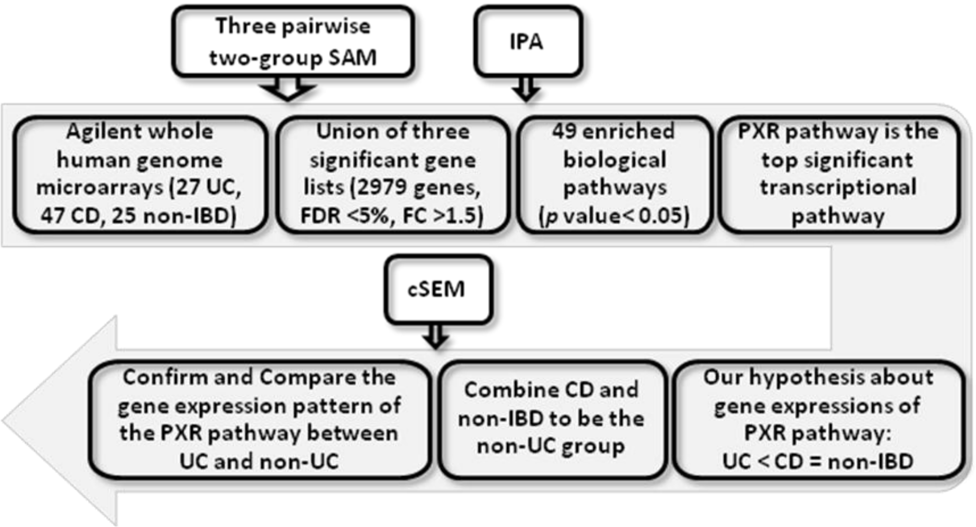

Three pairwise two-group SAM
IPA
Agilent whole human genome microarrays (27 UC, 47 CD, 25 non-IBD)
Union of three significant gene lists (2979 genes, FDR <5%, FC >1.5)
49 enriched biological pathways (p value< 0.05)
PXR pathway is the top significant transcriptional pathway
cSEM
Confirm and Compare the gene expression pattern of the PXR pathway between UC and non-UC
Combine CD and non-IBD to be the non-UC group
Our hypothesis about gene expressions of PXR pathway: UC < CD = non-IBD

Figure 12. Application of the proposed comparative pathway analysis pipeline to the IBD study.

4.3.4 Comparative pathway analysis by covariate SEM

Based on the hypothesized cSEM path diagram of the PXR pathway illustrated in Figure 11, we performed cSEM analysis comparing the UC and non-UC groups. The indicator variables and the corresponding structural equations are shown below.

Model: UC =1 for UC patient, UC =0 for non-UC (CD + non-IBD)

$$\begin{aligned}
CYP3A4 &= (\gamma_{1,0} + \gamma_{1,1}UC) * PXR + \zeta_1 \\
CYP3A7 &= (\gamma_{2,0} + \gamma_{2,1}UC) * PXR + \zeta_2 \\
CES2 &= (\gamma_{3,0} + \gamma_{3,1}UC) * PXR + \zeta_3 \\
GSTA1 &= (\gamma_{4,0} + \gamma_{4,1}UC) * PXR + \zeta_4 \\
GSTM4 &= (\gamma_{5,0} + \gamma_{5,1}UC) * PXR + \zeta_5 \\
SULT1A2 &= (\gamma_{6,0} + \gamma_{6,1}UC) * PXR + \zeta_6 \\
SULT2A1 &= (\gamma_{7,0} + \gamma_{7,1}UC) * PXR + \zeta_7 \\
SULT2B1 &= (\gamma_{8,0} + \gamma_{8,1}UC) * PXR + \zeta_8 \\
ABCB1 &= (\gamma_{9,0} + \gamma_{9,1}UC) * PXR + \zeta_9 \\
ABCC2 &= (\gamma_{10,0} + \gamma_{10,1}UC) * PXR + \zeta_{10} \\
ABCC3 &= (\gamma_{11,0} + \gamma_{11,1}UC) * PXR + \zeta_{11}
\end{aligned}$$

The estimated parameters, *t* values and *p* values are shown in Table 2, and the corresponding path diagram is shown in Figure 13. In Figure 13, fold changes of gene expressions between UC and non-UC were highlighted on the nodes, where green and grey indicated down-regulated expression and no change of expression using the threshold of 1.5 respectively (actual fold changes are tabulated in Table 1B). The majority of PXR downstream genes were decreased in UC patients compared to non-UC subjects, in accordance with previous results in colon by other researchers (Crotty 1994; Langmann, Moehle et al. 2004; Englund, Jacobson et al. 2007).

Table 2. Estimated coefficients of cSEM (UC vs. Non-UC), and the corresponding standard errors, *z* values and *p* values. *P* values in bold indicates the corresponding coefficient is significant at the significance level of 0.05 (one-sided)

Path coefficients	Estimate	Std Error	*Z* value	*p* value
$\gamma_{1,0}$	2.002	0.187	10.727	< **0.001**
$\gamma_{1,1}$	-0.371	0.355	-1.044	0.297
$\gamma_{2,0}$	1.802	0.212	8.481	< **0.001**
$\gamma_{2,1}$	-0.400	0.384	-1.042	0.298
$\gamma_{3,0}$	1.059	0.137	7.748	< **0.001**
$\gamma_{3,1}$	-0.232	0.280	-0.827	0.408
$\gamma_{4,0}$	1.550	0.230	6.738	< **0.001**
$\gamma_{4,1}$	-0.012	0.506	-0.024	0.981
$\gamma_{5,0}$	1.164	0.133	8.720	< **0.001**
$\gamma_{5,1}$	-0.491	0.289	-1.698	**0.090**
$\gamma_{6,0}$	0.774	0.129	5.981	< **0.001**
$\gamma_{6,1}$	0.116	0.206	0.561	0.575
$\gamma_{7,0}$	2.721	0.332	8.192	< **0.001**
$\gamma_{7,1}$	-0.909	0.726	-1.253	0.210
$\gamma_{8,0}$	1.142	0.192	5.945	< **0.001**
$\gamma_{8,1}$	-0.755	0.446	-1.690	**0.091**
$\gamma_{9,0}$	1.539	0.148	10.378	< **0.001**
$\gamma_{9,1}$	-0.544	0.379	-1.438	0.151
$\gamma_{10,0}$	2.295	0.248	9.256	< **0.001**
$\gamma_{10,1}$	-1.065	0.587	-1.815	**0.070**
$\gamma_{11,0}$	-0.186	0.172	-1.081	0.280
$\gamma_{11,1}$	0.547	0.332	1.647	**0.099**

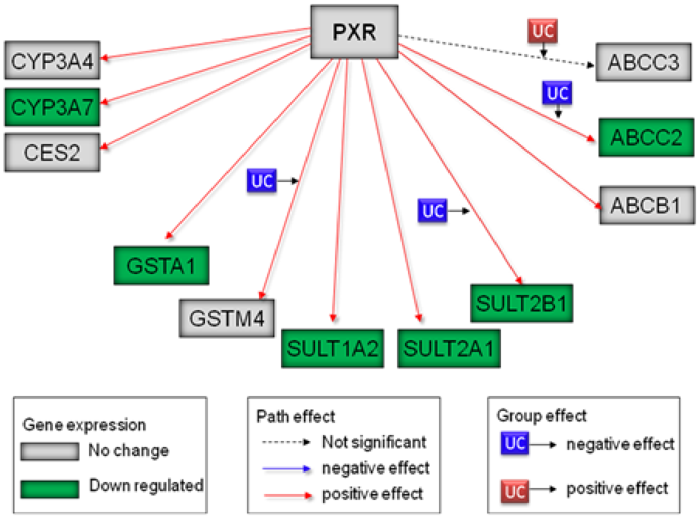


Figure 13. The fitted path diagram for the PXR pathway based on the microarray data. Significant changes in gene expressions between UC and non-UC were highlighted on the nodes,

where green and grey indicated down-regulated expression and no change of expression using SAM with the FDR set at 0.05 and the fold change at 1.5. Significant path and covariate effect are examined through cSEM at the significance level of 0.05 one-sided. Paths in red suggested positive relations between PXR gene and target genes involved in xenobiotic metabolism and homeostasis of endobiotics. Paths with significantly positive or negative UC effect (versus non-UC) are shown in red and blue respectively.

As shown in Figure 13, according to the first level cSEM analysis, all the paths are in red except the path between PXR and ABCC3. Paths in red suggested positive relation between PXR gene and target genes involved in xenobiotic metabolism and homeostasis of endobiotics. When PXR gene expression goes down, the expression of positive-related genes also go down, and vice versa. Although PXR did not down-express as significantly in UC as its downstream genes, we can infer from their positive relations that reduced PXR expression and activity will strongly down-regulate expression of genes encoding detoxification enzymes and transporters. As shown by the second level cSEM group analysis (UC versus non-UC), there are four negative UC group effect on PXR to GSTM4, SULT2B1 and ABCC2 paths, and one positive UC group effect on PXR to ABCC3 path. UC had no significant impact on the other paths of the PXR pathway indicating the relations remained similar between PXR to target genes in UC and non-UC subjects. Negative UC effect on the path, for example, from PXR to SULT2B1, indicates significant reduction of expressions in UC patients compared to non-UC subjects. This result is in accordance with the reduced expression pattern in PXR pathway in the ileum samples of UC patients, which resemble the findings based on the colon samples of UC patients (Langmann, Moehle et al. 2004).

To summarize, our comparative pathway analysis suggested some potentially interesting interactions, especially the effect on gene-gene interaction compared between groups (UC versus non-UC). This result seemed biologically justifiable since it is consistent with previous experimental evidence. Moreover, it provided novel information, for instance, the positive relation from PXR to SULT2B1 is probably influenced by the disease phenotype. The underlying pathway mechanism will help biologists design further experiments to validate the genetic network discovered through the proposed analysis pipelines.

4.4 Discussion

In this work, we proposed a pipeline consisting of SAM, IPA and cSEM for comparative gene expression pathway analysis utilizing both known biological database and newly available experimental data. As noted by the nation’s leading scientists, “The (traditional) reductionist approach has successfully identified most of the components and many of the interactions but, unfortunately, offers no convincing concepts or methods to understand how system properties emerge...the pluralism of causes and effects in biological networks is better addressed by observing, through quantitative measures, multiple components simultaneously and by rigorous data integration with mathematical models" (Sauer, Heinemann et al. 2007). It is our hope that more, and better, work in this direction will follow in the near future.

Chapter 5 Comparative Analysis of Microbiome Measurement Platforms Using Latent Variable Structural Equation Modeling

Culture-independent phylogenetic analysis of 16S ribosomal RNA (rRNA) gene sequences has emerged as an incisive method of identifying bacteria present in a specimen. Currently, multiple techniques are available to enumerate the abundance of bacterial taxa in specimens, including Sanger sequencing, pyrosequencing, and quantitative PCR. In this work we present a novel application of the latent variable structural equation modeling (SEM) to compare these different measurement platforms and to combine them for a unified analysis of the microbiome. This model treats the true relative frequency of a given bacteria in the tissue sample as the latent (unobserved) variable and estimates the reliabilities of, and similarities between different measurement platforms, and subsequently weighs these measures optimally for a unified analysis of the microbiome composition. The latent variable SEM contains the repeated measures ANOVA models as special cases and, as a more general and realistic modeling approach features superior goodness-of-fit as well as more reliable analysis results. We demonstrate the latent variable SEM approach through a microbiome study of the inflammatory bowel diseases (IBD) where the goal is to compare and integrate measurements for the bacterial taxa *Firmicutes/ Clostridium Group XIVa* from four modalities: Sanger, two windows of 454 pyrosequencing, and qPCR, and furthermore, to examine the impact of IBD (sub) phenotypes on the bacteria abundance.

5.1 Motivation

Complex microbial communities, like those of the human gastrointestinal (GI) tract and other environmental specimens, are gaining increased studies, thanks to the technological advances in culture-independent methods based on the amplification of 16S rRNA genes in recent years (Weisburg, Barns et al. 1991). Traditional phylogenetic analysis of a sample is performed by amplifying 16S rRNA genes, cloning, and sequencing by the Sanger method (Sanger and Coulson 1975). Advantage of this method is the sufficiency of single pass Sanger sequencing of 900-1000 bases for classifying bacteria. Disadvantages include annealing bias (Suzuki and Giovannoni 1996), potential cloning bias (Zoetendal, Akkermans et al. 1998), as well as time and expenses. The high cost associated with this approach has been prohibitive for in-depth sampling of complex microbial communities.

Next generation sequencing technology provides a promising alternative to quantifying the microbiome without the limitations of cloning/Sanger sequencing. One 454 sequencing run can produce 1.2 million sequences in 8 hours (Margulies, Egholm et al. 2005), which would require months or years of work with the older methods. The high throughput per run means the unit cost of the next gen. sequencing is only a tiny fraction of that for Sanger sequencing. The new technology also eliminates the cloning bias by directly sequencing the 16S rRNA genes generated by polymerase chain reaction (PCR). Therefore, high throughput sequencing is ideal if adaptable to meet the requirements needed for microbiome work. However, the main limitation of high throughput sequencing is read length. Reads from next generation sequencing technologies are considerably shorter than those from Sanger sequencing. Illumina's Solexa and Applied Biosystem's SOLiD platforms generate reads of about 25-100 bases, while 454 sequencing technology reads up to 400-500 bases per sequence. The concern is loss of classification accuracy with shorter sequence reads (Roesch, Fulthorpe et al. 2007; Dowd, Sun et al. 2008). Several strategies have been tried to maximize the information obtained from short sequences. One is to target certain hypervariable regions (HVR) that are most informative for a specific microbiome of interest (Chakravorty, Helb et al. 2007; Spear, Sikaroodi et al. 2008). As a comparison to the Sanger and the next generation sequencing methods, quantitative PCR (qPCR) approach employs primers specific for particular bacterium to detect and quantify bacteria. It is regarded as a reliable and accurate quantification measure for the absolute amount of 16S rRNA genes from one specific organism (Zemanick, Wagner et al. 2010). However, the accuracy of qPCR highly relies on proper designs of the primers (Rosey, Abachin et al. 2007).

To date, few attempts were made at systematically compare and combine different measurement platforms for microbiome analysis. Nossa *et al.* (2010) analyzed Sanger sequencing and 454 pyrosequencing for 16s rRNA gene sequences and compared the classification accuracies based on these platforms. Here we propose the latent variable structural equation modeling (SEM) for platform comparison and combination. The latent variable SEM includes the repeated measures ANOVA, both the univariate and the multivariate versions, as special cases, and is free from the rigid assumption of the latter approaches such as weighing each platform equally in the analysis regardless of their reliabilities and equal measurement error variances (Kline 1998). The latent variable SEM treats the true bacteria expression as the latent (unobserved) variable and estimates the reliabilities of, and relations between, different measurement platforms, and subsequently combines them for a joint analysis with each platform weighted by its reliability. Furthermore, like the repeated measures ANOVA, the latent variable SEM can easily incorporate covariates such as disease phenotypes and genotypes (Frank, St Amand et al. 2007; Frank, Robertson et al. 2011) to examine their influences on the underlying microbiome composition/bacteria expression.

The latent variable SEM approach is demonstrated through a microbiome study of the inflammatory bowel diseases (IBD) with several measurement platforms including Sanger sequencing, 454 pyrosequencing with different hypervariable regions (windows), and quantitative PCR (qPCR). Our goal is to identify the most reliable microbiome measurement platform, and furthermore, to examine the impact of covariates, especially the IBD disease phenotypes (Crohn's Disease and ulcerative colitis) on the enteric microbiota. This study was supported by the Human Microbiome Project (http://nihroadmap.nih.gov/hmp/) dedicated to uncovering the association between the human microbiome of various anatomical sites and related diseases.

5.2 Methods

5.2.1 Measurement model of latent variable SEM

In latent variable SEM, a latent variable refers to the unknown ground truth such as the true frequency of a certain bacteria in the microbiome. The latent variable is linked to its various measurements or indicators through a measurement model. Figure 14 describes a measurement model where the latent variable ξ (in terms of the IBD study, the true frequency of a certain bacteria in the tissue) is gauged through m measurements Y_i $(i = 1, \dots, m)$ (for the IBD study, measurements from four platforms including Sanger, two 454 windows, and qPCR). Let $\mathbf{Y} = \left(Y_1, Y_2, \cdots, Y_m\right)'$, the latent variable SEM model is a system of linear equations: $\mathbf{Y} = \mathbf{\Lambda}\xi + \boldsymbol{\varepsilon}$, where $\mathbf{\Lambda} = \left(\lambda_1, \lambda_2, \cdots, \lambda_m\right)'$ is the vector of path coefficients showing the expected number of unit changes in the observed variables/measurements for a one-unit change in the true level of ξ. Random errors for the measurements and the latent variable itself are denoted by $\boldsymbol{\varepsilon} = \left(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_m\right)'$ and ζ respectively. We further assume that all errors are normally distributed and independent with $Var(\xi) = \sigma_\zeta^2, Cov(\varepsilon_i, \xi) = 0, Cov(\varepsilon_i, \varepsilon_j) = 0$, and $Var(\varepsilon_i) = \sigma_{\varepsilon_i}^2$ $(i, j = 1, \dots, m,\ i \neq j)$. By convention, **Y** is usually centered about its mean and thus the intercept terms are eliminated.

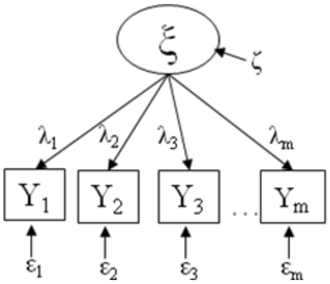


Figure 14. Path diagram for a latent variable SEM measurement model with one latent variable and *m* measurements (observaed variables).

Let θ be the vector of the model parameters including the path coefficients and the error variances, for the latent SEM model illustrated in Figure 14, the population covariance matrix $\Sigma(\theta)$ of Y implied by the SEM model is:

$$\begin{aligned}\Sigma(\boldsymbol{\theta}) &= E(\mathbf{YY}') = E\left[(\boldsymbol{\Lambda}\xi+\boldsymbol{\varepsilon})(\xi\boldsymbol{\Lambda}'+\boldsymbol{\varepsilon}')\right] \\ &= E\left[\boldsymbol{\Lambda}\xi^2\boldsymbol{\Lambda}'+\boldsymbol{\varepsilon\varepsilon}'\right] \\ &= \boldsymbol{\Lambda}E(\xi^2)\boldsymbol{\Lambda}'+E(\boldsymbol{\varepsilon\varepsilon}') \\ &= \boldsymbol{\Lambda}\sigma_\zeta^2\boldsymbol{\Lambda}'+\operatorname{cov}(\boldsymbol{\varepsilon})\end{aligned}$$

Given the multivariate normally distribution of Y, one can estimate the model parameters via the traditional maximum likelihood (ML) method that will eventually result in the minimization of the following ML fit function:

$$F_{ML} = \log\left|\Sigma(\theta)\right| + tr\left[S\Sigma^{-1}(\theta)\right] - \log\left|S\right| - m\,,$$

where S is the sample covariance matrix. This in turn reduces to minimizing the difference between S and $\Sigma(\theta)$, and thus the maximum likelihood estimators of θ are obtained by solving $\Sigma(\theta) = S$.

To fix ideas, we now illustrate the modeling and estimation of the latent variable SEM in details by setting *m* = 3 in Figure 14. The SEM equations are:

$$\begin{aligned}Y_1 &= \lambda_1\xi + \varepsilon_1 \\ Y_2 &= \lambda_2\xi + \varepsilon_2 \\ Y_3 &= \lambda_3\xi + \varepsilon_3,\end{aligned}$$

where $E(Y_i) = 0, E(\varepsilon_i) = 0, Var(Y_i) = \sigma_{y_i}^2, Var(\xi) = \sigma_\zeta^2, Var(\varepsilon_i) = \sigma_{\varepsilon_i}^2, Cov(\xi, \varepsilon_i) = 0$ and $Cov(\varepsilon_i, \varepsilon_j) = 0$.

The model implied covariance matrix (*its upper triangular portion is omitted in the matrix form due to symmetry) is:

$$\begin{aligned}\mathbf{\Sigma}(\boldsymbol{\theta}) &= E(YY') = E[(\Lambda\xi + \varepsilon)(\Lambda\xi + \varepsilon)'] \\ &= \Lambda E(\xi\xi')\Lambda' + E(\varepsilon\varepsilon') \\ &= \Lambda\sigma_{\zeta}^{2}\Lambda' + E(\varepsilon\varepsilon') \\ &= \begin{bmatrix} \sigma_{\zeta}^{2}\lambda_1^2 + \sigma_{\varepsilon 1}^{2} & & \\ \sigma_{\zeta}^{2}\lambda_2\lambda_1 & \sigma_{\zeta}^{2}\lambda_2^2 + \sigma_{\varepsilon 2}^{2} & \\ \sigma_{\zeta}^{2}\lambda_3\lambda_1 & \sigma_{\zeta}^{2}\lambda_3\lambda_2 & \sigma_{\zeta}^{2}\lambda_3^2 + \sigma_{\varepsilon 3}^{2} \end{bmatrix}\end{aligned}$$

Following convention for latent variable SEM estimation, we set one of the path coefficients to 1 to assign a scale to the latent variable (Bollen 1989). This seemingly arbitrary scale assignment has no consequence on the ensuing model estimation because the estimated standardized path coefficients, invariant to this arbitrary scale assignment, will be reported eventually. Thereby without loss of generality, we set $\lambda_1 \equiv 1$to obtain:

$$\mathbf{\Sigma}(\boldsymbol{\theta}) = \begin{bmatrix} \sigma_{\zeta}^{2} + \sigma_{\varepsilon 1}^{2} & & \\ \sigma_{\zeta}^{2}\lambda_2 & \sigma_{\zeta}^{2}\lambda_2^2 + \sigma_{\varepsilon 2}^{2} & \\ \sigma_{\zeta}^{2}\lambda_3 & \sigma_{\zeta}^{2}\lambda_3\lambda_2 & \sigma_{\zeta}^{2}\lambda_3^2 + \sigma_{\varepsilon 3}^{2} \end{bmatrix} \qquad (1)$$

The sample covariance matrix (*again the portion above the diagonal is omitted because of symmetry), on the other hand, is denoted by:

$$S = \begin{bmatrix} S_{11} & & \\ S_{21} & S_{22} & \\ S_{31} & S_{32} & S_{33} \end{bmatrix}$$

By equating $\mathbf{\Sigma}(\boldsymbol{\theta})$ and S, the estimators of the model parameters soon emerge as:

$$\begin{aligned}&\hat{\lambda}_2 = \frac{S_{23}}{S_{13}}, \hat{\lambda}_3 = \frac{S_{23}}{S_{12}}, \hat{\sigma}_{\zeta}^{2} = \frac{S_{12}S_{13}}{S_{23}}, \\ &\hat{\sigma}_{\varepsilon 1}^{2} = S_{11} - \sigma_{\zeta}^{2}, \hat{\sigma}_{\varepsilon 2}^{2} = S_{22} - \sigma_{\zeta}^{2}, \hat{\sigma}_{\varepsilon 3}^{2} = S_{33} - \sigma_{\zeta}^{2} \qquad (2)\end{aligned}$$

In order to evaluate the consistency of the measurement platforms, we adopt the concept of reliability originated from the classical test theory by assuming a true score underlies a measure (Allen and Yen 2002). In the latent SEM measurement model, $R_{y_i}^{2}$, the squared correlation coefficient between the latent variable ξ and its measure Y_i, is a good reliability measure representing the percentage of variance in a measure that is explained by the latent variable (true score). It is appropriate under very general conditions and in simple cases is equal to some of the traditional techniques such as Cronbach's alpha (Bollen 1989). For the latent SEM model illustrated in Figure 14, we have:

$$
\begin{aligned}
R_{y_i}^2 &= \rho_{y_i,\xi}^2 \qquad \text{(squared correlation between observed and latent variable)} \\
&= \frac{\operatorname{cov}^2(y_i,\xi)}{Var(y_i)Var(\xi)} = \frac{\operatorname{cov}^2(\lambda_i\xi+\varepsilon_i,\xi)}{Var(\lambda_i\xi+\varepsilon_i)Var(\xi)} \qquad (\because y_i = \lambda_i\xi+\varepsilon_i) \\
&= \frac{[\lambda_i Var(\xi)+\operatorname{cov}(\varepsilon_i,\xi)]^2}{[\lambda_i^2 Var(\xi)+\operatorname{cov}(\lambda_i\xi,\varepsilon_i)+Var(\varepsilon_i)]Var(\xi)} \\
&= \frac{[\lambda_i Var(\xi)]^2}{[\lambda_i^2 Var(\xi)+Var(\varepsilon_i)]Var(\xi)} \qquad (\because \varepsilon_i \perp \xi, \operatorname{cov}(\varepsilon_i,\xi)=0) \\
&= \frac{\lambda_i^2 Var(\xi)}{\lambda_i^2 Var(\xi)+Var(\varepsilon_i)} = \frac{\lambda_i^2 Var(\xi)}{Var(y_i)} = \frac{Var(\lambda_i\xi)}{Var(y_i)} = 1-\frac{Var(\varepsilon_i)}{Var(y_i)}
\end{aligned}
$$

The last term in the equation can be interpreted as the proportion of variance in the measure Y_i that is explained by the latent variable ξ. The estimated reliability is also closely related to correlations between observed measures. For example, the reliability of y_2 for the simple case of one latent variable with three measurements (Figure 14 with *m* = *3*) is computed from Equation (2) as:

$$
\begin{aligned}
\hat{R}_{y_2}^2 &= \frac{\hat{\lambda}_2^2\hat{\sigma}_\xi^2}{\hat{\sigma}_{y_2}^2} = \left(\frac{S_{23}}{S_{13}}\right)^2 \times \frac{S_{12}S_{13}}{S_{23}} \times \frac{1}{S_{22}} \\
&= \frac{S_{12}S_{23}}{S_{13}S_{22}} = \frac{S_{12}S_{23}\sqrt{S_{11}S_{33}}}{S_{13}S_{22}\sqrt{S_{11}S_{33}}} = \frac{S_{12}}{\sqrt{S_{11}S_{22}}}\frac{S_{23}}{\sqrt{S_{22}S_{33}}}\frac{\sqrt{S_{11}S_{33}}}{S_{13}} \\
&= \frac{r_{12}r_{23}}{r_{13}}
\end{aligned}
$$

Here r_{ij} is the sample Pearson product moment correlation coefficient between the observed variables Y_i and Y_j. Similarly, we have $\hat{R}_{y_1}^2 = \frac{r_{12}r_{13}}{r_{23}}$ and $\hat{R}_{y_3}^2 = \frac{r_{13}r_{23}}{r_{12}}$.

By now we have shown how to compute the R-square from the data, and furthermore, how the R-square is related to the correlations between the observed variables. Suppose the first two of the three measurement platforms are perfectly correlated ($r_{12} = 1$) while the third measure is poorly correlated to the first two with $r_{13} = r_{23} = 0.5$. Then we have$R_{y_1}^2 = R_{y_2}^2 = 1$, and $R_{y_3}^2 = 0.25$. That is, the first two measurements are deemed perfectly reliable on the strength of their perfect consistency, while the third one is considered very unreliable due to its poor correlation to the other measures.

The standardized path coefficients are defined as $\hat{\lambda}_i^* = \hat{\lambda}_i \frac{\hat{\sigma}_\zeta}{\hat{\sigma}_{y_i}}$. Together with the definition of the reliability $\hat{R}_{y_i}^2 = \frac{\hat{\lambda}_i^2 \hat{\sigma}_\zeta^2}{\hat{\sigma}_{y_i}^2}$, we can easily obtain that $\hat{R}_{y_i}^2 = \frac{\hat{\lambda}_i^2 \hat{\sigma}_\zeta^2}{\hat{\sigma}_{y_i}^2} = (\hat{\lambda}_i^*)^2$. Therefore, the standardized path coefficient $\hat{\lambda}_i^*$ is indeed the sample correlation between the observed measurement Y_i and the latent variable.

5.2.2 Latent variable SEM with covariates

While one advantage of the latent variable SEM is to simultaneously incorporate multiple measures for the same underlying latent variable in a measurement model as shown in the previous section, its other advantage is to integrate covariates for the latent variable in the same model. In the ensuing example of the inflammatory bowel diseases, this means one can simultaneously examine the potential influence of covariates such as disease phenotypes on the underlying bacteria expression while incorporating microbiome measures from multiple platforms as shown in Figure 15.

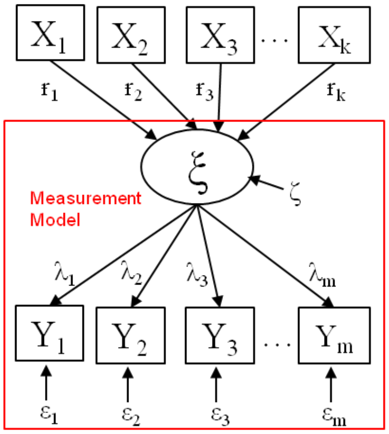


Figure 15. Latent variable SEM path diagram with one latent variable, *m* measurements and *k* covariates.

The SEM model for Figure 15 is:

$$\mathbf{Y} = \mathbf{\Lambda}\xi + \boldsymbol{\varepsilon}$$

$$\xi = \mathbf{\Gamma}'\mathbf{X} + \zeta$$

Here, **Y** is a vector of measurement variables for the latent variable ξ, and **X** is a vector of independent variables (covariates) affecting the latent variable ξ. Both **Y** and **X** have been centered about their means per SEM convention. In addition to notations in the measurement model, we have Γ = [ɼ$_1$,…, ɼ$_k$] representing the vector of path coefficients from the covariates to the latent variable. The estimation procedure is very similar to the measurement model as well. We can break the covariance matrix $\Sigma(\boldsymbol{\theta})$ into a block matrix as follows:

$$\Sigma(\boldsymbol{\theta}) = \begin{bmatrix} \Sigma_{YY}(\boldsymbol{\theta}) & \Sigma_{YX}(\boldsymbol{\theta}) \\ \Sigma_{XY}(\boldsymbol{\theta}) & \Sigma_{XX}(\boldsymbol{\theta}) \end{bmatrix}$$

Each block individually can be factored as follows.

$$\begin{aligned} \Sigma_{YY}(\boldsymbol{\theta}) = E(\mathbf{YY}') &= E\left[(\boldsymbol{\Lambda\Gamma}'\mathbf{X} + \boldsymbol{\Lambda}\zeta + \boldsymbol{\varepsilon})(\mathbf{X}'\boldsymbol{\Gamma\Lambda}' + \zeta'\boldsymbol{\Lambda}' + \boldsymbol{\varepsilon}')\right] \\ &= E\left[\boldsymbol{\Lambda\Gamma}'\mathbf{XX}'\boldsymbol{\Gamma\Lambda}' + \boldsymbol{\Lambda}\zeta^2\boldsymbol{\Lambda}' + \boldsymbol{\varepsilon\varepsilon}'\right] \\ &= \boldsymbol{\Lambda\Gamma}'E(\mathbf{XX}')\boldsymbol{\Gamma\Lambda}' + \boldsymbol{\Lambda}E(\zeta^2)\boldsymbol{\Lambda}' + E(\boldsymbol{\varepsilon\varepsilon}') \\ &= \boldsymbol{\Lambda}(\boldsymbol{\Gamma}'\mathrm{cov}(\mathbf{X})\boldsymbol{\Gamma} + \sigma_\zeta^2)\boldsymbol{\Lambda}' + \mathrm{cov}(\boldsymbol{\varepsilon}) \end{aligned}$$

$$\begin{aligned} \Sigma_{YX}(\boldsymbol{\theta}) = E(\mathbf{YX}') &= E\left[(\boldsymbol{\Lambda\Gamma}'\mathbf{X} + \boldsymbol{\Lambda}\zeta + \boldsymbol{\varepsilon})\mathbf{X}'\right] \\ &= E\left[\boldsymbol{\Lambda\Gamma}'\mathbf{XX}' + \boldsymbol{\Lambda}\zeta\mathbf{X}' + \boldsymbol{\varepsilon}\mathbf{X}'\right] \\ &= \boldsymbol{\Lambda\Gamma}'E(\mathbf{XX}') \\ &= \boldsymbol{\Lambda\Gamma}'\mathrm{cov}(\mathbf{X}) \end{aligned}$$

$$\Sigma_{XY}(\boldsymbol{\theta}) = E(\mathbf{XY}') = \left[\boldsymbol{\Lambda\Gamma}'\mathrm{cov}(\mathbf{X})\right]' = \mathrm{cov}(\mathbf{X})\boldsymbol{\Gamma\Lambda}'$$

$$\Sigma_{XX}(\boldsymbol{\theta}) = \mathrm{cov}(\mathbf{X})$$

Now we can assemble Σ(θ) as the following:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\Lambda}(\boldsymbol{\Gamma}'\mathrm{cov}(\mathbf{X})\boldsymbol{\Gamma} + \sigma_\zeta^2)\boldsymbol{\Lambda}' + \mathrm{cov}(\boldsymbol{\varepsilon}) & \boldsymbol{\Lambda\Gamma}'\mathrm{cov}(\mathbf{X}) \\ \mathrm{cov}(\mathbf{X})\boldsymbol{\Gamma\Lambda}' & \mathrm{cov}(\mathbf{X}) \end{bmatrix}.$$

Thus the parameters can be estimated through minimizing the ML fitting function, or equivalently, by equating $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and S, the sample covariance matrix for both X and Y.

5.2.3 Comparison to repeated measures ANOVA

The traditional approach to incorporate multiple repeated measures to the same underlying latent variable is the repeated measures ANOVA. Here we show that the latent variable SEM is a more general model with the repeated measures ANOVA, both the univariate and the multivariate analysis approaches, as special cases (Figure 16).

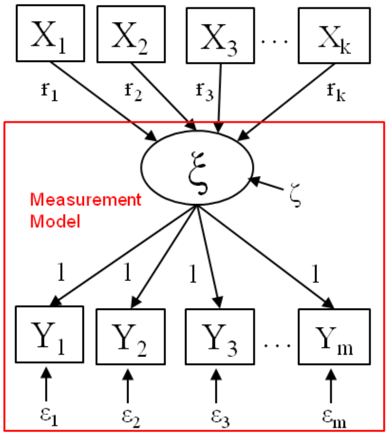


Figure 16. Path diagram for repeated measures ANOVA. In comparison to the latent variable SEM model (Figure 15), the repeated measures ANOVA assumes equal path coefficients for both the multivariate and univariate analysis approaches. In addition, for the univariate approach the measurement error variances are assumed to be equal as well.

The univariate repeated measures ANOVA model is: $Y_{ij} = \mu_j + Z_i + \varepsilon_{ij}$, where μ_j is the (fixed) effect of covariate $X_j, (j = 1, \dots, k)$, Z_i is the (random) effect of subject i, and ε_{ij} are independent and identically distributed random errors independent of Z_i . Let $\underline{Y_i} = (Y_{i1}, Y_{i2}, \cdots, Y_{im})'$, assuming the errors are independent and normally distributed, we have $\underline{Y_i} \overset{iid}{\sim} N_m(\underline{\mu}, \Sigma), i = 1, \cdots, n$, where $\underline{\mu} = (\mu_1, \mu_2, \cdots, \mu_m)'$ and omitting the upper triangle of the matrix by symmetry, we have

$$\Sigma = \begin{bmatrix} \sigma_z^2 + \sigma_\varepsilon^2 & & & \\ \sigma_z^2 & \sigma_z^2 + \sigma_\varepsilon^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_z^2 & \sigma_z^2 & \cdots & \sigma_z^2 + \sigma_\varepsilon^2 \end{bmatrix}.$$

This particular structure of the variance covariance matrix is called “compound symmetry”. The univariate repeated measures ANOVA can be obtained from the more general latent variable SEM shown in Figure 16 (A) by imposing equal measurement error variances and equal path coefficients from the measurements to the latent variable. That is, $\lambda_i = 1$ and $\sigma_{\varepsilon i}^2 = \sigma_\varepsilon^2 \ (i = 1, 2, ...m)$.

The multivariate approach for repeated measures ANOVA allows different measurement error variances but still imposes equal weights to path coefficients from the measurements to the latent variable, that is, $\lambda_i = 1, (i = 1, 2, ...m)$ as shown in Figure 16 (B). The resulting variance covariance matrix Σ for $\underline{Y_i} \overset{iid}{\sim} N_m\left(\underline{\mu}, \Sigma\right) \ (i = 1, \cdots, n)$ is:

$$\begin{bmatrix} \sigma_z^2 + \sigma_{\varepsilon_1}^2 & & & \\ \sigma_z^2 & \sigma_z^2 + \sigma_{\varepsilon_2}^2 & & \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_z^2 & \sigma_z^2 & \cdots & \sigma_z^2 + \sigma_{\varepsilon_m}^2 \end{bmatrix}$$

In summary, the repeated measures ANOVA models, both the univariate and the multivariate approaches, are special cases of the latent variable SEM with constraints on the error variances and path coefficients. The latent variable SEM is a more realistic, flexible and general model to evaluate the latent variable with several measurements, especially when the reliability of each measurement is unclear and the assumption of equal error variances is questionable.

5.3 Case study

5.3.1 Data and model descriptions

Inflammatory bowel diseases (IBD), including Crohn’s disease (CD) and ulcerative colitis (UC), represent the inflammatory conditions of the small intestine and/or the colon. The IBD study reported here includes 44 ileal CD patients, 53 UC patients, and 60 non-IBD control subjects.

The abundance of the bacterial taxa *Firmicutes/ Clostridium Group XIVa* from unaffected ileal samples collected from the proximal margin of resected ileum from each subject is obtained from four microbe measurement modalities: Sanger sequencing, 454 pyrosequencing with two windows: v1v3 and v3v5, and qPCR. For each sequencing platform, the relative frequency of this bacterial taxa was calculated and then subjected to the logit transformation. The qPCR data (dCT) are converted as $qPCR = logit(2^{dCT})$ to have the consistent transformation for other three measurements.

IBD phenotypes (CD and UC) are incorporated as two covariates into the SEM model for an association analysis as well. Path diagrams for the latent variable SEM measurement, and covariate models for *Clostridium GroupXIVa* are shown in Figure 17 (A) and (B) respectively.

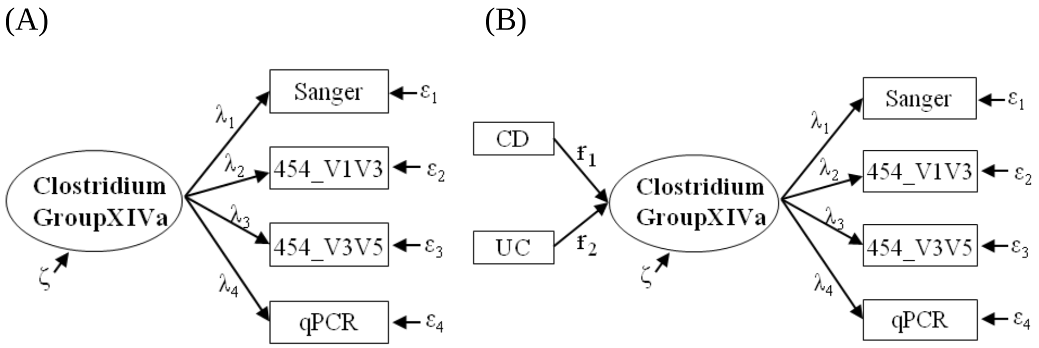


Figure 17. Latent variable SEM approach to incorporate multimodal microbiome measurements. (A) The measurement model with four measurements / indicators for the true (logit-transformed) relative frequency of *Clostridium GroupXIVa*. (B) The covariate model with two binary disease indicators: CD (= 1 for Crohn's disease, and 0 otherwise), and UC (= 1 for ulcerative colitis, and 0 otherwise).

5.3.2 Results

Consistency and reliability of different measurement modalities

Table 3 showed the Pearson correlation among the four measurement modalities for the logit transformed relative frequency of *Clostridium GroupXIVa*. The qPCR data have low correlations with all three sequencing measures, and the v3v5 pyrosequencing window is the best correlated among all modalities as expected since they are based on the same technique. The qPCR analyses were conducted for *F. prausnitzii* using previously established primers (Rinttila, Kassinen et al. 2004). While *F. prausnitzii* is a major species within the *Clostridium Group XIVa* category, nevertheless the target might be different from those in the sequencing modalities. Therefore, although qPCR is often treated as the gold standard for the quantification of nucleotide sequences, it is limited by its high dependency on the accurate specification of primers of targets.

Table 3. Pearson correlations among four different measurement modalities for the logit transformed relative frequency of *Clostridium GroupXIVa* (N = 157).

	Sanger	454_v1v3 (*p* value)	454_v3v5 (*p* value)	qPCR (*p* value)
Sanger	1	0.664 (<.001)	0.691 (<.001)	0.110 (0.173)
454_v1v3		1	0.877 (<.001)	0.206 (0.010)
454_v3v5			1	0.141 (0.078)
qPCR				1

The reliabilities of these measurement modalities are shown in the Table 4. The reliability stands for the squared correlation coefficient between the measurement and the latent variable. Thus it indicates the closeness of one measure with its true value. Again, the v3v5 pyrosequencing is found to be the most reliable with a reliability score of 0.902, and a correlation of 0.950 to the true underlying *Clostridium GroupXIVa* expression.

Table 4. Reliability of each measurement platform in the four-modality latent variable SEM measurement model, and its correlation to the latent variable (true relative frequency of *Clostridium GroupXIVa*).

	Four- modality measurement model			
	Sanger	454_v1v3	454_v3v5	qPCR
Reliability	0.524	0.853	0.902	0.031
Correlation to the latent variable	0.724	0.924	0.950	0.176

Since the reliability is closely related to the correlations among measurement modalities, and since the two 454 pyrosequencing windows feature the highest correlation (r = 0.877) as expected, we also ran a three-modality measurement model by only retaining v3v5, the more reliable pyrosequencing window for quantifying *Clostridium GroupXIVa* (Table 5). Again, the v3v5 pyrosequencing window emerged as the most reliable among the three modalities with an estimated reliability of 0.891 and an estimated correlation of 0.944 with the underlying *Clostridium GroupXIVa* frequency.

Table 5. Reliability of each measurement platform in the three-modality latent variable SEM measurement model, and its correlation to the latent variable (true relative frequency of *Clostridium GroupXIVa).*

	Three-measurement modality model		
	Sanger	454_v3v5	qPCR
Reliability	0.536	0.891	0.022
Correlation to the latent variable	0.732	0.944	0.148

Path diagrams for the measurement models with the estimated standardized path coefficients are shown in Figure 18 below. As shown before, the standardized path coefficient is indeed the correlation between each measurement and the latent variable.

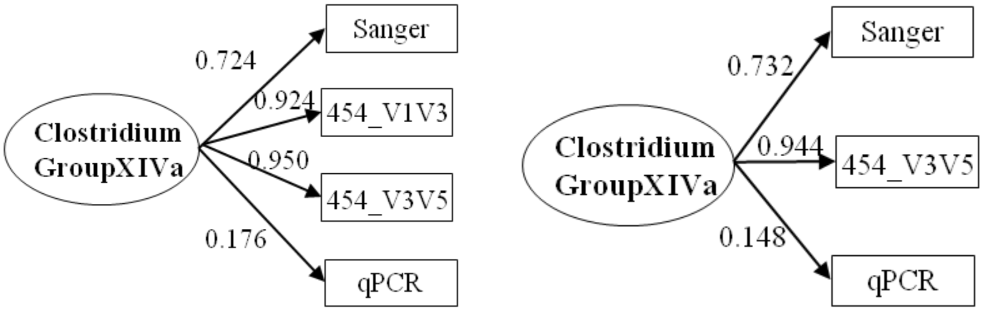


Figure 18. The estimated 4- and 3-modality latent variable SEM measurement models for a study of the inflammatory bowel diseases.

Comparison to repeated measures ANOVA

The model goodness-of-fit indices for the 4- and 3-modality latent variable SEM measurement models for *Clostridium GroupXIVa* are listed in Table 6, and compared to those for the repeated measures ANOVA in both the univariate and the multivariate analysis approaches. SEM relies on several statistical tests to determine the adequacy of model fit to the data. The chi-square test indicates the amount of difference between the expected and the observed covariance matrices. A chi-square value close to zero indicates little difference between the expected and observed covariance matrices. The root mean square error of approximation (RMSEA) is related to the residuals in the SEM model. The RMSEA values range from 0 to 1 with a smaller RMSEA value indicating better model fit. Acceptable model fit is indicated by an RMSEA value of 0.06 or less (Hu and Bentler 1999). The Comparative Fit Index (CFI) is equal to the discrepancy function adjusted for the sample size. That is, CFI = 1 – $d_{(\text{proposed model})}/d_{(\text{Null model})}$, where d equals to the corresponding chi-square minus the degrees of freedom of the model. The CFI ranges from 0 to 1 with a larger value indicating better model fit. Acceptable model fit is indicated by a CFI value of 0.90 or greater (Hu and Bentler 1999). As shown in Table 6, the latent variable SEM (model A) has significantly better Chi-square goodness-of fit index ($\chi^2 = 3.428$, $p = 0.180$) than model B and C representing the repeated measures ANOVA in the multivariate and univariate approaches respectively; model A also has acceptable RMSEA model fit index, while model B and C have poor fit in terms of the RMSEA index; For the CFI criterion, models A and B both provide good fit with CFI values above 0.9.

Table 6. Model goodness-of-fit comparison between latent variable SEM and repeated measures ANOVA approach of *Clostridium Group XIVa* based on four platforms (Sanger, V1V3, V3V5 and qPCR).

MODEL	MODEL CONSTRAINT	GOODNESS-OF-FIT	
A: Latent variable SEM	set $\lambda_1 = 1$	Chi-square	3.428 (df = 2) Pr > χ^2: 0.180
		RMSEA	0.068
		CFI	0.996
B: Equivalent to repeated measures ANOVA (multivariate approach)	set all indicator path coefficient $\lambda_j = 1$	Chi-square	26.562 (df = 5) Pr > χ^2: <0.001
		RMSEA	0.166
		CFI	0.936
C: Equivalent to repeated measures ANOVA (univariate approach)	set all indicator path coefficient $\lambda_j = 1$; set all indicator error variances to be equal, that is, let var (ε_i) $=\sigma^2$	Chi-square	352.835 (df = 8) Pr > χ^2: < .001
		RMSEA	0.526
		CFI	0.000

Estimation of the latent variable SEM model with covariates IBD phenotypes

In this section, we examine the impact of two IBD phenotypes, CD and UC, on the relative frequency of *Clostridium GroupXIVa* via the latent variable SEM simultaneously utilizing measurements of the given taxa from either all four platforms, or only three (minus the v1v3 window of the 454 pyrosequencing). The results are summarized in Figure 19.

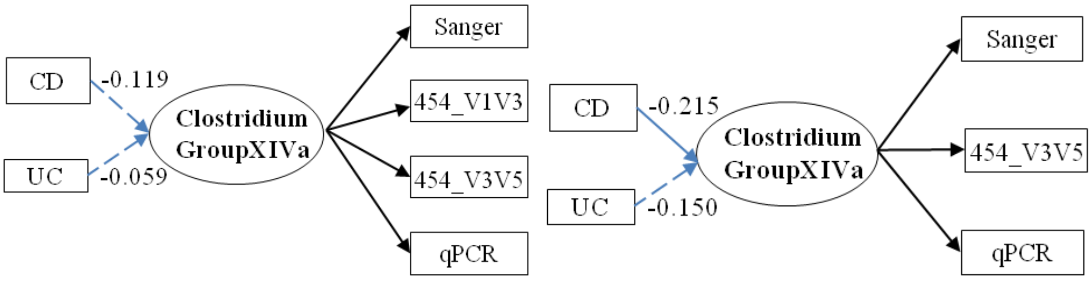


Figure 19. The estimated 4- and 3-modality latent SEM models examining the effect of two covariates: CD and UC phenotypes.

Subjects with Crohn's Disease (CD) was found to have a significantly lower relative abundance of *Clostridium GroupXIVa* (p = 0.023) in the three-modality (Sanger, 454 v3v5, and qPCR) latent variable SEM analysis. The four-modality latent variable SEM analysis utilizing all available measurements showed a trend of reduction among CD patients (p = 0.192) . The difference may lie in the decrease of model parameters for the three-modality model that renders it more powerful to detect the underlying difference than the four-modality model. The three-modality model also showed a trend of negative impact of the UC phenotype on the given bacteria taxa (p = 0.108).

To our knowledge, this is the first application of SEM modeling to studies of the human microbiome. Because human gastrointestinal microbial communities typically are complex and difficult to study in situ, multiple experimental modalities are required to provide a deep description of the dynamic microbe-microbe and microbe-host interactions within the gut. In this study we demonstrate that latent variable SEM can provide a robust means of integrating datasets derived from different experimental methodologies. Moreover, we show that SEM can be used to evaluate the relative merits of different measurement techniques, in this example, Sanger sequencing, pyrosequencing, and qPCR.

Chapter 6 Mixed variable SEM – Joint analysis of pathway with mixed continuous and categorical endogenous variables

6.1 Model estimation

6.1.1 A simple example

To fix ideas, we will start with a simple case as shown in Figure 20. In this example, we assume that *X* is a continuous exogenous variable, *Y* is a continuous endogenous variable, and *W* is a categorical endogenous variable. The continuous variables *X* and *Y* are centered to have mean 0.

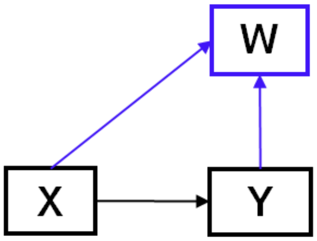


Figure 20. A simple example of a pathway with both a categorical variable (W) and a continuous variable (Y) as the endogenous variables (i.e. dependent or response variables) on the pathway.

Unlike the bivariate mixed variable analyses we reviewed in Chapter 2.3, here random variable Y is both as dependent variable to X and as independent variable to W, a simple scheme of pathway. Nevertheless, we are inspired by previous methods -- the factorization idea and GLLAMM -- here we propose to use a factorization method for the joint density distribution of variables. The main idea of the factorization method is to write the likelihood as the product of the marginal distribution of one outcome and the conditional distribution of the other outcome given the previous one. Cox and Wermuth (1992) discussed two possible factorizations for modeling a continuous and a binary outcome as functions of predictors. As our example, the joint density function has two ways of factorization $f(X,Y,W) = f(X,Y|W)f(W) = f(W|X,Y)f(X,Y)$. The way of factorization models represent a structure in variables, which the conditioning variable treated as intermediate variable and the conditioned variable as the ultimate response. The direction of conditioning is suggested by the direction of the arrows on the path diagram. Since we have the categorical variable W as the endogenous variable on the pathway, we naturally adopted the second way of factorization: $f(X,Y,W) = f(W|X,Y)f(X,Y)$. The distribution $f(X,Y)$ can be further written as conditional probability density function (PDF) of continuous endogenous variable Y given X: $f(Y|X)f(X)$.

The likelihood function of GLLAMM is also based on the product of distributions; however the setting of latent variable is mandatory, even with the only measurement, one has to specify the measurement error in order to make the model identifiable. The corresponding likelihood function of GLLAMM for the pathway in Figure 20 is:

$$L(\theta,\tau) = \prod_i \int f(W_i | u_i;\theta) f(Y_i | u_i;\theta) g(u_i;\tau) \mathrm{d}u_i$$

where u is the latent variable underlying the observed variable Y, and it will be integral out for the estimation. The parameter τ is the corresponding measurement error to be estimated. When there is only one observed measurement of Y, specification of τ is mandatory. Iterative Gauss-Hermite quadrature method is used to find the estimates in the form above. Therefore, GLLAMM is criticized by their very low computation efficiency especially for the model with all observed variables.

In addition, note that in GLLAMM, W and Y are assumed conditionally independent given the latent variable, and the distribution of exogenous variable X is not considered. It is atypical to conventional SEM which considers covariance structure of all variables on the pathway. Bollen's book (1989) discussed this question:

"The assumption that Y and X are sampled independently from a multinormal distribution can be replaced with either of two alternatives that lead to the same ML estimators, standard errors, and tests of significance for $\mathbf{B}$, Γ and Ψ as before. The first alternative assumes that X is a random variable distributed independently of error, whereas the second assumes that x is fixed in repeated samples. Both alternatives assume that error is multinormal with a covariance matrix of Ψ. Though these options are appealing they are not always appropriate. For example, in most non-experimental research, x is random rather than fixed and varies as new units are sampled..."

Therefore, we felt it is obliged to include distribution of X in the likelihood function for the estimation of the whole pathway. In a result, to the example in Figure 20, we adopted the derivation based on $f(X,Y,W)=f(W\,|\,X,Y)f(X,Y)$. We assume *(X, Y)* follows bivariate normal distribution and W given X and Y follows Bernoulli(π). The joint distribution $f((X,Y,W)\,|\,\theta)$ is expressed by conditional probability mass function (PMF) of *W* given (X, Y) and PDF of (*Y X*):

$$f((X,Y,W)\,|\,\theta)=f(X,Y;\theta_c)P(W\,|\,X,Y;\theta_b)$$

where θ_c,θ_b are two subsets of the parameter set θ. As indicated by the subscripts, θ_c contains parameters in the distribution of continuous variables $f(X,Y)$ and θ_c,θ_b contains parameters in the conditional distribution of binary variable $P(W\,|\,X,Y)$.

The first part is the distribution of (X, Y):

$$\begin{aligned}&f(X,Y;\theta_c)\\&=\frac{1}{2\pi}\left|\Sigma(\theta_c)\right|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}[Y\;X]\Sigma(\theta_c)^{-1}\begin{bmatrix}Y\\X\end{bmatrix}\right)\end{aligned}$$

Where $\Sigma(\theta_c)$ is the variance-covariance matrix of (Y X) implied by the model, which containing the parameter set $\theta_c:(\beta,\phi,\psi)$. It follows the traditional form for SEM with β being the coefficient of Y regressed on X; ϕ being the variance of X and ψ being the error variance.

$$\Sigma(\theta_c)=\begin{bmatrix}Var(Y) & Cov(Y,X)\\Cov(X,Y) & Var(X)\end{bmatrix}=\begin{bmatrix}\beta^2\phi+\psi & \beta\phi\\\beta\phi & \phi\end{bmatrix}.$$

The above expressions of $\Sigma(\theta_c)$ is then inserted into the density function.

The second part is the conditional PMF of dichotomous variable *W* given (Y X). Here we will derive this conditional PMF based on generalized linear model (GLM). Generalized linear modeling is the most common approach to model a wide range of response processes, including continuous, dichotomous, ordinal data, counts and durations, etc. The explanatory variables affect the response through the linear predictor *v_i* for unit *i*, $v_i=\mathbf{x}_i'\beta$ where $\mathbf{x}_\mathrm{i}$ is a vector of explanatory variables and β contains the corresponding regression parameters. Both continuous and categorical explanatory variables can be accommodated. For categorical variables such as colors, dummy variables would typically be specified. The response process is described by specifying the conditional probability of *y_i* given the linear predictor. The special case for dependent variable is continuous. A linear regression model $y_i = v_i + \varepsilon_i$ is usually specified, where the residual ε_i are independently normally distributed with zero mean and variance σ^2. The linear regression model can alternatively be defined by setting the conditional expectation of the response, given the linear predictor v_i , $\mu_i \equiv E(y_i \mid v_i) = v_i$, and specifying that the y_i are independently normally distributed with mean μ_i and variance σ^2.

All the generalized linear models have a common structure and can be defined by two components: (1) the function between the expectation of the response and the linear predictor $\mu_i = g^{-1}(v_i)$ or $g(\mu_i) = v_i$, where g(.) is a *link* function. Logit is the link function for logistic regression; (2) the conditional probability distribution of the responses is a member of the exponential family with expectation μ_i . The conditional probability distribution of the dichotomous response is Bernoulli distribution.

For dichotomous responses taking on values 0 or 1, the conditional probability of response 1, $\Pr(y_i = 1 \mid v_i)$, is just the conditional expectation μ_i of y_i. This can be modeled as a *logistic regression*

$$\mu_i = \frac{\exp(v_i)}{1+\exp(v_i)} \text{ or } \ln(\frac{\mu_i}{1-\mu_i}) = v_i .$$

Conditional on v_i, the y_i are independently Bernoulli distributed.

For the current example in Figure 20 the conditional PMF of dichotomous variable *W* given (Y X) which we assume follows a Bernoulli distribution with the logit link function:

$$\text{logit}(E(W \mid X\ Y)) = \alpha_0 + \alpha_1 X + \alpha_2 Y$$

Let $\pi = E(W \mid X\ Y)$. After taking *e* to both sides of the equation above, we get:

$$(\frac{\pi}{1-\pi}) = \exp(\alpha_0 + \alpha_1 X + \alpha_2 Y) .$$

Solve the equation we get

$$\pi = \frac{\exp(\alpha_0 + \alpha_1 X + \alpha_2 Y)}{1+\exp(\alpha_0 + \alpha_1 X + \alpha_2 Y)} \quad \textit{and} \quad (1-\pi) = \frac{1}{1+\exp(\alpha_0 + \alpha_1 X + \alpha_2 Y)}$$

Thus the PMF of W given (Y X) is as below:

$$
\begin{aligned}
P(W \mid X\ Y;\theta_b)) &= \pi^{W}(1-\pi)^{1-W} \\
&= (\frac{\pi}{1-\pi})^{W}(1-\pi) \\
&= \exp(W(\alpha_0+\alpha_1 X+\alpha_2 Y))\times\frac{1}{1+\exp(\alpha_0+\alpha_1 X+\alpha_2 Y)} \\
&= \frac{\exp(W(\alpha_0+\alpha_1 X+\alpha_2 Y))}{1+\exp(\alpha_0+\alpha_1 X+\alpha_2 Y)}
\end{aligned}
$$

The parameter set θ_b contains parameters $\alpha_0,\alpha_1,\alpha_2$, which are intercepts and coefficients of regressors X and Y.

By multiplying two parts together, we obtain the joint PDF of *X*, *Y* and *W*:

$$
\begin{aligned}
& f((X,Y,W)\mid\theta) \\
&= f(X,Y;\theta_c)P(W\mid X,Y;\theta_b) \\
&= \frac{1}{2\pi}\left|\Sigma(\theta_c)\right|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}[Y\ X]\Sigma(\theta_c)^{-1}\begin{bmatrix}Y\\X\end{bmatrix}\right)\times\frac{\exp(W(\alpha_0+\alpha_1 X+\alpha_1 Y))}{1+\exp(\alpha_0+\alpha_1 X+\alpha_1 Y)}
\end{aligned}
$$

Suppose there is a sample (x_1, y_1, w_1), (x_2, y_2, w_2), … , (x_n, y_n, w_n) of *N* iid observations, coming from a population with a joint distribution of (*X*, *Y*, *W*). Therefore, the joint PDF above is true for each independent observation/subject. We can obtain the corresponding likelihood function of this sample by their product:

$$
\begin{aligned}
& L(\theta\mid(X_1,Y_1,W_1),(X_2,Y_2,W_2),...,(X_N,Y_N,W_N)) \\
&= f((X_1,Y_1,W_1),(X_2,Y_2,W_2),...,(X_N,Y_N,W_N)\mid\theta) \\
&= \prod_{i=1}^{N} f((X_i,Y_i,W_i)\mid\theta)
\end{aligned}
$$

where θ is an appropriate set of parameters.

Hence,

$$
\begin{aligned}
& L(\theta\mid(X_1,Y_1,W_1),(X_2,Y_2,W_2),...,(X_N,Y_N,W_N)) \\
&= \prod_{i=1}^{N} f((X_i,Y_i,W_i)\mid\theta) \\
&= \prod_{i=1}^{N}\left\{ f(X_i,Y_i;\theta_c)P(W_i\mid X_i,Y_i;\theta_b)\right\} \\
&= \prod_{i=1}^{N}\left\{\frac{1}{2\pi}\left|\Sigma(\theta_c)\right|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}[Y_i\ X_i]\Sigma(\theta_c)^{-1}\begin{bmatrix}Y_i\\X_i\end{bmatrix}\right)\times\frac{\exp(W_i(\alpha_0+\alpha_1 X_i+\alpha_2 Y_i))}{1+\exp(\alpha_0+\alpha_1 X_i+\alpha_2 Y_i)}\right\}
\end{aligned}
$$

This is the kernel of the likelihood function to maximize. However, it is still cumbersome to differentiate and can be simplified a great deal further by taking its log. Since the logarithm is a monotonic function, any maximum of the likelihood function will also be a maximum of the log likelihood function and vice versa. Thus taking the natural log of the equation above yields the log likelihood function:

$$\begin{aligned}&\ln(L(\theta \mid (X_1,Y_1,W_1),(X_2,Y_2,W_2),\ldots,(X_N,Y_N,W_N)))\\&=\sum_{i=1}^{N}\left(-\ln(2\pi)-\tfrac{1}{2}\ln\left|\Sigma(\theta_c)\right|-\tfrac{1}{2}[Y_i\ X_i]\Sigma(\theta_c)^{-1}\begin{bmatrix}Y_i\\X_i\end{bmatrix}\right)\\&\quad+\sum_{i=1}^{N}[W_i(\alpha_0+\alpha_1X_i+\alpha_2Y_i)-\ln(1+\exp(\alpha_0+\alpha_1X_i+\alpha_2Y_i))]\end{aligned}$$

The log-likelihood function above contains parameters $\beta,\phi,\psi,\alpha_0,\alpha_1,\alpha_2$ to be estimated. There are three unknown parameters in the model implied covariance structure β,ϕ,ψ and we have three known information from the covariance matrix implied by the data: S_{YY},S_{XY} and S_{XX}. Thus the model is just identifiable. Let us call the first part of summation $g(\theta_c)$ and the second one $g(\theta_b)$. Note that two parts of the summation above have no common parameters, containing $\{\theta_c:\beta,\phi,\psi\}$ and $\{\theta_b:\alpha_0,\alpha_1,\alpha_2\}$, respectively.

$$g(\theta_c)=\sum_{i=1}^{N}\left(-\ln(2\pi)-\tfrac{1}{2}\ln\left|\Sigma(\theta_c)\right|-\tfrac{1}{2}[Y_i\ X_i]\Sigma(\theta_c)^{-1}\begin{bmatrix}Y_i\\X_i\end{bmatrix}\right)\text{ and}$$

$$g(\theta_b)=\sum_{i=1}^{N}[W_i(\alpha_0+\alpha_1X_i+\alpha_2Y_i)-\ln(1+\exp(\alpha_0+\alpha_1X_i+\alpha_2Y_i))]$$

where $g(\theta_c)$ can be further reduced as:

$$\begin{aligned}g(\theta_c)&=\sum_{i=1}^{N}\left(-\ln(2\pi)-\tfrac{1}{2}\ln\left|\Sigma(\theta_c)\right|-\tfrac{1}{2}[Y_i\ X_i]\Sigma(\theta_c)^{-1}\begin{bmatrix}Y_i\\X_i\end{bmatrix}\right)\\&=-N\ln(2\pi)-\tfrac{N}{2}\ln\left|\Sigma(\theta_c)\right|-\tfrac{1}{2}\sum_{i=1}^{N}\left\{[Y_i\ X_i]\Sigma(\theta_c)^{-1}\begin{bmatrix}Y_i\\X_i\end{bmatrix}\right\}&&(\because\text{ a scalar equals to its trace})\\&=-N\ln(2\pi)-\tfrac{N}{2}\ln\left|\Sigma(\theta_c)\right|-\tfrac{1}{2}\sum_{i=1}^{N}tr\left\{[Y_i\ X_i]\Sigma(\theta_c)^{-1}\begin{bmatrix}Y_i\\X_i\end{bmatrix}\right\}&&(\because tr(AB)=tr(BA))\\&=-N\ln(2\pi)-\tfrac{N}{2}\ln\left|\Sigma(\theta_c)\right|-\tfrac{1}{2}\sum_{i=1}^{N}tr\left\{\begin{bmatrix}Y_i\\X_i\end{bmatrix}[Y_i\ X_i]\Sigma(\theta_c)^{-1}\right\}\end{aligned}$$

$$
\begin{aligned}
&= -N\ln(2\pi) - \tfrac{N}{2}\ln\left|\Sigma(\theta_c)\right| - \tfrac{N}{2}\sum_{i=1}^{N} tr\left\{\tfrac{1}{N}\begin{bmatrix} Y_i \\ X_i \end{bmatrix}[Y_i\ \ X_i]\Sigma(\theta_c)^{-1}\right\} \\
&= -N\ln(2\pi) - \tfrac{N}{2}\ln\left|\Sigma(\theta_c)\right| - \tfrac{N}{2} tr\left\{\Sigma(\theta_c)^{-1}\sum_{i=1}^{N}\left(\tfrac{1}{N}\begin{bmatrix} Y_i \\ X_i \end{bmatrix}[Y_i\ \ X_i]\right)\right\} \\
&= -N\ln(2\pi) - \tfrac{N}{2}\ln\left|\Sigma(\theta_c)\right| - \tfrac{N}{2} tr\left\{\Sigma(\theta_c)^{-1}S^*\right\} \qquad\qquad S^* = \sum_{i=1}^{N}\left(\tfrac{1}{N}\begin{bmatrix} Y_i \\ X_i \end{bmatrix}[Y_i\ \ X_i]\right)
\end{aligned}
$$

Then we have the sample variance covariance matrix $S = \sum_{i=1}^{N}\left(\tfrac{1}{N-1}\begin{bmatrix} Y_i \\ X_i \end{bmatrix}[Y_i\ \ X_i]\right) = \dfrac{N}{N-1}S^*$, thus $S^* = \dfrac{N-1}{N}S$, finally we reduce the function $g(\theta_c)$ as:

$$
\begin{aligned}
g(\theta_c) &= -N\ln(2\pi) - \tfrac{N}{2}\ln\left|\Sigma(\theta_c)\right| - \tfrac{N}{2} tr\left\{\Sigma(\theta_c)^{-1}\frac{N-1}{N}S\right\}. \\
&= -N\ln(2\pi) - \tfrac{N}{2}\ln\left|\Sigma(\theta_c)\right| - \tfrac{N-1}{2} tr\left\{\Sigma(\theta_c)^{-1}S\right\}
\end{aligned}
$$

Hence, the resulting log likelihood function is reduced as:

$$
\begin{aligned}
&\ln(L(\theta\,|\,(X_1,Y_1,W_1),(X_2,Y_2,W_2),\ldots,(X_N,Y_N,W_N))) \\
&= -N\ln(2\pi) - \tfrac{N}{2}\ln\left|\Sigma(\theta_c)\right| - \tfrac{N-1}{2} tr\left\{\Sigma(\theta_c)^{-1}S\right\} \\
&\quad + \sum_{i=1}^{N}[W_i(\alpha_0+\alpha_1X_i+\alpha_2Y_i) - \ln(1+\exp(\alpha_0+\alpha_1X_i+\alpha_2Y_i))]
\end{aligned}
$$

Note that two parts of the summation above have no common parameters, containing $\{\theta_c:\beta,\phi,\psi\}$ and $\{\theta_b:\alpha_0,\alpha_1,\alpha_2\}$, respectively. Our goal is to find parameters that maximize the log-likelihood function. A necessary condition for finding the maximum of any function, say $f(\theta)$, is to set the first partial derivative of $f(\theta)$ with respect to each θ_i to zero and solve for θ_i. In differentiating the log likelihood function:

$$
\begin{aligned}
&\frac{\partial}{\partial\theta}\ln(L(\theta)) = \frac{\partial}{\partial(\theta_c,\theta_b)}\ln(L(\theta)) \\
&= \frac{\partial}{\partial(\theta_c,\theta_b)}\big(g(\theta_c)+g(\theta_b)\big) = \frac{\partial}{\partial(\theta_c,\theta_b)}g(\theta_c) + \frac{\partial}{\partial(\theta_c,\theta_b)}g(\theta_b) \\
&= \frac{\partial}{\partial\theta_c}g(\theta_c) + \frac{\partial}{\partial\theta_b}g(\theta_b)
\end{aligned}
$$

Since the terms in $g(\theta_c)$ do not depend on θ_b, thus $\frac{\partial}{\partial\theta_b}g(\theta_c)=0$, and $\frac{\partial}{\partial(\theta_c,\theta_b)}g(\theta_c)$ equals to $\frac{\partial}{\partial\theta_c}g(\theta_c)$. Similarly $\frac{\partial}{\partial(\theta_c,\theta_b)}g(\theta_b)$ equals to $\frac{\partial}{\partial\theta_b}g(\theta_b)$. According to the form of this first derivative of log-likelihood, we found interestingly that maximization of this log-likelihood function is equivalent to two separate maximizations for log-likelihood functions of one traditional SEM and one logistic regression. This equivalency largely simplified implementations of our mixed variable SEM. More importantly, it is ready to be extended to more general framework, for instance, the generalized linear model with other link functions accommodating other types of categorical variables.

6.1.2 General form of model estimation

In this section, we derive the general form for our mixed variable SEM. When there are multiple endogenous and exogenous variables with more complicated relations among them on the pathway, we can extend our ML function to the general case.

Assume we have endogenous variable vector ***X*** and exogenous variable vector ***Y*** with a general form: $\mathbf{Y}=\mathbf{BY}+\mathbf{\Gamma X}+\boldsymbol{\zeta}$, and one categorical variable W. Following the notations in the traditional SEM, here $\mathbf{B}$ is a matrix containing path coefficients where the entry $\mathbf{B}_{i,j}$ is the coefficient of the path from endogenous node *j* to endogenous node *i*. $\mathbf{\Gamma}$ is a matrix containing coefficients of paths from exogenous variables to endogenous variables. $\Gamma_{i,j}$ is the coefficient of the path from exogenous node *j* to endogenous node *i*. $\boldsymbol{\zeta}$ is a vector containing the error variables in the equations for the path diagram.

Assume **Y** and **X** are vectors of multivariate normally distributed variables. Let $\mathbf{Z}=\begin{pmatrix}\mathbf{Y}\\ \mathbf{X}\end{pmatrix}$. **Z** has length *p*+*q* (*p* is the number of endogenous variables and *q* is the number of exogenous variables in the model). Variables ***Z*** are centered so that all variables have mean 0. The resulting probability density function for Z is

$$f(Z;\theta_c)=(2\pi)^{-\frac{p+q}{2}}\left|\Sigma(\theta_c)\right|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}Z'\Sigma(\theta_c)^{-1}Z\right)$$

This is the standard form for multivariate normal distribution.

Furthermore we assume, for simplicity, that the categorical variable W conditioned on ***Z*** follows Bernoulli distribution. Then we will model the conditional PMF of dichotomous variable W given ***Z*** through the logit link function:

$$
\begin{aligned}
&\operatorname{logit}(E(W \mid \mathbf{Z})) = \mathbf{A}'\begin{bmatrix} 1 \\ \mathbf{Z} \end{bmatrix} \\
&= \begin{bmatrix} \alpha_0 & \alpha_1 & \dots & \alpha_{p+q} \end{bmatrix}\begin{bmatrix} 1 \\ \mathbf{Z} \end{bmatrix} \\
&= \alpha_0 + \alpha_1 Z_1 + \dots + \alpha_{p+q} Z_{p+q}
\end{aligned}
$$

in which here **A** is a vector containing path coefficients where α_0 represents the intercept and the entry α_j is the coefficient of the path from the node $\mathbf{Z}_j$ to *W*.

Accordingly, we can obtain the conditional PMF of W given Z:

$$
\begin{aligned}
P(W \mid Z;\theta_b)) &= \frac{\exp(W(\mathbf{A}'\begin{bmatrix} 1 \\ \mathbf{Z} \end{bmatrix}))}{1+\exp(\mathbf{A}'\begin{bmatrix} 1 \\ \mathbf{Z} \end{bmatrix})} \\
&= \frac{\exp(W(\alpha_0 + \alpha_1 Z_1 + \dots + \alpha_{p+q} Z_{p+q}))}{1+\exp(\alpha_0 + \alpha_1 Z_1 + \dots + \alpha_{p+q} Z_{p+q})}
\end{aligned}
$$

The resulting joint distribution $f((\mathbf{Z},W) \mid \theta)$ can be expressed by conditional PMF of *W* given **Z** and PDF of **Z**:

$$
\begin{aligned}
&f((Z,W) \mid \theta) \\
&= f(Z;\theta_c)P(W \mid Z;\theta_b) \\
&= (2\pi)^{-\frac{p+q}{2}} \left|\Sigma(\theta_c)\right|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} Z'\Sigma(\theta_c)^{-1} Z\right) \times \frac{\exp(W(\alpha_0 + \alpha_1 Z_1 + \dots + \alpha_{p+q} Z_{p+q}))}{1+\exp(\alpha_0 + \alpha_1 Z_1 + \dots + \alpha_{p+q} Z_{p+q})},
\end{aligned}
$$

where θ_c and θ_b are two subsets of the parameter set θ. As indicated by the subscripts, θ_c contains parameters in the distribution of continuous variables $f(Z)$ and θ_b contains parameters in the conditional distribution of binary variable $P(W \mid Z)$.

Suppose there is a sample ($y_{1,1}$, … , $y_{p,1}$, $x_{1,1}$, …, $x_{q,1}$, w_1), ($y_{1,2}$, … , $y_{p,2}$, $x_{1,2}$, …, $x_{q,2}$, w_2), … , ($y_{1,N}$, … , $y_{p,N}$, $x_{1,N}$, …, $x_{q,N}$, w_N) of *N* iid observations, coming from a population with a joint distribution of (*Z*, *W*).

The corresponding likelihood function of the model is:

$$
\begin{aligned}
&L(\theta \,|\, (Z_1, W_1), (Z_2, W_2), ..., (Z_N, W_N)) \\
&= \prod_{i=1}^{N} f((Z_i, W_i) \,|\, \theta) \\
&= \prod_{i=1}^{N} \left\{ f(Z_i; \theta_c) P(W_i \,|\, Z_i; \theta_b) \right\} \\
&= \prod_{i=1}^{N} \left\{ (2\pi)^{-\frac{p+q}{2}} \left| \Sigma(\theta_c) \right|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} Z_i{}' \Sigma(\theta_c)^{-1} Z_i \right) \times \frac{\exp(W_i \mathbf{A}' \begin{bmatrix} 1 \\ \mathbf{Z}_i \end{bmatrix})}{1 + \exp(\mathbf{A}' \begin{bmatrix} 1 \\ \mathbf{Z}_i \end{bmatrix})} \right\}
\end{aligned}
$$

In accordance, the log likelihood function is:

$$
\begin{aligned}
&\ln(L(\theta \,|\, (Z_1, W_1), (Z_2, W_2), ..., (Z_N, W_N))) \\
&= \sum_{i=1}^{N} \left(-\frac{p+q}{2} \ln(2\pi) - \tfrac{1}{2} \ln \left| \Sigma(\theta_c) \right| - \tfrac{1}{2} Z_i{}' \Sigma(\theta_c)^{-1} Z_i \right) \\
&\quad + \sum_{i=1}^{N} \left(W_i \mathbf{A}' \begin{bmatrix} 1 \\ \mathbf{Z}_i \end{bmatrix} - \ln(1 + \exp(\mathbf{A}' \begin{bmatrix} 1 \\ \mathbf{Z}_i \end{bmatrix})) \right) \\
&= -N \ln(2\pi) - \tfrac{N}{2} \ln \left| \Sigma(\theta_c) \right| - \tfrac{N-1}{2} tr \left\{ \Sigma(\theta_c)^{-1} S \right\} \\
&\quad + \sum_{i=1}^{N} \left(W_i \mathbf{A}' \begin{bmatrix} 1 \\ \mathbf{Z}_i \end{bmatrix} - \ln(1 + \exp(\mathbf{A}' \begin{bmatrix} 1 \\ \mathbf{Z}_i \end{bmatrix})) \right)
\end{aligned}
$$

The last equality is obtained by using a similar transformation from the previous section, where S represents the sample variance-covariance matrix of variables in Z. Parameter estimations can be obtained by maximizing the log-likelihood function above. As shown in the previous simple case, the log-likelihood function of this general form of our mixed variable SEM also features a summation of two components that do not share common parameters. Thus we can draw similar conclusion when maximizing this log-likelihood function: it is equivalent to perform one SEM model and one generalized linear modeling, separately. Because of this property, we can naturally derive our statistical inference and overall model fit based on the knowledge of traditional SEM and GLM.

6.2 Statistical inference of parameters

The estimation of standard errors of path coefficients has the same form for both SEM and MLE. It is described in the book Bollen (1989) and Agresti (2007), respectively. The maximum likelihood estimator, θ, of the parameter vector θ, is distributed as

$$N\left(\theta, \left\{-E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}\right).$$

Standard errors (square root of variances of parameter estimates) of the estimates can be calculated via the asymptotic covariance matrix (inverse of Fisher Information Matrix), with respect to the method through which the parameter estimates are obtained. By definition, the asymptotic covariance matrix of the ML estimator of arbitrary θ is

$$ACOV\left(\theta\right) = \left\{-E\left[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}.$$

This calculation depends on the expected Fisher Information Matrix, and the required partial derivatives of $\log L(\theta)$. Numerical methods for these procedures are described and adopted by software SAS and R.

6.3 Overall model fit

There is no existing model fit measure for our joint mixed variable SEM model. Also overall model fit is not computed in GLLAMM program (Rabe-Hesketh, Skrondal et al. 2003). Here we propose a Chi-square test statistic based on the ML estimates.

In the traditional SEM, the Chi-square test is developed as the overall fit measure to gauge whether $\Sigma = \Sigma(\theta)$ by measuring the departure of Σ from $\Sigma(\theta)$. Being population parameters, Σ and $\Sigma(\theta)$ are not directly observable, and are thus estimated by their sample counterparts S, the usual sample covariance matrix, and $\Sigma(\hat{\theta})$, the estimated implied covariance matrix where $\hat{\theta}$ is the MLE of the model parameters θ.

Similarly, for the generalized linear model, researchers utilized the deviance to test the closeness between the model of interest and the saturated model. The saturated model is defined to have a separate parameter for each observation, and thus provides a perfect fit to the data. The deviance is the likelihood-ratio statistic for the hypothesis that all parameters that are in the saturated model but not in model of interest should equal to zero. Because the saturated model has additional parameters, its maximized log likelihood is at least as large as the maximized log likelihood for the reduced model.

Looking at our mixed variable SEM, the joint null hypothesis to test the overall model fit should be that $\Sigma = \Sigma(\theta_c)$ and θ_b is the reduced parameter vector for the conditional distribution of W with the additional parameters set to zero. The overall model fit is then derived via the likelihood ratio test method. The likelihood ratio test statistic, by definition, is:

$$D = -2\ln \frac{\sup\left\{L(\theta \mid x) : \theta \in \Theta_0\right\}}{\sup\left\{L(\theta \mid x) : \theta \in \Theta\right\}}.$$

Under H_0, we have MLEs for all parameters $\{\hat{\theta} : \hat{\theta}_c, \hat{\theta}_b\}$. Then

$$\ln L_0(\theta) = -N\ln(2\pi) - \tfrac{N}{2}\ln\left|\Sigma(\hat{\theta}_c)\right| - \tfrac{N-1}{2} tr\left\{\Sigma(\hat{\theta}_c)^{-1} S\right\}$$
$$+\sum_{i=1}^{N}\left(W_i \hat{\mathbf{A}}'\begin{bmatrix}1\\ \mathbf{Z}_i\end{bmatrix} - \ln(1+\exp(\hat{\mathbf{A}}'\begin{bmatrix}1\\ \mathbf{Z}_i\end{bmatrix}))\right)$$

is the numerator for the likelihood ratio test.

To form the denominator of the test statistic, we must choose an alternative hypothesis, H_1, for which the value of the corresponding log-likelihood function is at its maximum. If we set Σ to be the sample covariance S and set $\pi(W)$ be the sample proportion $\pi(W=1)$, then $\log L_1$ is at its maximum value as in the SEM and GLM. Then, the likelihood function for H_1, $\log L_1$ is

$$\begin{aligned}\ln L_1(\theta) &= -N\ln(2\pi) - \tfrac{N}{2}\ln|S| - \tfrac{N-1}{2} tr\left\{S^{-1}S\right\} + \sum_{i=1}^{N}\left(W_i \ln\pi + (1-W_i)\ln(1-\pi)\right) \\ &= -N\ln(2\pi) - \tfrac{N}{2}\ln|S| - \tfrac{N-1}{2}(p+q) + N\pi\ln\pi + N(1-\pi)\ln(1-\pi)\end{aligned},$$

in which we make use of

$$f(W_i) = \pi^{W_i}(1-\pi)^{1-W_i},$$

$$f(W_1,...,W_N) = \prod_{i}^{N} \pi^{W_i}(1-\pi)^{1-W_i},$$

and $\ln(f(W_1,...,W_N)) = \sum_{i=1}^{N}\left(W_i \ln\pi + (1-W_i)\ln(1-\pi)\right)$

as well as that $N\pi$ equals to the number of observations having W = 1 and $N(1-\pi)$ equals to the number of observations having W = 0. Here the expression above is the standard of perfect fit to compare H_0 to.

The test statistic is distributed as chi-square when N is large. In this case,

$$\begin{aligned} D &= -2\ln\frac{L_0}{L_1} = -2\ln L_0 + 2\ln L_1 \\ &= N\ln\left|\Sigma(\hat{\theta}_c)\right| + (N-1)tr\left\{\Sigma(\hat{\theta}_c)^{-1}S\right\} - 2\sum_{i=1}^{N}\left(W_i\hat{\mathbf{A}}'\begin{bmatrix}1\\ \mathbf{Z}_i\end{bmatrix} - \ln(1+\exp(\hat{\mathbf{A}}'\begin{bmatrix}1\\ \mathbf{Z}_i\end{bmatrix}))\right) \\ &\quad -N\ln\left|S\right| - (N-1)(p+q) + 2N\pi\ln\pi + 2N(1-\pi)\ln(1-\pi) \end{aligned}$$

If we reorganize the log-likelihood ratio above, we can obtain the expression below:

$$\begin{aligned} D &= -2\ln\frac{L_0}{L_1} \\ &= \left[N\ln\left|\Sigma(\hat{\theta}_c)\right| + (N-1)tr\left\{\Sigma(\hat{\theta}_c)^{-1}S\right\} - N\ln\left|S\right| - (N-1)(p+q)\right] && (1) \\ &\quad + \left[-2\sum_{i=1}^{N}\left(W_i\hat{\mathbf{A}}'\begin{bmatrix}1\\ \mathbf{Z}_i\end{bmatrix} - \ln(1+\exp(\hat{\mathbf{A}}'\begin{bmatrix}1\\ \mathbf{Z}_i\end{bmatrix}))\right) + 2N\pi\ln\pi + 2N(1-\pi)\ln(1-\pi)\right] && (2) \end{aligned}$$

The expressions (1) and (2) are log-likelihood ratios for SEM and GLM, respectively. The corresponding degree of freedom (*df*) equals to the sum of two parts: the number of unique information in the covariance matrix ($\frac{1}{2}(p+q)(p+q+1)$) plus the number of observations (N) and then minus the number of model parameters. The *p* value is calculated as the right-tail probability above the observed test statistic value, based on chi-square distribution. If large test statistic and small *p* value is obtained, it provide strong evidence that the model poorly represents or fits the data.

6.4 Simulation study

6.4.1 Comparison to GLLAMM

The simple model described in section 6.1.1 is simulated by the formula below:

$$
\begin{aligned}
&X \sim N(0,0.5^2),\;\; \varepsilon \sim N(0,0.5^2)\\
&Y = \beta X + \varepsilon \qquad\qquad\qquad\qquad ,\\
&W \sim bin(1,\frac{\exp\left(\alpha_0+\alpha_1 X+\alpha_2 Y\right)}{1+\exp\left(\alpha_0+\alpha_1 X+\alpha_2 Y\right)})
\end{aligned}
$$

where we set $\beta = 2, \alpha_0 = 1, \alpha_1 = -2$ and $\alpha_2 = 4$. We generated the simulated data set with N = 1000. The estimates by our mixed variable SEM and those from GLLAMM are compared in Table 7.

Table 7. Comparison of point estimates between mixed variable SEM and GLLAMM

Parameters		**Mixed variable SEM**		**GLLAMM** set measurement error 0.001		**GLLAMM** set measurement error 0.1	
	True	Estimate (Std Error)	Z value (*p* value)	Estimate (Std Error)	Z value (*p* value)	Estimate (Std Error)	Z value (*p* value)
β	2	1.992 (0.031)	63.28 (<.001)	1.992 (0.031)	63.34 (<.001)	1.993 (0.032)	62.97 (<.001)
α_0	1	0.879 (0.110)	7.99 (<.001)	0.881 (0.110)	7.98 (<.001)	1.817 (0.494)	3.68 (<.001)
α_1	-2	-2.238 (0.492)	-4.55 (<.001)	-2.278 (0.496)	-4.59 (<.001)	-16.143 (4.871)	-3.31 (<.001)
α_2	4	4.092 (0.321)	12.74 (<.001)	4.119 (0.325)	12.68 (<.001)	14.337 (3.893)	3.68 (<.001)

GLLAMM treats endogenous variables as latent variables, thus when there is no multiple indicators for one latent variable, one has to specify a non-zero measurement error in order to make the model identifiable. This assumption is not flexible for simple example as in current case. However, specification of the measurement error would be arbitrary. We tried to specify a small measurement error variance to make these two methods equivalent. We can see point estimates and standard errors between two methods are very similar when measurement error variance is set to be 0.001. Nevertheless, if we set a smaller measurement error variance, like 0.0001, convergence of GLLAMM doesn't achieve; furthermore, if we set larger measurement error variance, like 0.1 as shown in Table 7, GLLAMM generated far deviated estimates compared to the true parameters. In a word, GLLAMM is not robust when there is only one measurement for endogenous variables. In addition, one drawback of GLLAMM, as complained by many users, is highly time-consuming. Even for the simplest pathway in the current example with only three variables, GLLAMM takes several minutes which is over ten times more than our mixed variable SEM.

Chapter 7 Application of mixed variable SEM

7.1 Diet and coronary heart disease model

We implemented our mixed variable model using data from GLLAMM example (Rabe-Hesketh, Pickles et al. 2003; Rabe-Hesketh, Skrondal et al. 2003). The dataset is from Morris, Marr and Clayton (1977). This study investigated the relationship between dietary fiber intake and coronary heart disease (CHD). In the experiment, 333 middle-aged men weighed their food intake over a 7-day period, allowing food constituents to be derived, and were then followed up for CHD. The model will estimate the effect of dietary fiber intake on CHD, considering for occupation as potential factor. The relevant variables are:

CHD: dummy variable for CHD (1: present; 0: absent)

Fiber: dietary fiber intake in the experiment

Bus: dummy variable for whether the man works for London transportation (1: London transportation; 0: bank staff)

Since the fiber intake has a skewed distribution, it has been log-transformed and centered as suggested in the paper. The results from GLLAMM and our model are shown in the table below.

Table 8. Mixed variable SEM and GLLAMM estimates for diet and coronary heart disease example

Parameters		**Mixed variable SEM**		**GLLAMM**	
		Estimate (Std Error)	Z value (*p* value)	Estimate (Std Error)	Z value (*p* value)
(Intercept)	$\boldsymbol{\alpha_0}$	-1.87 (0.24)	-7.64 (< .001)	-1.87 (0.25)	-7.52 (< .001)
bus -> CHD	$\boldsymbol{\alpha_1}$	-0.14 (0.33)	-0.43 (0.66)	-0.20 (0.34)	-0.59 (0.557)
diet -> CHD	$\boldsymbol{\alpha_2}$	-1.63 (0.54)	-2.99 (0.003)	-2.10 (0.72)	-2.92 (0.004)
bus -> diet	$\boldsymbol{\beta}$	-0.12 (0.03)	-3.54 (< .001)	-0.12 (0.03)	-3.54 (< .001)

As illustrated in the Table 8, for this real data analysis, our mixed variable SEM generated very similar estimates and corresponding statistical inferences compared to the GLLAMM results. However, our model is superior to GLLAMM in terms of better computing efficiency, which is also seen in the simulation study.

7.2 Crohn’s disease model

In this section, we will apply our mixed variable SEM on the Crohn’s disease model. We obtained the hypothesis from our gastroenterology expert as shown in the Figure 21. Related genotype, gene expression and changes of specific bacterial abundance in the microbiome are considered to lead to the change of risk of Crohn's disease. We have collected data from 79 observations, among whom 31 subjects are Crohn's disease patients, and others are control subjects.

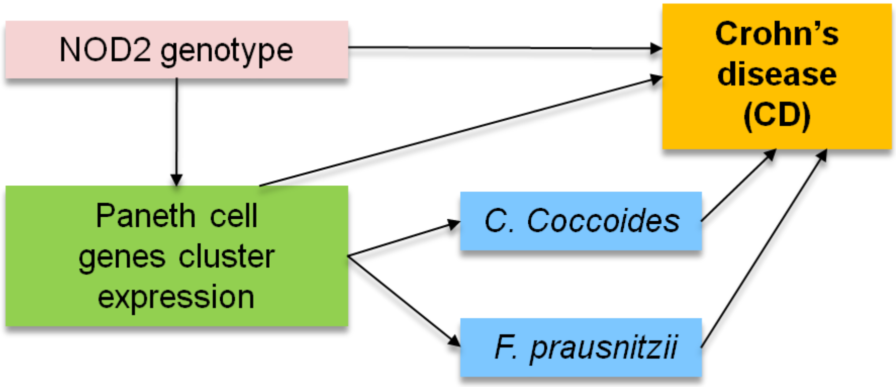


Figure 21. Path diagram for Crohn's disease model using mixed variable SEM.

In Figure 21, the orange box stands for the dummy variable for the presence of Crohn's disease (1: presence; 0: absence). The blue boxes stand for the abundance of two groups of bacteria in the gut which measured by qPCR technique. The changes of gut bacterial quantities are an influential factor to the risk of Crohn's disease. The green box represents the gene expressions related to Paneth cells which are of particular interests related to Crohn's disease. The gene expression values were generated by the Agilent whole human genome microarray. Normalization and pre-processing of the data to filter out undetectable gene-probes resulted in a total of 26,765 gene-probes. 2,979 significant genes were identified as significantly differentially expressed by the threshold of fold change greater than 1.5 and controlled FDR at 0.05. The hierarchical clustering on the 2,979 genes was then performed to further reduce the dimensions of the microarray data to 43 clusters (refer to Zhang et al.(2011) for more details). We found that genes clustered together tend to share the same biological functions. Particularly, we are interested in cluater 24. Cluster 24 is enriched for antimicrobial peptides that are expressed by Paneth cells. These cells are particularly abundant in the ileum and are implicated in host containment of the microbiome (Dieckgraefe, Stenson et al. 2000). Genes encoding Paneth cell signature markers, such as *DEFA5, DEFA6, ITLN1, REG3A, REG3G* had similar expression patterns, and thus were all included in the cluster 24. Therefore, in the current study, median expression values of genes in cluster 24 are used to represent the expression level of Paneth cell related pathway. The pink box represents the number of risk alleles located in the nucleotide oligomerization domain 2 (NOD2) gene. NOD2 encodes an intracellular bacterial sensor that is expressed in Paneth cells, macrophages, and dendritic cells. It likely plays an important role in host containment of gut bacteria in the ileum where Paneth cells are particularly abundant. Risk allele of NOD2 is associated with high risk of ileal Crohn's disease (CD) (Hamm, Reimers et al. 2010). Thus genotype NOD2, associated with the gene expressions of Paneth cell markers, as well as bacterial abundances are studied in the current Crohn's disease model.

On the path diagram, straight arrows demonstrated the directional relations from on variable to the other variable. For instance, Paneth cell gene expressions are supposed to have direct effect on the risk of Crohn's disease and also are hypothesized to indirectly affect through the changes in the abundances of *C. coccoides* and *F. prausnitzii*. Path coefficents on the pathway are estimated and tested the significance as shown in the Table 9.

Table 9. Mixed variable SEM estimates for Crohn's disease example

Path coefficients	**Estimate**	**Std Error**	**Z value**	***p* value**
NOD2 → Paneth cell	0.380	0.134	2.826	**0.005**
Paneth cell → *C. coccoides*	-0.760	0.461	-1.650	0.099
Paneth cell → *F. prausnitzii*	-1.209	0.554	-2.183	**0.029**
NOD2 → CD	1.516	0.645	2.349	**0.019**
Paneth cell → CD	1.788	0.645	2.774	**0.006**
C. coccoides → CD	-0.305	0.123	-2.474	**0.013**
F.prausnitzii → CD	-0.331	0.114	-2.916	**0.004**

Chi-square test statistic of the current model fit is 71.345 with df 77, and the corresponding *p* value is 0.660. Large *p* value of chi-square test suggested a good overall model fit to the observed data. For the seven estimates in the Table 9, we can treat them as two parts of model: the disease model based on the conditional logistic regression; and the structural model among variables except the disease variable. In the disease model, we have path coefficients from NOD2, Paneth cell gene expressions, *C. coccoides* and *F. prausnitzii* to CD. For example, the coefficient 1.788 of Paneth cell gene expression represents its estimated effect on risk of Crohn's disease. The corresponding odds ratio is exp(1.788) = 5.98. The large positive odds ratio indicates that elevated expression level tends to increase the risk of Crohn's disease. In the structural model, the relations among NOD2, Paneth cell, *C. coccoides* and *F. prausnitzii* are also evaluated, where Paneth cell doesn't appear to have an important effect on *C. coccoides* with the estimate of -0.760 and p value 0.099. The significances of path coefficients at level of 0.05 are demonstrated in Figure 22.

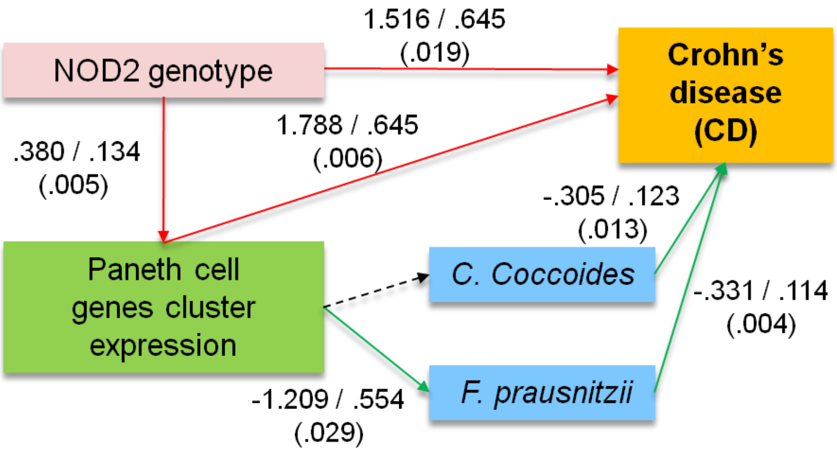


Figure 22. The fitted path diagram for Crohn's disease model by covariate mixed variable SEM. Paths determined to be insignificant are shown as dashed arrows. Arrows and factors highlighted are significant after a two-sided z-test with $\alpha = .05$. Red paths indicate a positive influence while green paths indicate a negative influence.

In Figure 22, red solid lines refer to significantly positive relations; while green solid lines refers to significantly negative relations; and the dashed line indicates the insignificant relation. It appears having NOD2 risk alleles and elevated Paneth cell gene expressions will lead to the increase of chance to have Crohn's disease, where NOD2 risk alleles has direct effect and indirect effect via Paneth cell gene expressions. On the other hand, amplification of two bacterial quantities seems have effect to reduce the chance getting Crohn's disease. Paneth cell gene expression has an alternatively indirect way to affect disease by decreasing quantity of bacterial group *F. prausnitzii*.

7.3 Bootstrap

As traditional SEM, the derivation of our mixed variable SEM is based on the assumption of the multivariate normal distribution among all variables besides the categorical endogenous one. As discussed in the chapter 6, normality of exogenous variable is not that restrictive in terms of estimation and statistical inference, while violation of multivariate normal assumption on those endogenous variables is problematic. In our previous example in Section 7.2, if we examined the normality of three continuous endogenous variables, Paneth cell gene expression, *F. prausnitzii* and *C. coccoides*, the Shapiro-Wilk tests showed *p* values <.001, <.001 and 0.032. It suggested they do not follow normal distribution individually, thus jointly they will not satisfy the assumption of multivariate normality. The corresponding density functions of three variables are shown in Figure 23.

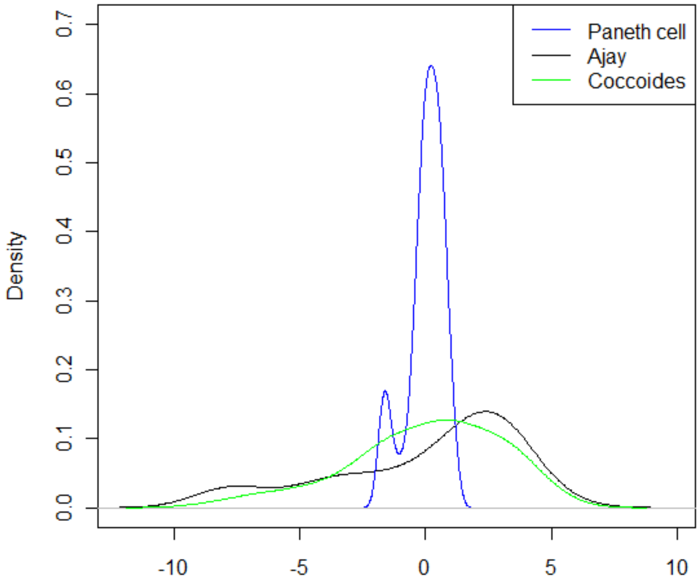


Figure 23. Distribution of three endogenous variables in Crohn's disease model

There are several strategies available when one appears to have nonnormal data. Suggested in Loehlin's book p.59 (2004), first and most obviously, one should check for outliers – extreme cases that represent errors; A second option, if one has some variables that are skewed, is to transform them to a scale that more nearly normal, such as logarithms or square roots of the original scores; A third option is to make use of a bootstrap procedure. The boot strap is to take repeated samples from one's own data, taken as representative of the population distribution, to see how much empirical variation there is in the results. Instead of computing the standard error of a given path coefficient based on multivariate normal distribution, one simply fits the model several hundred times in different re-sampled data from original observations. Then one can derive the confidence intervals of the estimates. With fair sized samples, bootstrapping can provide an attractive way of dealing with non-normal distributions. Bollen and Stine (1993), Yung and Bentler (1996), and Nevitt and Hancock (2001) included discussions of bootstrapping in SEM. Therefore, here we performed bootstrap procedures on Crohn's disease model since we had fairly large dataset. The mixed variable SEM analysis is performed 1000 times by resampling with replacement. The corresponding estimates and 95% confidence intervals are shown in Table 10.

Table 10. Mixed variable SEM estimates based on bootstrapping for Crohn's disease example

Path coefficients	**Estimate**	**95% CI**
NOD2 → Paneth cell	0.380	(0.180, 0.579)
Paneth cell → *C. coccoides*	-0.760	(-1.753, 0.240)
Paneth cell → *F. prausnitzii*	-1.209	(-2.395, -0.136)
NOD2 → CD	1.738	(0.223, 3.761)
Paneth cell → CD	3.448	(0.667, 10.214)
C. coccoides → CD	-0.307	(-0.714, 0.148)
F. prausnitzii → CD	-0.382	(-0.751, -0.107)

If CI contains zero inside, it suggests the null hypothesis about parameter equal zero should not be rejected. In Table 10, the 95% CIs of estimates from bootstrapping technique showed consistency compared to results in Table 9. Zero is only included in 95% CI of coefficient from Paneth cell to *C. coccoides*. Therefore, besides this parameter estimate, the rest estimates can be considered significant with zero outside the CIs.

Chapter 8 Mixed variable covariate SEM

Conventional SEM endeavors to determine the strength of links between nodes that are uniformly continuous variables (Figure 24 A), while in biological studies the pathway link strength could be modulated by other factors. This is especially true for complex diseases where the state of the cell often dictates how genetic variation can affect the final disease phenotype. We proposed to solve this problem by adding a node pointing to links, where the node stands for the potential important covariate to the relations on the pathway. This covariate SEM is based on the previous work from our group (Sharpe 2010), which, however, was limited to continuous variables on the pathway nodes and categorical variables as pathway covariates only (Figure 24 C). As shown in the previous chapters, biological pathways often involve mixed categorical and continuous variables as pathways nodes (Figure 24 B). In addition, it is conceivable that the pathway link strength may be affected by both categorical variables such as phenotypes, genotypes, etc., but also by continuous variables such as age, levels of certain enzymes, etc. In this dissertation, we extended the covariate cSEM method (Sharpe, 2010) for pathway comparisons to allow both continuous and categorical variables as pathway nodes as well as pathway covariates (Figure 24 D). Alternatively the effect of covariate on the link can be viewed as the effect from interaction term between covariate and source variable of the link to the target variable of the link.

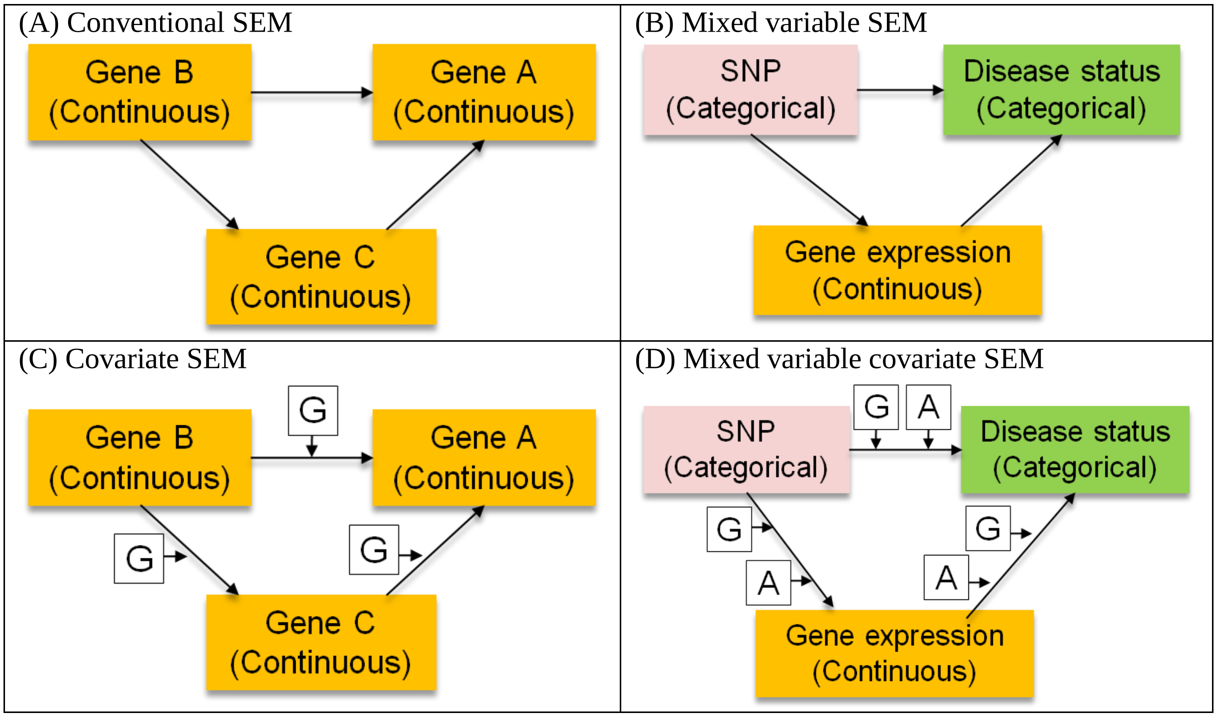


Figure 24. Illustration of structural equation models for different types of biological pathways. (A) Conventional SEM for the pathway with all continuous variables (gene expression values) as nodes. (B) Mixed variable SEM for the pathway with both continuous (gene expression values) and categorical (genotype and phenotype) nodes. (C) Covariate SEM with continuous nodes and categorical covariates (G, such as gender) (Sharpe, 2010). (D) Mixed variable covariate SEM where both the pathway nodes and covariates can be either categorical or continuous variables (G, categorical covariate such as gender; A, continuous covariate such as age).

8.1 Mixed variable SEM with categorical pathway covariate

It is natural thinking to extend previous covariate SEM to a unified mixed variable SEM frame work that incorporates categorical covariates. In this sense, the joint mixed variable SEM is able to handle all scenarios that categorical variables are considered as exogenous, endogenous as well as covariates on the pathway. GLLAMM by Rabe-Hesketh (2004) is not applicable to this complicated case with respect to addition of categorical covariates. However, this situation is ubiquitous and important in applications in biological and medical studies. Because often instead of being assumed a causal relation in the pathway, a categorical (group) variable is considered a covariate, and researchers would like to know model changes between groups. For instance, a disease model is constructed by several risk factors. Disease is the response and risk factors form the structural model among each other. One is interested in how the model differentiates in between covariate gender (male and female). In this case, rather than including gender as an exogenous node in the pathway, it is more reasonable to have it as a covariate pointing to existing paths and then test the additional path significance. Path diagram for a simple example of mixed variable covariate SEM is demonstrated in the Figure 25.

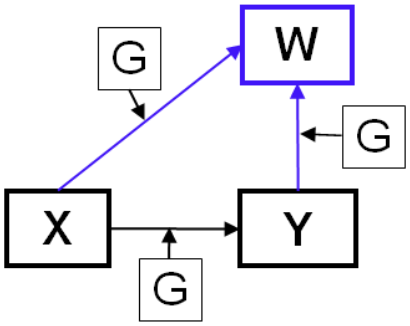


Figure 25. A simple example that demonstrated mixed variable covariate SEM. Here categorical variables W and G are an endogenous variable and a covariate, respectively on the pathway.

The derivation of jointly mixed variable covariate SEM is straightforward since the compatible ML based estimation used in both covariate SEM and mixed variable SEM. For the example in Figure 25, we assumed that W and G are dichotomous variables, W conditioned on X and Y follows Bernoulli distribution, X and Y are bivariate normal distributed under each group of G. By re-parameterization, covariate G can be incorporated into equations as:

$$g(E(W \mid X,Y)) = \alpha_0 + (\alpha_1 + \alpha_2 G)X + (\alpha_3 + \alpha_4 G)Y$$
$$Y = (\beta_1 + \beta_2 G)X + \varepsilon$$

The corresponding log-likelihood function for N observations can be easily derived as:

$$\begin{aligned} &\ln(L(\theta \mid (X_1,Y_1,W_1),(X_2,Y_2,W_2),...,(X_N,Y_N,W_N))) \\ &= -N\ln(2\pi) - \tfrac{N_0}{2}\ln\left|\Sigma(\theta_{c,0})\right| - \tfrac{N_1}{2}\ln\left|\Sigma(\theta_{c,1})\right| - \tfrac{N_0-1}{2} tr\left\{\Sigma(\theta_{c,0})^{-1}S_0\right\} - \tfrac{N_1-1}{2} tr\left\{\Sigma(\theta_{c,1})^{-1}S_1\right\} \\ &+\sum_{i=1}^{N}[W_i(\alpha_0+\alpha_1 X_i+\alpha_2 G_i X_i+\alpha_3 Y_i+\alpha_4 G_i Y_i)-\ln(1+\exp(\alpha_0+\alpha_1 X_i+\alpha_2 G_i X_i+\alpha_3 Y_i+\alpha_4 G_i Y_i)] \end{aligned}$$

Note that subscriptions 0 and 1 in $N_0, N_1, \Sigma(\theta_{c,0}), \Sigma(\theta_{c,1}), S_0, S_1$ indicate parameters in corresponding G = 0 and 1, respectively. N_0, N_1 are number of observations in G = 0 and 1, and $N_0 + N_1 = N$; $\Sigma(\theta_{c,0}), \Sigma(\theta_{c,1})$ are covariance matrixes containing exclusive parameters for G =

0 and 1; and S_0, S_1 are sample covariance matrixes of observations with G =0 and 1, respectively. Other parameters have similar interpretation in Chapter 6. Parameters can then be obtained by maximize the log-likelihood function above. The following section gives an example of mixed variable SEM with one categorical covariate.

8.2 Mixed variable SEM with categorical pathway covariate for Crohn's disease study

The same Crohn's disease data set is considered in this section. Here we consider one additional variable: immunomodulator (e.g. 6-mercaptopurine and azathioprine). Immunomodulator (IM) is a substance which has an effect on immune system, where cause of Crohn's disease might relate to the attack from autoimmune system (Marks 2011). In this example we treated IM (current under treatment or not) as a binary covariate to the original pathway, to test the hypothesis that the pathway differs under different IM status. Out of total 79 subjects, 51 have IM therapy and 28 do not. They are coded as 0 and 1, respectively.

The jointly mixed variable covariate SEM is fitted.

Table 11 showed the resulting estimates and statistical tests. For each directional link on the pathway, there are two coefficients. For example, the path from Paneth cell to *F. prausnitzii*, the path coefficient according to the effect of Paneth cell on *F. prausnitzii* in group IM = 0, and Immunomodulator coefficient is the change to the path coefficient when IM = 1. Paneth cell gene expression has significant negative relation on *F. prausnitzii* quantity with *p* value 0.018, while covariate IM appears to have no significant changes to this relation. The whole picture of the Crohn's disease model including immunomodulator is shown in Figure 26. The fitted pathway clearly shows that for some paths, the IM treatments of subjects do affect the strength of connectivity between nodes.

Table 11. Mixed variable covariate SEM estimates, standard errors and corresponding z values and p values (two-sided) for Crohn's disease example.

		Estimate	Std Error	Z value	*p* value
NOD2 → Paneth cell	Path coefficient	0.324	0.183	1.767	0.077
	Immunomodulator	0.160	0.266	0.603	0.547
Paneth cell → *C. coccoides*	Path coefficient	-0.553	0.508	-1.089	0.276
	immunomodulator	-0.745	0.997	-0.747	0.455
Paneth cell →*F. prausnitzii*	Path coefficient	-1.711	0.725	-2.362	**0.018**
	immunomodulator	1.190	1.098	1.084	0.279
NOD2 → CD	Path coefficient	-0.129	1.302	-0.099	0.921
	immunomodulator	8.648	3.422	2.528	**0.011**
Paneth cell → CD	Path coefficient	7.433	2.279	3.262	**0.001**
	immunomodulator	-4.310	2.822	-1.527	0.127
***C. coccoides* → CD**	Path coefficient	-0.171	0.160	-1.069	0.285
	immunomodulator	-1.606	0.633	-2.536	**0.011**
***F. prausnitzii* → CD**	Path coefficient	0.278	0.238	1.168	0.243
	immunomodulator	-1.814	0.762	-2.380	**0.017**

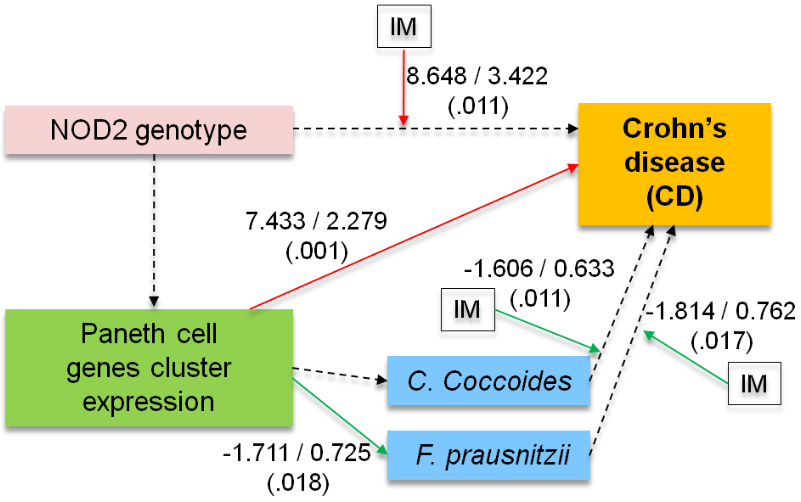


Figure 26. The fitted path diagram for Crohn's disease model by mixed variable covariate SEM. Immunomodulator is included in the pathway as the covariate. Paths determined to be insignificant are shown as dotted arrows. Arrows and factors highlighted are significant after a two-sided z-test with $\alpha = .05$. Red paths indicate a positive influence while green paths indicate a negative influence.

8.3 Mixed variable SEM with continuous pathway covariate

In the previous section, we have already extended covariate SEM (Sharpe 2010; Wu, Sharpe et al. 2011) to accommodate categorical endogenous variables. We illustrated the extended model in Crohn's disease mixed-variable example and successfully detected significant covariate effects on the strength of paths. However in real practice, covariates cannot always be treated as categorical as gender, race or immunomodulator. Taking Crohn's disease model for example, we have continuous covariates, like age and BMI. These covariates are not major focus of the study, but their effects on the disease model are potential concerns and of interests to biologists and medical doctors. In this case, current covariate SEM model by Dr. Sharpe (2010) is not applicable since its likelihood function is derived based on combination of likelihood functions that covariance structures are computed separately from each group of categorical covariates. When covariates are continuous, there is no straightforward separation of covariance structures of variables given the covariate. This motivated us to further generalize our mixed variable covariate SEM to incorporate continuous covariates and solve this complex model.

Look at the demonstration in Figure 27(A). In contrast to categorical covariate G in Figure 25, covariate C here has continuous values. Covariate C is hypothesized to have effect on strengths of paths.

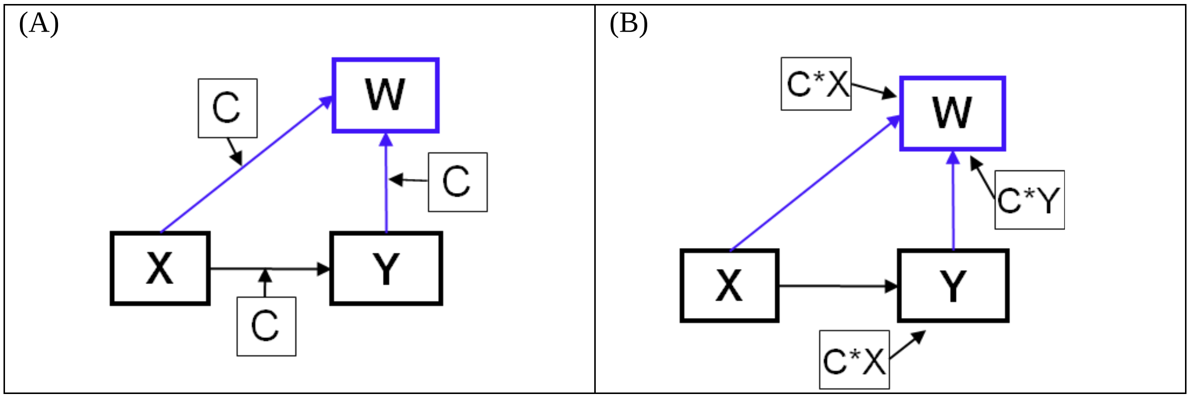


Figure 27. Mixed variable SEM with continuous covariate C. (A) On the pathway, endogenous variable W is categorical and covariate C is continuous. (B) An alternative illustration of mixed variable covariate SEM when covariate C is continuous. Here covariate C is incorporated, for each path, by adding a node to point to the corresponding endogenous variable. The added node can be viewed the interaction between covariate and exogenous variable.

The equations implied by the path diagram in Figure 27(A) are:

$$\begin{aligned}&\text{logit}(E(W \mid X, Y)) = \alpha_0 + \alpha_X X + \alpha_Y Y\\ &Y = \beta X + \varepsilon\end{aligned}$$

where, we also applied the idea of re-parameterization to re-write equations as:

$$\begin{aligned}&\text{logit}(E(W \mid X, Y))\\ &= \alpha_0 + (\alpha_1 + \alpha_3 C)X + (\alpha_2 + \alpha_4 C)Y\\ &= \alpha_0 + \alpha_1 X + \alpha_3 CX + \alpha_2 Y + \alpha_4 CY\end{aligned}$$

$$Y = (\beta_1 + \beta_2 C)X + \varepsilon = \beta_1 X + \beta_2 CX + \varepsilon$$

We interpret the coefficient, for each path, is consist of effect from the corresponding exogenous variable and from interaction between the covariate and the exogenous variable. Alternatively, it can be interpreted as, for each path, adding one node of interaction between the covariate and the exogenous variable to point to the corresponding endogenous variable (Figure 27 B).

Hence, in general, evaluation of continuous covariates on the paths is in accordance to adding several interaction terms as exogenous nodes to the model. The number of new nodes added to the model equals to the number of links on the pathway. Recall the derivation in Section 6.1, mixed variable SEM can be solved by factorization of two components: one conditional GLM model and one SEM model. For the conditional GLM, considering covariate effect is equivalent to adding interaction terms between covariate and original predictors to the equation. For SEM of all continuous variables, considering the case with p endogenous variables **Y** and q exogenous variable **X**. Assume, among p endogenous variables Y, p_{ex} variables are also as independent variables that have arrows out. To solve this model including one continuous variable C, we created a new exogenous variable vector $X^{*} = (X_1, \ldots, X_q, CX_1, \ldots CX_q, CY_1, \ldots, CY_{pex})'$, where we assumed that $Y_1, \ldots, Y_{pex}$ are the p_{ex} variables that have arrows out. Then we are ready to obtain the likelihood function of mixed variable SEM with continuous covariate by replace X with X^{*}.

Assume $Z^{*} = (Y\ X^{*})'$. For a sample of N *iid* observations coming from a joint distribution of $(Y_1,...,Y_p,X_1^{*},...,X_q^{*},W)$, the corresponding likelihood function of the model is:

$$
\begin{aligned}
&L(\theta \mid (Z_1^{*},W_1),(Z_2^{*},W_2),...,(Z_N^{*},W_N)) \\
&=\prod_{i=1}^{N} f((Z_i^{*},W_i)\mid \theta) \\
&=\prod_{i=1}^{N}\left\{ f(Z_i^{*};\theta_c)P(W_1 \mid Z_i^{*};\theta_b)\right\} \\
&=\prod_{i=1}^{N}\left\{ (2\pi)^{-\frac{p+2q+p_{ex}}{2}} \left|\Sigma(\theta_c)\right|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} Z_i^{*}\,'\Sigma(\theta_c)^{-1} Z_i^{*}\right) \times \frac{\exp(W_i\mathbf{A}'\begin{bmatrix}1\\ \mathbf{Z}_i^{*}\end{bmatrix})}{1+\exp(\mathbf{A}'\begin{bmatrix}1\\ \mathbf{Z}_i^{*}\end{bmatrix})} \right\}
\end{aligned}
$$

8.4 Mixed variable SEM with continuous pathway covariate for Crohn’s disease study

As an example, age of surgery, a continuous covariate, is incorporated into the Crohn’s disease model in Section 7.2. The mixed variable SEM with this continuous covariate is fitted and results are shown in Table 12 and Figure 28. The overall model fit chi-square is 94. 3 with degree freedom 79, and corresponding p value is 0.11 that indicating a good fit.

Table 12. Continuous-covariate mixed variable SEM estimates, standard errors and corresponding *z* values and *p* values (two-sided) for Crohn's disease example.

		Estimate	Std Error	Z value	*p* value
NOD2 → Paneth cell	Path coefficient	0.397	0.153	2.597	**0.009**
	Age of surgery	0.029	0.121	0.237	0.813
Paneth cell → *C. coccoides*	Path coefficient	-0.682	0.485	-1.407	0.159
	Age of surgery	0.302	0.597	0.506	0.613
Paneth cell →*F. prausnitzii*	Path coefficient	-1.075	0.582	-1.847	0.065
	Age of surgery	0.523	0.717	0.730	0.465
NOD2 → CD	Path coefficient	1.986	0.836	2.376	**0.017**
	Age of surgery	0.214	0.688	0.312	0.755
Paneth cell → CD	Path coefficient	2.683	0.970	2.767	**0.006**
	Age of surgery	-3.544	1.397	-2.538	**0.011**
***C. Coccoides* → CD**	Path coefficient	-0.314	0.157	-1.997	**0.046**
	Age of surgery	0.008	0.156	0.049	0.961
***F. prausnitzii* → CD**	Path coefficient	-0.459	0.160	-2.877	**0.004**
	Age of surgery	0.117	0.167	0.702	0.482

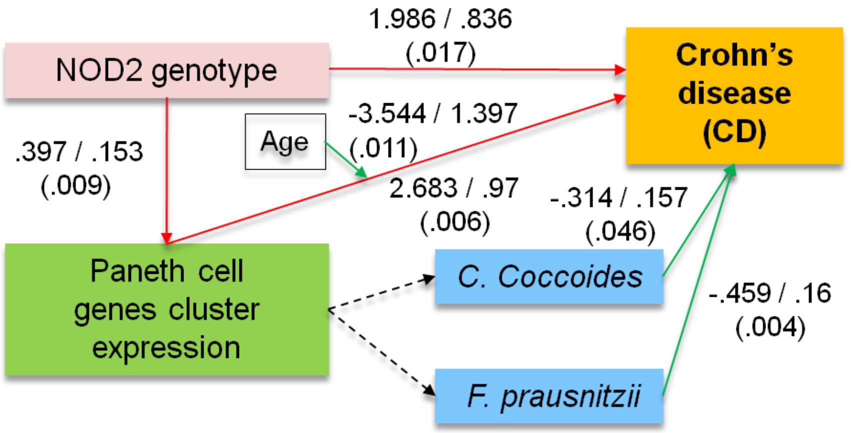


Figure 28. The fitted path diagram for Crohn's disease model by continuous-covariate mixed variable SEM. Paths determined to be insignificant are shown as dotted arrows. Arrows

and factors highlighted are significant after a two-sided z-test with $\alpha = .05$. Red paths indicate a positive influence while green paths indicate a negative influence.

We can see that one age of surgery has a negative association to the effect from Paneth cell related gene expressions to CD. The strength of increase of risk of CD by Paneth cell related gene expressions may be attenuated by older age of surgery. On the other hand, we noticed in our data set, control subjects, usually cancer patients but without Crohn's disease, are older than CD patients. Therefore, this significant covariate effect of age here might be due to the features of sampled data, thus it is suggestive and need to be further examined by larger data sets.

Part II Novel Bioinformatics Pipeline and Application

The second part of this thesis is based on my research experience in Dr. van der Lelie and Dr. Taghavi's lab in Brookhaven national laboratory. In contrast to the first part, the novel biological pathway analysis by structural equation modeling, this part focuses on a bioinformatics work flow for a systems biology project.

Biological systems such as cells, regulatory gene networks and protein interaction complexes cannot be understood from individual components (genes, mRNA, proteins etc) alone, but through considerations involving all components simultaneously. The use of systematic genomic, transcriptomic, metabolic and proteomic technologies to construct models of complex biological systems and diseases is becoming increasingly commonplace (Ideker 2004). Although the concept of systems biology has existed for a while, these approaches have recently become far more powerful because of the inventions of new technologies that are high-throughput, quantitative and large-scale (Zhu and Snyder 2002). In so doing, systems biology integrates data and knowledge from diverse biological components into models of the system as a whole. The biological knowledge is growing very rapidly. A list of existing biological knowledge-bases is summarized here (http://www.biochemweb.org/systems.shtml). In the bioinformatics area, tools have been developed and will be advanced to handle the rapidly growing amount of data in databases. Integration of the latest knowledge and tools, in a sense, becomes the most critical aspect in systems biology. This suggests the exciting but also challenging feature of systems biology: the requirement of multiple-disciplinary abilities in biology, statistics and computation, etc. As Dr. Ideker (2004) mentioned "students should beware of the universal curse of systems biology: as you quickly attain a breadth of knowledge in biology and mathematics, you risk losing, or fail to attain, depth in either. Jack of all trades, or master of one? The choice is yours." In the following chapters, a study of interactions between poplar and endophytic bacteria, with major efforts stitching a complete story via integration of biological knowledge, statistical and computational approaches, is presented.

Endophytic bacteria are bacteria that reside within the living tissue of their host plants without substantively harming it (Misaghi and Donndelinger 1990). They are ubiquitous in most plant species, residing or actively colonizing the tissues. The diversity of cultivable bacterial endophytes is exhibited not only in the variety of plant species colonized but also in the many taxa involved, with most being members of common soil bacterial genera such as *Enterobacter*, *Pseudomonas*, *Burkholderia*, *Bacillus*, and *Azospirillum* (Lodewyckx, J. Vangronsveld et al. 2002). There are several mechanisms by which endophytic bacteria can promote plant growth. These mechanisms are of great importance for the use of plants as feedstocks for biofuels and for carbon sequestration through biomass production. Moreover, this is vital when considering the aim of improving biomass production of marginal soils, thus avoiding competition for agricultural resources, which is one of the critical socioeconomic issues of the increased use of biofuels (Taghavi, Garafola et al. 2009). The aim of this study is to identify mechanisms by which the endophytic bacteria can regulate the growth of poplar, to quantify the degree to which major regulatory pathways of endophytic bacteria are involved, and the ultimate goal is to model these pathways to optimize the production of poplar biomass on marginal soils as a feedstock for bio-refineries. The objective of the proposed research is to characterize endophytic bacteria at the level of the genome, transcriptome and metabolome. It can be achieved by using a systems biology approach, including comparative genome analysis and mRNA sequencing analysis.

Figure 29, as seen in Chapter 1, demonstrated the analysis pipeline, and details will be given in the next chapter.

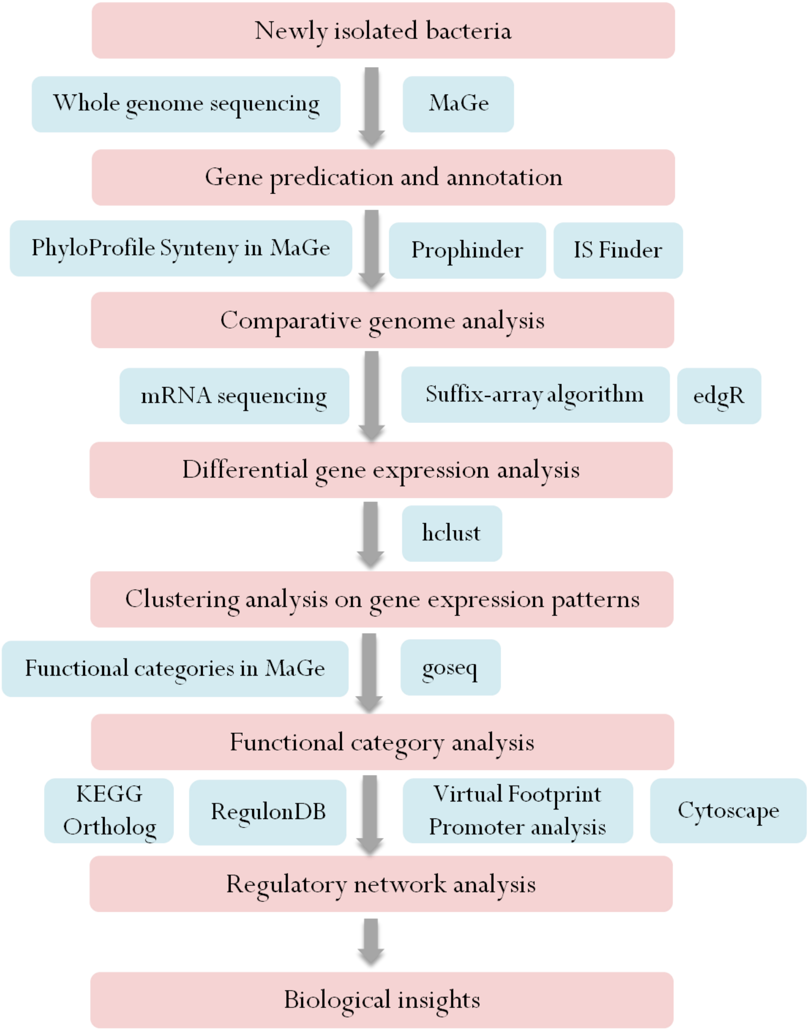

Newly isolated bacteria
Whole genome sequencing
MaGe
Gene predication and annotation
PhyloProfile Synteny in MaGe
Prophinder
IS Finder
Comparative genome analysis
mRNA sequencing
Suffix-array algorithm
edgR
Differential gene expression analysis
hclust
Clustering analysis on gene expression patterns
Functional categories in MaGe
goseq
Functional category analysis
KEGG Ortholog
RegulonDB
Virtual Footprint Promoter analysis
Cytoscape
Regulatory network analysis
Biological insights

Figure 29. Overview of bioinformatics analyses work flow from newly isolated bacteria to biology discoveries. The steps in the pipeline are in red boxes; the methodological components and software used are in blue boxes.

Chapter 9 Analysis workflow

9.1 Comparative genome analysis

9.1.1 Genome sequencing and annotation

Total DNA was isolated from bacteria as described according to the method of Bron and Venema (Bron and Venema 1972). Genome sequencings of bacteria were performed at the Joint Genome Institute (JGI) (Walnut Creek, California, USA). Putative CoDing Sequences (CDS) were identified by the Magnifying Genome (MaGe) annotation platform (http://www.genoscope.cns.fr/agc/mage/) (Vallenet, Labarre et al. 2006).

A number of annotation tools have been designed for an increasing demand for fast and accurate analysis of completely sequenced genomes. Many efforts have been made in terms of project management (i.e. complex biological data models and integrated databases), spectrum of bioinformatics tools applied (including multiple genome comparison-based annotation strategies), sophistication of the user interfaces (extensive visualizations, fully interactive graphical interfaces) and the presence of convenient features such as data editors. Examples of commonly used annotation platforms are given by commercial systems, such as ERGO (Overbeek, Larsen et al. 2003) or Pedant-Pro (successor of PEDANT), and open source systems, such as Artemis (Berriman and Rutherford 2003), GenDB (Meyer, Goesmann et al. 2003). In the study of microbial genomes, compared to these methods, MaGe is featured by the systematically integrated contextual analysis of genes and proteins, to detect functional constraints on genome evolution (Vallenet, Labarre et al. 2006; Vallenet, Engelen et al. 2009). Using MaGe platform, we performed automatic and manual genome annotations, as well as comparative genomics and functional analysis altogether.

On genomes of our bacteria, in MaGe system, all CDS identified were manually reviewed, and false CDS were flagged as “artifact”. The remaining CDS were then submitted to automatic functional annotation via BLAST searches against the UniProt databank in order to determine significant homology. Putative genes coding for enzymes were classified with the PRIAM software (Claudel-Renard, Chevalet et al. 2003), transmembrane domains were identified by TransMembrane Hidden Markov Model (TMHMM), and signal peptide were predicted using SignalP 3.0, all embedded in the MaGe software (Vallenet, Labarre et al. 2006).

9.1.2 Comparative genomics and functional analysis

Genome comparisons were performed using MaGe (Vallenet, Labarre et al. 2006). The "PhyloProfile Synteny" program was used to show the number of homologous genes in related bacteria. Genomic Islands (GI) were identified using the automated "Genomic Islands" tool, followed by a manual curation focusing on several GI properties (Mergeay, Monchy et al. 2009). These properties include the presence at one extremity of a site-specific recombinase, the preferential insertion of GI at tRNA sites, the presence of flanking insertion sequence elements, a base composition and/or phylogeny which differs from the bulk of the genome, a higher content in hypothetical genes than the neighboring regions, the presence of hot spots for mobile genetic elements (MGEs) including recombinase genes, IS elements, integrase and transposase genes, and the conservation of GI between different unrelated hosts together with their absence in related hosts. A region was considered as a genomic island if at least three criteria were met. Metabolic reconstructions were performed using both the PRIAM software, which is based on the KEGG database, and the MetaCyc/EcoCyc tools embedded into the MaGe platform. The identification of prophages was done using "Prophinder" (Lima-Mendez, Van Helden et al. 2008) (http://www.aclame.ulb.ac.be/Tools/Prophinder/). IS Finder (http://www-is.biotoul.fr/) was used for the classification into families of the identified IS elements.

9.2 mRNA sequencing (RNA-seq) analysis

9.2.1 RNA preparation, sequencing and reads mapping

Endophytic bacterium *Enterobacter sp.* 638 was grown at 30°C in minimal growth medium with sucrose or lactate as the sole carbon source to mimic conditions of plant association or free living, respectively. Total RNA was isolated and processed to remove rRNAs. The purified mRNAs were then transferred into cDNA, and sent to Cold Spring Harbor Laboratory for whole transcriptome sequencing via Illumina next generation sequencing.

There are three major platforms for next generation sequencing experiments: 454, Illumina and SOLiD. 454 sequencing is featured with high time-efficiency and longer length of each read; and SOLiD has advantages of higher throughput per run; and Illumina has reasonably good time efficiency and relatively low unit cost. Dr. Mardis (2008) provided a good review of these next generation sequencing techniques. Our experiment used the Illumina Genome Analyzer given its high throughput and low cost, as well as the convenient service close by.

Libraries were prepared for sequencing according to the manufacturer's instructions. 76 base-pair-long single-ended reads were obtained. For higher quality score, the trimmed reads of length 36 starting from the 5th position were used for subsequent mapping. The *k*-mer uniqueness for the genome of *Enterobacter sp*. 638 is shown in Figure 30. The uniqueness curve approaches 100% rapidly, hitting 99.7% at k = 30, which encourages read mapping using shorter subsequences. The obtained 36-bp reads were mapped to the *Enterobacter sp.* 638 genome by suffix-array lookup program, which allows for one substitution error.

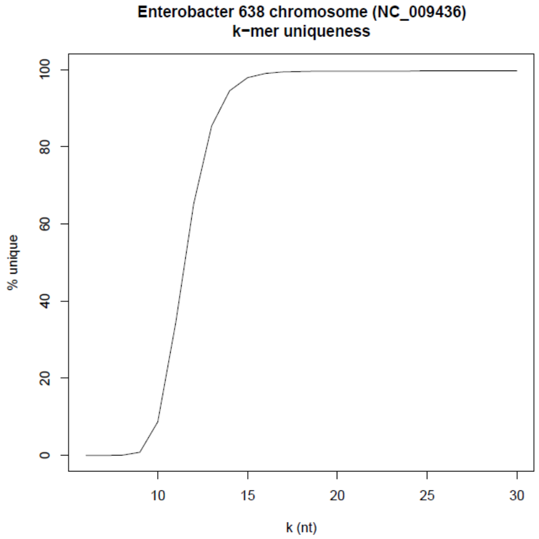


Figure 30. The *k*-mer uniqueness plot for *Enterobacter sp.* 638 chromosome.

9.2.2 Pre-processing of gene expression measures

Raw counts as expression level are summarized on the gene level. The number of reads falling into the boundary of one gene is recorded as the raw expression count. The *Enterobacter sp.* 638 reference annotations are requested from the MaGe manual annotation system. There are high spearman correlations between the replicates (all above 0.8), with majority of them over 0.9. Thus, we pooled the counts that from the same biological sample to increase the coverage of the transcriptome.

In order to derive gene expression level and compare values between conditions, one first needs to normalize read counts to adjust for varying lane sequencing depths and potentially other technical effects. It has been shown that normalization is an essential step in the analysis of RNA-seq data (Anders and Huber 2010). Two types of normalizations are necessary: between- and within-library normalizations.

Within-library normalization allows quantification and comparisons between genes in one sample. Because longer transcripts have more read counts falling into (at the same expression level), a common method for within-library normalization is to divide the summarized counts by the length of the gene (Marioni, Mason et al. 2008). The widely used RPKM (reads per kilobase of exon model per million mapped reads) accounts for both total read counts and gene length in within condition comparisons (Mortazavi, Williams et al. 2008). Some recent methods may improve the comparability within-sample by assuming non-uniform distributed reads inside gene boundary (Li, Jiang et al. 2010). However, performance of this method depends case by case. Here, we adopted the common and straightforward RPKM index for comparison between genes.

To compare expression changes of certain gene between samples, gene length bias will cancel out because the underlying sequence used for summarization is the same between samples. But between-sample normalization is still critical for comparing counts from different libraries relative to each other. Several between-sample normalization methods, including total reads, housekeeping gene normalizations, have been evaluated and concluded that the quantile normalization, inspired from microarray study (Bolstad, Irizarry et al. 2003; Irizarry, Hobbs et al. 2003) has the best performance (Bullard, Purdom et al. 2010). In current study, we applied quantile normalization to the pooled count data. The normalized data are then rounded to produce integer values as genes expression level for further differential expression analysis.

9.2.3 Differentially expressed gene analysis

The goal of a differentially expressed gene analysis is to highlight genes that have changed significantly across experimental conditions. In general, this means taking a table of normalized count data for each condition and performing statistical testing between samples of interest.

Many methods have been developed for the analysis of differential expression using microarray data, such as SAM (Tusher, Tibshirani et al. 2001). Microarray intensities are typically log-transformed and analyzed as normally distributed random variables. However, RNA-seq provides a discrete measurement for each gene. Even log-transformed, measures are not well approximated by continuous distributions, especially in the lower count and for small samples. Therefore, statistical models appropriate for count data are vital for data mining from RNA-seq experiment.

In general, the Poisson distribution forms the basis for modeling count data. In an early RNA-seq study using an Illumina GA sequencer, goodness-of-fit statistics suggested that the distribution of counts across lanes for majority of genes was indeed Poisson distributed (Marioni, Mason et al. 2008). However, biological variability is not well captured by the Poisson assumption (Robinson and Smyth 2007). Hence, Poisson-based analyses for datasets with biological replicates tend to have high false positive rates resulting from the underestimation of sampling error. To account for biological variability, the negative binomial distribution has been used as a natural extension of the Poisson distribution, requiring an additional dispersion parameter to be estimated. A few variations of negative-binomial-based analysis of count data have emerged, including common dispersion models (Robinson and Smyth 2008), sharing information over all genes using weighted likelihood (Robinson and Smyth 2007), and an empirical Bayesian implementation using equivalence classes (Hardcastle and Kelly 2010). Negative-binomial-based analysis is implemented in R package edgeR (Robinson, McCarthy et al. 2010). Here we identified differentially expressed genes using edgeR package, where the exact test that has strong parallels with Fisher's exact test, is used to test differential expressions and compute exact *p* values. The significance level is controlled by false discovery rate at 0.05 (Benjamini and Hochberg 1995).

9.2.4 Clustering and functional category analysis (GO analysis)

In many cases, a list of differential expressed genes is not the final step of the analysis; further biological insight can be gained by looking at the expression changes of sets of genes. In the current experimental design, four biological conditions are tested and thus clustering analysis is first performed to group genes with similar expression pattern across four conditions. Then hierarchical clustering was carried out with distance as (1 – spearman correlation)/2. Therefore, the distances range from 0 to 1 with smaller values indicating similar expression patterns.

Having each cluster with a distinct expression pattern leads to the next step -- functional category analysis – to identify enriched functions associated with specific expression patterns. Many tools focusing on gene set testing and knowledge databases have been proposed for microarray datasets as we introduced in chapter 2. However, RNA-seq is affected by biases not present in microarray data. For example, gene length bias is an issue in RNA-seq data, in which longer genes have higher counts (at the same expression level) (Oshlack and Wakefield 2009). These biases can dramatically affect the results of downstream analyses, such as testing Gene Ontology (GO) terms for enrichment among significant genes (Oshlack and Wakefield 2009). Bullard et al. (2010) suggested modifying a t-statistic by dividing by the square root of gene length to minimize the effect of length bias on tests. Alternatively, GO-seq is an approach developed specifically for RNA-seq data that can incorporate length or total count bias into gene set tests (Young, Wakefield et al. 2010). In the present study, GO-seq package 1.0.3 was used. Instead of GO terms that only well specified for model organisms, the manually curated functional categories in MaGe, bioprocess and biological roles, are used for our newly isolated bacteria.

9.3 Regulatory network analysis

There is wide scope for integrating the results of RNA-seq data with other sources of biological data to establish a more complete picture of gene regulation (Hawkins, Hon et al. 2010). For example, integration of expression data with genotype, transcription factor binding, RNA interference, histone modification and DNA methylation information has the potential for greater understanding of a variety of regulatory mechanisms (Montgomery, Sammeth et al. 2010). A few reports of these 'integrative' analyses have emerged recently (Ouyang, Zhou et al. 2009). Although our current experiment did not generate these additional types of biological data, some efforts were made to better understand regulatory networks as a basis of observed transcriptional changes.

The genome of *Enterobacter sp.* 638 is very close related to *Escherichia coli* K12. *E. coli* K12 is the best known annotated model organism for bacteria. Therefore, we proposed to first map orthologs from *E. coli* K12 to *Enterobacter sp.* 638 using KEGG ortholog database within MATLAB (http://www.genome.jp/kegg/soap/doc/keggapi_manual.html), and then infer regulatory relationships in *Enterobacter sp.* 638. The database RegulonDB ((http://regulondb.ccg.unam.mx/html/Database_summary.jsp)) recodes the most comprehensive and updated transcriptional network for *E. coli* K12. We resorted to regulatory networks of *E. coli* orthologs to give hints of ones in *Enterobacter sp.* 638. The resulting regulatory networks are customized and visualized in Cytoscape tool (Smoot, Ono et al. 2011).

Chapter 10 Results and biological discoveries

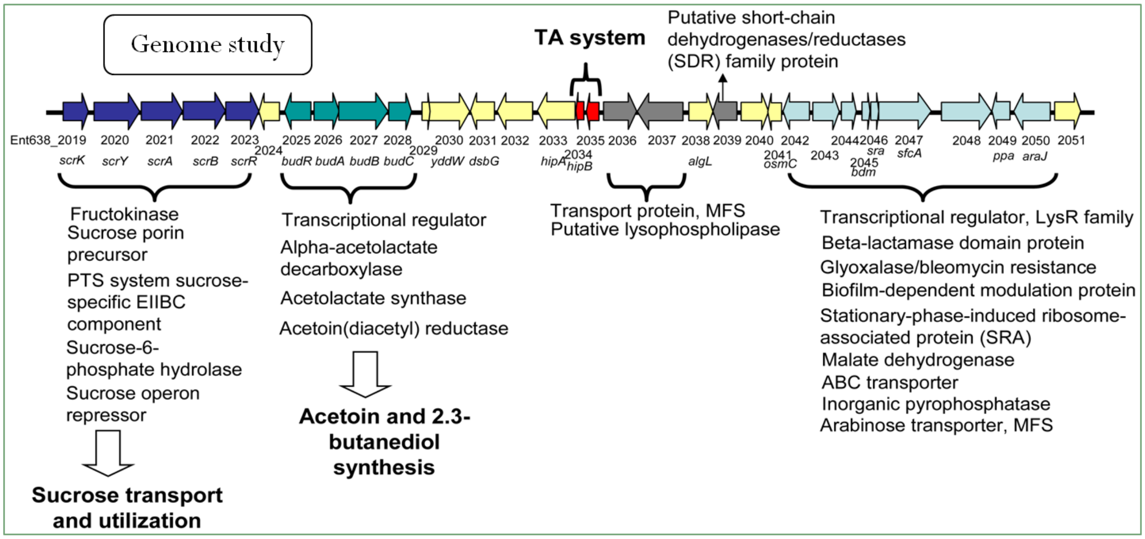
10. 1 Genome annotation and comparative genome analysis

As representatives for the dominant genera of endophytic gammaproteobacteria, we selected *Enterobacter sp.* 638, *Stenotrophomonas maltophilia* R551-3, *Pseudomonas putida* W619, and *Serratia proteamaculans* 568 for genome sequencing and analysis of their plant growth-promoting effects (Taghavi, Garafola et al. 2009). Gene sequencing and annotation allows the identification of the whole set of genes on the genome. Gene homologs were found in these endophytic bacteria that are putatively involved in phytohormone production and metabolisms of plant sugars and growth-regulating compounds, such as acetoin and 2, 3-butanedial synthesis, ACC metabolism and PTS sugar uptake systems. Among them, further *in silico* comparative genome analysis was applied to the genomes of *Enterobacter sp.* 638 (Taghavi, van der Lelie et al. 2010) and *Psedomonas putida* W619 (Wu, Monchy et al. 2011). Comparative genomics provided a powerful tool to gain new insights into the niche-specific adaption of bacteria. *Pseudomonas putida* W619 was compared to three other *P. putida* strains: the rhizospheric strain KT2440, the aromatic hydrocarbon-degrading strain F1 and the manganese-oxidizing strain GB-1. Many genes were suggested to be related to their adaptation to specific niches, including the ability to live in soils and sediments contaminated with high concentrations of heavy metals and organic contaminants (Wu, Monchy et al. 2011).

In addition to the finding of genes encoded for production of phytohormones acetoin and 2, 3-butanedial, in *Enterobacter sp.* 638, metabolite analysis as well as quantitative RT–PCR showed that, the production of acetoin and 2,3-butanediol is induced by the presence of sucrose in the growth medium. Interestingly, both the genetic determinants required for sucrose metabolism and the synthesis of acetoin and 2,3-butanediol are clustered on a genomic island. These findings point to a close interaction between *Enterobacter sp.* 638 and its poplar host, where the availability of sucrose, a major plant sugar, affects the synthesis of plant growth promoting phytohormones by the endophytic bacterium (Taghavi, van der Lelie et al. 2010). This interaction provided us an entry point to better understand the synergistic interactions between poplar and its growth promoting endophyte *Enterobacter sp.* 638. Subsequent transcriptome analysis was thus performed in terms of presence or absence of the plant sugar sucrose. My work in this project includes genome annotations and comparative genome analysis of *Psedomonas putida* W619, and RNA-seq data analyses of *Enterobacter sp.* 638 in the following sections.

10. 2 Experimental design and physiological observations

As a result, analysis of genome sequences pointed to a remarkable interaction between *Enterobacter sp.* 638 and its poplar host (Taghavi, van der Lelie et al. 2010). Particularly it showed the adjacency of two functional operons: sucrose utilization operon (*scrKYAB*) and acetoin / 2,3-butanediol synthesis operon (*budABC*) on the *Enterobacter sp.* 638 genome (Figure 31). It is possible that these two operons interact and play an important role in the crosstalk between the *Enterobacter* sp. 638 and its plant host. The presence of sucrose -- the major photosynthate -- is a signal of proximity with plants to bacteria, which was hypothesized to trigger the transcription of the *budABC* operon in *Enterobacter* sp. 638, resulting in the synthesis of the phytohormones acetoin and 2,3-butanediol. It is a convincing mechanism proved from the genomic, transcriptional and metabolic analyses (Taghavi, van der Lelie et al. 2010).



Genome study
TA system
Putative short-chain dehydrogenases/reductases (SDR) family protein
Ent638_2019 2020 2021 2022 2023 2024 2025 2026 2027 2028 2029 2030 2031 2032 2033 2034 2035 2036 2037 2038 2039 2040 2041 2042 2043 2044 2045 2046 2047 2048 2049 2050 2051
scrK scrY scrA scrB scrR budR budA budB budC yddW dsbG hipA hipB algL osmC bdm sra sfcA ppa araJ
Fructokinase
Sucrose porin precursor
PTS system sucrose-specific EIIBC component
Sucrose-6-phosphate hydrolase
Sucrose operon repressor
Sucrose transport and utilization
Transcriptional regulator
Alpha-acetolactate decarboxylase
Acetolactate synthase
Acetoin(diacetyl) reductase
Acetoin and 2.3-butanediol synthesis
Transport protein, MFS
Putative lysophospholipase
Transcriptional regulator, LysR family
Beta-lactamase domain protein
Glyoxalase/bleomycin resistance
Biofilm-dependent modulation protein
Stationary-phase-induced ribosome-associated protein (SRA)
Malate dehydrogenase
ABC transporter
Inorganic pyrophosphatase
Arabinose transporter, MFS

Figure 31. Schematic representation of one genomic region found on the chromosome of *Enterobacter sp.* 638. Putative open reading frames are indicated by arrows, below which the *Enterobacter sp.* 638 gene number and gene annotation are shown. The genes involved in sucrose transport and utilization, acetoin and 2,3-butanediol synthesis, the toxin-antitoxin (TA system), as well as other putative functions are also indicated.

However it might be a simplified scheme given the distinct phenomenon of this bacterium in sucrose or lactate medium in terms of the growth curve, pH, the extracellular structures and phytohormone productions, etc. There are necessarily more gene transcriptions and regulators involved to respond to presence of sucrose, and finally coordinate a chain of reactions that are as the basis for the strain's adaptation to its endophytic lifestyle. Therefore, we performed whole transcriptome analysis on Enterobacter sp. 638 grown in sucrose or the lactate at 6 and 12 hours, in order to gain insights in the differential gene expression profiles under these distinct conditions (Figure 32).

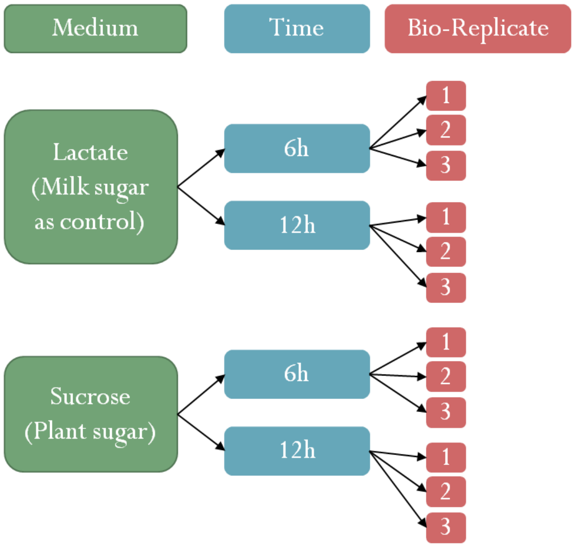


Figure 32. Experimental design of RNA-seq of *Enterobacter sp.* 638 strain. Four distinct conditions are compared with two growth media and two time points: growth media contain lactate or sucrose as sole carbon source, respectively, where sucrose is a plant sugar mimicking the presence of plant while lactate is a milk sugar as a control; two time points are 6 hours or 12 hours after growth. Triplicates of each condition are considered in this experiment.

Enterobacter sp. 638 was grown at 30°C in minimal growth medium with sucrose or lactate as the sole carbon source to mimic conditions of plant association or free living, respectively. When growing on lactate, strain 638 showed an exponential growth phase until the culture reaches an OD_{660} of 0.9 after approximately 24 hours. This growth pattern is in sharp contrast to the pattern observed for strain 638 when grown on sucrose (Figure 33). Cultures growing on sucrose initially grow faster than on lactate, but after reaching an OD_{660} of approximately 0.4, they transitioned into the stationary growth phase. After this transition the following changes in cell behavior were observed for the cultures growing on sucrose: the cells shifted from a planktonic lifestyle to the formation of bacterial aggregates and cell elongation as was observed by light microscopy. Furthermore, no increase in cell biomass was observed. One of the major problems caused by pathogenic bacteria is uncontrolled growth and blockage of the plants vascular tissue (Ryan, Vorholter et al. 2011), therefore control of cell density during endophytic colonization is very important to avoid pathogenic responses and induction of the plant's immune response.

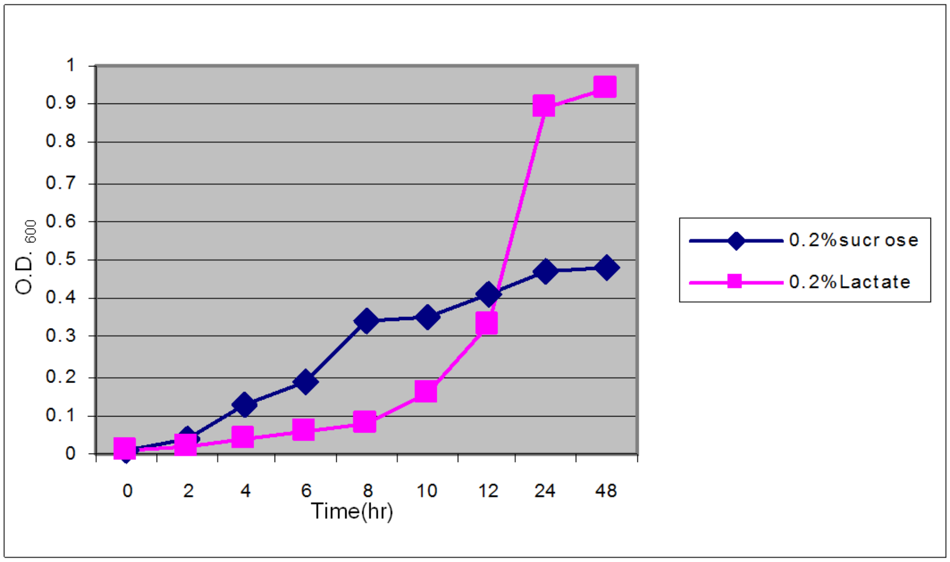


Figure 33. The OD_{600} of *Enterobacter sp.* 638 in growth medium of lactate (0.2%) and sucrose (0.2%), respectively.

10. 3 Summary of mRNA sequencing results

After applying suffix-array lookup algorithm, we aligned short reads from mRNA sequencing to the reference genome of *Enterobacter sp.* 638 (Table 13). Multiple mapped reads in Table 13 are mostly due to seven copies of ribosomal genes. Unmapped reads were examined and found to be patented constructs from Illumina (Lin, Wang et al. 2008). Therefore, uniquely mapped reads are our focus and used for further differential expression analysis.

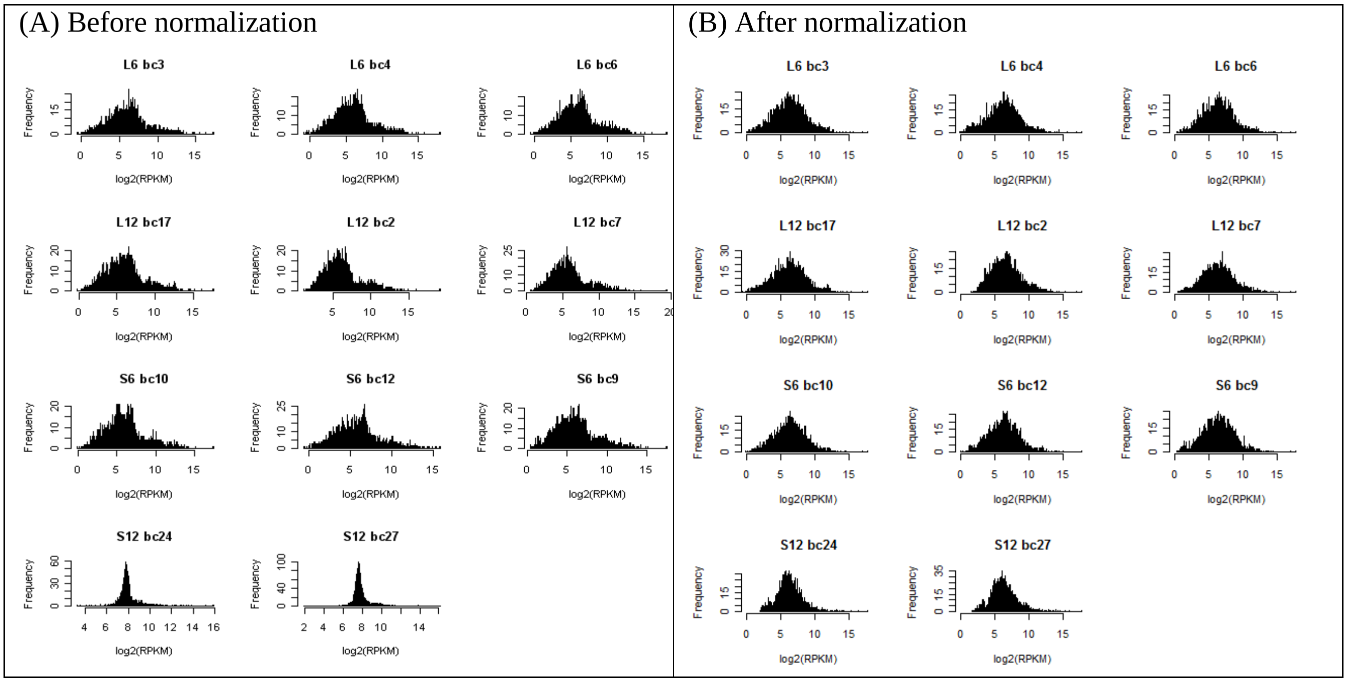
Table 13. Sequence mapping summary

	Uniquely mapped reads		Multiple mapped reads	Unmapped reads	Total reads
	Perfect Match	One mismatch			
Lactate 6hr	4,743,891	355,503	34,470,352	7,974,513	47,544,259
Lactate 12hr	1,447,897	68,559	30,734,301	17,559,946	49,810,703
Sucrose 6hr	2,147,743	169,491	37,532,613	8,670,633	48,520,480
Sucrose 12hr	1,977,053	89,197	24,152,632	2,780,478	28,999,360

10. 4 Differential gene expression analysis

Uniquely mapped reads were then summarized into raw gene expression values by counting the number of reads falling into each defined gene boundary. Raw gene expression values were processed through within normalization RPKM (Mortazavi, Williams et al. 2008) and quantile normalization (Bolstad, Irizarry et al. 2003) (Figure 34). Under the assumption that majority of transcriptome should remain similar expression level while only a small set of genes are responsive to conditions, the normalization generated similar quantile distribution for every sample to make values from different arrays comparable. In Figure 34A, before normalization two samples from sucrose 12 hours have overall higher expressions (log(RPKM) is around 8) than other samples (log(RPKM) is around 6), which might be due to variations from amounts of cDNA or systematic handling for each array. In this sense, many genes might be false-positively tested as significant when comparing sucrose 12 hours to other conditions. Thus normalization was applied to make all samples have similar distributions (Figure 34B).

Figure 34. Histogram of gene expression level log2(RPKM) (A) before normalization; (B) after quantile normalization. RPKM: Reads per Kilobase of gene per Million mapped total reads.



By normalized gene expressions, we performed four pair-wise comparisons -- sucrose verse lactate after 6 hours, sucrose verse lactate after 12 hours and time point 12 hours verse 6 hours for sucrose condition as well as for lactate condition. The differentially expressed genes were determined by biological significance fold changes as well as statistical significance *p* values. False discovery rate is also controlled for comparing thousands of genes simultaneously. The analog MA plots from microarray study are shown in Figure 35 for each comparison. There appears no bias on proportion of significantly differential genes regarding to the expression level of genes. The results are summarized in Table 14.

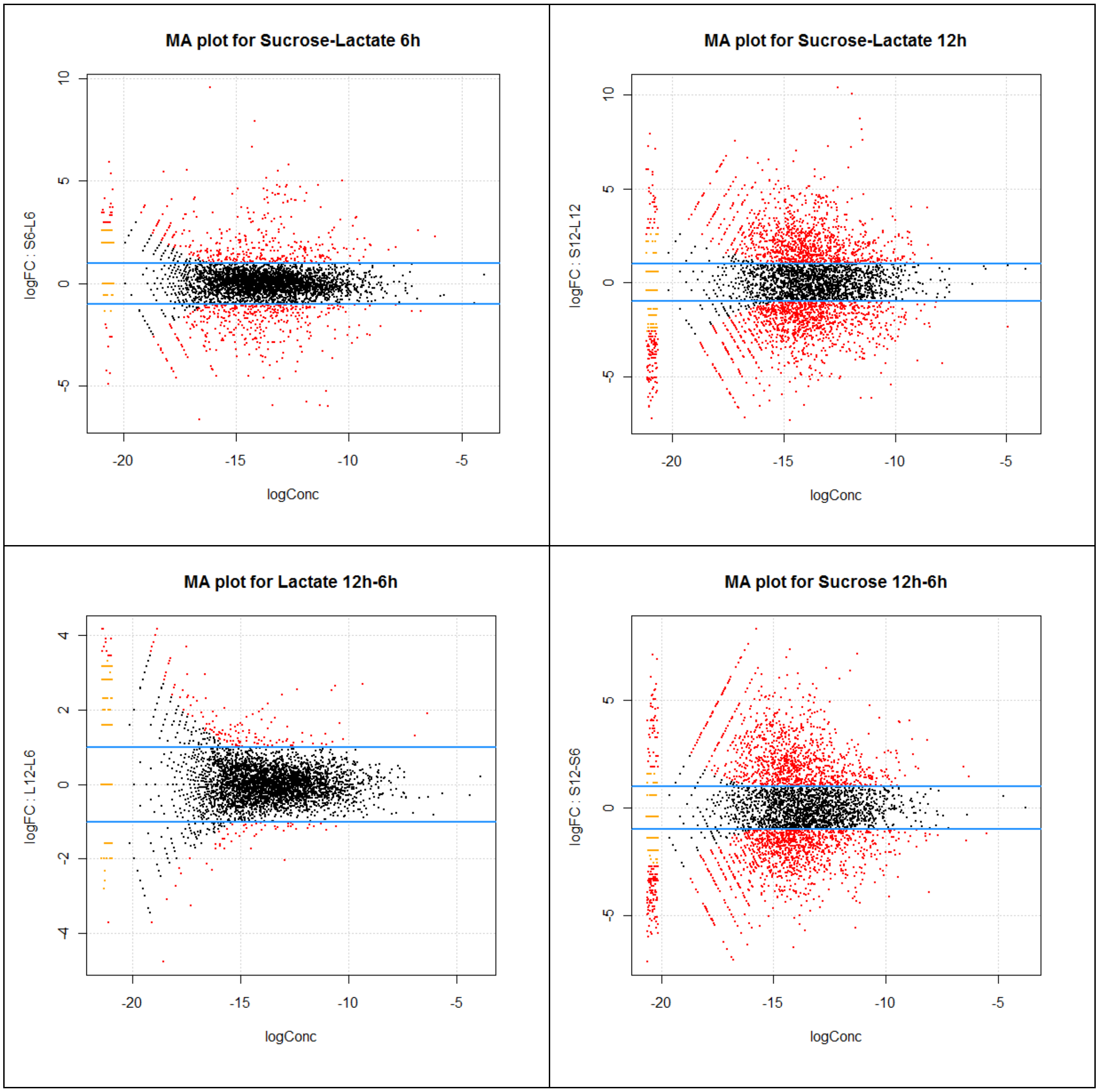


Figure 35. MA plots of four differential gene expression comparisons. Y axis: the log-fold change is plotted against x axis: the log-concentration for each gene. Concentration is defined as the proportion of reads of one gene among total reads in that sample. The genes with fold change greater than 2 and controlled FDR less than 0.05 are highlighted in red. A smear of points at the left-most edge of the plot represents genes which have zero counts in one of the conditions.

Table 14. Differentially gene expression analysis summary. The analysis has been done in R package edgeR.

Condition	*p* values < 0.01 (# of genes)	Controlling FDR < 0.05 (# of genes)	FDR < 0.05 & FC >2		
			# of genes (percentage)	Up-regulated	Down-regulated
Sucrose-Lactate (6h)	1403	1528	790 (17.4%)	391	399
Sucrose-Lactate (12h)	2369	2663	2392 (52.6%)	1166	1226
12h – 6h (Lactate)	538	385	208 (4.6%)	150	58
12h – 6h (Sucrose)	2745	3036	2423 (53.3%)	1160	1263

*FDR: controlled false discovery rate; FC: fold change.

10.4.1 Differential gene expressions linked to carbon source utilization

Sucrose or lactate as the sole carbon source

As expected, expression of genes for carbon source utilization and central metabolism are differently induced depending on the carbon source. Compared to growth on lactate, the operon encoding genes for sucrose uptake and metabolism (*scrKYAB*, located on genomic region 29 (Taghavi, van der Lelie et al. 2010)) is 6~200 fold more induced for cultures growing on sucrose; the *lldPRD* operon for lactate uptake and utilization is 11 ~ 70 fold more induction for cultures growing on lactate (Figure 36). This is the most direct genetic response of *Enterobacter* sp. 638 to different carbon and energy sources. Alternatively, *Enterobacter* sp. 638 can utilize sucrose by transporting it into the cell by specific permeases. Its subsequent metabolism can proceed via phophorolysis. For cultures growing on sucrose, the expression of the sucrose phosphorylase gene (EC 2.4.1.7, Ent638_2165) was up-regulated 76 fold after 12 hours compared to 6 hours, pointing towards phosphorolysis as the preferred pathway for sucrose metabolism over hydrolysis by the *scrKYAB* operon during the later stages of growth (Reid and Abratt 2005).

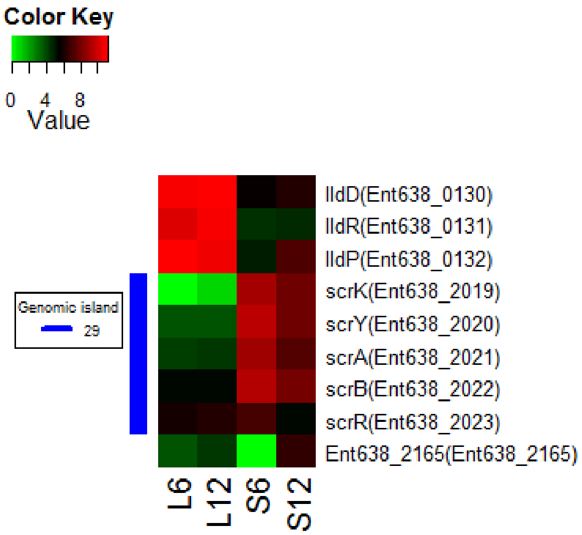


Figure 36. Expression patterns of genes directly involved in sucrose and lactate metabolisms. Log$_2$ of normalized gene expression values (RPKM) are plotted for each condition. L6, L12, S6 and S12 represent the condition lactate 6 hours, lactate 12 hours, sucrose 6 hours and sucrose 12 hours, respectively.

Energy metabolism

Among the various pathways of central energy metabolism in *Enterobacter* sp. 638, the Entner-Doudoroff and the pentose-phosphate pathway show similar levels of gene expression for both growth conditions. The tricarboxylic acid (TCA) cycle, however, shows after 12 hours growth on sucrose a reduction in gene expression. In particular, expression of the succinate dehydrogenase gene cluster (Ent638_1221-1229) decreases 7 fold compared to cultures grown on lactate. This might reflect the differences in growth phase of the sucrose and lactate cultures, as was shown in Figure 33: once entering the stationary growth phase, as is the case for growth on sucrose, *Enterobacter sp.* 638 should have lower energy requirement. Furthermore, pyruvate, the input for the TCA cycle, is also the precursor in the synthesis of various secondary metabolites, including acetoin, 2,3-butanediol and colanic acid that are important in the symbiotic relationship between *Enterobacter sp.* 638 and its plant host, and whose synthesis levels are significantly increased for cultures growing on sucrose.

10.4.2 Transcriptional patterns of genes involved in motility and biofilm formation

Motility and chemotaxis

In *E. coli*, the MqsR regulator acts through a two-component motility regulatory system QseBC (Gonzalez Barrios, Zuo et al. 2006) to transcriptionally regulate FlhDC, the master regulator of flagella and motility genes (Clarke and Sperandio 2005). The *Enterobacter sp.* 638 genome contains multiple flagellar biosynthesis operons as well as determinants involved in chemotaxis, including *flgNMABCDEFGHIJKL* (Ent638_1584-1597), *fliCDSTEFGHIJKLMNOPQR* (Ent638_2522-2541) *flhEAB* (Ent638_2445-2447) *cheZYBR* (Ent638_2452-2455), *tap tar* (Ent638_2456-2457), *cheWA motBA* (Ent638_2465-2468) and *flhCD* (Ent638_2469-2470). The expressions of these gene clusters were, compared to growth on lactate, reduced for cultures grown on sucrose, both after 6 and 12 hours, indicating a reduction in the motility of the cells. This was also confirmed by microscopic imaging (Figure not shown). On the contrary, the majority of genes associated with pili biosynthesis are up-regulated after 12 hours of growth on sucrose. Pili are primarily involved in adhesion to surfaces and are among the few factors known to affect endophytic colonization (Dorr, Hurek et al. 1998). This opposite expression pattern for genes involved in motility and adhesion suggests that *Enterobacter sp.* 638 becomes less motile under conditions that mimic the association with the plant host. The induced expression levels of pili biosynthesis genes is also consistent with the observed pili on the surfaces of cells grown on sucrose.

Curli fibers are another factor mediating host cell adhesion and invasion. However, except for the *csgG* gene (2.9-fold induction), no significantly different levels of gene expression were observed for the curli biosynthesis cluster (Ent638_1553-1559). It has been established that motility plays an important role in biofilm development of *E. coli* (Pratt and Kolter 1998; Wood, Gonzalez Barrios et al. 2006). Although studies on DNA microarrays of *E. coli* cells have not found to date a significant difference in flagella expression during biofilm development (Schembri, Kjaergaard et al. 2003; Beloin, Valle et al. 2004; Ren, Bedzyk et al. 2004), other studies have shown that motility genes are repressed in *Pseudomonas aeruginosa* in a 5-day old biofilm, and in *Bacillus subtilis* biofilms after 8, 12, and 24 h (Whiteley, Bangera et al. 2001; Stanley, Britton et al. 2003). On the other hand, the presence of conjugative plasmids increases biofilm formation (Ghigo 2001) independent of the expression of motility genes (Reisner, Haagensen et al. 2003). The transfer functions located on the *Enterobacter sp.* 638 plasmid pENT638-1 (Ent638_4285-4312) were found mostly to be up-regulated after 12 hours growth on sucrose (with average seven-fold change), which might suggest as a positive effect on biofilm formation.

Extracellular Poly Saccharide synthesis

In *E. coli*, MqsR induces the expression of the transcription factor McbR (Gonzalez Barrios, Zuo et al. 2006). McbR inhibits the expression of the periplasmic McbA protein in order to prevent the overproduction of colanic acid; excess colanic acid causes mucoidy, which inhibits biofilm formation (Zhang, Garcia-Contreras et al. 2008). In *Enterobacter sp.* 638, *mcbR* gene expression decreased for cultures grown on sucrose after both 6 and 12 hours. After 6 hours of growth on sucrose, the colanic acid biosynthesis operon (Ent638_2657-2676) became over-expressed 22-folds on average, but after 12 hours gene expression levels were back to those similar as observed for cultures growing on lactate. Previously, it was shown that the increased expression of the colanic acid operon is the genetic response underlying biofilm formation and colonization processes in *E. coli*(Prigent-Combaret, Vidal et al. 1999). In *E. coli* several other genes are induced during biofilm formation, including *ompC* (porin), the *proVWX* operon (high-affinity transport system of glycine betaine), *pepT* (tripeptidase), and *nikA* (nickel high-affinity transport system)(Prigent-Combaret, Vidal et al. 1999). Increased expression levels were observed after 12 hours growth on sucrose for all these genes, with the exception of the *pepT* gene that showed no significant change in expression levels.

Colanic acid is a polyanionic heteropolysaccharide containing a repeat unit with D-glucose, L-fucose, D-galactose, and D-glucuronate sugars that are nonstoichiometrically decorated with O-acetyl and pyruvate side chains. The subunits of colonic acid based EPS, such as pyruvate, D-glucose and L-fucose, are also substrates for the TCA cycle of central metabolism. The lower gene expression levels for genes constituting the TCA cycle as observed for growth on sucrose could be due to the deprival of these substrates by the synthesis of colonic acid.

Acid stress

Another common trend in the biofilm transcriptome studies is that stress genes are induced (Wood 2009). Gene *asr* encoding acid shock protein is highly induced in the sucrose condition at both 6 and 12 hours. It is also consistent with the pH level drop observed at the later growth stage of bacteria in the sucrose.

Colonization

One active colonization pathway of *Enterobacter sp.* 638 that described by Taghavi *et al*. (2010) is through pectin / pectate degradation. Pectin is a structural polysaccharide contained between plant cell walls that help cells bind together. Harbored on the genomic islands 11 and 29 (Taghavi, van der Lelie et al. 2010), genes were found to putatively encode enzymes to degrade demethylated pectin – pectate into oligogalacturonides, and uptake and finally to compounds of the general cellular metabolism. During the growth in the two different media, these genes were significantly induced in the presence of sucrose from 6 to 12 hours, but not in lactate. This pattern is also observed for the *uxaABC* genes, which encode enzymes for an alternative pathway of one middle step of the degradation of pectin – conversion from galacturonate into 2-dehydro-3-deoxy-D-gluconate. Interestingly, the *uxaB* gene (Ent638_2013, genomic island 29) located closely to the sucrose utilization operon (Ent638_2019-2023), and both also have a similar expression pattern – induced by the sucrose growth medium. A collaborative relation is suggested between the sucrose metabolism and the host invasion by *Enterobacter sp.* 638 by the genomic analysis as well as the transcriptome analysis.

10.4.3 Transcriptional patterns of genes involved in the phytohormone synthesis

Production of acetoin / 2,3-butanediol

Consistent *budABC* gene expression pattern from the previous RT-PCR study (Taghavi, van der Lelie et al. 2010) has also been seen in the current study. Cells in the sucrose medium after 12 hours showed a 1321, 1054 and 283 fold induction of *budABC* genes respectively compared to cells in the lactate. It further confirmed the co-regulated pattern between the sucrose utilization operon and the acetoin, 2,3-butanediol synthesis operon.

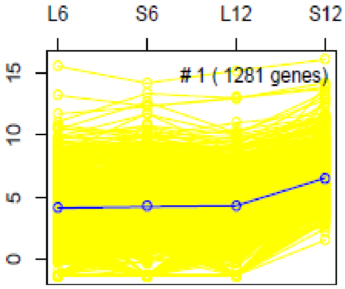
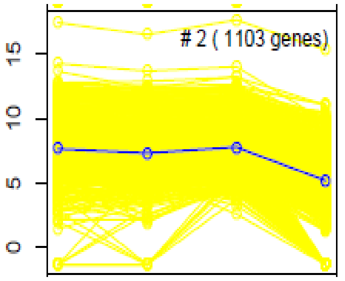
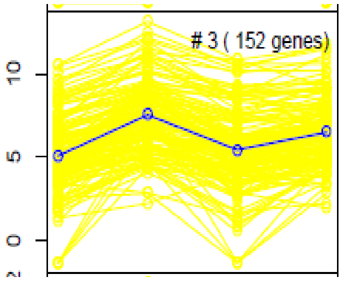
Production of indole acetic acid (IAA)

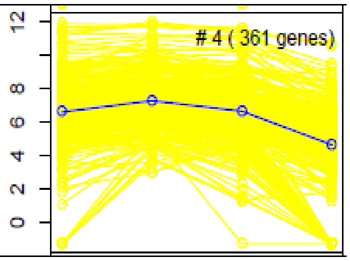
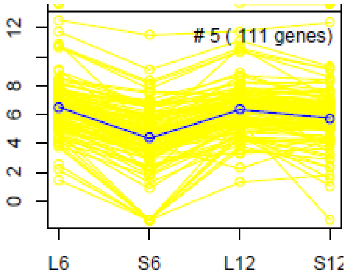
Alternative plant growth promoting mechanism of endophytic and rhizosphere bacteria is to synthesize the plant auxin indole acetic acid (IAA). The production of IAA by *Enterobacter sp*. 638 was experimentally demonstrated and is likely via indolepyruvate as an intermediate molecule by the tryptophan degradation pathway VII (Taghavi, Garafola et al. 2009; Taghavi, van der Lelie et al. 2010). The key enzyme of this pathway is indole-3-pyruvate carboxylase (IpdC, Ent638_2923), whose gene expression shows 8-fold induction in the sucrose condition compared to lactate at 12h. Gene expressions of two other enzymes (aminotransferase and aldehyde dehydrogenase) on the pathway appear not differentially significant between sucrose and lactate. However, these two enzymes are considered less important as IpdC, since they are usually present in most bacteria, including those that cannot produce IAA. Furthermore, experiments have shown that IpdC is solely responsible for the regulation of this pathway, and that the first enzyme on the pathway, L-tryptophan aminotransferase, operates very close to equilibrium (Koga, Syono et al. 1994; Koga 1995).

10. 5 Clustering and functional category analysis

To demonstrate gene expression patterns across four conditions, clustering analysis was performed on the expression data of *Enterobacter sp.* 638 at condition lactate 6, 12 hours and sucrose 6, 12 hours using spearman correlation as the distance metric. Five distinctive expression patterns have been found: high level in sucrose but low level in lactate (cluster 1 and 3), low level in sucrose but high level in lactate (cluster 2 and 5), and not much differential gene expressions (cluster 4) (Table 15). For genes in each cluster, we performed over-representative analysis to identify the associated functions with each expression pattern (Table 15). As discussed in the previous section, functions including TCA cycle, motility and chemotaxis are enriched in the cluster 2 which are featured by repressed expression level in the sucrose condition. Cluster 4 appears no significant enriched functions. The gene expression pattern of cluster 4 is flatter than other clusters and the top representative function category is unknown function, which might explain why no other specific functions significantly enriched for this group of genes. Biosynthesis of surface polysaccharides, colanic acid and pilus are over-representative functions in cluster 1 or 3, which represent high expression level in the sucrose condition. While strikingly, functions related to several nutrients uptake (N, Fe and siderophore) also showed up in the cluster 1 and 3 that are up-regulated in the sucrose condition even with no shortage of the corresponding nutrients in the medium.

Table 15. Clustering and functional categories analyses of significantly differential genes. The functional categories bioprocess and role are provided from the manual annotation in MaGe annotation system. The over-representative analysis is done by R package goseq 1.0.3.

Bioprocess	FDR	Roles	FDR	Cluster
Cluster 1				
Transport and binding proteins	0.000	The Major Facilitator Super family	0.000	
Prophage functions	0.000	Structural component	0.000	
Plasmid functions	0.000	plasmid transfer	0.000	
Carbohydrates, organic alcohols, and acids	0.011	Replication	0.004	
Scavenge(Catabolism)	0.020	DNA packaging, phage assembly	0.006	
Cations and iron carrying compounds	0.023	Fe aquisition	0.006	
		Pilus	0.008	
Cluster 2				
Chemotaxis and motility	0.000	Motility	0.000	
Ribosomal proteins	0.000	Ribosome	0.000	
Translation factors	0.000	Translation	0.000	
Surfacestructures	0.001	Flagella	0.000	
ATP-proton motive force inter-conversion	0.002	Cytoplasm	0.000	
TCA cycle	0.046	H+	0.000	
PTS	0.046	Periplasmic space	0.000	
		The H+/Na+-translocating F, V-and A-type ATPase Super family	0.000	
		Tricarboxylic acid cycle	0.024	
		Inhibition/activation of enzymes	0.025	
Cluster 3				
Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	0.000	Colanicacid(M antigen)	0.000	
Explore	0.000	molybdate	0.033	
Protect	0.000	Methionine	0.033	
Nitrogen metabolism	0.006	Nitrogen metabolism	0.039	
Sugar-nucleotide biosynthesis and conversions	0.026	Dessication	0.039	
Anaerobic	0.046	Glutamate biosynthesis I	0.039	
Siderophores	0.050	Enterochelin (enterobactin)	0.039	
		periplasmic binding component	0.040	

Cluster 4				
-	-	-	-	
Cluster 5				
Carbohydrates, organic alcohols, and acids	0.006	4.S.34:citrate/succinate	0.044	

Nitrogen metabolism

Although *Enterobacter sp.* 638 is unable to fix nitrogen, it has the genetic capacity for the dissimilatory and assimilatory nitrate reduction pathway. Gene clusters encoding multiple nitrate/nitrite transport and reduction pathways locate on the different positions on its genome (Taghavi, van der Lelie et al. 2010). These gene clusters showed consistently induced expression pattern in the sucrose medium at both 6 hours and 12 hours as cluster 3. Especially, gene *nasAB nrtCBA nasR*, on the putative genomic island 33, have average 20 fold more expression index in the sucrose medium than in the lactate at 12 hours.

Iron scavenging

Siderophore is one of the most efficient systems of bacteria to compete for the limited iron for their synergistic interaction with the host plant (Höfte and PAHM 2007). Similarly to *E.coli* K12, *Enterobacter sp.* 638 is able to synthesize the siderophore enterobactin, which allows it to capture iron more efficiently than plant pathogens and restrict their proliferation and thus protect plants (Walsh, Morrissey et al. 2001). In *Enterobacter sp.* 638, genes associated with siderophore majorly located within a large cluster (Ent638_1111-1128), including the siderophore biosynthesis genes (*entFCEBA*), secretion genes (*entS*), and an enterobactin esterase gene (*fes*), as well as two ABC transporter genes for iron uptake (*sitABCD* and *fepCGDB*). The overall gene expression level of this gene cluster in the sucrose medium is higher compared to the one in the lactate medium. There are several other transporters involved in the iron uptake having similar pattern that expressed more in the sucrose medium (for example, *fepBGDC*). Particularly, the *hmu* operon for hemin transport, on the genomic region 25, has been largely induced in the sucrose medium at 12 hours (over 100 folds).

Heavy metal tolerance and metal homeostasis

Metals are vital chemical species to the microorganisms. Although an excess of heavy metals are generally toxic to the cell, some of metals are essential to the life at a trace amounts. The genomic survey of *Enterobacter sp.* 638 showed possible mechanisms of metal(loid) homeostasis, tolerance (Taghavi, van der Lelie et al. 2010). Significant differential expressions were also observed for genes involved in the metal uptake, tolerance and balance. Up-regulated expressions at the sucrose condition at 12 hours have been observed by the P-type ATPAse gene *copA*, copper efflux operon *cusABCF*, nickel uptake operon *nikABCDE*, and P-type efflux ATPase gene *zntA* that involved in zinc/ cadmium/ cobalt resistance, etc. Nickel is an essential cofactor for urease (*ureABC*, Ent638_3464-3466), which is able to convert urea into ammonia (Dosanjh, Hammerbacher et al. 2007). Consistent with the expression change of nickel uptake genes *nikABCDE*, it is also seen the over-expression of gene *ureA* and *ureC* for the sucrose condition at 12 hours. Besides, *nikA* gene was shown to be related to the colonization and biofilm formation process (Prigent-Combaret, Vidal et al. 1999).

Inductions of these various uptake systems in this endophytic bacterium are solely observed under the sucrose growth condition. It may suggest us the well preparation of *Enterobacter sp.* 638 to compete for limited nutrient resources inside the plants that also serves as a strong anti-pathogen strategy for its commensal life style.

10. 6 Regulatory network of RcsAB

Furthermore, to identify one or more hub regulators responsible for our observed transcriptional changes, we resorted to the well-identified regulatory network in *E. coli*. The overview of regulatory network in *Enterobacter sp.* 638 is shown in Figure 37. This figure includes nodes and edges that showed consistent expression (S12 – L12) to the relationships recorded in RegulonDB of *E.coli* K12. The color of edges: red -- deactivation; green – activation; blue – dual regulation, which are implied from RegulonDB. There are 851 nodes in total. Among the whole regulatory network, one sub-network of RcsAB particularly is of our interest.

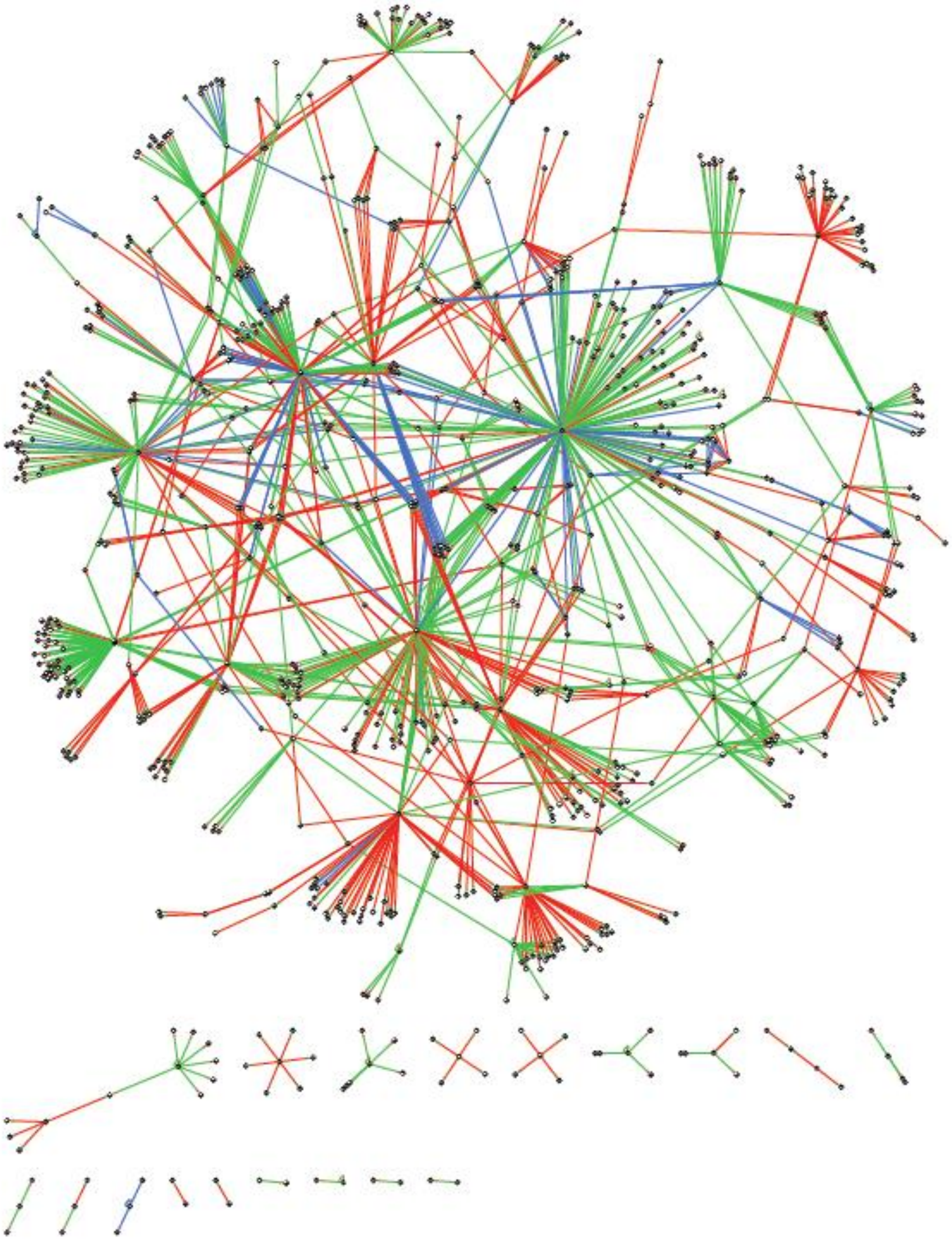


Figure 37. Overview of regulatory network in *Enterobacter sp.* 638. Nodes and edges are included here only when they showed consistent expression (S12 – L12) to the relationships recorded in RegulonDB of *E.coli* K12. The color of edges: red -- deactivation; green – activation; blue – dual regulation, which are implied from RegulonDB. The figure was generated using Cytoscape.

When searching orthologs of significant genes from our experiments in *E. coli* database, one regulator GadE attracted us. GadE is the immediate upstream regulator for targets involved in EPS synthesis and acid response as discussed in section 10.4.2. However, this regulator does not have a homolog on the genome of *Enterobacter sp.* 638. By searching for regulators that share the same targets of GadE, we found the dual regulators RcsA (Ent638_2542) and RcsB (Ent638_2797). These two genes encode a two-component system RcsAB, which are also involved in capsule biosynthesis and cell division. RcsA (Ent638_2542), the DNA-binding transcriptional activator, showed the highest expression level in sucrose condition after 6 hours. Their targets the *wca* operon showed very similar expression pattern compared to *rcsA*. While *rcsB* (Ent638_2797) didn't show much differential expressions across different conditions. The other operon yjbEFGH, as RcsAB targets, is also involved in the EPS production (Ferrieres, Aslam et al. 2007) and has consistent expression with its regulator RcsA in our experiment. RcsB is involved in a phosphorelay cascade (Majdalani and Gottesman 2005). In contrast, it is reported that RcsA, a DNA binding protein related to response regulators but not believed to be regulated by phosphorylation, binds with RcsB to activate transcription of target genes (Majdalani and Gottesman 2005). Hence, in accordance to the nature of two regulators, RcsA showed differential changes across conditions since it functions via transcriptional activation; while RcsB remained at a similar expressional level, because it is regulated through phosphorylation and thus without transcriptional changes.

Although RcsAB are not necessarily the direct substitute of the GadE regulator in *Enterobacter sp.* 638, they are involved in several bio-processes of great interests. We noticed that the *rcsAB* genes are not on the same location on genome, and there are two members of Rcs regulon rcsCD adjacent to rcsB but on a convergent operon. This organization is similar to that occurred in *E.coli*. The regulatory connections of RcsAB implied from *E. coli* are shown in Figure 38. On the diagram, four blocks under each gene represent expression levels of L6, L12, S6 and S12, respectively and with brighter color indicates higher expressions and blank block indicates not expressed. The RcsAB regulator appears to activate transcription of genes for capsular polysaccharide and repress genes for flagella synthesis (Majdalani and Gottesman 2005). This is consistent with our observations of activated gene expressions of EPS production but reduced ones for flagella synthesis. Therefore, it is suggested that in *Enterobacter sp.* 638, RcsAB regulators are also responsible for modulation of EPS production, flagellar and curli fiber synthesis. These outer membrane related activities are essential as part of an inevitable colonization process for bacteria entering host plants.

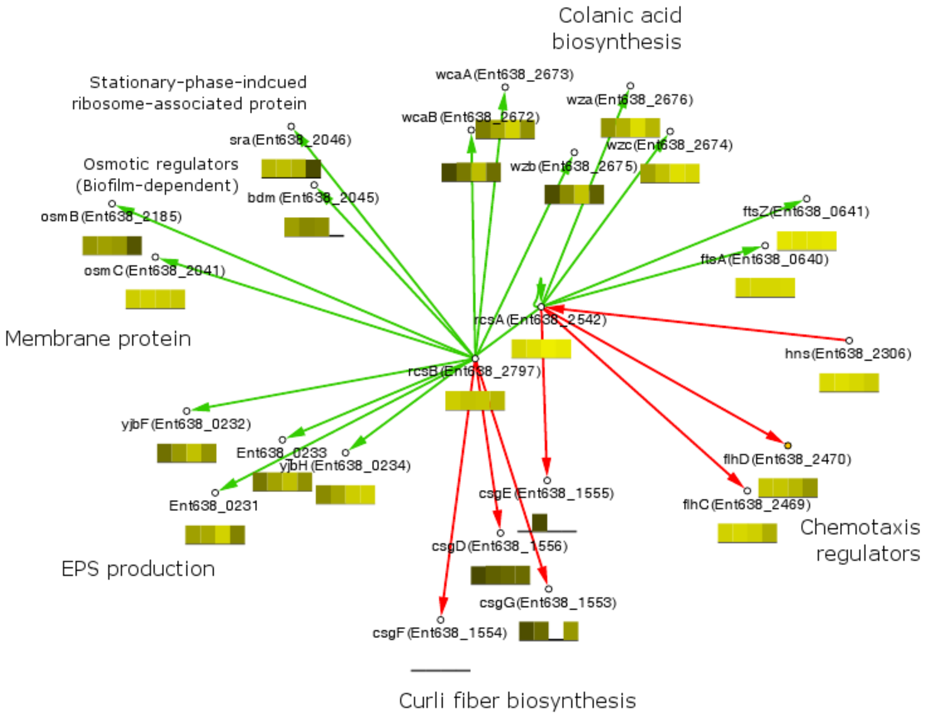


Figure 38. The regulatory network of RcsAB. The color of edges: red -- deactivation; green – activation, which are implied from RegulonDB. Four blocks under each gene represent expression levels of L6, L12, S6 and S12, respectively and with brighter color indicates higher expressions and blank block indicates not expressed. The figure was generated using VistaClara plugin (1.05) (Kincaid, Kuchinsky et al. 2008) in Cytoscape.

10. 7 Outlook

In this bioinformatics pipeline, we have outlined the major steps starting from newly sequenced organism to the final discovery of regulatory networks. In brief, it includes genome annotations, comparative genomics and functional analyses, differential gene expression analysis by mRNA sequencing. Further biological insight can be gained by exploring patterns of expression changes within clusters and associated functions, and ultimately integrates results to regulatory networks.

We achieved good biological interpretations from this pipeline; nevertheless there are spaces for further refinements. For example, many existing approaches may deserve further study in terms of their flexibility to accommodate various study designs and sample sizes. Furthermore so far, comparative genomic analysis and regulatory network analysis are limited by the mapped existing knowledge from the model organism *E. coli*, therefore lack unique links of exclusive genes in our bacteria. This incompletion of regulatory relations may be addressed by further Chromatin Immunoprecipitation (ChIP) experiment or consensus promoter analysis, etc. As this field is evolving fast, we expect many new methods and tools, from biology, computer science and statistics, for analyses at each level to emerge in the near future, and thus help better understand the complex biological pathways.

Bibliography

Adewale, A. J., I. Dinu, J. D. Potter, Q. Liu and Y. Yasui (2008). "Pathway analysis of microarray data via regression." J Comput Biol **15**(3): 269-277.

Agresti, A. (2007). An introduction to categorical data analysis. Hoboken, New Jersey, John Wiley & Sons, Inc.

Al-Shahrour, F., R. Diaz-Uriarte and J. Dopazo (2004). "FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes." Bioinformatics **20**(4): 578-580.

Al-Shahrour, F., R. Diaz-Uriarte and J. Dopazo (2005). "Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information." Bioinformatics **21**(13): 2988-2993.

Alekseev, O. M., R. T. Richardson, O. Alekseev and M. G. O'Rand (2009). "Analysis of gene expression profiles in HeLa cells in response to overexpression or siRNA-mediated depletion of NASP." Reprod Biol Endocrinol **7**: 45.

Allen, M. J. and W. M. Yen (2002). Introduction to Measurement Theory. Long Grove, IL, Waveland Press.

Allocco, D. J., I. S. Kohane and A. J. Butte (2004). "Quantifying the relationship between co-expression, co-regulation and gene function." BMC Bioinformatics **5**: 18.

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

Arminger, G. and U. Kusters (1988). Latent trait and latent class models. New York, Plenum Press**:** 51-73.

Barry, W. T., A. B. Nobel and F. A. Wright (2005). "Significance analysis of functional categories in gene expression studies: a structured permutation approach." Bioinformatics **21**(9): 1943-1949.

Baumgart, D. C. and S. R. Carding (2007). "Inflammatory bowel disease: cause and immunobiology." Lancet **369**(9573): 1627-1640.

Beissbarth, T. and T. P. Speed (2004). "GOstat: find statistically overrepresented Gene Ontologies within a group of genes." Bioinformatics **20**(9): 1464-1465.

Beloin, C., J. Valle, P. Latour-Lambert, P. Faure, M. Kzreminski, D. Balestrino, J. A. Haagensen, S. Molin, G. Prensier, B. Arbeille and J. M. Ghigo (2004). "Global impact of mature biofilm lifestyle on Escherichia coli K-12 gene expression." Mol Microbiol **51**(3): 659-674.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society. Series B (Methodological) **57**(1): 289-300.

Berriman, M. and K. Rutherford (2003). "Viewing and annotating sequence data with Artemis." Brief Bioinform **4**(2): 124-132.

Bollen, K. (1989). Structral Equations with Latent Variables. New York, Wiley-Interscience.

Bollen, K. and R. A. Stine (1993). Bootstrapping goodness-of-fit measures in structural equatjion models. Thousand Oaks, CA, Sage.

Bollen, K. A. (1989). Structural equations with latent variables, John Wiley & sons, Inc.

Bolstad, B. M., R. A. Irizarry, M. Astrand and T. P. Speed (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-193.

Boyle, E. I., S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry and G. Sherlock (2004). "GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes." Bioinformatics **20**(18): 3710-3715.

Bron, S. and G. Venema (1972). "Ultraviolet inactivation and excision-repair in Bacillus subtilis. IV. Integration and repair of ultraviolet-inactivated transforming DNA." Mutat Res **15**(4): 395-409.

Bullard, J. H., E. Purdom, K. D. Hansen and S. Dudoit (2010). "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments." BMC Bioinformatics **11**: 94.

Butte, A. J. and I. S. Kohane (2000). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." Pac Symp Biocomput: 418-429.

Caspi, A., K. Sugden, T. E. Moffitt, A. Taylor, I. W. Craig, H. Harrington, J. McClay, J. Mill, J. Martin, A. Braithwaite and R. Poulton (2003). "Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene." Science **301**(5631): 386-389.

Chakravorty, S., D. Helb, M. Burday, N. Connell and D. Alland (2007). "A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria." Journal of Microbiological Methods **69**(2): 330-339.

Cho, R. J., M. Huang, M. J. Campbell, H. Dong, L. Steinmetz, L. Sapinoso, G. Hampton, S. J. Elledge, R. W. Davis and D. J. Lockhart (2001). "Transcriptional regulation and function during the human cell cycle." Nat Genet **27**(1): 48-54.

Clarke, M. B. and V. Sperandio (2005). "Transcriptional regulation of *flhDC* by QseBC and sigma (FliA) in enterohaemorrhagic *Escherichia coli*." Mol Microbiol **57**(6): 1734-1749.

Claudel-Renard, C., C. Chevalet, T. Faraut and D. Kahn (2003). "Enzyme-specific profiles for genome annotation: PRIAM." Nucleic Acids Res **31**(22): 6633-6639.

Costanzo, M., A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R. P. St Onge, B. VanderSluis, T. Makhnevych, F. J. Vizeacoumar, S. Alizadeh, S. Bahr, R. L. Brost, Y. Chen, M. Cokol, R. Deshpande, Z. Li, Z. Y. Lin, W. Liang, M. Marback, J. Paw, B. J. San Luis, E. Shuteriqi, A. H. Tong, N. van Dyk, I. M. Wallace, J. A. Whitney, M. T. Weirauch, G. Zhong, H. Zhu, W. A. Houry, M. Brudno, S. Ragibizadeh, B. Papp, C. Pal, F. P. Roth, G. Giaever, C. Nislow, O. G. Troyanskaya, H. Bussey, G. D. Bader, A. C. Gingras, Q. D. Morris, P. M. Kim, C. A. Kaiser, C. L. Myers, B. J. Andrews and C. Boone (2010). "The genetic landscape of a cell." Science **327**(5964): 425-431.

Cox, D. R. and N. Wermuth (1992). "Response models for binary and quantitative variables." Biometrika **79**(3): 441-461.

Crotty, B. (1994). "Ulcerative colitis and xenobiotic metabolism." Lancet **343**(8888): 35-38.

de Leon, A. R. and K. C. Chough (2010). Mixed-outcome data.

Dieckgraefe, B. K., W. F. Stenson, J. R. Korzenik, P. E. Swanson and C. A. Harrington (2000). "Analysis of mucosal gene expression in inflammatory bowel disease by parallel oligonucleotide arrays." Physiol Genomics **4**(1): 1-11.

Dinu, I., J. D. Potter, T. Mueller, Q. Liu, A. J. Adewale, G. S. Jhangri, G. Einecke, K. S. Famulski, P. Halloran and Y. Yasui (2007). "Improving gene set analysis of microarray data by SAM-GS." BMC Bioinformatics **8**: 242.

Dorr, J., T. Hurek and B. Reinhold-Hurek (1998). "Type IV pili are involved in plant-microbe and fungus-microbe interactions." Mol Microbiol **30**(1): 7-17.

Dosanjh, N. S., N. A. Hammerbacher and S. L. Michel (2007). "Characterization of the Helicobacter pylori NikR-P(ureA) DNA interaction: metal ion requirements and sequence specificity." Biochemistry **46**(9): 2520-2529.

Dowd, S. E., Y. Sun, P. R. Secor, D. D. Rhoads, B. M. Wolcott, G. A. James and R. D. Wolcott (2008). "Survey of bacterial diversity in chronic wounds using pyrosequencing, DGGE, and full ribosome shotgun sequencing." BMC Microbiol **8**: 43.

Drton, M. and M. D. Perlman (2004). "Model selection for Gaussian concentration graphs." Biometrika(32): 407-499.

Dunson, D. (2000). "Bayesian latent variable models for clustered mixed outcomes." Journal of the Royal Statistical Society Series B-Statistical Methodology **62**(2): 35-366.

Edward, D. (2000). Introduction to Graphical Modelling. New York, Springer.

Efron, B. and R. Tibshirani (2007). "On testing the significance of sets of genes." Tech Report.

Englund, G., A. Jacobson, F. Rorsman, P. Artursson, A. Kindmark and A. Ronnblom (2007). "Efflux transporters in ulcerative colitis: decreased expression of BCRP (ABCG2) and Pgp (ABCB1)." Inflamm Bowel Dis **13**(3): 291-297.

Ferrieres, L., S. N. Aslam, R. M. Cooper and D. J. Clarke (2007). "The yjbEFGH locus in Escherichia coli K-12 is an operon encoding proteins involved in exopolysaccharide production." Microbiology **153**(Pt 4): 1070-1080.

Fisher, L. D. and G. van Belle (1993). Biostatistics: A methodology for Health Sciences., John Wiley and Sons, New York.

Frank, D. N., C. E. Robertson, C. M. Hamm, Z. Kpadeh, T. Zhang, H. Chen, W. Zhu, R. B. Sartor, E. C. Boedeker, N. Harpaz, N. R. Pace and E. Li (2011). "Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases." Inflamm Bowel Dis **17**(1): 179-184.

Frank, D. N., A. L. St Amand, R. A. Feldman, E. C. Boedeker, N. Harpaz and N. R. Pace (2007). "Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases." Proc Natl Acad Sci U S A **104**(34): 13780-13785.

Ghigo, J. M. (2001). "Natural conjugative plasmids induce bacterial biofilm development." Nature **412**(6845): 442-445.

Goeman, J. J., J. Oosting, A. M. Cleton-Jansen, J. K. Anninga and H. C. van Houwelingen (2005). "Testing association of a pathway with survival using gene expression data." Bioinformatics **21**(9): 1950-1957.

Goeman, J. J., S. A. van de Geer, F. de Kort and H. C. van Houwelingen (2004). "A global test for groups of genes: testing association with a clinical outcome." Bioinformatics **20**(1): 93-99.

Gonzalez Barrios, A. F., R. Zuo, Y. Hashimoto, L. Yang, W. E. Bentley and T. K. Wood (2006). "Autoinducer 2 controls biofilm formation in *Escherichia coli* through a novel motility quorum-sensing regulator (MqsR, B3022)." J Bacteriol **188**(1): 305-316.

Gueorguieva, R. V. and G. Sanacora (2006). "Joint analysis of repeatedly observed continuous and ordinal measures of disease severity." Stat Med **25**(8): 1307-1322.

Hamm, C. M., M. A. Reimers, C. K. McCullough, E. B. Gorbe, J. Lu, C. C. Gu, E. Li, B. K. Dieckgraefe, Q. Gong, T. S. Stappenbeck, C. D. Stone, D. W. Dietz and S. R. Hunt (2010). "NOD2 status and human ileal gene expression." Inflamm Bowel Dis **16**(10): 1649-1657.

Hardcastle, T. J. and K. A. Kelly (2010). "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data." BMC Bioinformatics **11**: 422.

Hawkins, R. D., G. C. Hon and B. Ren (2010). "Next-generation genomics: an integrative approach." Nat Rev Genet **11**(7): 476-486.

Hever, A., R. B. Roth, P. A. Hevezi, J. Lee, D. Willhite, E. C. White, E. M. Marin, R. Herrera, H. M. Acosta, A. J. Acosta and A. Zlotnik (2006). "Molecular characterization of human adenomyosis." Mol Hum Reprod **12**(12): 737-748.

Hofling, H. and R. Tibshirani (2009). "Estimation of sparse binary pairwise Markov networks using pseudo-likelihood." Journal of Machine Learning Research(10): 883-906.

Höfte, M. and B. PAHM, Eds. (2007). Competition for iron and induced systemic resistance by siderophores of plant growth promoting rhizobacteria. Microbial siderophores. Heidelberg, Springer Verlag.

Hosack, D. A., G. Dennis, Jr., B. T. Sherman, H. C. Lane and R. A. Lempicki (2003). "Identifying biological themes within lists of genes with EASE." Genome Biol **4**(10): R70.

Hu, L. and P. M. Bentler (1999). "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives." Structural Equation Modeling **6**(1): 1-55.

Ideker, T. (2004). "Systems biology 101--what you need to know." Nat Biotechnol **22**(4): 473-475.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics **4**(2): 249-264.

Johnston, J. (1972). Econometric Methods. St. Louis, McGraw-Hill.

Khatri, P. and S. Draghici (2005). "Ontological analysis of gene expression data: current tools, limitations, and open problems." Bioinformatics **21**(18): 3587-3595.

Kincaid, R., A. Kuchinsky and M. Creech (2008). "VistaClara: an expression browser plug-in for Cytoscape." Bioinformatics **24**(18): 2112-2114.

Klein, G. and E. Klein (1985). "Evolution of tumors and the impact of molecular oncology." Nature(315): 190-195.

Kline, R. B. (1998). Principles and Practice of Structural Equation Modeling. New York, The Guiford Press.

Koga, J. (1995). "Structure and function of indolepyruvate decarboxylase, a key enzyme in indole-3-acetic acid biosynthesis." Biochim Biophys Acta **1249**(1): 1-13.

Koga, J., K. Syono, T. Ichikawa and T. Adachi (1994). "Involvement of L-tryptophan aminotransferase in indole-3-acetic acid biosynthesis in Enterobacter cloacae." Biochim Biophys Acta **1209**(2): 241-247.

Langmann, T., C. Moehle, R. Mauerer, M. Scharl, G. Liebisch, A. Zahn, W. Stremmel and G. Schmitz (2004). "Loss of detoxification in inflammatory bowel disease: dysregulation of pregnane X receptor target genes." Gastroenterology **127**(1): 26-40.

Langmann, T. and G. Schmitz (2006). "Loss of detoxification in inflammatory bowel disease." Nat Clin Pract Gastroenterol Hepatol **3**(7): 358-359.

Lauritzen, S. L. (1996). Graphical Model. Oxford, U.K., Clarendon Press.

Lee, H. K., W. Braynen, K. Keshav and P. Pavlidis (2005). "ErmineJ: tool for functional analysis of gene expression data sets." BMC Bioinformatics **6**: 269.

Li, J., H. Jiang and W. H. Wong (2010). "Modeling non-uniformity in short-read rates in RNA-Seq data." Genome Biol **11**(5): R50.

Lima-Mendez, G., J. Van Helden, A. Toussaint and R. Leplae (2008). "Prophinder: a computational tool for prophage prediction in prokaryotic genomes." Bioinformatics **24**(6): 863-865.

Lin, B., J. Wang and Y. Cheng (2008). "Recent Patents and Advances in the Next-Generation Sequencing Technologies." Recent Pat Biomed Eng **2008**(1): 60-67.

Lodewyckx, C., J. Vangronsveld, F. Porteous, E. R. B. Moore, S. Taghavi, M. Mergeay and D. v. d. Lelie (2002). "Endophytic bacteria and their potential applications." Critical Reviews in Plant Sciences(21): 583-606.

Loehlin, J. C. (2004). Latent variable models: an introduction to factor, path, and structural equation analysis. Mahwah, Lawrence Erlbaum Associates.

Majdalani, N. and S. Gottesman (2005). "The Rcs phosphorelay: a complex signal transduction system." Annu Rev Microbiol **59**: 379-405.

Man, M. Z., X. Wang and Y. Wang (2000). "POWER_SAGE: comparing statistical tests for SAGE experiments." Bioinformatics **16**(11): 953-959.

Mansmann, U. and R. Meister (2005). "Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach." Methods Inf Med **44**(3): 449-453.

Mardis, E. R. (2008). "The impact of next-generation sequencing technology on genetics." Trends Genet **24**(3): 133-141.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley and J. M. Rothberg (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

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

Marks, D. J. (2011). "Defective innate immunity in inflammatory bowel disease: a Crohn's disease exclusivity?" Curr Opin Gastroenterol **27**(4): 328-334.

Mergeay, M., S. Monchy, P. Janssen, R. Van Houdt and N. Leys (2009). Megaplasmids in *Cupriavidus* genus and metal resistance. Berlin, Springer.

Meyer, F., A. Goesmann, A. C. McHardy, D. Bartels, T. Bekel, J. Clausen, J. Kalinowski, B. Linke, O. Rupp, R. Giegerich and A. Puhler (2003). "GenDB--an open source genome annotation system for prokaryote genomes." Nucleic Acids Res **31**(8): 2187-2195.

Misaghi, I. J. and C. R. Donndelinger (1990). "Endophytic bacteria in symptom free cotton plants." Phytopathology(80): 808-811.

Montgomery, S. B., M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo and E. T. Dermitzakis (2010). "Transcriptome genetics using second generation sequencing in a Caucasian population." Nature **464**(7289): 773-777.

Mootha, V. K., C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler and L. C. Groop (2003). "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes." Nat Genet **34**(3): 267-273.

Morris, J. N., J. W. Marr and D. G. Clayton (1977). "Diet and heart: a postscript." Br Med J **2**(6098): 1307-1314.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.

Nevitt, J. and G. R. Hancock (2001). "Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling." Structural Equation Modeling(8): 353-377.

Nossa, C. W., W. E. Oberdorf, L. Yang, J. A. Aas, B. J. Paster, T. Z. Desantis, E. L. Brodie, D. Malamud, M. A. Poles and Z. Pei (2010). "Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome." World J Gastroenterol **16**(33): 4135-4144.

Oshlack, A. and M. J. Wakefield (2009). "Transcript length bias in RNA-seq data confounds systems biology." Biol Direct **4**: 14.

Ouyang, Z., Q. Zhou and W. H. Wong (2009). "ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells." Proc Natl Acad Sci U S A **106**(51): 21521-21526.

Overbeek, R., N. Larsen, T. Walunas, M. D'Souza, G. Pusch, E. Selkov, Jr., K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, A. Bhattacharyya, H. Burd, W. Gardner, P. Hanke, V. Kapatral, N. Mikhailova, O. Vasieva, A. Osterman, V. Vonstein, M. Fonstein, N. Ivanova and N. Kyrpides (2003). "The ERGO genome analysis and discovery system." Nucleic Acids Res **31**(1): 164-171.

Parts, L., O. Stegle, J. Winn and R. Durbin (2011). "Joint genetic analysis of gene expression data with inferred cellular phenotypes." PLoS Genet **7**(1): e1001276.

Pavlidis, P., J. Qin, V. Arango, J. J. Mann and E. Sibille (2004). "Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex." Neurochem Res **29**(6): 1213-1222.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference, Cambridge University Press.

Pehkonen, P., G. Wong and P. Toronen (2005). "Theme discovery from gene lists for identification and viewing of multiple functional groups." BMC Bioinformatics **6**: 162.

Peng, J., P. Wang, N. Zhou and J. Zhu (2009). "Partial Correlation Estimation by Joint Sparse Regression Models." J Am Stat Assoc **104**(486): 735-746.

Podolsky, D. K. (2002). "Inflammatory bowel disease." N Engl J Med **347**(6): 417-429.

Pomerance, A., E. Ott, M. Girvan and W. Losert (2009). "The effect of network topology on the stability of discrete state models of genetic control." Proc Natl Acad Sci U S A **106**(20): 8209-8214.

Pratt, L. A. and R. Kolter (1998). "Genetic analysis of Escherichia coli biofilm formation: roles of flagella, motility, chemotaxis and type I pili." Mol Microbiol **30**(2): 285-293.

Prigent-Combaret, C., O. Vidal, C. Dorel and P. Lejeune (1999). "Abiotic surface sensing and biofilm-dependent regulation of gene expression in Escherichia coli." J Bacteriol **181**(19): 5993-6002.

Rabe-Hesketh, S., A. Pickles and A. Skrondal (2003). "Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation." Statistical Modelling **3**(3): 215-232.

Rabe-Hesketh, S., A. Skrondal and A. Pickles (2003). "Maximum likelihood estimation of generalized linear models with covariate measurement error." The Stata Journal **3**(4): 386-411.

Reid, S. J. and V. R. Abratt (2005). "Sucrose utilisation in bacteria: genetic organisation and regulation." Appl Microbiol Biotechnol **67**(3): 312-321.

Reisner, A., J. A. Haagensen, M. A. Schembri, E. L. Zechner and S. Molin (2003). "Development and maturation of Escherichia coli K-12 biofilms." Mol Microbiol **48**(4): 933-946.

Ren, D., L. A. Bedzyk, S. M. Thomas, R. W. Ye and T. K. Wood (2004). "Gene expression in Escherichia coli biofilms." Appl Microbiol Biotechnol **64**(4): 515-524.

Renz, H., E. von Mutius, P. Brandtzaeg, W. O. Cookson, I. B. Autenrieth and D. Haller (2011). "Gene-environment interactions in chronic inflammatory disease." Nat Immunol **12**(4): 273-277.

Rinttila, T., A. Kassinen, E. Malinen, L. Krogius and A. Palva (2004). "Development of an extensive set of 16S rDNA-targeted primers for quantification of pathogenic and indigenous bacteria in faecal samples by real-time PCR." J Appl Microbiol **97**(6): 1166-1177.

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

Robinson, M. D. and G. K. Smyth (2007). "Moderated statistical tests for assessing differences in tag abundance." Bioinformatics **23**(21): 2881-2887.

Robinson, M. D. and G. K. Smyth (2008). "Small-sample estimation of negative binomial dispersion, with applications to SAGE data." Biostatistics **9**(2): 321-332.

Roesch, L. F., R. R. Fulthorpe, A. Riva, G. Casella, A. K. Hadwin, A. D. Kent, S. H. Daroub, F. A. Camargo, W. G. Farmerie and E. W. Triplett (2007). "Pyrosequencing enumerates and contrasts soil microbial diversity." ISME J **1**(4): 283-290.

Rosey, A. L., E. Abachin, G. Quesnes, C. Cadilhac, Z. Pejin, C. Glorion, P. Berche and A. Ferroni (2007). "Development of a broad-range 16S rDNA real-time PCR for the diagnosis of septic arthritis in children." J Microbiol Methods **68**(1): 88-93.

Ryan, R. P., F. J. Vorholter, N. Potnis, J. B. Jones, M. A. Van Sluys, A. J. Bogdanove and J. M. Dow (2011). "Pathogenomics of Xanthomonas: understanding bacterium-plant interactions." Nat Rev Microbiol **9**(5): 344-355.

Sammel, M., X. Lin and L. Ryan (1999). "Multivariate linear mixed models for multiple outcomes." Stat Med **18**(17-18): 2479-2492.

Sammel, M., Ryan LM and L. JM. (1997). "Latent Variable Models for Mixed Discrete and Continuous Outcomes." Journal of the Royal Statistical Society Series B-Statistical Methodology **59**: 651-658.

Sanger, F. and A. R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase." J Mol Biol **94**(3): 441-448.

Sauer, U., M. Heinemann and N. Zamboni (2007). "Genetics. Getting closer to the whole picture." Science **316**(5824): 550-551.

Schafer, J. and K. Strimmer (2007). " A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics." Statistical Applications in Genetics and Molecular Biology **4**(1): Article 32.

Schembri, M. A., K. Kjaergaard and P. Klemm (2003). "Global gene expression in Escherichia coli biofilms." Mol Microbiol **48**(1): 253-267.

Sharpe, K. (2010). Structural Equation Modeling for Mixed Designs. Applied Mathematics and Statistics. Stony Brook, State University of New York at Stony Brook.

Shi, L., R. G. Perkins, H. Fang and W. Tong (2008). "Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential." Current Opinion in Biotechnology **19**(1): 10-18.

Shipley, B. (2000). Cause and correlation in biology. New York, Cambridge University Press.

Simon, H. (1953). Causal ordering and identifiability. New York, Wiley.

Skrondal, A. and S. Rabe-Hesketh (2004). Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models, Chapman & Hall/CRC.

Smith, E. N. and L. Kruglyak (2008). "Gene-environment interaction in yeast gene expression." PLoS Biol **6**(4): e83.

Smoot, M. E., K. Ono, J. Ruscheinski, P. L. Wang and T. Ideker (2011). "Cytoscape 2.8: new features for data integration and network visualization." Bioinformatics **27**(3): 431-432.

Spear, G. T., M. Sikaroodi, M. R. Zariffard, A. L. Landay, A. L. French and P. M. Gillevet (2008). "Comparison of the diversity of the vaginal microbiota in HIV-infected and HIV-uninfected women with or without bacterial vaginosis." J Infect Dis **198**(8): 1131-1140.

Stanley, N. R., R. A. Britton, A. D. Grossman and B. A. Lazazzera (2003). "Identification of catabolite repression as a physiological regulator of biofilm formation by Bacillus subtilis by use of DNA microarrays." J Bacteriol **185**(6): 1951-1957.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-15550.

Suzuki, M. T. and S. J. Giovannoni (1996). "Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR." Appl Environ Microbiol **62**(2): 625-630.

Taghavi, S., C. Garafola, S. Monchy, L. Newman, A. Hoffman, N. Weyens, T. Barac, J. Vangronsveld and D. van der Lelie (2009). "Genome survey and characterization of endophytic bacteria exhibiting a beneficial effect on growth and development of poplar trees." Appl Environ Microbiol **75**(3): 748-757.

Taghavi, S., D. van der Lelie, A. Hoffman, Y. B. Zhang, M. D. Walla, J. Vangronsveld, L. Newman and S. Monchy (2010). "Genome sequence of the plant growth promoting endophytic bacterium *Enterobacter* sp. 638." PLoS Genet **6**(5): e1000943.

Tibshirani, R. (2006). "A simple method for assessing sample sizes in microarray experiments." BMC Bioinformatics **7**: 106.

Tomfohr, J., J. Lu and T. B. Kepler (2005). "Pathway level analysis of gene expression using singular value decomposition." BMC Bioinformatics **6**: 225.

Tusher, V. G., R. Tibshirani and G. Chu (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-5121.

Vallenet, D., S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli and C. Medigue (2009). "MicroScope: a platform for microbial genome annotation and comparative genomics." Database (Oxford) **2009**: bap021.

Vallenet, D., L. Labarre, Z. Rouy, V. Barbe, S. Bocs, S. Cruveiller, A. Lajus, G. Pascal, C. Scarpelli and C. Medigue (2006). "MaGe: a microbial genome annotation system supported by synteny results." Nucleic Acids Res **34**(1): 53-65.

van den Donk, M., M. van Engeland, L. Pellis, B. J. Witteman, F. J. Kok, J. Keijer and E. Kampman (2007). "Dietary folate intake in combination with MTHFR C677T genotype and promoter methylation of tumor suppressor and DNA repair genes in sporadic colorectal adenomas." Cancer Epidemiol Biomarkers Prev **16**(2): 327-333.

Viswanathan, G. A., J. Seto, S. Patil, G. Nudelman and S. C. Sealfon (2008). "Getting started in biological pathway construction and analysis." PLoS Comput Biol **4**(2): e16.

Walsh, U. F., J. P. Morrissey and F. O'Gara (2001). "Pseudomonas for biocontrol of phytopathogens: from functional genomics to commercial exploitation." Curr Opin Biotechnol **12**(3): 289-295.

Wang, P., D. L. Chao and L. Hsu (2011). "Learning oncogenic pathways from binary genomic instability data." Biometrics **67**(1): 164-173.

Weisburg, W. G., S. M. Barns, D. A. Pelletier and D. J. Lane (1991). "16S ribosomal DNA amplification for phylogenetic study." J Bacteriol **173**(2): 697-703.

Whiteley, M., M. G. Bangera, R. E. Bumgarner, M. R. Parsek, G. M. Teitzel, S. Lory and E. P. Greenberg (2001). "Gene expression in Pseudomonas aeruginosa biofilms." Nature **413**(6858): 860-864.

Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics. Chichester, U.K., Wiley.

Wood, T. K. (2009). "Insights on *Escherichia coli* biofilm formation and inhibition from whole-transcriptome profiling." Environ Microbiol **11**(1): 1-15.

Wood, T. K., A. F. Gonzalez Barrios, M. Herzberg and J. Lee (2006). "Motility influences biofilm architecture in Escherichia coli." Appl Microbiol Biotechnol **72**(2): 361-367.

Wright, S. S. (1921). "Correlation and causation." Journal of Agricultural Research **20**: 557-585.

Wu, X., S. Monchy, S. Taghavi, W. Zhu, J. Ramos and D. van der Lelie (2011). "Comparative genomics and functional analysis of niche-specific adaptation in Pseudomonas putida." FEMS Microbiol Rev **35**(2): 299-323.

Wu, X., K. Sharpe, T. Zhang, H. Chen, E. Li, S. Taghavi, D. van der Lelie and W. Zhu (2011). Comparative genetic pathway analysis using structural equation modeling. IEEE International Conference on Computational Advances in Bio and medical Sciences Orlando.

Xiong, M., J. Li and X. Fang (2004). "Identification of genetic networks." Genetics **166**(2): 1037-1052.

Yi, M., J. D. Horton, J. C. Cohen, H. H. Hobbs and R. M. Stephens (2006). "WholePathwayScope: a comprehensive pathway-based analysis tool for high-throughput data." BMC Bioinformatics **7**: 30.

Young, M. D., M. J. Wakefield, G. K. Smyth and A. Oshlack (2010). "Gene ontology analysis for RNA-seq: accounting for selection bias." Genome Biol **11**(2): R14.

Yu, H., N. M. Luscombe, J. Qian and M. Gerstein (2003). "Genomic analysis of gene expression relationships in transcriptional regulatory networks." Trends Genet **19**(8): 422-427.

Yung, Y. F. and P. Bentler (1996). Bootstrapping techniques in analysis of mean and covariance structures. Mahwah, NJ, Lawrence Erlbaum Associates.

Zeeberg, B. R., W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett and J. N. Weinstein (2003). "GoMiner: a resource for biological interpretation of genomic and proteomic data." Genome Biol **4**(4): R28.

Zemanick, E. T., B. D. Wagner, S. D. Sagel, M. J. Stevens, F. J. Accurso and J. K. Harris (2010). "Reliability of quantitative real-time PCR for bacterial detection in cystic fibrosis airway specimens." PLoS One **5**(11): e15101.

Zhang, B., D. Schmoyer, S. Kirov and J. Snoddy (2004). "GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies." BMC Bioinformatics **5**: 16.

Zhang, T., R. DeSimone, H. Chen, C. Hamm and J. Yuan (2011). Cluster analysis of genome-wide expression differences in disease-unaffected ileal mucosa in inflammatory bowel diseases. IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences, Orlando, FL.

Zhang, X. S., R. Garcia-Contreras and T. K. Wood (2008). "*Escherichia coli* transcription factor YncC (McbR) regulates colanic acid and biofilm formation by repressing expression of periplasmic protein YbiM (McbA)." ISME J **2**(6): 615-631.

Zhu, H. and M. Snyder (2002). ""Omic" approaches for unraveling signaling networks." Curr Opin Cell Biol **14**(2): 173-179.

Zoetendal, E. G., A. D. L. Akkermans and W. M. De Vos (1998). "Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria." Applied and Environmental Microbiology **64**(10): 3854-3859.