

<http://researchspace.auckland.ac.nz>

ResearchSpace@Auckland

Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage.

<http://researchspace.auckland.ac.nz/feedback>

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the [Library Thesis Consent Form](#) and [Deposit Licence](#).

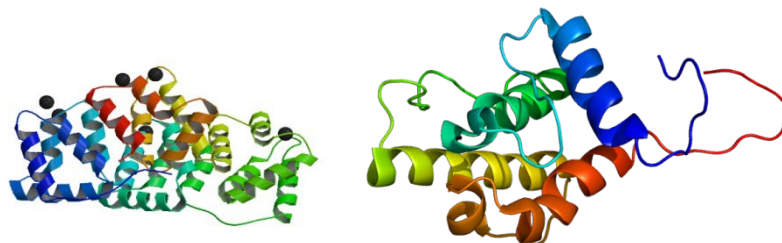
Statistical methods in clinical proteomic studies

2007-2014

-A protein concerto-

PhD candidate: Irene Suilan Zeng

**A thesis submitted in fulfilment of
the requirements for the degree of Doctor of Philosophy in Statistics,
The University of Auckland,
2014**



Abstract

Clinical proteomics is a subject of systems biology that investigates large numbers of protein biomarkers associated with human disease. Like the other “omics”, proteomics use systems biology techniques to identify proteome-wide markers simultaneously. Unlike genomics that has been established for decades, proteomics is still in its infancy. The current biotechnologies have limited power to discover all the existing 20,000s proteins from the human body. Biologists have not been able to understand the molecular functions of lots of those identified proteins. Statistical techniques become essential in proteomics research because clinical proteomic studies generate a large amount of quantitative information through systems biology techniques to investigate proteins’ molecular activities. The complexities of clinical study and proteomic experiments also require statistical inputs to achieve valid and unbiased inferences.

This PhD research firstly proposed a new method to assess the reproducibility in clinical proteomic studies when a new device or new tissue is being used for a proteomic experiment. The reproducibility assessment utilizes a dimensions reduction technique and permutation method to extend the evaluation from a single feature scale to a proteome-wise scale. It secondly proposed algorithms to optimize the study design for a multiple stage study which bridges the biomarker discovery to clinical utility. The optimal design algorithms utilized a hybrid simulated annealing approach to finding the design parameters that achieve a maximal number of discoveries, under the constraints of cost and number of false discoveries. These algorithms were realized via a R package named “proteomicdesign”. Finally, a multivariate multilevel model has been proposed for the analysis of proteomic data. The non-random missing data presented in proteomic mass spectrometric experiments were estimated under a Bayesian framework. The proposed analytical method was tested in a simulated study and used in two real life clinical proteomic studies.

Acknowledgements

Like forming a musical team, our quartet started with Sharon Browning, Ralph Stewart, Kathy Ruggiero and me in 2007. Sharon bravely accepted to take the supervision role for the challenging research on proteomics and worked with a student she barely knew. Ralph worked with me in the hospital and influenced on me as a mentor, adviser and clinical supervisor. Through giving him statistical supports in his research proposals and also for the other experienced researchers like him, I inadvertently learnt how to write a research proposal by reading their works. Sharon and I always had delightful meeting on a late afternoon where I have started to learn the genomic association study, permutation, MA plot from her. We decided to publish the first paper in order to receive comments from the international reviewers and I could learn how to write a paper as a first author. In 2010, Sharon took a position at the University of Washington with family back to Seattle. Coincidentally, Professor Thomas Lumley joined the statistics department from the University of Washington Seattle at the same year. He replaced Sharon to lead the team. His joining was as if a gift from the science gods.

Kathy, Thomas and I then started our joyful discussions in our fortnightly meetings in Thomas' office. Like a quartet, when we played our instruments together, we enjoyed the different inputs from each other. Kathy knew about proteomics experiments, having her like adding the harmonic in the music. Thomas is very knowledgeable. I needed to do loads of reading after every meeting. He also shared the beauty of the data with us, and taught me how to make a good poster and presentation. It was a pity that Ralph could not join us simultaneously as he had busy clinical duties in the hospital, but he had made great efforts to come to my presentations and meet with Thomas. I took the message from different meetings with Thomas and Kathy and then meeting with Ralph to discuss our work. His inner mathematician with his clinical research knowledge always helped me. I remembered one day while I was on the way to the hospital, I had been thinking that the mass of the molecule may be related to the intensity. After talking to him, I realized that it maybe the mass to charge ratio that actually had the influences. At the end, this variable was proved to be a very important covariate in the two proteomic case studies, for the first time.

In the research journey, I also could not be able to finish it without the supports of biochemists Martin Middledith and See-tarn Woon. Martin worked in the Centre for

Genomics and proteomics in the school of biological science, See-tarn worked in labPLUS in Auckland hospital. They shared their knowledge in proteomic science with me and we learnt from each other in clinical proteomic research. I appreciated their patience, robustness, and being open minded in scientific research. Their main inputs in the laboratory methods are used in the two cases studies.

As a mother and a busy biostatistician working in the hospital, these 7 years of PhD research have not been easy. **I would like to give my gratitude to Ralph, my first teacher of medical research; Sharon, my first teacher of statistical research; Kathy, my teacher of statistical writing; Thomas, my academic father who directed me to a pathway of a scholar.** My gratitude is also given to colleagues (Lisa Davies, Patricia Loft, Cathy Anderson, Christin Choomorasamy, Sam Everitte and Kathryn Askerlund); the administration team, the post graduate advisory team, Professor Chris Triggs of the statistics department and my family.

Because of these wonderful people, I was able to maintain my physical and mental strengths to finish this thesis.

My acknowledgments are also given to A+ trust, Green Lane Research Education which provided the funding to the two proteomic studies; the eResearch facility and their staffs that provided the computing cluster facility and consultation; Dr Mia Julig, Dr Jocelyn Benatar, Dr Patrick Gladding and Dr Rohan Ameratunga who have collaborated with me in the proteomic projects. Lastly I would also like to acknowledge Howard Gilbert who proofread my thesis and gave me useful suggestions.

Like a musician who hopes to make good music for the audiences and to create a new piece of works, I hope my PhD thesis contributes to the literatures in both proteomic and statistical sciences.

Table of Contents

Chapter 1. Introduction and overall review	1
1.1. Motivation.....	1
1.2. Overview	2
1.3. General review on mass spectrometry	5
1.4. General review on statistical methods in proteomic studies	11
Chapter 2. A multi-feature reproducibility assessment of mass spectral data in clinical proteomic studies	13
2.1. Introduction.....	14
2.2. Method	18
2.3. Results	22
2.4. Discussion.....	26
Chapter 3. Two optimization strategies of multi-stage design in clinical proteomic studies	35
3.1. Introduction and motivation.....	35
3.2. Statistical strategies in the three-stage design	39
3.3. Case studies.....	54
3.4. A comparison between using grouping information and not using grouping information Discussion	62
3.5. Discussion.....	63
3.6. Software	65

Chapter 4. A multivariate multilevel model for analysing clinical proteomics data with non-random missingness	73
4.1. Introduction.....	74
4.2. The analytical methods.....	81
4.3. The missing mechanism for the iTRAQ data –a Bayesian approach	92
4.4. A simulation study	111
Chapter 5. A cardiac proteomics study-case study I.....	132
5.1. Description of the study.....	132
5.2. The laboratory methods (a brief summary of the clinical laboratory sample preparation and the iTRAQ experiment in the University lab)	132
5.3. The analytical methods of the study.....	135
5.4. Discussion.....	146
Chapter 6. An immunology proteomic study-Case study II.....	156
6.1. Description of the study.....	156
6.2. The laboratory methods (a brief summary of the clinical laboratory sample preparation and the iTRAQ experiment in the University lab)	156
6.3. The reproducibility assessment	159
6.4. The analytical methods for the discovery section	160
6.5. Results	167
6.6. Discussion.....	173
Chapter 7. Discussion	194
7.1. Overall review	194
7.2. Reproducibility assessment for high throughput devices	195

7.3. Multi-stage design is the necessary strategy in clinical proteomic study.....	201
7.4. Using Bayesian methods will be an advance for analysing proteomics studies.....	206
7.5 Conclusion	210
Bibliography	212

Co-Authorship Form

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 2 is extracted from a published paper:

A Multi-feature Reproducibility Assessment of Mass Spectral Data in Clinical Proteomic Studies

Clinical Proteomics 5(3):170-177. DOI:10.1007/s12014-009-9039-y

Nature of contribution by PhD candidate

Initiated the ideas, developed the method, wrote the paper and revised the chapter

Extent of contribution by PhD candidate (%)

95%

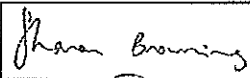



CO-AUTHORS

Name	Nature of Contribution
Sharon Browning	Supervised the development of the statistical methods, reviewed the writing and edited the paper.
Patrick Gladding	Initiated the cardiac proteomic study with RS,IZ,MJ and MM, being the Principal investigator of the cardiac proteomic study.
Mia Julig	Being a co-investigator of the cardiac proteomic study, conducted part of the biochemist analysis for the cardiac proteomic study, provided feedbacks and communications to the first author, edited the paper.
Martin Middleditch	Being a co-investigator of the cardiac proteomic study, conducted part of the biochemist analysis for the cardiac proteomic study, provided feedbacks and communications to the first author, edited the paper.
Ralph Stewart	Supervised the research method, initiated the cardiac proteomic research idea with PG and IZ, reviewed the writing and edited the paper.

Certification by Co-Authors

The undersigned hereby certify that:

- ❖ the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- ❖ in cases where the PhD candidate was the lead author of the work that the candidate wrote the text.

Name	Signature	Date
Sharon Browning		24/03/2014
Patrick Gladding		Click here 27/3/14
Mia Julig		Click here 3/4/14
Martin Middleditch		Click here 03/4/14

Ralph Stewart

Bruner

Click here *7/7/14*

Click here

Co-Authorship Form

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 3 was extracted from a published paper:

Two optimization strategies of multi-stage design in clinical proteomic studies

Statistical applications in genetics and molecular biology, May, 2013

Nature of contribution by PhD candidate

Initiated the ideas, developed the method, wrote the paper and revised the chapter

Extent of contribution by PhD candidate (%)

80%

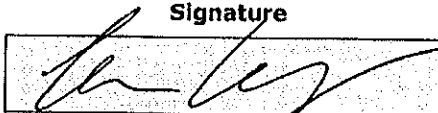

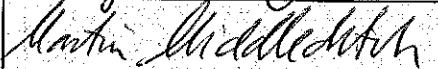
CO-AUTHORS

Name	Nature of Contribution
Thomas Lumley	Supervised the development of the statistical methods and the R package for the proteomic design, reviewed the writing and edited the paper
Kathy Ruggiero	Co-supervised the proteomic experimental design, reviewed the writing and edited the paper
Martin Middleditch	Being a co-investigator of the immunology proteomic study, conducted the proteomic analysis for the immunology proteomic study, provided feedbacks and communications to the first author, edited the immunology case study in the paper.
See-Tarn Woon	Being a collaborator of the immunology proteomic study, initiated the immunology study with IZ and MM, conducted the lab preparation for the immunology proteomic study, provided feedbacks and communications to the first author, edited the immunology case study in the paper.
Ralph A.H. Stewart	Supervised the research method, provided support for the immunology proteomic research with STW and IZ, reviewed the writing and edited the background and discussion of the paper.

Certification by Co-Authors

The undersigned hereby certify that:

- ❖ the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- ❖ In cases where the PhD candidate was the lead author of the work that the candidate wrote the text.

Name	Signature	Date
Thomas Lumley		Click here 1/4/14
Kathy Ruggiero		Click here 1/4/14
Martin Middleditch		Click here 3/4/14

See-Tarn Woon

Ralph A.H. Stewart

Handwritten signature

Handwritten signature

Click here 7/4/14

Click here 7/4/14

Click here

CHAPTER 1

Introduction and overall review

1.1. Motivation

In the past five decades, scientists have made tremendous progress in understanding the human biological system at a genetic and genomics level. The current genetic and genomic literature extensively reports studies about understanding, treating and even preventing certain diseases (especially cancer). However, genes are only the “recipes” of the cell: the proteins encoded by the genes are ultimately the biological units driving the pathology and physiology of the disease (Annual report of National Cancer Institute 2007). The protein expression changes reflect the mutation from DNA, the therapeutic effects from drugs and the environmental changes in the human body. For these reasons, using advanced proteomics technology to study proteins systematically will improve mankind’s understandings of human disease and provide information at the molecular level for the prevention and invention of cures for the disease (Greef et al., 2007).

Compared to genomics, proteomics is still in its infancy. The modern biotechnologies enable us to discover thousands of proteins simultaneously, but the high cost and the complexities of protein discovery and quantification remain the limitations for the whole-proteome study. Given the complexities of the proteomic studies, a lack of statistical methods to guide the study design, to manage and analyze the high throughput proteomic data is one of the obstacles to advancing proteomics in clinical research and ultimately in practice. Few of the current proteomic studies reach the last clinical validation stage. This problem reflects the future demands on advanced statistical and computational methods for the proteomics discovery and validation in the coming decades. As a biostatistician working in medical research, under the direction of a clinical supervisor who has been proactively working in biomarker research, my visionary and research-oriented mind encouraged me to initiate this PhD study of statistical methods in clinical proteomics.

1.2. Overview

The National Cancer Institute in the United States is considered to be a pioneer of proteomic research. In 2007, it advocated standardizing the process from sampling collection to final data analysis in future proteomic research. In particular, clinical validation is considered to be the most important step in the entire process.

“We do not suffer a lack of reported cancer biomarkers: The literature reports upwards of 1,200 protein biomarkers, though very few of these have been validated, and even fewer have found their way into clinical practice. It has become increasingly clear that this dichotomy can be traced-in large part-to several levels of confounding variables.”

- John E, Niederhuber, M.D. Director, National Cancer Institute 2007.

According to this advocate, a five-year collaboration program CPTC (The clinical proteomic technology assessment for cancer), which aimed to identify, quantify and ultimately reduce sources of variability in the current proteomics workflows, was established. The long-term goals for the collaborative program are the generation of standard materials, including samples, antibodies, data, and protocols to be made available to the community for little or no cost. The first NCI biomarker discovery workflow suggested using defined proteomic platform performance characteristics (i.e. standard operating procedures and reference materials) at each step starting from sample collection up to data analysis of the biomarker discovery pipeline. The second workflow suggested is to use a three-stage process from unbiased discovery using 10's samples, to targeted verification using 100's samples and finally to clinical validation by using 1000's samples (Figure 1.1a).

This PhD research aims to investigate and establish statistical methods for clinical proteomic studies on the multi-dimensional scales. It comprises three parts: Part I) Reproducibility assessment for proteomics studies using mass spectrometry; Part II) Optimal design for a multistage clinical proteomic study; Part III) Analytical method to cope with laboratories and patients' variability for the proteome-wide data analysis, including methods for handling missing data.

1.2.1 Overview for part I: Multi-dimensional reproducibility assessment in clinical proteomic study

In this first section, a global reproducibility assessment method is proposed as an alternative to the traditional approach that only assesses one single feature (peak) in proteomic studies. The global reproducibility assessment method works on multiple features (peaks) of the high throughput data from the mass spectrometers. It utilizes principal component analysis to reduce the dimensions of the data. It then derives the Tracy-Widom distributed test statistics for the k largest Eigenvalues of the data matrix to distinguish the underlying structure of the high dimensional data from the noise. The proposed global assessment adopts the multivariate permutation method on the identified significant principal component subspaces, to assess if there are significant deviations between the mass spectrometry technical replicates, by using two proposed multivariate test statistics. Chapter 2 of this thesis embeds a paper describing the proposed method published in a proteomics journal. More detailed statistical reviews on the relevant methodologies can also be found in the discussion chapter 7.

1.2.2 Overview for part II: Multi-stage optimal design of proteomic study

In this second section, a multi-stage optimal design in proteomic study is investigated. The optimization design concept comes from traditional multistage sampling theory (Skol et al., 2007) and optimal experimental design theory. In the multi-stage sampling theory, n_j samples are selected from j specified clusters. The total number of clusters m and the total number of candidates n are chosen as to minimize the overall variance given a set budget and cost of the survey. m and n are also chosen as to maximize the inverse information matrix (fisher information matrix). In recent years, gene association studies adopt this similar idea to optimize the multi-stage design (mainly two stages) where genetic markers are selected from the first stage and validated in the second or later stages. The optimization criteria commonly are the overall cost (Moerkerke and Goetghebeur, 2008), the overall power (Zehetmayer et al., 2008; Kitamura et al., 2009) or the False Discover Rate (FDR)(Kitamura et al., 2009). In proteomic studies, the number of proteins identified from the current technology is much less than the number of identified genes. The number of markers is normally between several hundred to several thousand in proteomic studies compared to hundreds of thousands in a

genetic study. However, the design problem is similar. These studies have limited research budgets to test a large sample size for the identification, and the laboratory technology is expensive. Optimization in allocations of the number of screened candidates and number of samples is necessary. In the past decade, many genome-wide association studies showed that two-stage design is more cost-effective than one stage design (Jaya M. Satagopan and Elston, 2003; Zuo et al., 2008). This part of the PhD research aims to assess methods used in genetic studies and develop a multi-stage optimization method in proteomic studies. The optimization option being investigated is to maximize power when cost is a limit for a three-stage design.

Hyper simulated annealing algorithm is used to find the optimal solution for a simulated function and an approximated analytical function of the estimated number of true discoveries (power), under a three-stage design as outlined in NCI's workflow (figure 1.1a). The utilization of biological grouping information to assist the design is also assessed, and recommendations of when to use grouping information are given in this section.

Chapter 3 of this thesis presents a paper describing the proposed method published in "Journal of statistical application in genetic and molecular biology". More detailed statistical reviews on the relevant methodologies are also described in the discussion chapter 7.

1.2.3 Overview for part III: A multivariate multilevel model for analyzing clinical proteomic data with non-random missingness

In this third section, multivariate multilevel methods are investigated for the analysis of the data from clinical proteomic studies using the mass spectrometer. The high-throughput data from multiple runs of mass spectrometry experiments brings challenges to the data analysis. These challenges originate from the hierarchical levels of the quantification of protein abundance, the complexities of the clinical study and the experiment, the large amount of information, and the non-random missingness of the intensity data.

A multivariate multilevel model is proposed to analyze the hierarchical protein expression data, taking into account different types of variations from the experimental factors such as the physical features of the assays, the molecular feature of proteins, and the experimental effects. The identified non-random missingness of the protein expression data is proposed to be modeled under the Bayesian hierarchical framework. Two different posterior sampling

methods, Gibbs sampling and Hamiltonian Monte Carlo using No U Turn sampling are evaluated in this section.

In this section, a simulated study and two cases studies are included using the proposed analytical method. One case study evaluated the coronary proteomic profile changes after percutaneous coronary intervention in patients with ischemic heart disease. The other case study aims to evaluate the reproducibility for the proteomics analysis using the lymphocyte cells, and to identify if there are any outstanding cellular proteins markers for patients with Common Variable Immune Deficiency.

1.3. General review on mass spectrometry

1.3.1. History of proteomic development

The proteomics study discovers proteins and quantifies the expression of protein in live creatures under different conditions. Clinical proteomics identifies and validates disease related proteins from human tissues, mainly from patients' tissues. Healthy humans normally participate in the study as controls for comparison with patients. The original proteomic research began in 1970-1980. In the 1990s, Wilkins formalized it into a discipline and formally gave it the name "proteomics" as a term to represent "the protein complement of genome" (<http://www.proteome.org.au/History-of-Proteomics/default.aspx>). Although proteomics has become a mainstream discipline since this time, Leigh Anderson and others firstly investigated it using two-dimensional gel electrophoresis technology in the 1970s (L. Anderson, 2005). Anderson used this highly revolutionary technique to separate the proteins in blood and leukocyte. While the limitations in its reproducibility hindered its expansion, the introduction of immobilized PH gradient in 1980s improved its production and kept it as one of the important techniques for protein quantification. Meanwhile, mass spectrometry ionization technology was developed to quantify peptides, and this allowed protein identification and quantification to be performed on a larger scale. After the mid-1990s, mass spectrometry became the mainstream technique in systems biology for protein identification and quantification.

Systems biology is a systematic approach that integrates bio-analytical platforms with biostatistical and bioinformatic platforms to investigate complementary measurement modalities, it includes transcriptomics, proteomics and metabolomics. In life science, it is applied particularly for innovative drug discovery and development, and biomarker discovery (Greef et al., 2007). Proteomics belongs to systems biology. It is a part of network biology that enables us to study the proteins behaviors of an entire biological system (i.e. cell, subfraction of the cell, plasma, serum, etc.). The main application of proteomic will be identifying diseased related multiple proteins using the systematic quantification technologies (2D gel, mass spectrometry) and verifying the finding by a targeted approach. The systematic approach will allow the direct selection of optimal biomarker candidate proteins, skipping over the long laboratory process from non-systematic techniques. The targeted approaches will validate the results in a much larger size of diseased samples. The current commonly used targeted approaches are immunoassays and mass spectrometry for candidate protein quantification (L. Anderson, 2005). The former approach to using antibody arrays has the advantage of high sensitivity and specificity for quantifying the specific protein, but it is limited by the available antibody arrays for constituting the new marker. The latter approach serves to evaluate candidate biomarkers prior to the big investment of immunoassay (L. Anderson, 2005).

1.3.2 The systematic approach using different types of protein quantification by mass spectrometry (MS) ionization

Mass spectrometry is one of the most important physical methods in analytical chemistry today. An outstanding advantage of MS, compared with other molecular spectroscopies, is its high sensitivity for quantification in trace amounts of chemicals. A mass spectrometer is designed to perform three basic functions (Chapman, 1996):

1. Provide gas-phase ions from sample molecules. Methods for ionizations are electron ionization (EI), Chemical ionization (CI), Fast-Atom bombardment (FAB), Matrix-assisted Laser desorption/Ionization (MALDI), Electrospray (ES), Ion-spray Ionizations (ESI), and Atmospheric pressure chemical Ionization (APCI).
2. Separate the gas-phase ions according to their mass-to-charge ratio (m/z).
3. Detect and record the separated ions.

The process of using mass spectrometry to separate ions is called ionization. There are two types of ionization: MALDI and ESI. In MALDI, ionization is realized by transferring laser energy to make the analyte molecules charged and accelerated before entering into the analyzer. ESI is an atmospheric ionization technique, where the ions are emitted from a droplet into the gas phase under atmospheric pressure. There are four types of mass analyzer commonly used and summarized as followed (Chapman, 1996):

MS-MS is a two-stage mass spectrometry. At the first stage, selected particular ions undertake collision induced dissociation (CID). At the second stage, the resultant fragment ions are subsequently measured using a second mass analyzer.

LC-MS/MS is liquid chromatography combined with a two-stage mass spectrometry. Liquid chromatography is an analytical chemistry technique that enables the separation of different compounds (i.e. peptides) from complex samples and thus assists the protein identifications (Palagi et al., 2007). The LC-MS/MS method combines reversed-phase high-pressure liquid chromatographic separation with electrospray ionization (ESI) in two-stage mass spectrometry. This technique known as peptide mapping separates and provides molecule weight information on the peptides resulting from digestion of the protein.

MALDI –TOF MS uses a time of flight (TOF) spectrometer together with MALDI (Matrix Assist Laser Desorption/Ionization) technique. It absorbs energy at the laser wavelength and isolates analyte molecules within some form of solid solution. It is easy to use and is considered as a good solution for clinical diagnosis due to its high automation, high throughput and better tolerant to salt and impurity samples than ESI (Chapman, 1996; Palmblad, 2009).

SELDI-TOF MS is another laser ionization technique combined with a time of flight analyzer, which SELDI stands for Surface Enhance Laser Desorption Ionization. It is similar to MALDI MS, with protein adsorption, partition, electrostatic interaction or affinity chromatography on a solid-phase protein chip surface. The laser ionizer samples have been co-crystallized with a matrix on a target surface. Unlike the MALDI MS, the protein chip chromatographic surfaces in SELDI are uniquely designed to retain proteins from complex mixtures according to their properties. SELDI MS can be used for targeted study (Issaq et al., 2003).

Among the different types of mass spectrometers, the most common strategies for protein identification are through MALDI-TOF MS and Tandem MS/MS. When identification using MALDI-TOF, it is the first step to generate peptide mass fingerprints (PMF's) for the enzymatically digested protein samples. However, not all proteins can be identified directly by PMF's, for instances, when the protein sequence does not present in the database or the spectrum only contained limited number of peptides. Additionally, MALDI-TOF is less effective for the analysis of complex protein mixtures as only the most abundant proteins are identified. Small proteins (20kda or less) may also prove difficult to analyze as these tend to generate fewer appropriately sized tryptic peptides for matching than big proteins do. When identification uses tandem MS/MS, the separation elements are physically separated and distinct. The process has multiple steps over time (Chapman, 1996).

The emphasis of proteomics is changing from a misfocussing high-throughput approach where the results are of limited value, to a more focusing approach that involves detailed analyses of the protein samples. The proteins of interest can be enriched by a cell fractionation method, and/or an affinity based protein purification strategy (Chapman, 1996).

1.3.3 The candidate approach

After the panel of candidate proteins is identified from the high-throughput mass spectrometer, a candidate or target approach is employed to verify and validate the identification. The targeted approach, which emerged as using immunoassay to identify a disease-associated marker in 1950, has a longer history than the systematic approach. It has produced most of the protein markers for diagnosis now (L. Anderson, 2005). Further laboratory improvement in the antibody specificity enables multiplex proteins to be tested in the immunoassay one at a time. In the 1990s, the Targeting finger print MS approach emerged as an alternative candidate approach that saved cost of producing a large amount of new immunoassay. It has been employed before a large amount of investment in the new antibody and immunoassay.

The two aforementioned candidate approaches, multiplex antibody and targeting finger print MS are considered to be an optimal strategy in the verification and validation stage of a clinical proteomic study.

Figure 1.1a Assessing the performance of Key Process Steps in the Candidate Biomarker Pipeline. This figure is published on the NCI website and recreated by the author. It illustrates a proposed multi-stage process in proteomic research with features and sample size suggested in each stage.

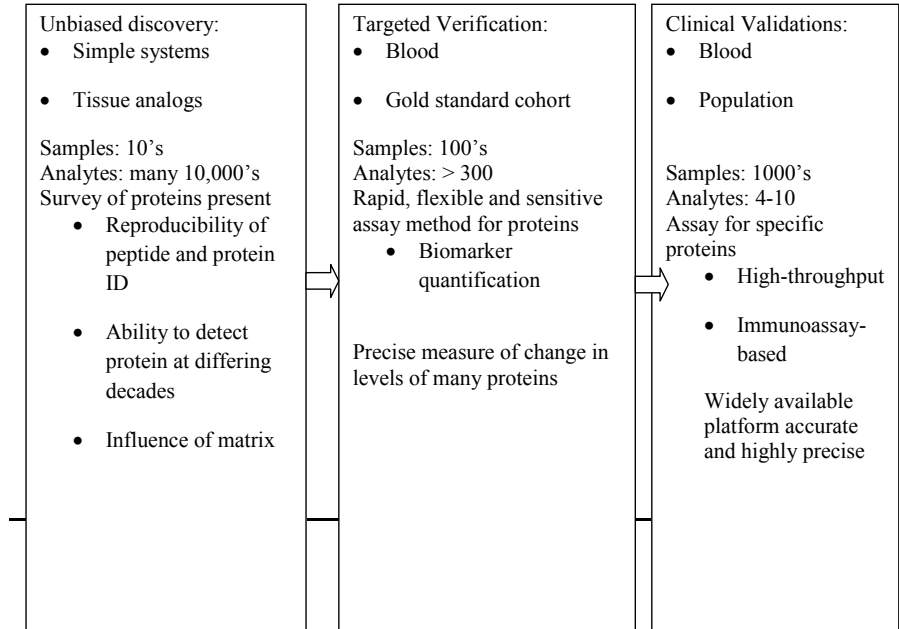


Figure 1.1b An optimized multistage design in clinical proteomic research proposed in this research

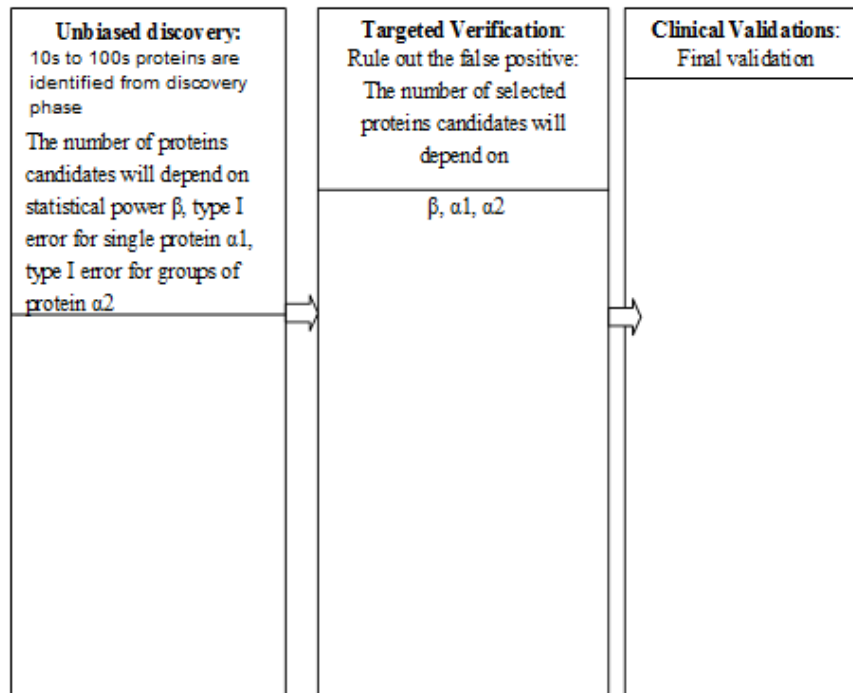
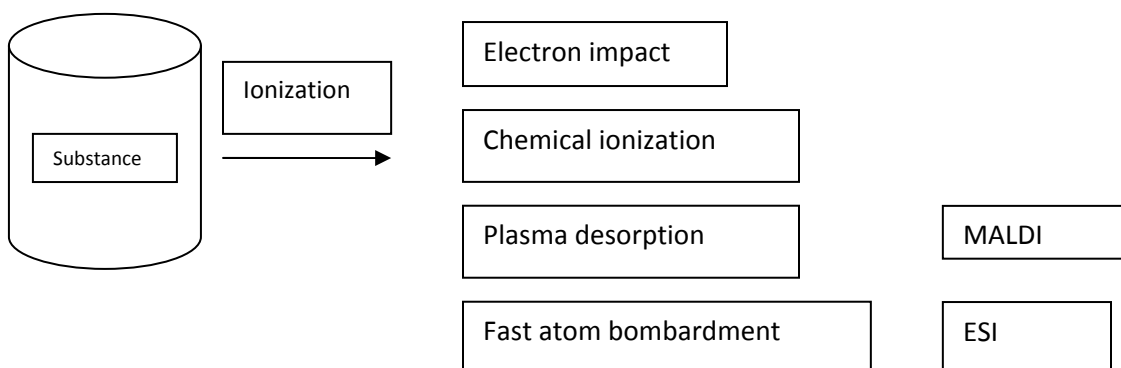


Figure 1.2 Ionization process



1.4. General review on statistical methods in proteomic studies

At the beginning of this PhD research, there were not many publications available discussing the statistical methods for clinical proteomic studies and proteomic studies..

Reproducibility was considered to be a potential issue in late 1990s–early 2000 when mass spectrometry assay were not stabilized yet. A couple of papers discussed how to use advanced statistical techniques to assess the reproducibility of the mass spectrometer data. One paper is from Semmes (2005), and the other paper is from Pelikan and Bigbee (2007). In Semmes’ paper, a decision algorithm used boosting logistic regression and boosting decision tree were introduced in the reproducibility assessment of the inter-laboratory platform outputs from surface-enhanced laser desorption (SELDI) instruments and assays. Boosting logistic regression and boosting decision tree are two machine learning classifiers used in bioinformatics. Pelikan and Bigbee introduced a Euclidian distance measure for calculating the similarity between the pair spectra. They also proposed a discriminating score to measure the dissimilarity between the spectra from disease cases and healthy controls. The dissimilarity measure is an index to assess the reproducibility in the classifications of cases and controls in the replicated mass spectrometry analysis.

Reviewing methods in the multi-stage study designs, it is found that although there are many publications in the literature discussing multi-stage design in gene association studies, there are not many for proteomic studies. In the literature of multi-stage gene association studies, Satagopan and Zuo published several papers in multi-stage design in the early 2000s (J. M. Satagopan et al., 2002; J. M. Satagopan et al., 2004; Zou et al., 2006). In one of these papers, Jaya M. Satagopan and Elston (2003) proposed a cost-optimized method utilizing the Monte Carlo grid search to find the optimal solution for a two-stage genotyping association study under the constraints of an overall type I error rate and statistical power. Zuo et al. (2008) also proposed an optimization method for maximizing the statistical power given the total costs. In Zuo's paper, a mixed integer nonlinear program (MINLP) was introduced to solve the numerical integration problem.

In the area of statistical methods for proteomic data analysis, Hill et al. (2008) and Oberg and Mahoney (2012) are the early advocates of using analysis of variances (ANOVA) as an alternative to the protein ratio approach for analyzing data from mass spectrometry experiments. They proposed using ANOVA to estimate the differences in protein abundances between disease and healthy subjects, and to take into account variances brought into the intensity data from the mass spectrometry experiments. Luo et al. (2009) recently suggested using the Markov Chain Monte Carlo method to draw inferences from the estimates of the ANOVA model, which also takes into account the protein abundances related missingness of the data. Luo's method has similarities to the analytical method proposed in chapter 4 of this thesis. Nevertheless, the analytical method presented in chapter 4 takes further improvements by introducing a multivariate multilevel structure and by including the instrumental influences of the mass spectrometer in a Bayesian model.

More literature reviews are given in the corresponding chapter for each topic.

CHAPTER 2

A multi-feature reproducibility assessment of mass spectral data in clinical proteomic studies

Abstract

Background

Use of mass spectrometry to investigate disease-associated proteins among thousands of candidates simultaneously creates challenges with the evaluation of operational and biological variation. Traditional statistical methods, which evaluate reproducibility of a single feature, are likely to provide an inadequate assessment of reproducibility. This chapter proposes a systematic approach for evaluation of the global reproducibility of multi-dimensional mass spectral data at the post- identification stage.

Methods

The proposed systematic approach combines dimensional reduction and permutation methods to assess and summarize the reproducibility. First, principal component analysis is applied to the mean quantities from identified features of the replicated samples. An eigenvalue test is used to identify the number of significant principal components that reflect the underlying correlation pattern of the multiple features. Second, a simulation-based multivariate permutation test is applied to the resultant principal components scores. As the byproduct of the analysis, a modified form of Bland Altman or MA Plot is produced to visualize the discordance among the replicates on the projected principal components subspaces.

Results

Application of this method to data from both a cardiac LC-MS/MS experiment with iTRAQ labeling and simulation experiments derived from an ovarian cancer SELDI-MS experiment demonstrate that the proposed global reproducibility test is sensitive to the simulated systematic bias when the sample size is above 15. The two proposed test statistics (max t statistics and a sign score statistic) for the permutation tests are shown to be reliable.

Conclusion

The methodology presented in this chapter provides a systematic approach for global measurement of reproducibility in clinical proteomic studies.

2.1. Introduction

2.1.1 Motivation and general review in the reproducibility assessments for proteomics studies

Mass spectrometry and liquid chromatography are standard tools used to profile and quantify thousands of proteins simultaneously in clinical proteome research. To obtain reliable results, a high level of reproducibility in both proteins identification and quantification are needed (Hale et al., 2003; Mcguire et al., 2008). Possible sources of variation may be technical or biological in origin. Technical sources of variation can alter the quantification of measured proteins due to small differences from sample preparation, chromatography, the condition of the ion source, and the overall performance of the mass spectrometer. Biological sources of variation include differences between individuals within a population and physiological variation in individuals from one time to another. It is therefore important when conducting clinical proteomic study to include an assessment of technical and clinical reproducibility in many circumstances, such as when the biological tissue or disease has not been studied before.

Standard statistical methods used for evaluating reproducibility include the Bland Altman coefficient of reproducibility, the limit of agreement, the correlation coefficient, and linear regression. However, these assessments are generally limited to single measurements. In proteomics studies, reproducibility assessments are usually performed for a randomly selected sample of peaks or for candidate peaks of interest. The coefficient of variation, the correlation coefficient, the intra class correlation coefficient or the limits of agreement are determined for one peptide or protein at a time. Few studies have evaluated reproducibility of mass spectral data at a multivariate level.

Some proteomic studies have borrowed statistical methods from those developed for genomic studies because of similarities in the properties of the data. In micro-array reproducibility studies (Lyne et al., 2003; Chen et al., 2007), correlation coefficients and auto-correlations have been used to assess the association between replicates of micro-array

data. The percentage of overlapping genes is used to assess the proportion of common identification between replications (Chen et al., 2007). McShane et al. (2002) introduced two global measures of reproducibility in the high-dimensional space of micro-array data. They employed a cluster-specific robustness index (R index) and a discrepancy index (D index) to assess the reproducibility of components of interest formed by cluster analysis in the original data and the noised perturbed replicates. The R index estimates the proportion of pair specimens in replicates that form the same cluster as the original data. The D index estimates the number of discrepancies between the clusters from the original data and the best-matching cluster from the replicates.

Statistical methods applied to assess the reproducibility of mass spectral data have shown similarity to those used in micro-array studies. In an early study (Semmes, 2005), inter-laboratory reproducibility was assessed by four measures: (1) coefficient of variation, (2) resolution, (3) signal to noise ratio and (4) normalized intensity for three chosen diagnostic peaks. They also assessed the classification agreement across laboratories by applying boosted logistic regression and boosted decision trees. The pre-processing of the data was standardized by a robotic system. The m/z values of peaks were controlled to within $\pm 0.2\%$. The coefficient of variation (CV) for the intensity of the three peaks used in the assessment was 15%-36%. Four out of the six labs obtained perfect agreement in the classification of patients and controls. The study was well designed with standardization and blind controls.

A study by (Pelikan and Bigbee, 2007) introduced methods to assess the multivariate reproducibility of proteomics studies. This study simulated the sequential features of clinical proteomic data from multiple time intervals (sessions). The authors assessed the reproducibility of signal, discriminative features and multivariate classification models between replicates from different sessions. They suggested a signal difference score to assess the reproducibility of profile signals. This signal difference score measures the average Euclidean distance d_E between all pairs of spectra, with smaller values indicating more similarity. Both the real signal (peak) and the noise were included in the measurement of similarity between spectra. They also suggested a differential expression score to assess the reproducibility of discriminative features. The differential score quantifies the difference observed in a single profile feature between the case and control groups. It is similar to the

Fisher-like score $\left| \frac{\mu^+ - \mu^-}{\sigma^+ + \sigma^-} \right|$, where μ and σ represent the mean and variance of the sample

respectively, while + and – represent patients and controls respectively.

Chong et al. (2006) conducted a reproducibility study of LC-MS/MS iTRAQ™ data. In this study, the authors used three different model organisms as well as a double database search strategy, which aimed to minimize the false positive rate. They also employed multiple LC-MS/MS analyses to achieve better reproducibility. The CV was the only measure used to quantify precision. The iTRAQ quantification was highly reproducible with an average CV of 0.09 (range 0.04 to 0.14).

Of these proteomic reproducibility studies, only Pelikan's group introduced a global measure to assess reproducibility in mass spectral signal data. They tried to minimize the information loss by using the whole range of the spectrum, but at the cost of increased noise. It is therefore difficult to distinguish poor reproducibility (real changes in the quantities of peaks) from noise. This chapter proposes a permutation method to assess the global reproducibility of multiple features (proteins or peaks) in the dimension-reduced principal component space simultaneously, and a discordance index based on cluster analysis methodology to summarize the bias between replicated samples.

2.1.2 The random matrix theory and high dimensional data

Random matrix theory (RMT) has been introduced in multivariate statistics analysis by T. W. Anderson (1984), Marida et al. (1980), and Muirhead (1982). The largest eigenvalues of the Wishart distributed sample covariance matrix are the center of the RMT research for multivariate analysis. Onatski (2008) extended Karoui (2007)'s theory from the non-singular Wishart complex matrices ($n > p$) to singular Wishart matrices. In a paper by Onatski (2008), they proved that the joint distribution of the first m scaled and centralized eigenvalue of a complex Wishart matrix weakly converged to the m -dimensional joint Tracy-Widom distribution, when n and p approach infinity but n/p is within the compact subset of $(0, +\infty]$. Such convergence takes places in both $n < p$ and $n \geq p$ cases. They further applied the extended theory to a sequential test that there are m significant largest eigenvalues in a high dimensional $n \times p$ data. The largest m eigenvalues theory establishes a new inferential

framework for high dimensional data applicable in “omic”, imaging and financial problems, where the underlying number of dimensions are usually believed to be much smaller than the observed data.

The proposed reproducibility method described below utilizes Onatski’s theory to test the underlying structure of the mass spectral data, in order to identify the major signal information from the noise.

2.1.3 The multivariate permutation

Permutation method provides a simulation method for estimating the population parameters, in contrast to the analytical method for approximating population parameters. There are three types of permutation test: 1) exact permutation tests; 2) moment permutation test; and 3) Monte Carlo permutation test (Berry et al., 2011). The exact permutation test generates all possible arrangements from the observed data with equal probability. The Monte Carlo permutation tests only generate a large number (i.e 10000) of the arrangements from the observed data when enumeration of the observed data is not feasible.

Classical multivariate test statistics, which assume the data is multivariate normal distributed under the MANOVA framework, cannot be applied in high dimensional data when the number of observation n is small than the number of dimensions p . The distribution-free permutation methods become the alternatives for these data, especially after the advent of high-speed computers between 1990-2000.

Permutation method will be an ideal tool to facilitate the global test of the multidimensional pair data. The permutation test relies on the a single assumption that the paired data are exchangeable (Good, 2005). It therefore can be used in reproducibility algorithm to derive a test statistics that takes into account the correlated structure of mass spectral data.

The following proposed method utilizes the permutation approach in the inferential analysis for the reproducibility assessment of the high dimensional proteomic data.

2.2. Method

2.2.1 *The clinical study design and experimental design for a clinical proteomic study to assess the reproducibility*

As for the classical reproducibility assessment in a clinical study, two to four biological or technical replicates will be prepared for each participant. In the case of testing unknown tissues, both patients and normal controls are expected to be included in the study.

If the assessment is for a labeling experiment, i.e. iTRAQ, the replicates can be randomly assigned into a 4 plex (label) or an 8 plex (label) assay, using completed random block design or row-to-column design to achieve the orthogonality for eliminating potential label effect.

2.2.2 *Types of data and pre-processing of the data*

The format of feature quantification from different types of MS experiments can be either the actual or relative intensity such as the area of peaks, or other derived quantities. Most of the peak identification algorithms include baseline subtraction and normalization for pre-processing raw MS data. Normalization reduces the variation among identified proteins.

2.2.3 *Global reproducibility testing*

A global permutation reproducibility test based on all identified features (proteins or peaks) is proposed. This reproducibility assessment tests the hypothesis that there is no significant difference in the paired quantities of multiple features projected in the dimensional reduced subspace. In this assessment, firstly the averages of all paired quantities are projected into the p dimensional principal component (PC) space, where p equals the number of features minus 1. Secondly an Eigenvalue test is used to verify how many of these p PC dimensions explain significant amounts of variance of the quantification data. The resultant m significant PC dimensions form the PC space for the further analysis. Thirdly, two global multivariate test statistics, the maximum T statistic and the sign score statistic, are proposed for a global permutation test in the principal component space. The empirical permuting distributions of these two test statistics are simulated using the Monte Carlo permutation for comparison with the observed sample test statistics. This post hoc assessment is expected to identify systematic bias between paired quantifications. Each step is described in more detail below.

Step I: Principal component analysis and limit of agreement in the first principal component subspace

To begin, the quantification format of data for analysis needs to be determined. Based on the determined quantification, the common features (proteins or peaks) from all individual spectra are identified for dimensional reduction. A high proportion of common features identified from each experiment (run) indicates good reproducibility of the identification process. Experiment and run are used interchangeably in the following sections. A data matrix $\mathbf{M}_{n \times p}$ is constructed by averaging the quantities of p features among all replicates $\mathbf{I}_{j \in [1, q]}$ in n biological samples, where q equals the total number of replicates. Principal component analysis (PCA) is applied to the data matrix $\mathbf{M}_{n \times p}$ to create the orthogonal principal unit projection vectors \mathbf{v}_j for p PC dimensions. The resultant PCA unit projection vector \mathbf{v}_j is used to project each individual replicate $\mathbf{I}_{j \in [1, q]}$ separately onto the PC space.

The first principal component scores derived from the PCA explain the highest percentage of the variance from the data and have the largest eigenvalue; an assessment of the agreement between replicates using the first principal component scores provides an initial estimate for the global reproducibility. A visualization tool, namely the First Principal Component (FPC) plot which is modified from the Bland Altman plot (Bland and Altman, 1986) is produced as the byproduct in this global assessment. The First Principal Component plot also has features similar to those of the MA plot in a micro-array study. It is a scatter plot with the x-axis being the first principal component scores $\boldsymbol{\eta}_0$, which is derived by projecting the data matrix of the **averaged** quantities of the replications $\mathbf{M}_{n \times p}$ onto the first principal component subspace (i.e. $\mathbf{M}_{n \times p} \times \mathbf{v}_1$), and the y-axis being the difference between $\boldsymbol{\eta}_j$ and $\boldsymbol{\eta}_0$, where $\boldsymbol{\eta}_j$ is the first principal component scores for individual replicate.

Step II: Eigenvalue testing

The proteomic profile of each sample contains proteins that are correlated and may belong to the same functional group. Principal component analysis projects these correlated data into independent PC dimensions to identify groups of proteins. While the collected data is a

sample from the population of interest, the PC space formed by the principal components (eigenvectors) may vary from sample to sample. In principal component analysis, a positive eigenvalue of the principal component reflects how much variance is explained by this component. The first principal component with the largest eigenvalue explains the largest percentage of the data variance, while a small positive eigenvalue could result from random noise. The eigenvalue test proposed by Onatski (2008) provides evidence of how many of the observed positive eigenvalues from the sample are not due to chance.

The eigenvalues from principal component analyses are random variables with their own distribution (BeJan, 2005). In a matrix with $n < p$, where n is the number of observations and p is the number of dimensions, the number of positive eigenvalues is $n-1$. Based on random matrix theory, Onatski (2008) extended the theory for the distribution of the largest eigenvalue to the distribution of several largest eigenvalues. He further applied this theory for the asymptotic test statistics $\text{MAX}_{k_0 < i \leq k_{\max}} (\lambda_i - \lambda_{i+1}) / (\lambda_{i+1} - \lambda_{i+2})$, where λ_i is the i th largest eigenvalue of the sample covariance matrix, and $\lambda_{i+1}, \lambda_{i+2}$ is the consecutive largest eigenvalue following λ_i , k_0 denotes the number of significant eigenvalues in the null hypothesis and k_{\max} denotes the maximum number of significant eigenvalues known a priori. He proved that $\text{MAX}_{k_0 < i \leq k_{\max}} (\lambda_i - \lambda_{i+1}) / (\lambda_{i+1} - \lambda_{i+2})$ equals the distribution of $\text{MAX}_{0 < i \leq k_{\max} - k_0} (\mu_i - \mu_{i+1}) / (\mu_{i+1} - \mu_{i+2})$, where $\mu_1, \dots, \mu_{k_{\max} - k_0}$ have the joint $(k_{\max} - k_0)$ -dimensional Tracy-Widom distribution,. This eigenvalue assessment tests the null hypothesis of k_0 significant eigenvalues against the alternative hypothesis that the significant number of eigenvalue k is greater than k_0 and less than $k_{\max}+1$. This will be equivalent to testing the null hypothesis of $k = 0$ against the alternative hypothesis that $k = k_{\max} + 1 - k_0$.

Using this test statistics sequentially, eigenvalues are assessed from the largest to the smallest until a non-significant positive eigenvalue is identified. The m identified significant eigenvectors which correspond to eigenvalue λ_1 to λ_m , are used to derive the $n \times m$ dimension principal components for permutation. This is an alternative approach to the subjective approach of Scree plots (Rencher, 2002) that have been widely used to determine the number of significant dimensions .

Step III: Permutation to test global reproducibility.

The permutation method has been widely used to simulate the empirical distributions of test statistics for comparing quantities between two groups (Good, 2005; Neubert and Brunner, 2007; Wheldon et al., 2007). In the context of a proteomic reproducibility study, we propose the permutation method to test whether there are significant differences in the paired multiple-feature quantities in m significant PC dimensions. Permutation provides the empirical distributions of the global test statistics. The observed global test statistics are compared with these empirical distributions to derive the permutation p values.

We propose both a parametric and a non-parametric test statistic. The parametric statistic is

the maximum t statistic $\text{Max}_{1 \leq i \leq m} T_i$ of the m paired PC differences. Set $T_i = \frac{(u_i - 0)}{std_i}$,

where u_i is the mean difference and std_i is the standard deviation of the difference in the i^{th}

resultant PC scores. The non-parametric statistic is a two dimensional sign score, $\log\left(\frac{P_+}{P_-}\right)$,

where P_+ is the total number of positive differences in m PCs of n samples, that is,

$P_+ = \sum_{j=1}^n \sum_{i=1}^m g_{ij}$, where $g_{ij} = 1$ when the difference between the two replicates is positive and

zero otherwise, and P_- is the total number of negative differences in m PCs of n samples, P_-

$= \sum_{j=1}^n \sum_{i=1}^m f_{ij}$, where $f_{ij} = 1$ when the difference between the two replicates is negative and zero

otherwise.

Let $Z_{m,n}$ represents the data matrix of the differences derived from m paired PC scores by n samples. In each Monte Carlo permutation (Good, 2005), the sign of each element of the $Z_{m,n}$ matrix is independently switched with probability 0.5. Equivalently, for each i and j ($1 \leq i \leq m$, $1 \leq j \leq n$), the original and replication values are independently permuted. One thousand Monte Carlo permutations provide the empirical distributions of the two proposed

global test statistics T_i and $\log\left(\frac{P_+}{P_-}\right)$. The permutation p value is the proportion of

permutations in which the absolute observed test statistics is equal to or greater than the absolute permutation test statistic.

2.2.2. Summary statistics for agreement with multiple features

In addition to detecting bias in the reproducibility, a global index of reproducibility is formed by applying cluster analysis to the data, fixing the number of clusters at the sample size. Ideally, each sample should cluster with its replicate. The discordance index measures the proportion of samples that fail to cluster with their replicates.

2.3. Results

Two different types of quantification data (SELDI-MS and LC-MS/MS with iTraq™ labeling) were used to demonstrate the proposed method. In the SELDI-MS experiment, common peaks were identified with the PROcess algorithm (Li, Xiaochun <http://bioconductor.org/packages/2.4/bioc/html/PROcess.html>), where the local maxima of intensities in each identified peak region were used as the analyzed quantity. In the LC-MS/MS labeling experiment, peptides identified by ProteinPilot™ with “used” indicator =1 were filtered by confidence score and aligned across different runs; For the purpose of this reproducibility analysis, the weighted averages of reporter ion peak areas were calculated for peptides that multiple observations in a single protein summary. The resultant peptide areas are summed for each protein that they belonged to. Within each run, median normalization was applied to the summed areas across labels on the natural log scale. After pre-processing, a relative protein quantity was derived for each sample. This pre-processing corrects for the iTRAQ labeling effects.

2.3.1 Case study

Coronary plasma blood samples of eight ischemic patients before and after an angioplastic procedure were collected from the Greenlane Cardiovascular Service of Auckland City Hospital and analyzed by LC-MS/MS with iTRAQ™ labeling at the Centre for Genomics and Proteomics, University of Auckland. Prior to the LC-MS/MS analysis, a depletion

process was used to exclude the ten most highly abundant proteins. The depletion process is to make sure the relative quantities of the low abundant proteins that are normally relevant to the disease of interest can be precisely measured during the LC-MS/MS analysis.

For the purposes of demonstrating the proposed statistical method for the reproducibility evaluation, it is hypothesized that there are no changes in the proteomic expression before and after the angioplasty procedure, so the post-procedure samples are treated as the replications of the baseline sample for the demonstration. Peptide profiles from 4 different runs of ProteinPilotTM were aligned and the areas under the peaks were log transformed and normalized by the median within each run. One hundred and twelve common peptides from the 4 different runs were used to construct the relative intensity of proteins for the reproducibility assessment. Principal component analysis was performed on the quantities of the 24 proteins found in all four runs.

Both the eigenvalue test and scree plot (Table 2.1 and figure 2.1) indicated that the first eigenvalue was significant, and the corresponding first eigenvectors explained 84% of the total variance. The first Principal Component plot (FPC plot) in Figure 2.3(a) shows a significant difference in the relative protein quantities between the post- and pre-angioplasty samples; the PC of post- procedure samples tends to be lower than the pre-angioplasty samples overall. This trend is consistent with the pattern in the second plot, where the differences in the relative quantity of the pre- and post-procedure for all proteins are plotted. Details of the post-angioplasty expression change are reported in chapter 6.

Table 2.1 Results of the Eigenvalue test

i	1	2	3	4	5	6	7
λ_i	12.7	0.93	0.48	0.42	0.26	0.12	0.06
$(\lambda_i - \lambda_{i+1}) / (\lambda_{i+1} - \lambda_{i+2})$	26.2	7.5	0.4	1.1	2.3	1.0	
	$K_0=1$	$K_0=2$	$K_0=3$	$K_0=4$	$K_0=5$	$K_0=6$	
$\text{MAX}_{K_0 < i \leq 7} (\lambda_i - \lambda_{i+1}) / (\lambda_{i+1} - \lambda_{i+2})$	26.2	7.5	2.3	2.3	2.3	1.0	
critical value	8.3	8.0	7.5	7.0	6.5	5.7	

*This row provides the maximal values of $(\lambda_i - \lambda_{i+1}) / (\lambda_{i+1} - \lambda_{i+2})$ when i ranges between 1-7, 2-7, 3-7 and up to 6-7. Seven is assumed to be the maximal number of eigenvalues.

2.3.2 Simulation experiments

A simulation experiment was used to investigate the sensitivity of the proposed method. Different types of noise, with different distributions and parameters were added to the relative peak quantities of 30 ovarian cancer patients to simulate different replicates from the MS experiment. The mass spectral intensity data is a random sample from a large ovarian proteomic experiment, downloaded from the proteomic databank of the Center for Cancer Research (<http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>). Sixty-one common peaks were identified from these 30 subsamples using the PROcess bioconductor package. Preprocessing procedures, including baseline subtraction by Loess and normalization, were applied. A relative quantity was derived for each identified peak.

Twelve distributions, with different parameters simulating systematic bias (parameterized by the mean, μ) and noise (parameterized by the standard deviation, σ), were generated and added into the relative quantity data. The distributions were normal distributions $\{(\mu=0/2/4, \sigma$

=2/4), $(\mu=0/2/4, \sigma =2/4)$ }, exponential distributions $\{(\mu= \sigma =0.5), (\mu=\sigma =1)\}$ and bimodal distributions (mixture of two normal distributions with different means and standard deviations at different sections of m/z values). The FPC plots from two simulations are shown in Figure 2.2a and 2.2b. In the FPC plot of the normally distributed differences with $(\mu=0, \sigma=4)$, the differences in the first principal component scores between the sample and its simulated replicate are randomly scattered above or below 0. In the FPC plot of the exponentially distributed differences with $\mu= \sigma =0.5$, the differences in the first principal component scores tend to be significantly above 0.

The eigenvalues of the sample matrices were tested before the permutation test basing on 1000 Monte Carlo permutations were proceeded. In the principal component analysis, quantities were normalized. The permutation tests were applied to differing sizes of samples (8, 15 and 30 samples), and to different distributions with different parameters in the simulated replicates. The comparison results are shown in Tables 2.1a and 2.1b.

2.3.2.1 Results of global permutation reproducibility testing (Tables 2.2a and 2.2b)

Sample size and sensitivity of the test

When the sample size was equal to 8, the permutation tests using maximum t statistics failed to detect the simulated bias; the permutation tests using the sign score statistics successfully identified bias for the normal distribution $(\mu=2, \sigma=4)$ and exponential distribution $(\mu=\sigma=1$ and $\mu=\sigma=0.5)$ but failed for the other simulated distributions.

When the sample size was equal to 15, the permutation tests using maximum t statistics successfully detected the bias with the normal distribution $(\mu=2, \sigma=2)$, the exponential distribution with $\lambda=1(\mu=\sigma=1)$ and $\lambda=2(\mu=\sigma=0.5)$, and the bimodal distribution $(\mu=1, \sigma=1$ ($m/z<1000$) $\mu=2, \sigma=2$ ($m/z>1000$)). However, it failed to detect the bias with the normal distribution $(\mu=2, \sigma=4)$, the bimodal distribution $(\mu=1, \sigma=2$ ($m/z<1000$) $\mu=2, \sigma=4$ ($m/z>1000$)) and the bimodal distribution $(\mu=2, \sigma=4$ ($m/z<1000$) $\mu=4, \sigma=8$ ($m/z>1000$)). The tests using the sign score statistics successfully detected all of the simulated biases.

When the sample size was equal to 30, both test statistics successfully identified all the simulated biases. The sensitivity of the reproducibility is affected by the sample size.

Variance and sensitivity

When the variation of the sample increased, the sensitivities of both test statistics were weakened. In the simulations, when the coefficient of variation of a normally distributed difference was greater than 1, the permutation test using maximum t statistics was not sensitive with sample sizes < 30 .

Discordance index and median percentage change

In the bias assessment, all 30 samples, including both replicates, were entered in a cluster analysis and 30 clusters were formed by the Ward method (Ward 1963) (Rencher, 2002). Table 2.3 summarizes sample details and grouping of replicates in the same cluster. The bias with distribution $\mu=2, \sigma=4$ ($m/z \leq 1000$) $\mu=4, \sigma=8$ ($m/z > 1000$) resulted in the largest discordance index and % of differences between the simulated replicates and the original data.

A high discordance index can be caused by a high degree of bias with high variation. The simulation results show that the discordance index is not sensitive to bias with small magnitude and large variation. However, the discordance index is an interesting way to summarize the data and provides extra information about outlying samples.

2.4. Discussion

This chapter proposes a method to assess the global reproducibility of mass spectral data rather than focusing on the reproducibility of single selected candidate proteins or peptides. A multivariate reproducibility assessment is useful to assess overall performance and identify problematic candidate proteins or peaks. Using principal component analysis, high dimensional correlated spectral data are reduced to lower dimensions and projected into orthogonal principal component space. Random matrix theory provides a basis for testing the underlying correlation pattern of proteins to eliminate non-significant principal components from further analysis. A permutation reproducibility test can be used to identify systematic bias and adjust for multiple testing. If bias is identified, further analysis of the principal component scores can identify problematic proteins or peaks by using a maximum t

test statistic or sign score statistic. The strategy of combining dimension reduction with permutation testing utilizes all the information effectively.

From the simulation experiments, it was found that a sample size of 30 will have greater statistical power to detect simulated bias than a sample size of 15 or 8. The size and variation of samples have significant impacts on the sensitivity of the assessment.

A large-scale reproducibility study using LC-MS/MS that assesses the real day-to-day operations and patient variations is needed. This study would be important before applying the proteomic technology in daily clinical laboratory practice. The reproducibility assessment in a clinical proteomic experiment is complex. It involves early phase assessment for reproducibility of laboratory technique and the late clinical phase assessment for reproducibility of patients' day-to-day physiological conditions. For the examples used in this study the reproducibility of quantification post-protein identification was assessed. However, the proposed method can be applied to specific sources of variation including intra/inter run reproducibility and day-to-day variability.

A limitation of the current study is that the sensitivity of eigenvalue testing is affected by the sample size. When the sample size is small, the eigenvalue test combined with the traditional Scree Plot may be a better way to identify the main pattern of protein profiles.

In conclusion, this chapter suggests extensions of reproducibility methods from the single-dimension assessment to a higher-dimension assessment and demonstrates that this systematic approach to reproducibility is useful and workable.

The proposed method was also applied in chapter 6- the immunology case study.

Table 2.2a Simulate replicates by adding bias and noise with different distributions

Data presented are median (P25,P75)					
Distribution of the systematic bias	Size of the samples and its replicates	Parametric version		Non Parametric version	
		Distribution of the permuted p values	Distribution of the test statistics	Distribution of the permuted p values	Distribution of the test statistics
Normal					
$\mu=2 \sigma=4$	30	0.02(0.01,0.06)	3.76 (3.42,4.19)	-	-
$\mu=0 \sigma=4$	15	0.51 (0.30, 0.67)	2.18 (1.95, 2.57)	0.59 (0.41, 0.82)	0.00 (-0.08,0.08)
$\mu=2 \sigma=2$		0.02 (0.009, 0.04)	4.01 (3.63, 4.65)	0.001 (0.001,0.001)	-0.84 (-0.96,-0.71)
$\mu=2 \sigma=4$		0.15 (0.06, 0.34)	2.94 (2.50, 3.47)	0.005 (0.001,0.03)	-0.42 (-0.48,-0.32)
$\mu=0 \sigma=4$	8	0.36 (0.13, 0.57)	2.34 (1.92,3.17)	0.64 (0.42,0.90)	0.00 (-0.14,0.14)
$\mu=2 \sigma=2$		0.18 (0.10, 0.31)	2.97 (2.42,3.54)	0.02 (0.004, 0.08)	-0.65 (-0.81, -0.50)
$\mu=2 \sigma=4$		0.32 (0.19,0.53)	2.50 (1.98,3.01)	0.50 (0.24,0.64)	-0.21 (-0.35, -0.07)
Exponential					
$\lambda=1(\mu=\sigma=1)$	15	0.01 (0.005, 0.03)	4.48 (3.88, 4.98)	0.001 (0.001,0.001)	-1.03 (-1.14, -0.89)
$\lambda=2(\mu=\sigma=0.5)$		0.007 (0.002,0.02)	4.70 (4.24, 5.37)	0.001 (0.001,0.001)	-1.17 (-1.29, -1.01)
$\lambda=1(\mu=\sigma=1)$	8	0.12 (0.05,0.19)	3.21 (2.86,4.22)	0.01 (0.001,0.03)	-0.73 (-1.01, -0.63)
$\lambda=2(\mu=\sigma=0.5)$		0.07 (0.04,0.13)	3.61 (3.16,4.56)	0.002 (0.001,0.005)	-0.98 (-1.17, -0.81)

Table 2.2b Simulate replicates by adding bias and noise with different distributions

Data presented are median (P25,P75)		Parametric version		Non Parametric version	
Distribution of the systematic bias	size of the samples and its replicates	Distribution of the permuted p values	Distribution of the test statistics	Distribution of the permuted p values	Distribution of the test statistics
Bimodal					
$\mu=1, \sigma=2$ ($m/z \leq 1000$) $\mu=2,$ $\sigma=4$ ($m/z > 1000$)	30	0.03 (0.008,0.08)	3.68 (3.33, 4.27)	-	-
$\mu=2, \sigma=4$ ($m/z \leq 1000$) $\mu=4,$ $\sigma=8$ ($m/z > 1000$)		0.03 (0.01,0.08)	3.68 (3.28, 4.02)	-	-
$\mu=1, \sigma=1$ ($m/z \leq 1000$) $\mu=2,$ $\sigma=2$ ($m/z > 1000$)	15	0.03 (0.01,0.08)	3.68 (3.28, 4.02)	0.001(0.001,0.001)	-0.96 (-1.08,-0.80)
$\mu=1, \sigma=2$ ($m/z \leq 1000$) $\mu=2,$ $\sigma=4$ ($m/z > 1000$)		0.12 (0.06,0.32)	3.07 (2.55, 3.51)	0.004(0.001,0.01)	-0.42 (-0.52,-0.34)
$\mu=2, \sigma=4$ ($m/z \leq 1000$) $\mu=4,$ $\sigma=8$ ($m/z > 1000$)		0.14 (0.06, 0.29)	3.00 (2.61, 3.48)	0.005(0.001,0.03)	-0.40 (-0.50,-0.31)
$\mu=1, \sigma=1$ ($m/z \leq 1000$) $\mu=2,$ $\sigma=2$ ($m/z > 1000$)	8	0.18 (0.08,0.37)	2.88 (2.39,3.56)	0.02 (0.003,0.14)	-0.65 (-0.81, -0.43)
$\mu=1, \sigma=2$ ($m/z \leq 1000$) $\mu=2,$ $\sigma=4$ ($m/z > 1000$)		0.26 (0.14, 0.55)	2.67 (1.95, 3.21)	0.48 (0.23,0.65)	-0.21 (-0.35,-0.07)
$\mu=2, \sigma=4$ ($m/z \leq 1000$) $\mu=4,$ $\sigma=8$ ($m/z > 1000$)		0.36 (0.19,0.54)	2.32 (2.03, 2.87)	0.32 (0.23,0.62)	-0.28 (-0.35, -0.14)

Table 2.3 Summary of discordance index and median % change from simulated bias

Distribution of bias with different parameter	Number of samples-replicates grouped in the same cluster	Discordance index (% of samples-replicates failed to group in the same cluster)	Median % changes between simulated replicate and original data Among all features (peaks)
<u>Normal n=30</u>			
$\mu=0 \sigma=4$	27	0.10	0.4% [-24%, 21%]
$\mu=2 \sigma=2$	27	0.10	67% [16%,288%]
$\mu=2 \sigma=4$	23	0.23	78.2% [17.1%, 313.2%]
<u>Exponential n=30</u>			
$\lambda=1(\mu=\sigma=1)$	30	0.0	43% [8.6%,137%]
$\lambda=2(\mu=\sigma=0.5)$	30	0.0	21% [4.3%, 68%]
<u>Bimodal n=30</u>			
$\mu=1,\sigma=2(m/z\leq 1000)$ $\mu=2,\sigma=4(m/z>1000)$	26	0.13	65% [8.8%,191%]
$\mu=2,\sigma=4(m/z\leq 1000)$ $\mu=4,\sigma=8(m/z>1000)$	10	0.67	130% [18%, 381%]

Figure 2.1 Scree plot for the cardiac case

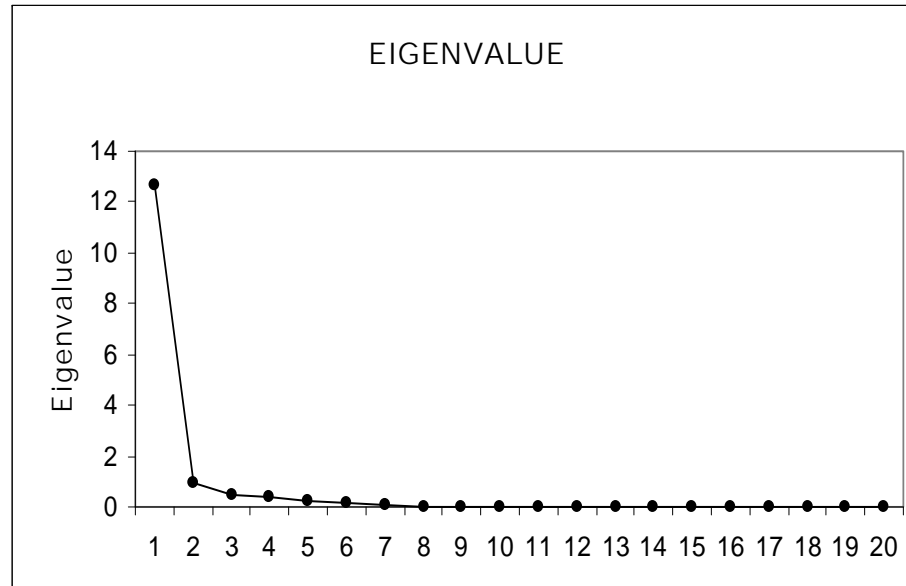


Figure 2.2a FPC plots of simulated systematic bias (exponential distributed).

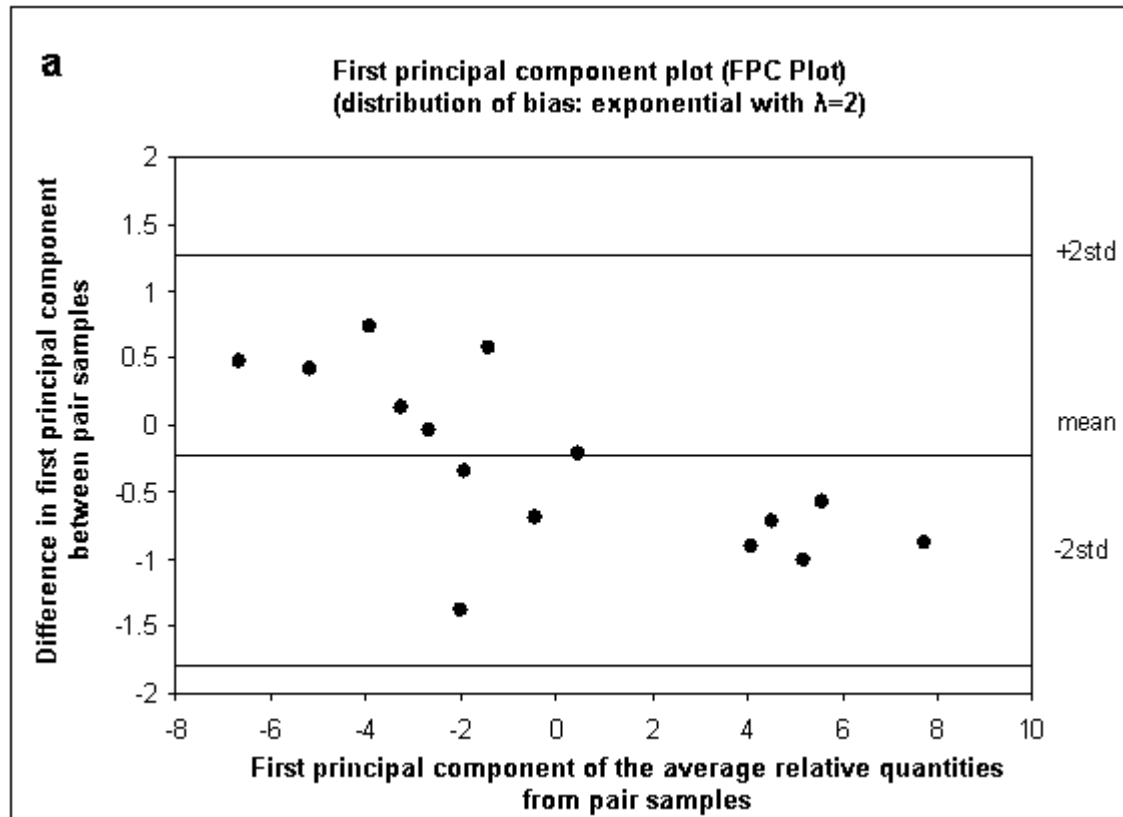


Figure 2.2b FPC plots of simulated systematic bias (Normal distributed).

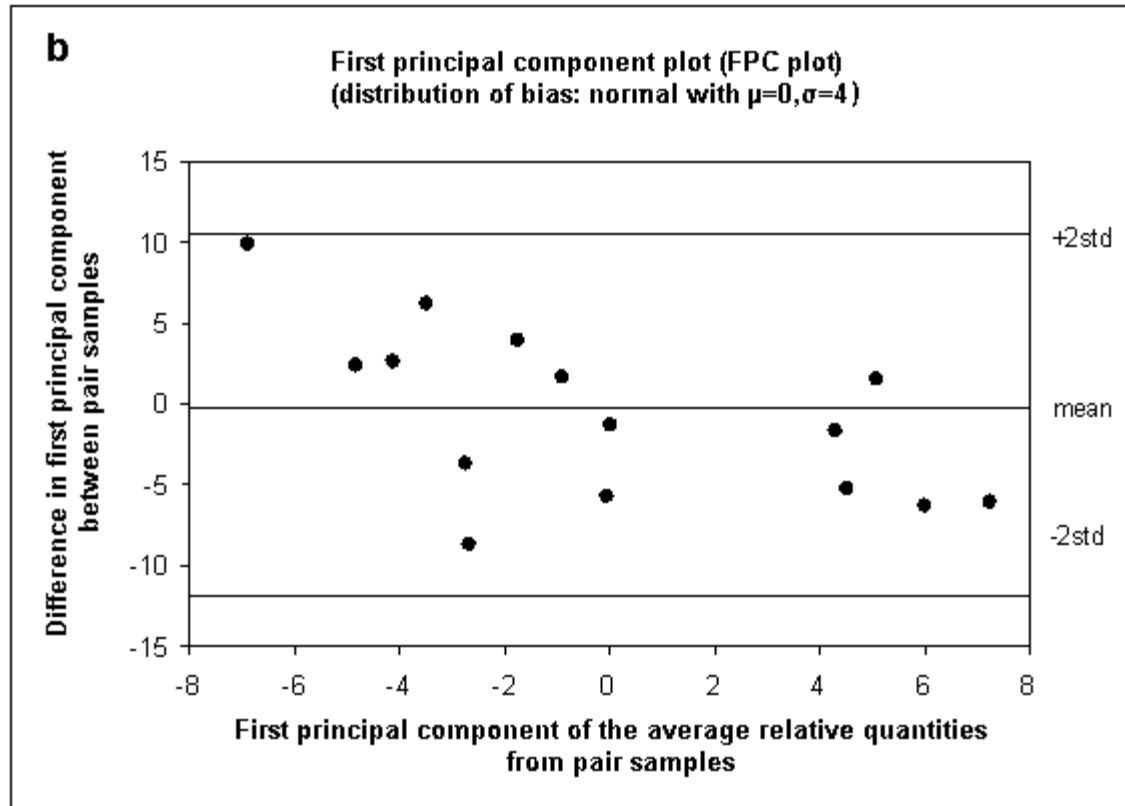
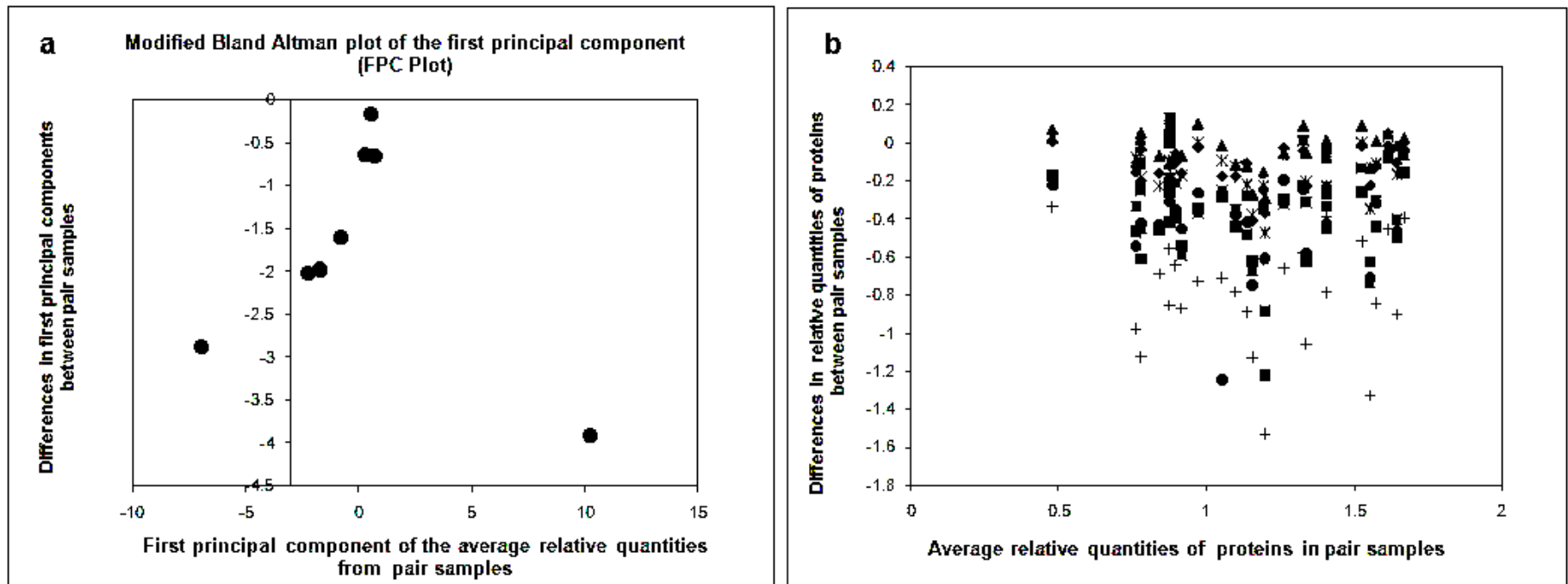


Figure 2.3

(a) PC plot of first dimension (FPC plot) from the cardiac LC-MS/MS ITRAQ™ data

(b) Difference in log (area) between replicates from all proteins vs. average of all log (area) from cardiac patient's LC-MS/MS ITRAQ™ data. A unique plotting symbol is used for each patient.



CHAPTER 3

Two optimization strategies of multi-stage design in clinical proteomic studies

ABSTRACT

We evaluated statistical approaches to facilitate and improve multi-stage designs for clinical proteomic studies which plan to transit from laboratory discovery to clinical utility. To find the design with the greatest expected number of true discoveries under constraints on cost and false discovery, the operating characteristics of the multi-stage study were optimized as a function of sample sizes and nominal Type-I error rates at each stage. A nested simulated annealing algorithm was used to find the best solution in the bounded spaces constructed by multiple design parameters. This approach is demonstrated to be feasible and lead to efficient designs. The use of biological grouping information in the study design was also investigated using synthetic datasets based on a cardiac proteomic study, and an actual dataset from a clinical immunology proteomic study. When different protein patterns presented, performance improved when the grouping was informative, with little loss in performance when the grouping was uninformative.

3.1 Introduction and motivation

Most laboratory-based biomarker discoveries do not reach clinical use. One reason may be the lack of connection between laboratory and clinical proteomics studies, so that laboratory selections and the clinical validation of the protein markers are separate processes in study design (Patterson et al., 2010). In addition, there is a risk that false discoveries are introduced by technical artifacts with different proteomic platforms. In 2007, the National Cancer Institute (NCI) suggested a three-stage workflow to link laboratory discovery to clinical utility in proteomic studies (National Cancer Institute, 2007). The stages are: (1) unbiased

discovery using tens of samples, followed by (2) targeted verification using hundreds of samples, and finally (3) clinical validation using thousands of samples. The whole process integrates knowledge on proteomic infrastructure, systematic study design and health economics. It thus requires a systematic design to optimize the number of discoveries under constraints of cost and false discovery.

3.1.1 Multi-stage design in gene-association and proteomic studies

In genetic association studies, Jaya M. Satagopan and Elston (2003) proposed a two-stage design excluding markers with little evidence of association in the first stage of the study and selecting only promising markers for the second stage. They used Monte Carlo grid search to obtain combinations of one-stage design parameters, i.e. power and type I error, and then applied numerical integration to find the solution of the two-stage design parameters that minimized the study cost subject to an overall type I error rate and a statistical power. Wang et al. (2006) expanded the two-stage approach from candidate-gene to genome-wide scale. Zuo et al. (2008) proposed an optimal resource allocation that maximized the overall power for a fixed total cost. They also investigated the impact of genotyping errors. They derived the joint distribution of the test statistic in the first and second stage, and converted the objective function to a mixed integer nonlinear programming problem (MINLP) with only two parameters under a series of constraints. Skol et al. (2007) described a similar approach to Zou et al. but using a different joint test statistic. Moerkerke and Goetghebeur (2008) added that genetic markers should be selected and ranked in order of evidence that balanced false positive rate and false negative rate at the first stage. At the second stage, more samples are selected and data from both stages combined. They proposed a gain function using the weighted sum of the false positive and false negative rates as the objective function for maximization.

Originally multi-stage designs were suggested for gene association studies for which budgetary considerations needed to be balanced against statistical power. However, the falling cost of genotyping and the economies of scale available from off-the-shelf SNP chips made these designs less useful (Spencer et al., 2009). In a study for genome-wide interaction analysis (GWIA), Steffens (2010) also argued against the adoption of a two-stage strategy and suggested that multi-stage screening will prevent the detection of pure epistatic effects.

In contrast, proteomic studies still have substantial per-protein marginal costs, especially in the final stage, so that a multistage design is substantially more affordable. The multistage design also provides built-in technical validation, with protein abundance being measured using different assays at each stage. The multistage design may still be weak for assessing pure interaction without main effects, but this is not currently a major focus of proteomic research.

3.1.2 Similarities and differences between multi-stage gene association studies and multi-stage clinical proteomic studies

A proteomic study using a systems biology approach to identify disease-related proteins has similarities to a multi-stage gene-disease association study. It starts with systematic identification and screening of hundreds or thousands of proteins. It then uses a targeted candidate quantification approach to verify and/or validate the findings in the same or a separate group of subjects. The decision on the proteins selected at the identification stage for further study is as vital as that in the screening stage of a genome-wide association study (GWAS). The optimization problem in a multi-stage proteomic study also has similar parameters to a multi-stage GWAS. A common problem in both gene and protein association studies is to search for a design that maximizes power with an acceptable false positive rate and cost, or which minimizes cost with fixed power and false positive rate.

However, there are important differences between proteins and genes relevant to association studies: the number of proteins measurable with current technologies is much less than the number of genes. Proteins are highly changeable and have a wider dynamic range: their abundance in a cell ranges from less than 500 to 2×10^7 copy numbers (Beck et al., 2011). The most abundant plasma proteins such as albumin and IgE are usually not disease-specific, or of primary interest. Depletion of high-abundance proteins can reduce the problem, but large differences in abundance remain after the depletion.

Current biotechnologies allow identification of thousands of proteins. To achieve an unbiased discovery, NCI advocated a second technical verification using a candidate-based platform. Thus, false discovery due to random error and/or technical artifacts needs to be considered in

the design. The statistical method used to adjust for multiple tests in gene association studies may not be the optimal solution for proteomic screening. With the upcoming advances in the biotechnologies, tailor-made study design for large-scale protein research is, therefore, a timely objective.

3.1.2 The potential value of using biological information in protein groups

Bioinformatics profiling information is often used to enrich the design of proteomic studies. Proteins are commonly studied in groups defined by function, structure, and localization (Greenbaum et al., 2003). For instance, a therapy or drug may target the proteins in the same disease related pathway. Hence, it is useful for biologists to study each molecule (protein, metabolite) with others that belong to the same signaling pathway (Meani et al., 2009) or biological function. For example, Hoorn et al. (2005) and Chornoguz et al. (2010) successfully identified proteins and their related pathways or networks that associated with disease or a physiological intervention. Hoorn suggested that the combination of pathway analysis and proteomic analysis both facilitated the interpretation of proteins' relationships and made it possible to identify low abundant proteins which otherwise would escape from the proteomic analysis. Meani et al. (2009) considered the understanding of protein signaling pathways in diseased and normal tissues to be the first step in cancer molecule characterization and personalized therapy. An optimal design using pathway or protein network information may increase the likelihood of candidate proteins of an important group being selected from stage I, and thereby improve biologically and clinically relevant discoveries.

3.1.4 Objectives of the proposed study

We investigate optimal designs under the NCI three-stage workflow, and explore extra options that utilize biological information of proteins via bioinformatics approaches or pathway analysis to enrich the study design. The approaches proposed for genetic association studies are expanded, focusing on validation of the discovery via candidate-based platforms. A range of design problems is investigated, starting from the simplest scenario that proteins

are selected separately to the comprehensive option of utilizing protein grouping information, but without consideration of the correlation structure across groups. Our main intention is to provide different options with computing algorithms to achieve robust designs when research resources are constrained.

3.2 Statistical strategies in the three-stage design

This section describes optimization strategies for a three-stage study from discovery to verification, and validation using different or the same platforms for different independent samples. In the discovery phase, peptides are identified systematically via mass spectrometry (MS) or 2D gel. The discovered peptides are used to identify and quantify proteins through reverse peptide sequence database and bioinformatics software (i.e. ProteinPilot™).

In the second verification stage, multiple-reaction monitoring (MRM) mass spectrometry is applied to verify the changes in abundance that were observed for multiple proteins in the discovery phase. Since the 1990s, MRM-based assays have emerged as an alternative candidate approach to enzyme-linked immunosorbent assays (ELISAs). These mass spectrometry assays eliminate the cost of producing a large number of new immunoassays at an early stage of research, allowing the development of antibodies to be deferred until the final stage.

In the third and final stage, new antibodies and immunoassays are developed and used in larger samples of patients for validation. Multiplex ELISA is one type commonly used in clinical laboratories. A novel alternative is the new mass spectrometry-based quantification for candidate peptides smaller than 10kDa (L. Anderson, 2005).

The proposed statistical methods for the optimization of multi-stage studies assume that a known set of p_1 proteins are discovered from the stage I process from which a subset of p_2 of these are then selected using a statistical significance threshold based on information from either individual proteins or both individual and groups of proteins. A subset p_3 of these proteins is then selected based on a second selection criterion at the verification stage. Finally, these p_3 candidate proteins are validated at the last stage. The sample sizes of the

second stage, n_2 , and the last stage, n_3 , and the stage-wise false positive rates (i.e. type I error) are selected to maximize the power of discovery under the constrained study cost and the overall number of false positives.

Before any proof-of-concept pilot experiment, prediction of the number of discoveries at stage I is difficult because of its dependence on various uncontrollable factors, such as the performance of the mass spectrometer, types of biological tissues and other technical artifacts. Given the limited prior information on stage I design parameters, the stage I sample size is not included in the objective function for optimization. We choose to start the optimization from the selection of p_2 from p_1 discovered proteins so that the optimal solution is not influenced by the number of discoveries at stage I. The stage I discoveries will also provide information (i.e. means and standard deviations) for the design parameters to be used in the optimization.

We demonstrate the optimization in the context of studies involving paired samples at each stage, such as 1:1 matched case-control or before-after intervention studies. This method can be generalized to parallel group studies, with or without paired samples. In the paired sample design, the analytical units will be the log-transformed relative intensities. The detectable mean differences between paired samples are determined based on either prior information and/or clinically or biologically relevant differences. The prior information can be obtained from the literature or prior experiments; it is not limited to the stage I discovery study. The standard deviations can be estimated from the stage I discovery study and/or obtained from prior experiments. In the computations for seeking the optimal design solution, the means and standard deviations of the differences are assumed to be constant across stages.

The optimization assumes the budget is fixed. The assay costs at stages II and III, the cost of recruitment and the stage I sample size are known. A solution of stage I/II nominal false positive rates (decision thresholds) and stage II/III sample sizes is derived to maximize the number of discoveries at the final stage. The following sections describe two algorithms for the optimization with and without biological grouping information.

3.2.1 The simplest scenario: proteins are selected individually

In the simplest scenario, selection is carried out independently for each protein, based on single-protein test statistics. Student's paired sample t -test is used to assess the differences in the log-transformed fold change between paired samples. p_2 proteins are selected from p_1 proteins based on p_1 individual tests at stage I. p_3 proteins are selected based on p_2 individual tests at stage II and finally p_3 protein candidates are validated at the final stage based on the individual tests.

3.2.1.1 Using Simulated Annealing (SA) to seek optimized solution in the multi-stage design: the algorithm SA-a

The proposed method maximizes the expected number of proteins with true effects discovered from a three-stage study under a cost constraint. The expected number of true effects is derived from an objective function which has four design parameters: the stage I type I error rate, α_1 , the stage II type I error rate, α_2 , the sample size at stage II, n_2 , and the sample size at stage III, n_3 . The values of these parameters were divided into small intervals within defined ranges (i.e. α_1 ranged between 0.005-0.50 with interval size 0.025; α_2 ranged between 0.005-0.25 with interval size 0.025; n_2 ranged between 100-1000 with interval size 10; n_3 ranged between 100-5000 with interval size 100). The combinations of knots at these intervals form the solution space of the objective function in the optimization.

Simulated annealing (SA) is used to determine the optimal design parameters in stages II and III for a specified sample size and number of proteins at the first stage. It is a stochastic optimization method that does not require the objective function to be smooth, and is capable of finding global optima even in problems where many local optima exist (Nikolaev and Jacobson, 2010). In the current problem, lack of smoothness and multiple optima result from the constraint and using Monte Carlo averages to approximate the expected number of discoveries. In contrast to 'hill-climbing' approaches that attempt to find a higher value of the objective function at each iteration, and so cannot escape a local minimum, SA will sometimes step down. At each iteration, the current solution is compared to the next candidate solution. A superior solution will be accepted with 100% probability; an inferior

solution will be accepted with a probability based on the current ‘temperature’ which is a predefined constant number decreasing as the algorithm progresses.

3.2.1.2 Definition of the SA-a algorithm

The solution space, Ω , bounded by the acceptable limit of each design parameter. Let the vector of design parameters $\omega = (n_2, n_3, \alpha_1, \alpha_2)$ be a solution in Ω , where n_2 and n_3 are the stage II and III sample sizes, respectively, and α_1 and α_2 are the stage I and II type I error rates, respectively. Ω contains all the possible combinations of these parameters which are categorized by small intervals within their bounded ranges.

Objective function. Let $f(\omega): \Omega \rightarrow \mathbb{R}$ be the objective function of the solution space, where f is the expected number of proteins that are discovered at stage III associating with the disease being investigated. It is in the range of $0, 1, \dots, p_1$, where p_1 is the number of proteins discovered in stage I and being considered for inclusion in stages II and III of the study.

The proposal neighborhood selection function. The proposed neighborhoods are constructed by M arbitrarily bounded and possibly overlapping solution subspaces, Ω_i ($i= 1, 2, \dots, M$). The Ω_i are formed by firstly selecting a point ω_i (*the centre of Ω_i*) according to either a uniform or Beta distributed jumping length from the previous centre point ω_{i-1} , and secondly selecting a uniformly distributed radius R_i with probability 0.5 for each direction from the selected center ω_i . Each candidate point can then be assigned within each Ω_i , according to a uniform distributed probability.

This nested SA starts with a uniformly random assignment of a solution ω in the radius R_1 bounded neighborhood Ω_1 , and then a local SA with k iterations is used to seek the global minimum of Ω_1 . After the first local SA, a new address is assigned as the centre of the next solution subset Ω_2 and the second local SA is repeated. This procedure repeats for up to M subsets; the solution from each local SA will be updated if it is better than the previous one.

The temperature cooling schedule. The logarithmic cooling schedule is defined as,

$$T_k = \frac{temp}{\log\left(\left[\frac{t-1}{t_{\max}}\right] \times t_{\max} + \exp(1)\right)},$$

where t is the current iteration, $temp$ is the starting temperature for the cooling scheme and t_{\max} is the number of function evaluations at each temperature (Belisle, 1992).

The acceptance probability. The Metropolis function, i.e.

$$\begin{cases} \exp\left(-\frac{f(\omega') - f(\omega)}{T_k}\right), & f(\omega') - f(\omega) > 0, \\ 1, & f(\omega') - f(\omega) \leq 0 \end{cases},$$

is used to derive the acceptance probability.

The objective function for SA-a. Let pr_i be the probability of protein i being discovered at stage III ($i = 1 \dots p_1$, where p_1 is the number of proteins selected from stage I). The objective function is then given by

$$f(\alpha_1, \alpha_2, n_2, n_3) = E\left(\sum_{i=1}^{p_1} pr_i\right) = \sum_{i=1}^{p_1} E(pr_i),$$

where α_1 and α_2 are the significance levels at stages I and II, respectively, and n_2 and n_3 are the sample sizes at stages II and III, respectively.

Now, let $c_1 = Pt^{-1}(1 - \alpha_1/2, df_1)$, $c_2 = Pt^{-1}(1 - \alpha_2/2, df_2)$ and $c_3 = Pt^{-1}(0.975, df_3)$ be the t quantiles corresponding to the type I error rates at stage I, II and III respectively, where Pt^{-1} is the quantile function for Student's t -distribution, and df_1 , df_2 and df_3 are the corresponding degrees of freedom at stages I, II and III, respectively.

Let $\beta_{1,i}$, $\beta_{2,i}$ and $\beta_{3,i}$ denote the paired t -test type II error rates at stages I, II and III, respectively, for protein i . It follows that $(1 - \beta_{j,i})$ is the power at each corresponding stage, j ($j = I, II, III$). The expected number of true discoveries (power) is expressed as a function of the cumulative density of t -statistics for the i th protein at each stage, i.e.

$$E(pr_i) = (1 - \beta_{1,i})(1 - \beta_{2,i})(1 - \beta_{3,i}),$$

where

$$\beta_{1,i} = P\left(\frac{\bar{x}_i - \theta_0}{\delta_i / \sqrt{n_1}} < c_1\right) = P\left(\frac{\bar{x}_i - \theta_i}{\delta_i / \sqrt{n_1}} < c_1 + \frac{\theta_0}{\delta_i / \sqrt{n_1}} - \frac{\theta_i}{\delta_i / \sqrt{n_1}}\right) = P\left(T_{df_1} < c_1 + \frac{\theta_0 - \theta_i}{\delta_i / \sqrt{n_1}}\right)$$

and n_1 represents the known stage I sample size. If $\theta_0 = 0$, this simplifies

to $1 - \beta_{1,i} = 1 - Pt\left(c_1 - \frac{\theta_i}{\delta_i / \sqrt{n_1}}\right)$, where θ_i is the absolute difference between the matched

diseased and normal groups under the alternative hypothesis for protein i and Pt is the cumulative paired sample Student's t -distribution function. Analogously, the objective functions for $1 - \beta_{2,i}$ and $1 - \beta_{3,i}$ are given by

$$f(\alpha_1, \alpha_2, n_2, n_3) = \sum_{i=1}^{p_1} \left(1 - Pt\left(c_1 - \frac{\theta_i}{\delta_i / \sqrt{n_1}}\right)\right) \left(1 - Pt\left(c_2 - \frac{\theta_i}{\delta_i / \sqrt{n_2}}\right)\right) \left(1 - Pt\left(c_3 - \frac{\theta_i}{\delta_i / \sqrt{n_3}}\right)\right)$$

The cost function is defined as $n_2 \times p_2 \times t_2 + n_3 \times p_3 \times t_3 + (n_2 + n_3) \times R$, where t_2 and t_3 are the assay costs and p_2 and p_3 are the numbers of proteins being tested at stages II and III respectively, and R is the recruiting cost. This cost function is used in the following simulation study; it may vary based on different cost structures.

The actual objective function of SA-a computes the expected number of positive findings by using the Monte Carlo average of 1000 simulations. Additionally, technical differences between the Stage I and Stage II assays can be simulated by multiplying each θ_i by a random 'technical artifact' adjustment, λ_i , in the Stage I calculations. Our simulations below incorporate this adjustment.

3.2.1.3 Comparison of nested neighborhood selection with single-step selection

Instead of using single-step SA, algorithm SA-a employs a nested-search strategy on subsets of the solution space determined by both the jumping length from one centre to another and the radius of the search space. Comparing the single-step method with the nested-search method, the latter constructs a local structure of the global search surface. This strategy is shown to be more efficient with shorter computation time and without losing effectiveness in finding a good solution. In a case study to identify the global solution of a function with known maxima under inequality constraints, the computing time of using the single-step search was about twice of that using the nested search. The discovery rate for the known maximum from 100 experiments using 10000 iterations in the global search was 54%. Compared to an equivalent nested-search of 100 subsets x 100 iterations, the discovery rates were 64%, 58% and 97% for uniform, Beta($\alpha=4$, $\beta=6$)-, and Beta($\alpha=4$, $\beta=20$)-distributed jumping lengths, respectively. When the global search used 100000 iterations and, equivalently, 100 subsets of 1000 iterations in the nested-search, the discovery rate of the known global maximum from 100 experiments were all 100%.

The convergence of SA-a can be proved by theorem 1 of both Belisle (1992) and Hajek (1988). Belisle's theorem 1 is a special case of Hajek's result in which the state space is discrete and finite. SA-a is defined over subsets of \mathbb{R}^d , with a temperature scheme converging in probability to 0. Its transition probability from one candidate to another is positive. When M (the number of subsets) is sufficiently large, it can naturally deduce that SA-a converges in probability to the global minimum of the bounded space Ω .

3.2.2 An enrichment design: using protein group information and protein selection by group and individual

Under this more complex scenario, proteins are analyzed in biological groups. Selection of proteins at stages I and II is based on the combined criteria of group and individual hypothesis tests. A protein is selected if the single-protein test statistic exceeds the threshold of a corresponding type I error rate for the t -test or if the group test statistics exceeds the threshold of a corresponding type I error rate for the Hotelling's T -test. The

validation/selection of proteins at the final stage is only based on t -tests for the individual protein.

The following paragraph describes a simulated annealing algorithm SA-b, which is used to optimize and simulate the three-stage design when grouping information for each discovered protein is available in a paired sample study. Utilizing the additional grouping information, nested simulated annealing with Beta-distributed jumping lengths is used to find the optimal design solution. The selection criteria combine decision thresholds of Hotelling's T -squared statistics for the groups and the t -statistics for the individual proteins.

Apart from using grouping information, compared to SA-a, several improvements have also been made in SA-b. The first is to convert the inequality cost constraint into an equality cost constraint by using a series of slack terms (Nocedal and Wright, 1999). The second is the reduction in the dimension of the design problem by using the fact that the cost constraint will always bind. Instead of searching the entire interval of the stage III sample size, n_3 , now n_3 is derived from the current cost constraint and other chosen design parameters from the early stages. Because the cost function is monotonic with all the design parameters, this change reduces the computing time used to search those n_3 s with inferior solutions. The third improvement is to add an overall false-positive constraint in the algorithm.

3.2.2.1 Definition of the simulated annealing algorithm SA-b using grouping information

The solution space Ω bounded by the acceptable limit of each design parameter. Let the vector of design parameters $\omega = (n_2, n_3, \alpha_{t_1}, \alpha_{t_2}, \alpha_{f_1}, \alpha_{f_2})$ be a solution in Ω , where n_2 and n_3 are the Stage II and III sample sizes, α_{t_1} and α_{t_2} are the stage I and stage II type I error rates for the individual tests and α_{f_1} and α_{f_2} are the type I error rates for the group tests. Ω contains all the possible combinations of these parameters categorized into small intervals within the bounded ranges.

Objective function. Let $f(\omega): \Omega \rightarrow \mathbb{R}$ be the objective function of the solution space, where f is the expected number of proteins detected at stage III. The expected number of detected

proteins with true effects is subject to first- and second-stage type I error rates of the group Hotelling's T -tests, the individual t -tests, second-stage sample size and third-stage sample size. In the optimization, this objective function is constrained by: 1) cost and 2) the number of false positives. The selection criteria of the multi-stage design are:

- Stage I: (group test p-value $< \alpha_{f_1}$) or (individual test p-value $< \alpha_{t_1}$ & group test p-value < 0.05), i.e.

$$\left(T^2 > F^{-1}_{df(p_1), df(n_1-p_1)}(1-\alpha_{f_1})\right) \cup \left(T > Pt^{-1}_{df(n_1)}(1-\alpha_{t_1}/2) \cap T^2 > F^{-1}_{df(p_1), df(n_1-p_1)}(0.95)\right)$$

- Stage II: (group test p-value $< \alpha_{f_2}$ & individual test p-value < 0.05) or (individual test p-value $< \alpha_{t_2}$), i.e.

$$\left(T^2 > F^{-1}_{df(p_2), df(n_2-p_2)}(1-\alpha_{f_2}) \cap Pt^{-1}_{df(n_2)}(0.975)\right) \cup \left(T > Pt^{-1}_{df(n_2)}(1-\alpha_{t_2}/2)\right)$$

- Stage III: $T > Pt^{-1}_{df(n_2)}(0.975)$

In the above, α_{t_1} and α_{t_2} are the significance levels of individual tests at stages I and II; α_{f_1} and α_{f_2} are the significance levels of the group tests at stages I and II; T^2 is the F -distributed Hotelling's T -squared statistic with degrees of freedom determined by the number of proteins and the sample size at each stage; T is the Student t -statistic; F^{-1} is the quantile function for the F -statistic.

The configuration of the objective function is described in section 2.2.2.

A similar cost function as described in 2.1.2 is defined as $n_2 \times p_2 \times t_2 + n_3 \times p_3 \times t_3 + (n_2 + n_3) \times R - S$, where t_2 and t_3 denote the assay costs at stages II and III respectively, R is the recruiting cost, S is the slack term of the total budget, and p_2 and p_3 are the numbers of proteins being tested at stages II and III, respectively.

The false-discovery constraint controls the expected number of false discoveries and is defined as $m \times 2Pt(c_1) \times 2Pt(c_2) \times 2Pt(c_3)$, where m represents the total number of proteins with true effects.

The actual objective function in SA-b computes the expected number of positives by using the Monte Carlo average of 1000 simulations with adjustment for technical artifacts. To utilize the grouping information and according to requirements from the subject area, the first-stage criterion is set to select groups with a variable significance level that will be changed for different solutions in the optimization, and proteins with a variable significance level but belonging to a group significant at the fixed 0.05 level. The second-stage criterion is set to select proteins with a variable significance level that will be changed in the optimization, and proteins significant at the fixed 0.05 levels but belonging to groups with a changeable significance level. The third-stage selection is based only on the individual tests being significant at the 0.05 levels.

In SA-b, the proposal neighborhood selection function, temperature cooling schedule, and acceptance probability are set to be the same as those of SA-a. The algorithm of SA-a is summarized in table 3.1.

Table 3.1: The SA-b algorithm

Step 1. Assign study parameters: cost constraint, ‘technical artifact’ adjustment vector λ , mean difference and its standard deviation for each protein, and cost functions for stages II and III.

Step 2. Initialize number of iterations, simulated annealing parameters and solution. The simulated annealing parameters include ranges of stage I t test p values, F test p values, stage II t test p values, stage II F test p values, stage II sample size and the slack term. The solution includes the stage I & II t test and F test p values thresholds, and stage II sample size.

Step 3. Initialize the sequences of slack term, S_i , for the cost constraint; i ranges from 1 to J .

Step 4. While the number of iterations $< M$, repeat the following steps:

4.1 Randomly select an address as the centre of the local search neighborhood using a uniformly or Beta distributed jumping length

4.2 Activate simulated annealing for the local search with k iterations

The simulated annealing local search algorithm contains three functions: 1. the objective function, which uses Monte Carlo simulation to calculate the expected number of detected positives at the final stage; 2. the proposal neighborhood function, which determines the next searching subset of new candidate points; and 3. the cost-sample size function that calculates the stage III sample size according to the inequality cost constraint, slack term S_i , cost functions and the currently chosen design parameters.

4.3 Compare the local maximum with the best solution from the past. If the current solution is better, then replace the previous best solution with the current one.

4.4 Start next neighborhood search and repeat Step 3.

4.5 Repeat Step 2 using the next slack term S_{i+1} .

3.2.2.2 Use of analytical approximation to compute the analytical objective function for SA-b

In SA-b, using the Monte Carlo average to estimate the expected number of true discoveries prolongs the optimization process. To simplify the optimization, we investigated using an approximated analytical function to replace the Monte Carlo average. The expected number of true discoveries is given by

$$\sum_{i=1}^p (1 - \beta_{1,i})(1 - \beta_{2,i})(1 - \beta_{3,i}),$$

where $\beta_{1,i}$, $\beta_{2,i}$ and $\beta_{3,i}$ represent the nominal type II error rates at stages I, II and III, respectively, for the i th protein. Under the selection criteria for this multi-stage design utilizing the protein group information, described in section 2.2.1, the analytical function for the type II error, $\beta_{1,i}$, of the i th protein at stage I is equivalent to the probability that *the group containing the i th protein is not selected at the current group test decision threshold* (event A), and *either the i th protein is not selected at the current individual test decision threshold* (event B) or *the group is not selected at the 0.05 level* (event C).

The probability of the i th protein not being selected at stage I is, therefore, be expressed as $pr(A \cap (B \cup C))$, and can be expanded to

$$pr((A \cap B) \cup (A \cap C)) = pr(B) \times pr(A|B) + pr(A \cap C) - pr(B) \times pr((A \cap C)|B).$$

Analytically, $\beta_{1,i}$ is a function of the cumulative density function of the t -statistic and the cumulative density function of the group Hotelling's T -squared statistic which is F distributed after the transformation and is conditional on the individual t -statistic for each protein. It can be decomposed as follows.

Let $pr(B)$ denote the probability that the i th protein is not selected at the current t -test threshold. It can be expressed as $pr(B) = Pt(t < c_1 + t_i)$, described in 2.1.2, where c_1 is the threshold for the corresponding type I error of the t -test; and t_i is the t -statistic for the i th protein. Now, let $pr(A|B)$ denote the probability that the group containing the i th protein is not selected at the current group test decision threshold, given that the i th protein is not

selected at the current t -test threshold. This can be expressed as $pr(A|B) = F(T_i^2 < d_1 | t < c_1 + t_i)$, where T_i^2 represents the scaled F distributed Hotelling's T -squared statistic of the group containing the i th protein; and d_1 represents the F -statistic for p -value $<$ the decision threshold of Hotelling's T -test for the group.

$pr(A \cap C)$ is the probability that the group of i th protein is not being selected under the combination of the group test statistic thresholds ($d_{0.05}$ and d_1) and can be expressed as $pr(A \cap C) = F(T_i^2 < \min(d_1, d_{0.05}))$, where $d_{0.05}$ represents the F statistic for p -value $<$ 0.05 in the group test. Finally, let $pr((A \cap C)|B)$ denote the conditional probability of $A \cap C$ given the i th protein is not selected.

The conditional cumulative F density, defined as $pr(A|B) = F(T_i^2 < d_1 | t < c_1 + t_i)$, is equivalent to the marginal distribution of the cumulative F density with respect to the t -statistic for the i th protein, i.e.

$$F(T_i^2 < d_1 | t < c_1 + t_i) = \int_{-\infty}^{c_1 + t_i} F(T_i^2 < d_1) \times pt(t) dt, (a)$$

where $pt(t)$ represents the density function of the t -statistic, and Hotelling's T -squared statistic is given by

$$T_i^2 = \frac{\lambda_i (X_i - u_i)^T (X_i - u_i)}{S_i},$$

where

$$\lambda_i = \frac{n_1 - p_{1,i}}{p_{1,i}(n_1 - 1)}$$

denotes the scale factor which transforms Hotelling's T -squared statistic into an F -statistic; X_i and u_i are the observed and null-hypothesis means for all proteins in the group containing

the i th protein; S_i is the variance-covariance matrix of this group, n_i is the stage I sample size, and $p_{1,i}$ is number of proteins included in stage I for the group to which the i th protein belongs.

The integrand in equation (a) is approximated by $F(\tilde{T}_i^2 < d_1 - \lambda \times t_{1,i}^2) \times pt(t)$, i.e.

$$F(T_i^2 < d_1 | t < c_1 + t_i) \approx \int_{-\infty}^{c_1+t_i} F(\tilde{T}_i^2 < d_1 - \lambda_i t_{1,i}^2) \times pt(t) dt ,$$

where the group test, Hotelling's T -squared statistic T_i^2 , is approximated by the sum of \tilde{T}_i^2 and the t -statistic for the i th protein (i.e. $T_i^2 \approx \tilde{T}_i^2 + \lambda_i t_{1,i}^2$), and \tilde{T}_i^2 is T_i^2 excluding the mean effect of the i th protein (i.e. $\tilde{T}_i^2 = S^{-1} \lambda_i (X_{-i} - u_{-i})^T (X_{-i} - u_{-i})$).

Finally, we approximate $pr((A \cap C) | B)$ by $pr(A \cap C)$, which would be exact if B were independent of A and C .

A similar approximation is also applied to the stage II nominal type II error $\beta_{2,i} = Pr(B \cap (A \cup D))$, where D denotes the event that *a protein is not selected at the 0.05 significance level*. $\beta_{2,i}$ is expanded as

$$Pr((B \cap A) \cup (B \cap D)) = pr(B) \times pr(A | B) + pr(B \cap D) - pr(B \cap D) \times pr(A | (B \cap D)).$$

Table 3.2 summarized the algorithm of SA-b using the analytical approximation. The difference between Table 1a and Table 1b is notified by italicizing the texts. When using analytical approximation, the range of stage III sample size is included as a searching parameter.

The approximated analytical objective function for SA-b was implemented in several synthetic datasets for comparing with its Monte Carlo simulated function. The computing times of using the analytical approximation were shown to be between 20-100 times faster than using the Monte Carlo average in SA-b. The design parameters and solutions were also

shown to be similar to the results utilizing the Monte Carlo simulated objective function. More discussions are provided in the following immunology case study.

Table 3.2: The SA-b algorithm, with analytical approximation

Step 1. Assign study parameters: cost constraint, ‘technical artifact’ adjustment vector λ , mean difference and its standard deviation for each protein, and cost functions for stages II and III.

Step 2. Initialize number of iterations, simulated annealing parameters and solution. The simulated annealing parameters include ranges of stage I t test p values, F test p values, stage II t test p values, stage II F test p values, stage II sample size *and stage III sample size*. The solution includes the stage I & II t test and F test p values thresholds, stage II and stage III sample size.

Step 3. While the number of iterations $< M$, repeat the following steps:

3.1 Randomly select an address as the centre of the local search neighborhood using a uniformly or Beta distributed jumping length

3.2 Activate simulated annealing for the local search with k iterations

The simulated annealing local search algorithm contains three functions: 1. the objective function, which uses *analytical approximation function* to calculate the expected number of detected positives at the final stage; 2. the proposal neighborhood function, which determines the next searching subset of new candidate points; and 3. *the cost calculating function that calculates the cost based on the currently chosen design parameters*.

3.3 Compare the local maximum with the best solution from the past. If the current solution is better, then replace the previous best solution with the current one.

3.4 Start next neighborhood search and repeat Step 3.

3.3. Case studies

3.3.1 *An immunology study*

The lymphocyte proteome was analyzed in 17 Common Variable Immunity Deficient (CVID) patients and 34 normal controls. Common Variable Immunodeficiency Disorder (CVID), also known as acquired hypogammaglobulinemia, is the most common primary immunodeficiency disorder encountered in clinical practice (M. A. Park et al., 2008). CVID patients have low levels of immunoglobulin G, A and M; and also are susceptible to recurrent infections because of their inability to produce antibodies. Much of the past research has focused on deciphering the genetic basis of CVID (M. A. Park et al., 2008). However, the genetic causes of this disease are complex and still not fully understood. We hypothesize that proteomic characterization of CVID cases (beyond the gross immunoglobulin deficiencies) will be an alternative approach to reveal genetic causes and mechanisms. This study aimed to identify proteins with differentiated expression in CVID patients compared to the matched controls.

Patients and controls were matched by age group, ethnicity and gender. All patients and controls are Caucasian. Lymphocytes were isolated from blood using Ni-NTA agarose (Invitrogen) in an accredited lab (LabPLUS, Auckland City Hospital). The proteome of the lymphocytes cell lysates were then analyzed and quantified by liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) at the Center for Genomics and Proteomics, University of Auckland (MM). Samples were all analyzed using the tagged proteomics technique iTRAQ, where 8 samples were allocated as one batch of the multiplex assay. A reproducibility pilot study was performed before the main discovery study. Since the reproducibility of the experiments was shown to be satisfactory, the main study was performed.

This mass spectrometry-based approach identified peptides from 289 proteins and provided the relative quantification for each peptide. The log-transformed relative quantity of the peptide was used to derive the natural log transformed protein ratio for patients and normal controls. The proteins were grouped into 20 rudimentary classes according to their biological

function by the Biochemist (Dr. Woon), while the overlapped functions of some proteins were not presented.

Grouping of proteins

The 289 proteins discovered were grouped according to their functions: namely immunity, metabolic, tumor, protein synthesis/degradation, nuclear metabolic, cell migration, ER membrane, protein structure, signaling function, mitochondrial, blood protein, DNA repair/structure, trafficking/secretory, inflammation, apoptosis, autoantibody, ER/membrane, angiogenesis, transcription, neuro protection and redox. Nine groups were noted to contain protein candidates with significant fold changes between COVID patients and normal controls. Analyzing protein one by one, fifty-two proteins in total were considered initially for inclusion in the stage II verification study. We used the SA-a, SA-b, and the SA-b with analytical approximation to identify the optimal solutions for the second and third stages of this study.

Demonstration of code and results for three-stage design using SA-a, SA-b and approximation for SA-b

The cost structure used in this study is different to that described in 2.2.1. At stage II, the cost per protein for peptide synthesis is \$280 and per biological sample for proteomic analysis is \$1015. At stage III, the cost is assumed to be \$200 per protein per biological sample for laboratory analysis. The recruitment cost is set to be \$100 per biological sample. The assay cost functions in the R language for stages II and III are defined as $\text{assaycost2}=\text{function}(n,p)\{280*p+1015*n\}$ and $\text{assaycost3}=\text{function}(p)$, respectively, where p is the number of proteins selected at the nominal stage and n is the sample size.

The programs were run in the computer clusters of NeSI: <http://www.nesi.org.nz/>, where each program was assigned to 4GB memory within a cluster.

The codes used in the R function to utilize group information and analytical approximations are:

```
> optim.two.stage.appr (budget=6e6, protein=protein, N1=30,  
  
  artifact=rep(1,52),iter.number=10,assaycost2.function=assaycost2,  
  assaycost3.function=assaycost3,   recruit=100,   a1.t.min=0.01,   a1.t.max=0.25,  
  a1.f.min=0.01,   a1.f.max=0.25,   a1.step=0.025,   a2.t.min=0.01,   a2.t.max=0.05,  
  a2.f.min=0.05,   a2.f.max=0.05,   a2.step=0.025,   n2.min=100,   n2.max=1000,  
  n2.step=100, n3.min=100, n3.max=1000, n3.step=100) ,
```

where `artifact` records the vector of artifact adjustment factors of 52 proteins; `assaycost2.function` records the cost function for assay used at stage II; `assaycost3.function` records the cost function for assay used in stage III; `recruit` records the cost for recruitment per sample; `a1.t.min` and `a1.t.max` records the range of p values for t test at stage I; `a1.f.min` and `a1.f.max` record the range of p values for f test at stage I; `a1.step` and `a2.step` record the step size of p values thresholds in the searching; `n2.min` and `n2.max` record the range of sample size at stage II; `n2.step` record the step size of sample size at stage II. The stage II p values and stage III sample size are also recorded accordingly.

Table 3.3 The optimal design parameters for a given budget using three different algorithms for the multi-stage CVID proteomic study

Objectives	Method		
	SA-a	SA-b	SA-b, with analytical approximation
Full discovery of 52 proteins Cost=\$6×10⁶ n₂ between 100-1000	<i>pt₁,pt₂, n₂, n₃</i> : 0.10,0.04,500,517 cost stage II: 572,060 cost stage III: 5,426,940 time: 12.7 hours	<i>pt₁,pf₁, pt₂, pf₂, n₂, n₃</i> : 0.22,0.10,0.04,0.05,500,517 cost stage II: 572,060 cost stage III: 5,426,940 time: 20.0 hours	<i>pt₁,pf₁, pt₂, pf₂, n₂, n₃' (100-1000)</i> : 0.22, 0.22, 0.01, 0.05,365,100 Monte Carlo objective function used to derive n ₃ : 532 cost stage II:421,535; cost stage III: 5,577,465 time: 56 min
Full discovery of 52 proteins Cost=\$1.2×10⁶ n₂ between 30-100	<i>pt₁,pt₂, n₂, n₃</i> : 0.18,0.01,86,104 cost stage II: 110,450 cost stage III: 1,088,550 time: 11.7 hours	<i>pt₁,pf₁, pt₂, pf₂, n₂, n₃</i> : 0.11, 0.18, 0.01, 0.05,86,104 cost stage II: 110,450 cost stage III: 1,088,550 time: 19.0 hours	<i>pt₁,pf₁, pt₂, pf₂, n₂, n₃' (100-1000)</i> : 0.04,0.15,0.01,0.05,90,100 Monte Carlo objective function used to derive n ₃ : 104 cost stage II:114,910; cost stage III: 1,084,090 time: 53 minutes
Discovery of 5 most interesting proteins Cost=\$5×10⁵ n₂ between 100-1000	<i>pt₁,pt₂, n₂, n₃</i> : 0.20,0.01,330, 118 cost stage II: 369,350 cost stage III: 129,650 time: 3.2 hours	<i>pt₁,pf₁, pt₂, pf₂, n₂, n₃</i> : 0.05,0.20,0.01,0.05,330,118 cost stage II: 369,350 cost stage III: 129,650 time:7.0 hours	<i>pt₁,pf₁, pt₂, pf₂, n₂, n₃' (100-1000)</i> : 0.02,0.04,0.01,0.05,100,200 Monte Carlo objective function used to derive n ₃ : 351 cost stage II: 112,900; cost stage III: 386,100 time: 5 minutes
Discovery of 5 most interesting proteins Cost=\$5x10⁵ n₂ between 30-100	<i>pt₁,pt₂, n₂, n₃</i> : 0.01,0.01,60,392 cost stage II: 68,300 cost stage III: 430,700 time: 3.3 hours	<i>pt₁,pf₁, pt₂, pf₂, n₂, n₃</i> : 0.06,0.01,0.01,0.05, 60, 392 cost stage II: 68,300 cost stage III: 430,700 time: 8.3 hours	<i>pt₁,pf₁, pt₂, pf₂, n₂, n₃' (100-1000)</i> : 0.04,0.01,0.01,0.05,74,100 Monte Carlo objective function used to derive n ₃ : 378 cost stage II: 83,910; cost stage III: 415,090 time: 7 minutes

Of note * pt_1 : stage I t -test p-value; pt_2 : stage II t -test p-value; pf_1 : stage I F test p-value; pf_2 : stage II F test p-value; n_2 : stage II sample size; n_3 : stage III sample size derived from using the Monte Carlo simulated objective function which assume the solution use the entire allocated budget minus a slack term; n_3' is the solution from the algorithm using the analytical approximating objective function and this n_3' did not assume used the entire allocated budget.

*Stage III sample size, n_3 , was re-derived using the Monte Carlo simulated objective function, for the purpose of comparison. The solution from this Monte Carlo simulated function assumes the study used up the budget minus the slack term. The stage I sample size equals to the number of controls in this study. It needs to be greater than the number of proteins in each group.

The minimal stage II sample size also needs to be greater than the number of proteins in each group.

The approximation programs had run times within an hour. The group program SA-b had run times between 7-10 hours for the 5 proteins examples and between 20-30 hours for the 52 proteins examples.

Different budget ranges determined by the known health funding agents were tested for this case study. Three different budgets with the solutions are presented for the verification/validations of the top 5 proteins of interest, and the targeted 52 proteins in Table 3. Considering the relatively low prevalence of COVID, all budgets were assessed by different ranges of stage II sample size. To verify and validate the top 5 proteins, using ranges of 30-100 and 100-1000 for the stage II sample size are demonstrated to be feasible. The 1.2 million dollar budget was shown to be insufficient for a sample size between 100-1000 at stage II, 1000-5000 at stage III, but is sufficient for a sample size in the range of 30-100 at stage II and 100-1000 at stage III.

The solutions using the approximated analytical objective functions are shown to be in a similar range to the results from their Monte Carlo simulated objective functions, except for the first and third scenarios presented in Table 3.3. In these two scenarios, while both solutions from the analytical function achieve the same numbers of discoveries as their Monte Carlo simulation, they contain two smaller stage II sample sizes as the design parameters and thus results in different resource allocations. This discrepancy indicates the existences of multiple global optimal solutions for the objective function.

Despite the similar results (sample sizes and costs) from using and not using grouping information in this study, due to the large fold changes for all included proteins, proteins' functional group information is still considered essential for biologists to assess the discoveries and assist in the decision making in the protein selections from stage I.

3.3.2 Using simulated protein datasets

3.3.2.1 Data

To assess the performance of the SA-b algorithm and to investigate the factors that are associated with the efficiency of the program, we simulated different protein patterns from

synthetic datasets that were generated from a cardiac proteomic study (I.S.L. Zeng et al., 2009). The cardiac proteomic study collected coronary plasma blood samples of eight ischemic patients before and after an angioplasty procedure, and used LC-MS/MS with iTRAQ™ labeling to discover and quantify the proteins. The simulated datasets were created by using mean differences and ranges of variances in the relative quantity on the log scale between these two time points. Different patterns were simulated by setting the mean difference to zero or by doubling the variances of some proteins. The factors being investigated included the grouping property of proteins, number of proteins with non-zero mean differences, variations in the protein effect, and budgets. The grouping property focuses on the co-regulation of proteins in the same biological functional group, which are believed *a priori* to act in concert with one another.

Each synthetic dataset comprises 50 proteins of which 44, 18 or 6 have non-zero mean difference, which we will refer to as ‘true effects’. These are either clustered in a few groups or scattered across different groups, with some proteins either in overlapping or non-overlapping groups.

The expected number of discovered true effects (true positives, power) is affected by multiple factors. These factors include cost, significance thresholds at stages I and II, sample sizes at stage II and III, and the effect size (mean difference/standard deviation) of each protein. Results from some of these datasets are shown in Tables 4a, 4b and 4c.

3.2.2.2 Results using SA-b for a multi-stage design in different simulated protein datasets

Computation time and number of true effects. The simulations were implemented using computer clusters with 16 CPUs of 1 GB per CPU. Computation time is shown to increase with the number of true effects.

Budget, numbers of true effects and design parameters. In the simulated data of 44 true effects among 50 proteins (Table 3.4a), the budget of \$10 million results in 90% discovery. In the simulation with 18 (Table 3.4b) or 6 (Table 3.4c) true effects among the 50 proteins of interest, \$5 million is sufficient for 95% discovery in the 18 true-effects scenario and 100% discovery for the 6 true-effects scenario. The budget of \$1 million achieves 100% discovery

in the 6 true-effects scenario. All simulations use the same stage I sample size of 60 and the same cost function as described in sections 2.2.1 and 2.2.2 and footnotes of Tables 4c.

In scenarios where the \$10 million budget cannot achieve 100% discovery of all true effects, we note that the optimal stage I F -test decision threshold for selection is close to the upper bound of the parameter space. This phenomenon indicates that, the default 0.05 threshold would be far from optimal given the small sample size at stage I and the budget constraint. Both of the decision thresholds for the stage I F - and t -test are greater than 0.05. Conversely, in Table 4c, where a \$1 million budget can achieve a 100% discovery for the 5 true effects, the optimal stage I t -test decision threshold is smaller than 0.05.

The relations between the cost ratio of stage III-to-stage-II and the p-value of the stage I individual t -test, the cost ratio of stage III-to-stage-II and the p-value of the stage I group test were investigated using the 44 true effects data. The stage II sample size was fixed at 100, and the budget at \$5 million. The p-values of the stage I t -tests were set between 0.001 and 0.25, and the p-values of the stage I F -tests were set between 0.01 and 0.25. When using SA-a, the cost ratio is shown to decrease with a higher p-value threshold for the t -test (Figure 3.1b). When using SA-b, although a similar relation between the cost ratio and the p-value threshold for the groups' F -tests is observed, the p-value of the t -test does not influence the cost ratio within the same band of the F -test p-values.

Effect size and number of detectable true effects. In the synthesized datasets, there are several proteins with extremely small effect sizes that cannot be detected. The detection of these proteins are hindered by the sample size and significance thresholds at stages I and II. Under the unconstrained optimization, 100% discovery was achieved for the case of 44 true effects with a second stage sample size of 670 and third stage sample size of 2800 when the stage I individual test p-value < 0.36 and the second stage individual test p-value < 0.16 , given that the stage I sample size was 60. When there are no multiple stage selections, a sample size of 4751 can detect the protein with the smallest effect size (mean difference = 0.1, standard deviation = 2.3) with 85% statistical power and 5% type I error rate. This indicates that the detection of proteins with small effect size may be restricted using the systematic approach due to the step-wise type I error rate control and the constrained monetary resource in a proteome-wide study.

Convergence. SA-b is restricted to a smaller solution space, in which only those n_3 meeting the cost constraint are included. Thus, the convergence of both algorithm SA-b is better than SA-a when the same number of iterations is applied.

Overlapping groupings. When utilizing biological information, a protein may belong to several functional groups (Whitford, 2005). It is known that there are overlapping protein complexes sharing several proteins within biological networks. For example, in the TNF/NF- κ B signaling pathway, proteins p100, 1KKa, 1KKb and 1KKc are shared by several functional groups in this pathway (Zotenko et al., 2006). When utilizing SA-b, the overlapping proteins can be included in the group statistic for every group to which they belong.

3.4. A comparison between using grouping information and not using grouping information

Simulations using different synthetic protein datasets were conducted and used to investigate the influence of different protein patterns in optimizations using SA-a (without group information) and SA-b (with group information). When the budget is under a tight constraint and the grouping is informative, SA-b results in more proteins being selected from stage I given that the number of proteins in each group is less than the sample size. SA-a results in fewer proteins being selected from stages I, but larger sample size in stage III.

Comparison of protein discovery rates between SA-a and SA-b within the same ranges of design parameters shows that, SA-b has more favorable results in the protein-wise discovery when there is informative grouping. Informative grouping information increases the individual protein discovery rate and the average number of true discoveries. Uninformative grouping information does not make a meaningful difference to the discovery rate and cost allocation. The benefit of using grouping information is greater when the budget is under a tight constraint for detecting a large number of true effects. Under this condition, SA-b tends to allocate more resources to verifying more proteins at stage II. With respect to CPU running time, SA-b uses about twice to three times more system time than SA-a.

Table 3.5 provides scenarios of when to use SA-b and SA-a. Since SA-b with the analytical approximation runs much faster than the other two methods, it should be used firstly to assess whether a fixed budget will yield a good design solution to verify/validate the proteins of interest.

Table 3.5 Different scenarios to use SA-a and SA-b

When to use SA-a	When to use SA-b
<ol style="list-style-type: none"> 1. There is a small number of proteins that are of interest (i.e. <5). 2. The fixed budget will be more than sufficient for the verification/validation of all proteins of interests. 3. All proteins in the same group have a large effect size. 4. All proteins belong to a single group. 	<ol style="list-style-type: none"> 1. There is a large number of proteins that are of interest. 2. There is informative group information (i.e. some proteins have a large effect size and are clustered in the same group). 3. A number of proteins of interest have small effect size and cluster with proteins of large effect size in the same group.

3.5 Discussion

Proteomic techniques used to investigate large numbers of proteins simultaneously are comparable to genomic platforms used to investigate gene-disease associations, and have similar challenges in experimental design and data analysis (Greef et al., 2007). In this chapter, we used simulated annealing to simultaneously optimize the design for a multi-stage proteomic study comprising discovery, verification and clinical validation phases, taking into account the resource constraints for maximizing the number of true discoveries.

We investigated two different strategies for the design of a multi-stage clinical proteomic study, and recommend considering biological grouping information in the optimization of the design. Multi-stage designs are cost-effective because non-promising candidates can be eliminated after the first stage, leaving only promising candidates to be validated in later

stages. While, with the falling cost of genotyping, multi-stage designs are no longer commonly used in genome-wide association studies, they remain appealing for proteomic studies given the substantial per-protein cost of clinical validation. As suggested by the NCI, verification using a candidate-based platform and validation in large-scale clinical samples will improve the discoveries of disease related proteins and their final translation to utilization. A systematic approach to design optimization allows resources to be allocated efficiently across the different stages of the study. Further, using integrated biological information enriches the design for laboratory discovery and clinical application and thereby optimizes the solution. From simulations of different protein datasets in the current study, we discovered that using protein grouping information improves the optimization results when the grouping information is informative.

We also found that a structured two-step search was more efficient than a one-step global search and that using a Beta distribution for jump lengths in the two-step search further improved the speed.

A design based only on individual-protein tests could be optimized more easily because the objective function is smooth and can be calculated analytically, but individual-protein tests do not make full use of available biological information. Using a combination of individual-protein and group tests gives an objective function that has no simple analytical form, and for this reason Monte Carlo estimation and simulated annealing is necessary.

An important limitation of the current group algorithm is that Monte Carlo estimation prolongs the computing time required for the optimization process. However, the computations that form the main computing load can be easily parallelized, and the code made more efficient by using a faster programming language. Greater gains are also shown to be achievable from an analytical function to approximate the objective function. The current algorithms are conditional on the stage I discovery design parameters (sample size and number of discoveries). This limitation reflects a common problem in the funding process that many biomedical researchers currently face. Before significant funding can be sought for a multiple-phase study, pilot data from a stage I discovery is often needed as proof-of-concept; the stage I sample size is, therefore, determined by the available funds at this pilot stage. In general, the pilot study has a small available budget. As recommended in

the current practice, the stage I sample size is in the range of 10-100. However, some of our simulations showed that a larger stage I sample size (>100) leads to a smaller cost allocation in the stage II verification, and increase the statistical power at stage I. This suggests that a bigger range of sample size at stage I may need to be considered in some cases. This will be one topic of our future research.

3.6 The software

The R functions `optim.two.stage.single` (SA-a), `optim.two.stage.group` (SA-b) and `optim.two.stage.app` (SA-b using analytical approximation) performing the methods described in this paper are contained in the R package `proteomicdesign` 2.0. This package is available from the CRAN website: <http://www.r-project.org>. The R functions have been assessed and tested on multiple synthetic datasets (parts of these results were shown in this chapter), and an actual case study dataset at the desktop and the computer cluster. The R package manual is provided as the appendix following the final page of the thesis.

Table 3.4a Optimal design for a given budget in scenario: dataset comprises of 50 identified proteins of interest, and of which 44 proteins with true effects distribute in 7 of the 10 protein groups (proteins with true effects are clustering within groups; each group has more than one protein with true effect-informative grouping).

Budget	\$1 million		\$5 million		\$10 million	
	SA-b	SA-a	SA-b	SA-a	SA-b	SA-a
Expected number of true effects for the final optimized solution	No acceptable solution for a full discovery		40.8	40.5	41.4	41.1
Design parameters of the optimized solution	NA		$pt_1, pf_1, pt_2, pf_2, n_2, n_3$ 0.10,0.25,0.01,0.05,100,138	pt_1, pt_2, n_2, n_3 0.25,0.01,100,156	$pt_1, pf_1, pt_2, pf_2, n_2, n_3$ 0.135,0.225,0.05,0.05,200,274	pt_1, pt_2, n_2, n_3 0.25,0.05,200,306
Results (budget allocation, false negative rates, discovery rate, and computing time) for the current optimized solution						
False negative rates for different proteins of true effects	NA		Protein no. 100: 0.5% Protein no. 104: 0.2% Protein no. 137: 15.2% Protein no. 139: 9.3% Protein no. 142: 0.1% Protein no. 144: 0.6% Protein no. 146: 3% Protein no. 148: 0% Protein no. 149: 0.2% Protein no.s 105,121,145: 100%	Protein no. 100: 3.7% Protein no. 104: 1.7% Protein no.137: 26.3% Protein no. 139: 16.5% Protein no. 142: 0.1% Protein no. 144: 0.9% Protein no. 146: 4.3% Protein no. 148: 0% Protein no. 149: 0.3% Protein no.s 105,121,145: 100%	Protein no. 100: 0% Protein no. 104: 0% Protein no. 137: 0.1% Protein no. 139: 0.1% Protein no. 142: 0% Protein no. 144:0% Protein no. 146: 0% Protein no. 148: 0% Protein no. 149: 0% Protein no.s 105,121,145: 100%	Protein no. 100: 0.9% Protein no. 104: 0.2% Protein no. 137: 5.8% Protein no. 139: 4% Protein no. 142: 0% Protein no. 144: 0.2% Protein no. 146: 1.3% Protein no. 148: 0% Protein no. 149: 0.1% Protein no. 105,121,145: 100%

Budget	\$1 million		\$5 million			
	SA-a	SA-b	SA-a	SA-b	SA-a	SA-b
Discovery rates for different proteins of true effects	NA		100% for others (excluding proteins recorded above)			
Costs at stage II and III	NA	NA	Stage II: 3,727,840 Stage III: 1,271,160	Stage II: 3,585,760 Stage III: 1,413,240	Stage II: 7,437,120 Stage III: 2,561,880	Stage II: 7,171,520 Stage III: 2,827,480
Computation time	172,876 sec	69,787 sec	171,847 sec	69,530 sec	174,224 sec	97,464 sec

- The above program used ranges of stage I t -test p-value (pt_1) between 0.01 and 0.25 with step size 0.025; stage II t -test p-value (pt_2) between 0.01 and 0.05 with step size 0.025; stage I F test p-value (pf_1) between 0.01 and 0.25 with step 0.025; stage II F test p-value (pf_2) 0.01 and 0.05 with step 0.025; n_2 from 100 to 1000 with step 100; False positive rate < 0.01 . The final stage used t -test with $>85\%$ power at 0.05 significance level.
- Table summarized results used 9x1000 Hybrid Simulated Annealing search; all results were verified by 19x1000 SA search. The technical artifact λ is set to be (1, 1, 0.8, repeat (1, for 45 times), 0.9, 0.8). The assay cost is set to (N\$800, N\$200) with recruitment cost of N\$1000.00 and slack term cost of N\$1000.00.

Table 3.4b Optimal design for a given budget in scenario: dataset comprises of 50 identified proteins of interest, and of which 18 proteins with true effects distribute in 18 of the 18 protein groups (only one protein has true effect in each group- non informative grouping).

Budget	\$1 million		\$5 million		\$10 million	
	SA-b	SA-a	SA-b	SA-a	SA-b	SA-a
Expected number of true effects	No acceptable solution for a full discovery		16.9	16.8	16.9	17.0
Design parameters of the optimized solution	NA		$pt_1, pf_1, pt_2, pf_2, n_2, n_3$ 0.25,0.25,0.05,0.05,100,411	pt_1, pt_2, n_2, n_3 0.11,0.01,200,345	$pt_1, pf_1, pt_2, pf_2, n_2, n_3$ 0.25,0.25,0.05,0.05,200,820	pt_1, pt_2, n_2, n_3 0.25,0.05,200,1288
Results (budget allocation, false negative rates, discovery rate, and computing time) for the current optimized solution						
False negative rates for different proteins of true effects	NA		Protein no. 100: 4.1% Protein no. 137: 12.2% Protein no. 142: 0.1% Protein no. 105: 100%	Protein no. 100: 4.5% Protein no. 137: 11.1% Protein no. 142 :0.1% Protein no. 105: 100%	Protein no. 100: 3.3% Protein no. 137: 8.5% Protein no. 142 :0.1% Protein no. 105: 100%	Protein no. 100: 1.2% Protein no. 137: 4.7% Protein no. 142: 0% Protein no. 105:100%
Discovery rates for different proteins of true effects	NA		100% for others (excluding proteins recorded above)			
Costs at stage II and III	NA	NA	Stage II: 3,124,000 Stage III: 1,875,000	Stage II: 3,492,000 Stage III: 1,507,000	Stage II: 6,248,000 Stage III: 3,751,000	Stage II: 4,242,880 Stage III: 5,756,120
Computation time	219,787 sec	54,065 sec	225,338 sec	71,590 sec	226,110 sec	76,367 sec

- The above program used ranges of stage I t -test p-value (pt_1) between 0.01 and 0.25 with step size 0.025; stage II t -test p-value (pt_2) between 0.01 and 0.05 with step size 0.025; stage I F test p-value (pf_1) between 0.01 and 0.25 with step 0.025; stage II F test p-value (pf_2) 0.01 and 0.05 with step 0.025; n_2 from 100 to 1000 with step 100; False positive rate < 0.01. The final stage used t -test with >85% power at 0.05 significance level.
- Table summarized results used 9x1000 Hybrid Simulated Annealing search; all results were verified by 19x1000 SA search. The technical artifact λ is set to be (1, 1, 0.8, repeat (1, for 45 times), 0.9, 0.8). The assay cost is set to (N\$800, N\$200) with recruitment cost of N\$1000.00 and slack term cost of N\$1000.00.

Table 3.4c Optimal design for a given budget in scenario: dataset comprises of 50 identified proteins of interest, of which 6 proteins with true effects distribute in 2 of the 10 protein groups (informative grouping).

Budget	\$1 million		\$5 million		\$10 million	
	SA-b	SA-a	SA-b	SA-a	SA-b	SA-a
Expected number of true effects	No acceptable solution for a full discovery	5.9	6.0	6.0	6.0	6.0
Design parameters of the optimized solution	NA	pt_1, pt_2, n_2, n_3 0.01,0.01,100,176	$pt_1, pf_1, pt_2, pf_2, n_2, n_3$ 0.10,0.25,0.01,0.05, 100,1464	pt_1, pt_2, n_2, n_3 0.20,0.01,200, 1091	$pt_1, pf_1, pt_2, pf_2, n_2, n_3$ 0.10,0.25,0.01,0.05, 100,3687	pt_1, pt_2, n_2, n_3 0.09,0.04,472, 2585
Results (budget allocation, false negative rates, discovery rate, and computing time) for the current optimized solution						
False negative rates for different proteins of true effects	NA	Protein no. 144: 1% Protein no. 145: 1% Protein no. 146: 1% Protein no. 147: 1.2% Protein no. 148: 10% Protein no. 149:0.3%	Protein no. 149: 0.1%	Protein no.149: 0.1%	Protein no. 149: 0.1%	Protein no. 149: 0.1%
Discovery rates for different proteins of true effects	NA	100% for others (excluding proteins recorded above)				
Costs at stage II and III	NA	Stage II: \$580,000 Stage III: \$385,349	Stage II: \$1,300,000 Stage III: \$3,220,114	Stage II: \$2,760,000 Stage III: \$2,398,333	Stage II: \$1,300,000 Stage III: \$8,109,918	Stage II: \$4,248,000 Stage III: \$5,686,083
Computation time	111,288sec	39,585 sec	106,202sec	39,553sec	111,032sec	39,858 sec

- The above program used ranges of stage I *t*-test p-value (pt_1) between 0.01 and 0.25 with step size 0.025; stage II *t*-test p-value (pt_2) between 0.01 and 0.05 with step size 0.025; stage I *F* test p-value (pf_1) between 0.01 and 0.25 with step 0.025; stage II *F* test p-value (pf_2) 0.01 and 0.05 with step 0.025; n_2 from 100 to 1000 with step 100; False positive rate < 0.01. The final stage used *t*-test with >85% power at 0.05 significance level.
- Table summarized results used 9x1000 Hybrid Simulated Annealing search; all results were verified by 19x1000 SA search. The technical artifact λ is set to be (1, 1, 0.8, repeat (1, for 45 times), 0.9, 0.8). The assay cost is set to (N\$800, N\$200) with recruitment cost of N\$1000.00 and slack term cost of N\$1000.00.

Figures 3.1: The associations between cost ratios and test decision thresholds in scenarios of using vs. not using biological group information

Legend for Figure 1a: The 6 graphs represent the associations between cost ratios and stage I t -test, when the group F test p-values are in different ranges. The 6 graphs arrange in a descending order of the group test p-values, starting from the bottom left corner to the upper right corner. The protein dataset has 44 true effects among 50 proteins and is the same one to that used in table 4a.

Legend for figure 1b: The graph represents the association between cost ratios and stage I t -test p-values with a same range as that in figure 2a. The protein dataset has 44 true effects among 50 proteins discovered at stage I and is the same one to that used in table 4a.

Figure 3.1a Associations between cost-ratio-of-stage III-to-stage II, p-values of stage I individual *t*-test, and p-values of stage I group *F*-test

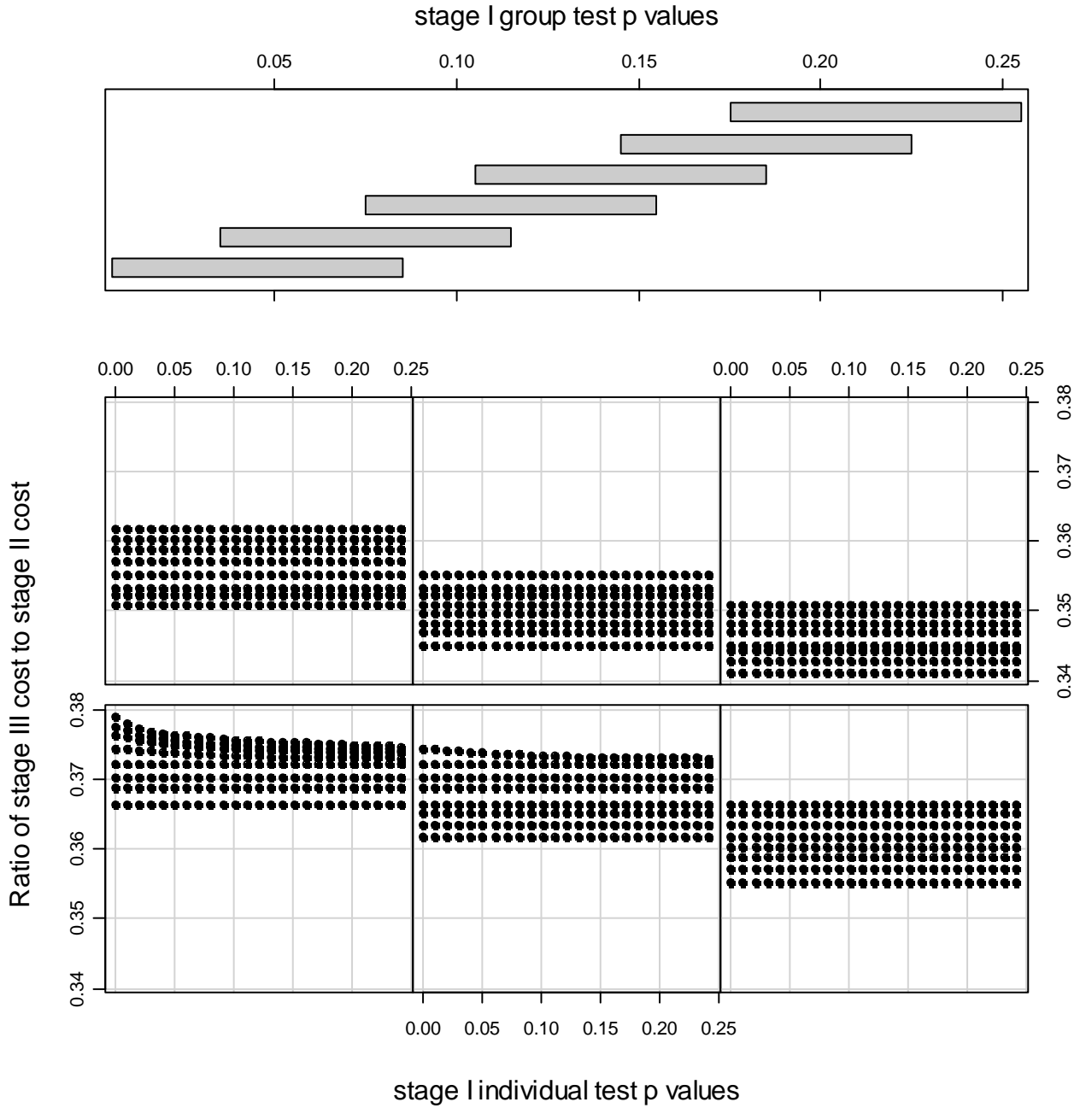
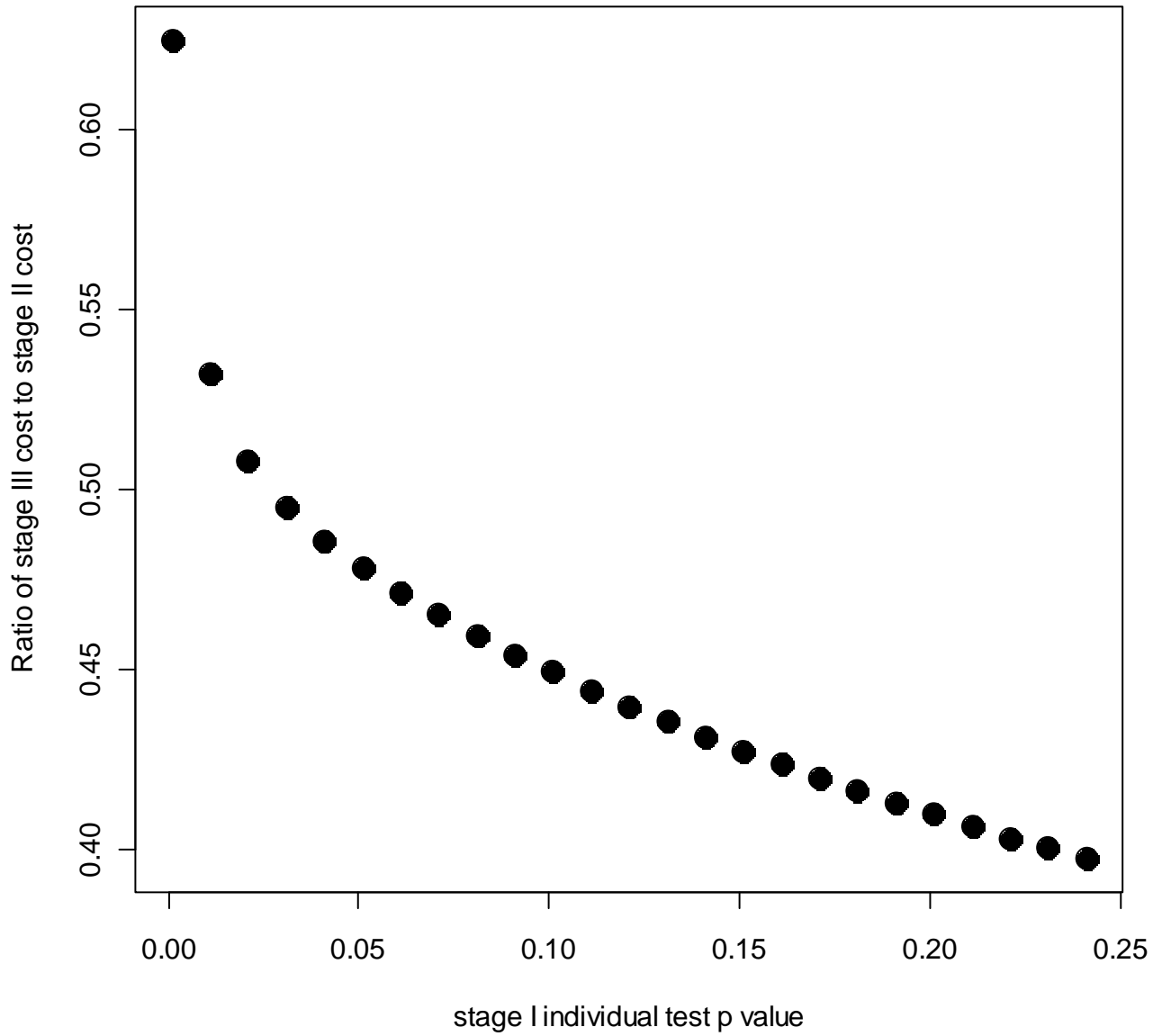


Figure 3.1b Associations between ratio-of-stage III-to-stage II-cost and stage I individual *t*-test



CHAPTER 4

A multivariate multilevel model for analyzing clinical proteomics data with non-random missingness

ABSTRACT

Introduction

Proteomics is emerging as a new stream in medical studies for investigating hundreds and thousands of molecular biomarkers simultaneously. The high-throughput data from proteomic study brings challenges to data analysis. The challenges originate from the hierarchical levels of the protein abundance data, the complexity of the experiment, the large amount of information, and the non-random missingness of the peptide intensity.

Method

We use multivariate multilevel models to analyse the hierarchical protein expression data. This proposed method takes into account the different types of variations from the experimental factors such as the physical features of the quantitative Mass Spectrometer, labels of ITRAQ, and potential run effects. It is demonstrated to be reliable for deriving the study parameters at the protein level comparing to using unadjusted protein ratio. Under this multivariate philosophy, a Bayesian hierarchical approach was used to handle the abundance-dependent missingness of the protein expression data and to provide shrinkage of overly-variable estimates. Gibbs sampling and Hamiltonian MC/No U-Turn Sampling were compared for evaluating the posterior joint distributions of the study parameters.

Results

The proposed methods were assessed in a simulated proteomic study and two clinical proteomics studies. The proposed multivariate multilevel model and the missing data approach enable us to cope with the large heterogeneity in the relative peptides intensity, from which the protein intensities are derived. It is shown to be an improvement compared to the protein ratios approach. The multivariate protein model utilizes experimental

information across all proteins and this enabled those proteins with small number of peptide information to be derived while adjusting for experimental effects. The HMC/NUTS sampler was substantially more efficient, as expected for a smooth, high-dimensional posterior distribution.

4.1 Introduction

Proteomics is emerging as a new stream in medical studies for investigating hundreds and thousands of molecular biomarkers simultaneously. It belongs to the family of system biology, including metabolomics, transcriptomics and genomics that study the interacting networks of different molecules using biochemical, mathematical, statistical and computer science methods inter-disciplinarily. Like aforementioned “omics”, proteomics utilizes biotechnology platforms that can systematically identify proteins and quantify their abundances. One of the popular platforms used for protein discovery is mass spectrometry, which can accurately determine the molecular mass of ions for peptide/protein sequencing and quantitation (Silva et al., 2005; Stephens et al., 2005; Vitzthum et al., 2005; Rodriguez et al., 2009). Through different ionization and detection processes, mass spectrometry (MS) coupled with other techniques (i.e. chromatography) (MacCoss and Matthews, 2005; Maddalo et al., 2005; Pelzing and Neususs, 2005; Shen and Smith, 2005; Palmblad, 2009; Kline and Sussman, 2010) enables the separation of different peptide species from the biological sample and produces large amounts of ions intensity data. These intensity data are used to inform the abundances of the peptides which are eventually used to construct the abundances of the proteins. The hierarchical structure of forming a protein’s abundance originates from the biochemical mechanism of the mass spectrometer proteomic experiment. In a mass spectrometer experiment, proteins of the tissues samples are tryptically digested into polypeptides before they are sent to the chromatography column for the initial separations when a chromatography device is used. These separated peptides are vaporized and ionized via different methods as introduced in chapter one (i.e. by electron spray or MALDI) in the mass spectrometry chamber, they are further broken into smaller charged molecules in the fragmentation chamber. The resultant charged molecules fly through the chamber, and hit on the ion detector which separates the molecule ions according to their mass-to-charge ratios (m/z), at last produces the ion intensity data. Since one peptide can be

fragmented into different combinations of smaller molecules, for the detection of peptides, the mass-to-charge ratios (m/z) of the constituent molecules will need to match against the protein library which has the known mass and charge information of different amino acids sequence. Finally, a further matching algorithm is used to derive the corresponding protein through the protein library.

When a proteomic experiment involves more than one biological sample, utilizing the biochemical peptide labeling system will enable the simultaneous analysis of multiple biological samples in a single experiment, or *run*, of the mass spectrometer through assay batching, or *multiplexing* (Kiyonami et al., 2005; Shadforth et al., 2005; Chong et al., 2006; D'Ascenzo et al., 2008). A multiplex comprises a batch of multiple samples for one single mass spectrometry experiment. Each sample is labeled in the multiplex for its identity in a mass spectrometry experiment. When multiple runs are involved, the biological samples will be allocated by labels and runs according to the experimental design.

The statistical analysis of the high-throughput data resulting from multiple runs of these multiplex-assays is challenging to perform at the protein level, due to the abundance data being in the form of peptide, instead of protein. The proteins' abundances must be derived from their observed constituent peptides' abundances that are derived from the ions intensity data. More complexities add in the data analysis due to the fact that, 1) each protein is not being identified by the same set of peptides in each run, 2) the experimental factors could have different impacts across different peptides/proteins. The confounding effects from a mass spectrometry experiments hence are multileveled. These non-ignorable multilevel confounding effects on the hierarchical structure of a protein's intensity become a significant cause of the challenges for data analysis. The other complexities include the sophisticated clinical study, unbalanced experimental design, large amount of information, and the non-random missingness of the peptide abundance data.

We use multivariate multilevel models to analyze the hierarchical protein abundance data in the clinical proteomics study, which takes into account different levels of variations in the study such as the experimental factors including physical features of the molecule, the labels of the multiplex, the biological features of proteins and the physiological features of the biological samples. Compared to the existing statistical methods in analyzing proteomic data, the novelty of our method is introducing the multilevel framework to the proteomic data

analysis so that the covariates can be distinguished and defined at the peptide, protein and biological sample levels, and the variance-covariance structures of protein, biological sample can also be defined accordingly. The multilevel framework also allows the clinical study design and experimental design be taken into account simultaneously for studies with multiple runs. Furthermore, we also used Bayesian hierarchical methods to handle the non-random missing intensity data, and made further improvements in modeling the non-random missingness by adding the mass-to-charge ratio (m/z) dependency and by separating the censoring missing from the other missing.

4.1.1. Quantification methods in Mass Spectrometers (MS) using iTRAQ labeling

4.1.1.1 Data structure and sources of variation of an iTRAQ experiment

The protein expression data generated from the mass spectrometer (MS) in proteomics experiments can either be the relative or absolute abundance of proteins. The relative abundances of proteins are measured by the intensity of the ions from the mass spectrometry, as opposed to absolute abundance of the protein which is measured by the concentration (Corthals and Rose, 2007). The relative quantitation of protein abundance utilizes different types of biochemical labeling techniques, including chemical, biological, metabolic and enzymatic incorporation, in a multi-samples proteomics experiment (Kiyonami et al., 2005; Shadforth et al., 2005; Corthals and Rose, 2007). Both of our case studies used iTRAQ, an enzymatic approach that is widely adopted in proteomics laboratories. iTRAQ is a biochemical reagent comprising a reactive group, a tag (label) with a mass balance group and a reporter group (figure 4.1)(Shadforth et al., 2005). In iTRAQ experiments, proteins from the multiple mixed biological tissue samples are digested into their constituent peptides by enzymatic reagents (i.e. trypsin) at the targeted c-termini in sample preparations (i.e. arginine or lysine in the case of trypsin) (Hamdan and Righetti, 2002; Sechi and Oda, 2003; Kiyonami et al., 2005; Boehm et al., 2007; Corthals and Rose, 2007; Wiese et al., 2007; D'Ascenzo et al., 2008). The cleaved peptides of each biological tissue sample are tagged by the peptide reactive group which connects to one of the label group in the 4-plex iTRAQ reagent assay (or an 8-plex assay). The tagged samples of a multiplex assay will also be purified and fractionated in order that the unbounded reagents are cleaned up. The purification is to make sure each peptide has a label. The purified multiple samples will then be mixed for the MS analysis simultaneously in a single experiment. The reporter groups of the iTRAQ

reagent, which have different molecular mass-to-charge ratios (m/z) ranging from 114.1-117.1 in the 4-plex assay and 113.1-121.1 in the 8-plex assay, are used to distinguish the multiple samples in the same assay. When the sample mixtures are ionized and charged, the reporter groups are knocked off in the fragmentation chamber. The molecular ions of the different reporter groups hit to the ions detector in different regions corresponding to their mass-to-charge ratios (m/z) which is the x axis in the diagram of intensity vs. mass-to-charge ratios. The abundance of the tagged peptide of a labeled sample can thus be measured by the intensities of these reporter group ions in the iTRAQ mass spectrometry analysis, and are used to inform the relative abundance for the protein (Leitner and Lindner, 2004). The current study focuses on the relative protein quantities from iTRAQ experiments, but the model can be generalized to other MS experiments.

4.1.1.2 Protein Ratios as the quantitation method (pros and cons)

In a single multiplex iTRAQ run, up to 4 or 8 samples can be analyzed simultaneously. For a 1:1 case-control experiment, a common approach is to place equal numbers of diseased cases and normal controls in one single run. One analytical method is to derive the distributions of the ratios for the relative abundances of the peptides between the cases and controls. The distribution of the peptide ratios informs the ratio of the protein abundance between cases and controls. The central tendency statistics (mean or median) of the peptides' ratios, which is a summary for the protein ratio, will indicate if the protein is deregulated when it significantly deviates from 1.

The advantages of using ratios are:

- 1) In a single run experiment, there is no need to control for between-run variation because the cases and controls are placed in the same run.
- 2) It is a relative measure of the magnitudes of interest (i.e. the difference between treatment and control) which eliminate potential label bias to some degree.

The disadvantages of using ratios are:

- 1) It only compares one single case with one single control in the same assay; different selections of the pairs of case and control will result in different distributions of the ratios.
- 2) It cannot take into account the variations from the experiment and the biological samples.
- 3) It is hard to use in clinical proteomic studies that consist of multiple runs (Oberg and Mahoney, 2012).

4) The threshold of protein ratios that is considered as being biological significant varies between studies (Hill et al., 2008), Seshi (2006) used > 1.2 or < 0.8 , Ross et al. (2004) used ± 1 standard deviation of the global ratios.

Instead of using a directly derived protein ratio, statisticians started using an analysis of variance model to derive the protein ratios from the model and account for the experimental variation across multiple runs (Hill et al., 2008). Whereas, a recent study by Breitwieser et al. (2011) proposed a method to derive protein ratios while taking into account usage of different pairs of biological samples in a single-run study under different experimental conditions (i.e. patient vs. control). They applied the multiplicative model to derive the variance of the relative peptide quantities and summarized the protein ratios weighted by their inverse-variances. The random protein ratios derived from their constituent peptides in the single run follow a Cauchy distribution; they are used to derive the p -values in experiments with biological replicates of different experimental conditions. They showed that this method outperforms the ANOVA model in handling the heteroscedasticity of the relative peptide quantity data in single-run experiments. Although there are other methods to derive protein ratios, Breitwieser's method is the only one that takes into account the biological variation in a single-run study.

In the proteomic literature, concerns have been raised that the ratios are shrunken towards 1 and leads to underestimation when using ratios in experiments of multiple runs (Zhanhua et al., 2005; Boehm et al., 2007; Corthals and Rose, 2007) (Boehm et al., 2007; Corthals and Rose, 2007; Karp et al., 2010). In multiple runs experiments, within-run normalization is inadequate for correcting run-to-run variation as the normalization does not adjust for between runs variation, and the most advanced pooled ratios approach cannot account for variations from different levels. By using an analysis of variance model, when the intensity measures are logarithm-transformed, the ratio estimate of each protein can be derived indirectly in the model. The precision of the derived ratio will also be adjusted for different sources of variation (i.e. from the experiments and biological variations) in the model. In a clinical proteomic study that needs to use advanced clinical study design and experimental design due to coexisting confounders between the biological samples, to decompose different sources of variations contributing to the intensity measures in a model is an essential approach. The model needs to cope with multiple experimental and biological factors, the hierarchical data structure, and the various structures of experimental and study design

simultaneously. It is also used to assess how influential each confounding factor is on the intensity measurement.

4.1.2 The instrumental feature of the mass spectrometer and the quantification

Apart from the hierarchical structure of the protein abundance data, it is observed that the physical instrumental features originating from the mass spectrometer influence the variation in the ions intensity. We discovered from the two proteomic iTRAQ studies that, the logarithmic intensity decreases with the mass-to-charge ratio (m/z) and the molecular mass (chapters 5 and 6). This indicates that the heavy molecules more likely miss the hits on the ion detector than light molecules. The identified variations in the intensity measurements caused by the physical nature of the mass spectrometer in our case studies are also consistent with the other studies (Breitwieser et al., 2011; Hrydziuszek and Viant, 2012). Breitwieser et al. (2011) observed more variability for the lower level signals and illuminated a funnel shaped association between the peptide ratio and the logarithmic intensity measurement. These evidences indicate that the relative abundance of proteins is likely to be underestimated by their heavy peptide components. The quantification of the protein needs to take the instrumental variations into account. The proposed multilevel model includes the mass-to-charge ratios (m/z) as one source of instrumental variations that contributes to measurements of the ion intensity, and is shown to increase the precision of the estimations for the unknown parameters including the protein ratios.

4.1.3 The missing mechanism in data from mass spectrometers

Missing data is commonly observed in medical studies. The missing mechanisms for abundance data from mass spectrometry are complex, comprising both random and non-random missingness components. The non-random missingness mechanism of proteins is driven by their peptides' abundances, masses, electrical charges, and ionization channel (Wells et al., 2011; Hrydziuszek and Viant, 2012); the likelihood of non-random missingness thus is associated with the protein abundance measured by the intensity value, the observable mass-to-charge ratio, and the reporter ion labels.

In an iTRAQ experiment, a peptide species observed in one labeled sample may not in any of the other samples within the same run, and/or it may be observed in only some runs but not others. The observed peptides with low signals have a greater number of missing intensity

values (Hrydziuszko and Viant, 2012). The intensity values of heavy peptides are more likely to be missed because their constituent molecular ions prone to failures of reaching the ion detector. The sensitivity of the mass spectrometer, which relates to the detectable level of the intensity in the device, will also have an impact on the data's completeness. A recent publication (Hrydziuszko and Viant, 2012) in metabolomics coincided with my postulation that the probability of having missing intensity value is a function of the signal level and the mass-to-charge ratio. Their data showed that higher probabilities of missingness were associated with the observed lower abundance peptides with lower peaks on the chromatogram and, furthermore that there is a curvilinear relationship between the probability of missingness and the mass-to-charge ratio. The probability is about 1 when the mass-to-charge ratio is less than 50 Th; it decreases with the mass-to-charge ratio in the mid-range of 200-300 Th, and then increases with the mass-to-charge ratios exceeding 300 Th. Luo et al. (2009) also reported that the efficiencies of peptide ionization and fragmentation affected the peptide intensity.

We propose a mixture model approach to take into account missingness mechanisms utilizing the physical properties of the peptides and the mass spectrometers; where logistic regression is used to estimate the effect of missing from m/z and intensity values.. We model the non-random missingness as either censoring below a threshold or completely missing. Censoring occurs when there are insufficient ions for a mass to be centroid to give a meaningful intensity value(Wells et al., 2011). In this situation, a peptide's intensity is below the instrument's detectable limit resulting in a zero intensity being recorded. Completely missing data correspond to those peptides without an intensity value (Wells et al., 2011). The probability of missingness (either by censoring or completely missing) is predicted by the observed peptides' abundances and mass-to-charge ratios. The regression coefficient parameters of the model for predicting missingness are treated as unknown in the proposed model, and are jointly estimated with the other parameters of the multilevel model under the Bayesian framework. The Bayesian approach enables us to utilize the prior information learned from the other studies (Hrydziuszko and Viant, 2012) and enrich the missing value imputation. Details of how the Bayesian multilevel model enhances the missing values imputation are explained in section 4.3.

4.2 The analytical methods

4.2.1 *The multilevel framework for the iTRAQ data*

Hill et al. (2008) proposed an ANOVA model including peptide as a categorical factor and peptides are nested within proteins and within a same run. Their multiplicative ANOVA model includes eleven variables to describe the variations of the observed peptide intensities. Through logarithmic transformation of the intensities, the multiplicative model simplifies to an additive model.

We proposed two sets of multilevel models that belong to the same statistical school as Hill's, and on the peptide intensity scale. One set of mixed models analyzes proteins individually, named as single protein model. The second set of mixed models analyzes a functional group or all of the proteins simultaneously, named as multiple proteins model. Both sets of models assume that peptides be nested within proteins and include the peptide as a covariate using its corresponding m/z ratio. The association between a peptide's m/z ratio and its relative abundance, which is the reporter ions peak area from the iTRAQ experiment, is taken into account in these models. In the multiple proteins model, proteins are treated as the secondary level units, peptides are treated as the first level unit. The response of the multivariate multilevel model is the logarithmic reporter ion peak area. The protein level estimates are derived from the information provided at the peptide level and allow the same peptide sequence to be used to identify different proteins. The protein level estimates of the regression coefficients are derived by their best linear unbiased predicted (BLUP) values in the mixed effect models. For proteins with a small number of biological and/or peptide samples, when the experiment has uniform influences on the peptide intensity, the multiple protein model will utilize the information available for all proteins and obtain better precision in their protein level estimates. . However, these two sets of models cannot distinguish the variance caused by post-translational modifications due to the absent information. An expanded model could be used when information on post-translational modification are available. Details of the single protein and multiple protein models are described in the following paragraphs.

4.2.2 The single protein model

The single-protein hierarchical multi-level model is defined using a two-level structure. Peptides nest within protein and are the level one unit. Since protein is analyzed one by one, the protein unit does not exist; we can equivalently treat peptides as if they nest within biological samples (subjects), which are the level two units. The level one of the model describes the relationship between the relative abundance of the peptide which is represented by the reporter ion peak area (intensity) and the experimental factors. The level two of the model describes the relationship between the random intercepts of subjects and those effects at the subject level such as demographics and condition (diseased or normal).

Define level one of the 2-level model:

In the following equations, fixed effect coefficients use the Greek letters, and random coefficients use the Roma letters.

$$y_{i,l} = b_{0,l} + \beta_1 mz_{i,l} + \sum_{h=1}^w \beta_{2,h} label_{h,i,l} + \sum_{r=1}^v \beta_{3,r} run_{r,i,l} + \sum_{h=1}^w \beta_{4,h} mz_{i,l} \times label_{h,i,l} + \sum_{r=1}^v \beta_{5,r} mz_{i,l} \times run_{r,i,l} + \sum_{k=1}^{wv} \beta_{6,k} label_{h,i,l} \times run_{r,i,l} + e_{i,l} \quad (1)$$

where

$y_{i,l}$ denotes the logarithmic transformed reporter ion peak area for peptide i and subject l , i has a range of 1 and n , l has a range of 1 and m ;

$b_{0,l}$ denotes the random intercept for subject l ;

$mz_{i,l}$ denotes the m/z ratio for peptide i and subject l ;

β_1 denotes the regression coefficient for m/z ratio;

$label_{h,i,l}$ is a dummy variable (0,1) for iTRAQ label h of the response $y_{i,l}$ for peptide i subject l ;; h is an integer number ranged between 1 and w ;

$\beta_{2,h}$ denotes the regression coefficient for label h ;

$run_{r,i,l}$ is a dummy variable (0,1) for the identity of run r that peptide i is identified for subject l ; r is an integer number with range between 1 to v ;

$\beta_{4,h}, \beta_{5,r}, \beta_{6,k}$ denote the regression coefficients for the interactions terms of m/z ratio & label, m/z ratio & run, and label &run;

$e_{i,l}$ denotes the unexplained residual error for peptide i and subject l .

Of note, although the same peptide sequence can be used for different subjects, the sequence itself is not included in the model but its m/z ratio is, as such the response $y_{i,l}$ represents the intensity value of a unique reporter ion in the model. The aforementioned Peptide i in essence is referred to the observed reporter ion i .

At level one of the model, the response is the logarithmic transformed reporter ion peak area $y_{i,l}$. The explanatory variables include the experimental factors as fixed effects, namely the iTRAQ $label_{i,l}, run_{i,l}, mz_{i,l}$, and their two-way interaction terms $mz_{i,l} \times label_{i,l}, mz_{i,l} \times run_{i,l}, label_{i,l} \times run_{i,l}$. It also includes a random effect, a random intercept $b_{0,l}$ for different subject l . Equation (1) defines an intercept term that is different across subjects, other terms such as run, label, m/z ratio and their two-way interactions that are the same for all *subjects*.

Define the level two of the 2-level model:

The level two of the model defines the effects of variations at the subject level through the random intercept $b_{0,l}$ at level one. The response is the random intercept $b_{0,l}$. The explanatory variables of this level include the condition (i.e. diseased vs. normal) and other subject level variables such as demographics.

$$b_{0,l} = \gamma_{0,0} + \sum_{b=1}^q \gamma_{1,b} \times z_{l,b} + \sum_{c=1}^g \gamma_{2,c} \times condition_{l,c} + u_{0,l},$$

(2)

where,

$\gamma_{0,0}$ represents the fixed intercept term;

$z_{l,b}$ denotes the covariate b of the subject l ;

$\gamma_{1,b}$ represents the regression coefficient for the subject level covariates $z_{l,b}$;

$condition_{l,c}$ is a dummy variable denoting the biological or psychological conditions (i.e. different interventions, disease vs. normal state) for the condition c and subject l ;

$\gamma_{2,c}$ represents the regression coefficient for the condition c ;

$u_{0,l}$ represents the random residual term of subject l .

Substituting equation (2) into equation (1) yields

$$y_{i,l} = \left[\begin{array}{l} \gamma_{0,0} + \sum_{b=1}^q \gamma_{1,b} \times z_{l,b} + \sum_{c=1}^g \gamma_{2,c} \times condition_{l,c} + \beta_1 m z_{i,l} + \sum_{h=1}^w \beta_{2,h} label_{h,i,l} + \sum_{r=1}^v \beta_{3,r} run_{r,i,l} + \\ \sum_{h=1}^w \beta_{4,h} m z_{i,l} \times label_{h,i,l} + \sum_{r=1}^v \beta_{5,r} m z_{i,l} \times run_{r,i,l} + \sum_{k=1}^{wv} \beta_{6,k} label_{h,i,l} \times run_{r,i,l} \end{array} \right] + [u_{0,l} + e_{i,l}] \quad (3)$$

where,

Random variable $u_{0,l} \sim N(0, \tau_{0,0}^2)$ represents the between-subject variation, which is normal distributed with mean 0 and variance $\tau_{0,0}^2$, and random variable $e_{i,l} \sim N(0, \sigma^2)$ represents the conventional residual error term of reporter ion intensity and is also normal distributed with mean 0 and σ^2 .

The fixed and random effects are grouped by separate square $[\cdot]$ brackets. The model defined in equation (3) assumes the regression coefficients $\beta_1, \dots, \beta_{6, w \times v}$ of experimental factors at the

peptide level are constant across subjects. The run effects can be treated as a random variable when it is hypothesized that there are variations introduced by different runs.

The variance-covariance matrix for the random effect at level two can be represented as \mathbf{G} , and the variance covariance matrix for the level one random residual is defined as \mathbf{R} . Since model defined in equation (3) only has one random residual variable at level two and one random residual at level one. The level two variance-covariance matrix \mathbf{G} only has one variance term $\tau_{0,0}^2$. The level one variance-covariance matrix \mathbf{R} for level one residual also only has one variance term σ^2 .

The variance-covariance matrix for the response $y_{i,l}$ is different from the variance-covariance matrix for the random effects.

When we assume that the level one error terms are independent, the variance and covariance for the response are derived as follows:

$$\begin{aligned} \text{var}(y_{il} | \beta_0, \dots, \beta_{w \times v}, u_{0,l}, X_{il}) &= \text{var}(u_{0,l} + e_{i,l}) = \tau_{0,0}^2 + \sigma^2 \\ \text{cov}(u_{0,l} + e_{i=j,l}, u_{0,l} + e_{i=k,l}) &= \text{cov}(u_{0,l}, u_{0,l}) = \tau_{0,0}^2 \end{aligned} \quad , \quad (4)$$

where $e_{i=j,l}$ and $e_{i=k,l}$ are the error term for peptide j and k respectively. The covariance in equation (4) only has one term.

When the covariance term for the paired residual term within a subject is not zero, in another word, when the level one error terms are not independent, the covariance term for the response as defined in equation (4) will become $\text{cov}(u_{0,l} + e_{i=j,l}, u_{0,l} + e_{i=k,l}) = \text{cov}(u_{0,l}, u_{0,l}) + \text{cov}(e_{i=j,l}, e_{i=k,l}) = \tau_{0,0}^2 + \delta_{j,k}$, Where $\delta_{j,k}$ is the covariance between the paired residuals of the reporter ion intensities.

The block diagonal matrix for the variances-covariance matrix $\mathbf{V}_{n \times n}$ of the response $y_{i,l}$ is

$$\begin{bmatrix} A_{l=1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & A_{l=m} \end{bmatrix}, \text{ where } A_l = \sigma^2 \otimes I_s + \mathbf{B} \otimes J_s, \text{ } s \text{ is the number of peptide observations of}$$

subject l , I_s represents the $s \times s$ identity matrix, and J_s represents the $s \times s$ matrix of ones. \mathbf{B} is the $s \times s$ matrix of $\{\tau_{0,0}^2 + \delta_{j,k}\}$, $\delta_{j,k}$ equals to σ^2 when $j=k$.

\mathbf{V} expands to a full matrix as

$$\begin{bmatrix} \tau_{0,0}^2 + \sigma^2 & \cdots & \tau_{0,0}^2 + \sigma_{j=1,k=s} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ \tau_{0,0}^2 + \sigma_{j=1,k=s} & \cdots & \tau_{0,0}^2 + \sigma^2 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \ddots & & \\ \vdots & \vdots & \cdots & 0 & \tau_{0,0}^2 + \sigma^2 & \cdots & \tau_{0,0}^2 + \sigma_{j=1,k=s} \\ \vdots & \vdots & \cdots & 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \tau_{0,0}^2 + \sigma_{j=1,k=s} & \vdots & \tau_{0,0}^2 + \sigma^2 \end{bmatrix},$$

where all diagonal terms are $\tau_{0,0}^2 + \sigma^2$ and off diagonal terms within the block sub-matrices are $\tau_{0,0}^2 + \sigma_{j,k}$, all the other off diagonal terms outside the block sub-matrices are zeros. For estimating the covariance term, \mathbf{V} can be further decomposed as $\mathbf{V} = \mathbf{Z}_{n \times q} \mathbf{G}_{q \times q} \mathbf{Z}'_{q \times n} + \mathbf{R}_{n \times n}$, Where \mathbf{Z} denotes the design matrix for the random effects excluding level one error term and q denotes the number of random effects at the subject level.

A sub-matrix of the block diagonal variance-covariance matrix of response, for example, for a subject with 3 peptide observations can be defined as

$$\begin{bmatrix} \tau_{0,0}^2 + \sigma^2 & \tau_{0,0}^2 + \sigma_{1,2} & \tau_{0,0}^2 + \sigma_{1,3} \\ \tau_{0,0}^2 + \sigma_{1,2} & \tau_{0,0}^2 + \sigma^2 & \tau_{0,0}^2 + \sigma_{2,3} \\ \tau_{0,0}^2 + \sigma_{1,3} & \tau_{0,0}^2 + \sigma_{2,3} & \tau_{0,0}^2 + \sigma^2 \end{bmatrix}, \text{ which has three extra unknown covariance terms.}$$

There are various covariance structures for the estimation of the unknown covariance terms $\sigma_{1,2}$, $\sigma_{1,3}$ and $\sigma_{2,3}$. Each kind of structure is determined by the assumptions of the within subject errors.

4.2.3 Group of protein mode

In comparison to the single protein model, the multiple protein models analyze multiple proteins in a multivariate model. Through the use of a multilevel model, the multivariate model can be analyzed as a univariate model where the different proteins are treated as a level 2 unit in the data hierarchy. Similar to the single protein model, the multiple protein model has two levels. Reporter ions representing peptides are the level one unit nested within proteins and subjects. Since the same protein is expected to be observed across different subjects, proteins and subjects are defined as cross factors and both are the level 2 units. Although in the real experiment, one protein may not be observed in every subject, the multilevel model allows unbalanced design and random missing because it does not require equal number of observations across the second level units. The level one of the model is similar as that is defined in equation (1). It defines the relation between the reporter ion intensity and the experimental factors. The level two of the model for the subject is also similar to equation (2) to describe the relation between the random intercept for subjects and the subject level variables. In addition, the level two of the model at the protein level selects regression coefficients that vary across different proteins as the random variables, and describes the relation between the selected protein level random coefficients and the explanatory variables.

Define level one of the 2-level model:

A full model with interactions at level one is defined as:

$$\begin{aligned}
 y_{i,l,p} = & \phi_{0,l} + \beta_{0,p} \times protein_{i,l,p} + \beta_{1,p} mz_{i,l,p} + \sum_{h=1}^w \beta_{2,p,h} label_{h,i,l,p} + \sum_{r=1}^v \beta_{3,p,r} run_{r,i,l,p} \\
 & + \sum_{h=1}^w \beta_{4,h} mz_{i,l,p} \times label_{h,i,l,p} + \sum_{r=1}^v \beta_{5,r} mz_{i,l,p} \times run_{r,i,l,p} + \sum_{k=1}^{wv} \beta_{6,k} label_{h,i,l,p} \times run_{r,i,l,p} + e_{i,l,p}
 \end{aligned} \tag{5}$$

where

$y_{i,l,p}$ denotes the logarithmic transformed reporter ion peak area for peptide i , subject l and protein p , i has a range between 1 and n , l has a range between 1 and m , and p has a range between 1 and z ;

$\phi_{0,l}$ denotes the random intercept for subject l ;

$protein_{i,l,p}$ denotes the protein identity for $y_{i,l,p}$ specifying that it is an intensity value for peptide i , subject l and protein p ;

$\beta_{0,p}$ denotes the random coefficient (intercept) for protein p ;

$mz_{i,l,p}$ denotes the m/z ratio for peptide i , subject l and protein p ;

$\beta_{1,p}$ denotes the random regression coefficient of m/z ratio for protein p ;

$label_{h,i,l,p}$ is a dummy variable (0,1) for iTRAQ label h of response $y_{i,l,p}$, h is an integer ranged between 1 and w ;

$\beta_{2,p,h}$ denotes the regression coefficient for label h and protein p ;

$run_{r,i,l,p}$ is a dummy variable (0,1) for the identity of run r that peptide i of protein p is identified from subject l , r is an integer with range between 1 and v ;

$\beta_{3,p,r}$ denotes the regression coefficient for run r and protein p ;

$\beta_{4,h}, \beta_{5,r}, \beta_{6,k}$ denote the regression coefficients for the interactions terms of m/z ratio & label, m/z ratio & run, and label & run respectively, it is not assumed to vary across proteins;

$e_{i,l,p}$ denotes the unexplained residual error for peptide i , protein p and subject l .

At this level of the model, it hypothesizes that intercept may vary across different proteins, the m/z ratio, runs and labels may have different effects on the peptide relative abundance (reporter ion intensity) across different proteins, and intercept may vary across different biological samples. The interactions term are assumed to be constant across proteins.

Define the level two of the 2-level model:

The random regression coefficients at the protein level can be described as the second level of the model,

$$\beta_{0,p} = \sum_{c=1}^g b_{1,p,c} \times condition_{l,c} + b_{0,p} ,$$

$$\beta_{1,p} = \mu_1 + b_{2,p} \quad \beta_{2,p,h} = \mu_{2,h} + b_{3,p,h} \quad \beta_{3,p,r} = \mu_{3,r} + b_{4,p,r} \quad (6)$$

where

$condition_c$ represents a dummy variable (0,1) indicating the physiological condition c (i.e. different interventions, disease vs. normal states) for subject l ;

$b_{0,p}$ represents a random residual of protein p for the random intercept varied across proteins;

$b_{1,p,c}$ represents a random residual of protein p and condition c for the random condition effect varied across proteins;

μ_1 represents the fixed intercept for m/z ratio, it is constant across proteins;

$b_{2,p}$ represents the random residual of protein p for the random slope of m/z varied across proteins;

$\mu_{2,h}, \mu_{3,r}$ represent the fixed intercepts for label h and run r respectively, they are constant across proteins;

$b_{3,p,h}, b_{4,p,r}$ represent the random residuals of protein p for the label and run respectively, they are varied across proteins.

The level 2 protein level model defines the random regression coefficients of intercept, m/z ratio, label run and subject's physiological condition, it is assumed that they are different across proteins. There are no fixed terms for the physiological conditions at the protein level as they are separately defined at the subject level in equation (7) shown below.

The random intercepts at the subject level can be described separately as

$$\phi_{0,l} = \gamma_{0,0} + \sum_{b=1}^q \gamma_{1,b} \times z_{l,b} + \sum_{c=1}^g \gamma_{2,c} \times \text{condition}_{l,c} + c_{0,l}, \quad (7)$$

where,

Covariates of the subjects, namely total-amount of proteins in the tissue, age, gender, etc., are represented by z_1, \dots, z_q ;

$\gamma_{0,0}$ represents the fixed intercept;

$\gamma_{1,b}, \gamma_{2,c}$ represents the fixed effect coefficients for the covariates b and the physiological condition c respectively, they are constant across subjects and proteins. $c_{0,l}$ represents the residual terms for subject l .

Of note, the $\gamma_{2,c}$ is the fixed effect term for physiological condition c and $b_{1,p,c}$ is the random effect term for condition c varied across proteins.

The level 2 subject level model defines the random intercept $\phi_{0,l}$ for each subject sample which is estimated by the combination of the fixed subject level covariates, z_1, \dots, z_q , the fixed $\text{condition}_{l,c}$ of l , and the unexplained random error term $c_{0,l}$.

Substituting equations (6)-(7) into equation (5) and grouping the random effects and fixed effects by separate square $[\cdot]$ brackets, we have the following equation (8),

$$\begin{aligned}
y_{i,l,p} = & \left[\left(\gamma_{0,0} + \sum_{b=1}^q \gamma_{1,b} \times z_{l,b} + \sum_{c=1}^g \gamma_{2,c} \times condition_{l,c} \right) + \right. \\
& \mu_1 mz_{i,l,p} + \sum_{h=1}^w \mu_{2,h} label_{h,i,l,p} + \sum_{r=1}^v \mu_{3,r} run_{r,i,l,p} \\
& \left. + \sum_{h=1}^w \beta_{4,h} mz_{i,l,p} \times label_{h,i,l,p} + \sum_{r=1}^v \beta_{5,r} mz_{i,l,p} \times run_{r,i,l,p} + \sum_{k=1}^{wv} \beta_{6,k} label_{h,i,l,p} \times run_{r,i,l,p} \right] \\
& + \left[\left(\sum_{c=1}^g b_{1,c} \times condition_{l,c} \right) \times protein_{i,l,p} + b_{2,p} \times mz_{i,l,p} + \sum_{h=1}^w b_{3,p,h} \times label_{h,i,l,p,h} + \sum_{r=1}^v b_{4,p,h} \times run_{r,i,l,p} \right] \\
& + b_{0,p} \times protein_{i,l,p} + c_{0,l} + e_{i,l,p}
\end{aligned} \tag{8}$$

Let \mathbf{B} be the vector of random effects for proteins, \mathbf{C} be the vector of random residuals for subjects, and \mathbf{e} be the vector of the random errors. We define the distribution of these random effect vectors as follows:

$$\begin{aligned}
\mathbf{B} &= (b_{0,p}, b_{1,p}, b_{2,p,1}, \dots, b_{2,p,w}, b_{3,p,1}, \dots, b_{3,p,v}, b_{4,p}) \sim MVN(0, \Phi), \\
\mathbf{C} &= (c_{0,1}, \dots, c_{0,l}, \dots, c_{0,n}) \sim MVN(0, \mathbf{G}), \quad \mathbf{e} \sim MVN(0, \mathbf{R}),
\end{aligned}$$

where Φ be the variance-covariance matrix for parameters at the protein level and \mathbf{G} be the variance-covariance matrix for parameters at the subject level, and \mathbf{R} is the variance-covariance matrix for the random residual errors at the peptide level;

The variance-covariance matrix for all the random terms is a block diagonal matrix

$$Var \begin{bmatrix} \mathbf{B} \\ \mathbf{C} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \Phi & 0 & 0 \\ 0 & \mathbf{G} & 0 \\ 0 & 0 & \mathbf{R} \end{bmatrix}, \text{ since } \mathbf{B} \text{ and } \mathbf{C} \text{ are the crossed random effects in the model. We}$$

further define the symmetric matrix Φ as follows, to simplify the illustration assuming there are two physiological conditions, two runs and two labels:

$$\Phi = \begin{bmatrix} \sigma_{b,0}^2 & \sigma_{b,0}\sigma_{b,1} & \sigma_{b,0}\sigma_{b,2} & \sigma_{b,0}\sigma_{b,3} & \sigma_{b,0}\sigma_{b,4} \\ \sigma_{b,0}\sigma_{b,1} & \sigma_{b,1}^2 & \sigma_{b,1}\sigma_{b,2} & \sigma_{b,1}\sigma_{b,3} & \sigma_{b,1}\sigma_{b,4} \\ \sigma_{b,0}\sigma_{b,2} & \sigma_{b,1}\sigma_{b,2} & \sigma_{b,2}^2 & \sigma_{b,2}\sigma_{b,3} & \sigma_{b,2}\sigma_{b,4} \\ \sigma_{b,0}\sigma_{b,3} & \sigma_{b,1}\sigma_{b,3} & \sigma_{b,2}\sigma_{b,3} & \sigma_{b,3}^2 & \sigma_{b,3}\sigma_{b,4} \\ \sigma_{b,0}\sigma_{b,4} & \sigma_{b,1}\sigma_{b,4} & \sigma_{b,2}\sigma_{b,4} & \sigma_{b,3}\sigma_{b,4} & \sigma_{b,4}^2 \end{bmatrix},$$

where the diagonal terms $\sigma_{b,0}^2, \dots, \sigma_{b,4}^2$ represent variance of the protein intercept, physiological condition, slope of m/z, label and run of protein p respectively, the off diagonal terms represent their pair-wised covariance.

\mathbf{G} and \mathbf{R} will be the same as defined in the simple protein model. We assume the random effects for subjects are independent of the random effects for the proteins, and are independent of the random errors at the peptide level. The variance-covariance matrix \mathbf{V} of response $y_{i,l,p}$ can be decomposed as $\mathbf{V} = \mathbf{Z}_{0,n \times 5} \Phi_{5 \times 5} \mathbf{Z}'_{0,5 \times n} + \mathbf{Z}_{1,n \times q} \mathbf{G}_{q \times q} \mathbf{Z}'_{1,q \times n} + \mathbf{R}_{n \times n}$. The random effect matrix $\mathbf{Z}_{0,n \times 5}$ for protein has n rows and 5 columns, and the random effect matrix $\mathbf{Z}_{1,n \times q}$ for subjects has n rows and q columns, n is the total number of reporter ion intensities of all proteins. The diagonal term of \mathbf{V} is the sum of variances term of protein level parameters, subject level parameters and the random error residuals. The off diagonal term of \mathbf{V} can be derived from the aforementioned decomposition equation.

4.3 The missing mechanism for the iTRAQ data- a Bayesian approach

4.3.1 A Bayesian approach

In the iTRAQ experiment, two types of missing peptide data from mass spectrometers are observed; one is identified as zero and the other is identified as blank. The zeros intensities are reported when there are not enough ions for the mass of the reporter ion to be centroid. The blanks are reported when the signal statistics are missing (Wells et al., 2011). The zeros can be defined as censored missing with values lower than a detectable threshold, and the blank missing are missed intensity due to a weak signal. Both types of missing values are considered to be related to the abundance of the proteins (Hrydziuszko and Viant, 2012).

Through the Bayesian framework, we can learn the missing data information from the other studies and incorporate them through the priors for the models. For example, the missing

probability of the ions intensity can be estimated using its associations with the instrumental features of the mass spectrometer reported in the other studies. We now propose a multivariate multilevel Bayesian model that can estimate the missing probability and missing values as the unknown variables, using the observed data and prior information learned from the other studies. We describe a hierarchical two-level model that includes a dummy variable for one physiology condition without interaction terms assuming that the interactions between the label and run, m/z and run, and m/z and label are not significant in equation (9). Equation (9) below defines the joint distribution of the observed and the missing logarithmic intensity values of the peptides conditional on the observed explanatory variables. It also outlines the relation between the peptide intensity and the explanatory variables including m/z ratio, run, label, and condition.

$$\left(\gamma_{i,l,p}^{obs}, \gamma_{i,l,p}^{miss} \mid mz_{i,l,p}, run_{r,i,l,p}, label_{h,i,l,p}, condition_{i,l,p}, \mathbf{U}_{0,p}, \mathbf{U}_{1,p}, \mathbf{U}_{2,p}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \right) \sim MVN(\boldsymbol{\gamma}_{i,p}, \boldsymbol{\Omega})$$

$$\boldsymbol{\gamma}_{i,p} = \mathbf{U}_{0,p} + \mathbf{U}_{1,p} \mathbf{mz}_{i,p}^T + \mathbf{U}_{2,p} \mathbf{condition}_{i,p}^T + \boldsymbol{\beta}_0 \mathbf{subject}_{i,l}^T + \boldsymbol{\beta}_1 \mathbf{label}_{i,l}^T + \boldsymbol{\beta}_2 \mathbf{run}_{i,l}^T$$

(9)

where

$\gamma_{i,l,p}^{obs}$ represents the observed logarithmic intensity values, and $\gamma_{i,l,p}^{miss}$ represents the missing logarithmic intensity values of protein p and the biological sample l ;

$\mathbf{U}_{0,p}, \mathbf{U}_{1,p}, \mathbf{U}_{2,p}$ are regression coefficients for the p protein, $\mathbf{U}_{0,p}$ is the vector of intercept, $\mathbf{U}_{1,p}$ is the vector of regression coefficients for the m/z ratio, and $\mathbf{U}_{2,p}$ is the vector of regression coefficients for the condition;

$\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ represent vectors of regression coefficients for different subjects, different iTRAQ labels and different runs respectively, They are assumed to be the same across different proteins;

Ω is a diagonal variance-covariance matrix for $\gamma_{i,l,p}^{obs}$ with a common unknown variance term σ^2 across diagonal terms;

$\gamma_{i,p}$ denotes the mean intensity for subject l , $\gamma_{i,p}$ is predicted by the covariates at the protein level, the subject level and the peptide level;

$\mathbf{mz}_{i,p}^T$, $\mathbf{condition}_{i,p}^T$, $\mathbf{subject}_{i,l}^T$, \mathbf{label}^T and \mathbf{run}^T represent the row vectors for m/z, physiological condition, subject, label and run respectively.

The variable $(\gamma_{i,l,p}^{obs}, \gamma_{i,l,p}^{miss})$ denoting the completed data of peptide's intensity on the logarithmic scale is assumed to be multivariate normal distributed with mean $\gamma_{i,p}$ and variance-covariance matrix Ω . For the purpose of explanation, equation (9) only includes intercept term as the random variable at the subject level. It can be easily extended to a model with other covariates at the subject level.

Let the matrix of regression coefficients be $\mathbf{B} = (\mathbf{U}_{0,p}, \mathbf{U}_{1,p}, \mathbf{U}_{2,p}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, the following equation (10) sets up prior distributions of the regression coefficients at the protein level,

$$\begin{pmatrix} \mathbf{U}_{0,p} \\ \mathbf{U}_{1,p} \\ \mathbf{U}_{2,p} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \alpha_{0,p} \\ \alpha_{1,p} \\ \alpha_{2,p} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{21} & \tau_{22} & \tau_{23} \\ \tau_{31} & \tau_{32} & \tau_{33} \end{pmatrix} \right),$$

(10)

where

$(\mathbf{U}_{0,p}, \mathbf{U}_{1,p}, \mathbf{U}_{2,p})$ is multivariate normal distributed with mean $\boldsymbol{\alpha} = (\alpha_{0,p}, \alpha_{1,p}, \alpha_{2,p})$ and variance-covariance $\boldsymbol{\Sigma}$, assuming $\boldsymbol{\Sigma}$ is the same for all subjects l ;

The hyper-prior of $\boldsymbol{\alpha}$ is multivariate normal distributed with mean \mathbf{f} and variance-covariance \mathbf{G} , $\boldsymbol{\alpha} = (\alpha_{0,p}, \alpha_{1,p}, \alpha_{2,p}) \sim MVN(\mathbf{f}, \mathbf{G})$;

The hyper-prior for Σ is inversed Wishart distributed $\Sigma^{-1} \sim \text{Wishart}(\mathbf{R}, d = 3, \nu_0)$, and \mathbf{R} is a $d \times d$ positive scaled definite matrix, $\nu_0 > d - 1$. We can assign non-informative or informative hyper-priors for vector \mathbf{f} , matrices \mathbf{G} and \mathbf{R} .

Priors for the residual variance σ^2 and the random intercept for subject β_0 and regression coefficients β_1, β_2 are set up as follows:

$$\sigma^2 \sim IG(a, b), \beta_1 \sim N(0, d_1), \beta_2 \sim N(0, d_2), \beta_0 \sim N(0, d_3),$$

where σ^2 is inversed Gamma distributed with shape a and scale b , $\beta_0, \beta_1, \beta_2$ are normal distributed with means 0 and variance d_1, d_2 and d_3 respectively, non-informative priors can be assigned to a, b, d_1, \dots, d_3 if no prior information is available.

After defining the model for the completed peptide intensity, the parameters involved for estimating the missing components (missing values and probability of missing) needs to be incorporated in the model. Let **miss** be a binary indicator variable denoting if the intensity has a missing value (1: missing, 0: non-missing). Based on the prior knowledge, we know that the probability of missing associated with the intensity and the m/z ratio. If we assume that the probability of intensity value being missed $\mathbf{pm} = \{pm_{i,l,p}\}$ (for $miss_{i,l,p} = 1$) is Bernoulli distributed, the joint distribution of the data and the unknown parameters including parameters for estimating the missing components can be constructed as equation (11),

$$\begin{aligned} & f(\gamma^{obs}, \gamma^{miss}, \mathbf{X}, \mathbf{miss}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \Sigma) \\ &= f_1(\gamma^{obs}, \gamma^{miss}, \mathbf{X}, \mathbf{miss} | \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \Sigma) \times g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \Sigma) \quad , \\ &= g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \Sigma | \gamma^{obs}, \gamma^{miss}, \mathbf{X}, \mathbf{miss}) \times f_1(\gamma^{obs}, \gamma^{miss}, \mathbf{X}, \mathbf{miss}) \end{aligned} \tag{11}$$

where

$$\gamma^{obs} = \{\gamma^{obs}_{i,l,p}\}, \gamma^{miss} = \{\gamma^{miss}_{i,l,p}\};$$

$\mathbf{X} = \{\mathbf{X}^{obs}, \mathbf{X}^{miss}\}$ is the data matrix that has the number of rows equivalent to the total number of observed and missing peptide quantities, and the number of columns equivalent to the total number of explanatory variables, the other variables have already been defined in equation (9) and (10);

$f_1(\gamma_{i,l,p}^{obs}, \gamma_{i,l,p}^{miss}, \mathbf{X}, \mathbf{miss} | \mathbf{B}, \sigma^2, \mathbf{a}, \mathbf{\Sigma})$ represents the likelihood function conditional on the unknown parameters;

$g_1(\mathbf{B}, \sigma^2, \mathbf{a}, \mathbf{\Sigma})$ represents the joint density function for the priors of all unknown parameters;

$g_1(\mathbf{B}, \sigma^2, \mathbf{a}, \mathbf{\Sigma} | \gamma_{i,l,p}^{obs}, \gamma_{i,l,p}^{miss}, \mathbf{X}, \mathbf{miss})$ represents the conditional posterior density of the unknown parameters given the observations;

$f_1(\gamma_{i,l,p}^{obs}, \gamma_{i,l,p}^{miss}, \mathbf{X}, \mathbf{miss})$ represents the marginal likelihood function for the observed and missing intensity values, and it is equivalent to $f_1(\gamma^{obs}, \gamma^{miss}, \mathbf{X})$ because \mathbf{miss} is an indicator variable derived from $\gamma_{i,l,p}^{miss}$.

The marginal distribution $f_1(\gamma^{obs}, \gamma^{miss}, \mathbf{X})$ is subject to the regression likelihood, it can also be denoted as $f_1(\gamma^{obs}, \gamma^{miss} | \mathbf{XB}, \delta^2)$, where the design matrix \mathbf{X} is defined as

$$(X_{obs}, X_{miss}) = \begin{pmatrix} subjects & proteins & run_1 & \dots & run_v & label_1 & \dots & label_w & condition \\ x_{i=1,l,p} & x_{i=1,l,p} & x_{i=1,l,p} & \dots & x_{i=1,l,p} & x_{i=1,l,p} & \dots & x_{i=1,l,p} & x_{i=1,l,p} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i=n,l,p} & x_{i=n,l,p} & x_{i=n,l,p} & \dots & x_{i=n,l,p} & x_{i=n,l,p} & \dots & x_{i=n,l,p} & x_{i=n,l,p} \end{pmatrix},$$

where v is an integer for the identity of the last run, w is an integer for the identity of the last label, run_{1-v} and $label_{1-w}$ are dummy variables to record the corresponding label and run for the response –the peptide intensity.

(of note: in the following sections, functions related to the unknown parameters will be denoted as g with different subscripts, and functions related to the likelihood function for the observations will be denoted as f with different subscripts).

Based on equation (11), we can derive the conditional posterior distribution of the unknown parameters $g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma} | \boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss}, \mathbf{X}, \mathbf{miss})$ as follows,

$$\begin{aligned} & g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma} | \boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss}, \mathbf{X}, \mathbf{miss}) \\ &= \frac{f_1(\boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss}, \mathbf{X}, \mathbf{miss} | \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \times g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma})}{f_1(\boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss}, \mathbf{X}, \mathbf{miss})}. \end{aligned}$$

The marginal distribution $f_1(\boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss}, \mathbf{X}, \mathbf{miss})$ with respect to all the possible values of the unknown parameters is considered to be constant, therefore the joint conditional posterior distribution of the unknown parameters is proportional to the product of the priors of the unknown parameters and the likelihood for the data observed:

$$\begin{aligned} & g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma} | \boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss}, \mathbf{X}, \mathbf{miss}) \\ & \propto f_1(\boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss}, \mathbf{X}, \mathbf{miss} | \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \times g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \\ & = f_1(\boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss} | \mathbf{X}, \mathbf{miss}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \times f_2(\mathbf{X}, \mathbf{miss} | \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \times g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \end{aligned} \quad (12a)$$

Since the probability of missing is independent of $\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}$, the right hand side of (12a) is simplified to

$$f_1(\boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss} | \mathbf{X}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \times f_2(\mathbf{miss} | \mathbf{X}) \times g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}),$$

with further factorization for the first term $f_1(\boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss} | \mathbf{X}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$ as shown below,

$$\begin{aligned} & f_1(\boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss} | \mathbf{X}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \\ & = f_1(\boldsymbol{\gamma}^{obs} | \boldsymbol{\gamma}^{miss}, \mathbf{X}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \times f_1(\boldsymbol{\gamma}^{miss} | \mathbf{X}^{miss}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \times g_3(\boldsymbol{\gamma}^{miss}) \end{aligned}$$

Then substitute the above in the right hand side of equation (12a), (12a) finally becomes

$$\begin{aligned} & g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma} | \boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss}, \mathbf{X}, \mathbf{miss}) \\ & \propto f_1(\boldsymbol{\gamma}^{obs} | \boldsymbol{\gamma}^{miss}, \mathbf{X}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \times f_1(\boldsymbol{\gamma}^{miss} | \mathbf{X}^{miss}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \times g_3(\boldsymbol{\gamma}^{miss}) \times f_2(\mathbf{miss} | \mathbf{X}) \times g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \end{aligned}$$

(12b)

where,

$f_1(\boldsymbol{\gamma}^{obs} | \boldsymbol{\gamma}^{miss}, \mathbf{X}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$ represents the regression likelihood for the observed intensity values and their corresponding \mathbf{X}^{obs} ;

$f_1(\boldsymbol{\gamma}^{miss} | \mathbf{X}^{miss}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$ represents the regression likelihood for the missing intensity values and their corresponding \mathbf{X}^{miss} ;

$g_3(\boldsymbol{\gamma}^{miss})$ represents the density function of the prior for the missing intensity values;

$g_2(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$ represents the density function of the priors for the unknown parameters;

$f_2(\mathbf{miss} | \mathbf{X})$ represents the probability density function of **miss** conditional on the observations.

According to observations from our case studies and the published study (Hrydziuszko and Viant, 2012), we assumed $pm_{i,l,p}$ has a relation with m/z ratio and the intensity,

$$\text{logit}(pm_{i,l,p}) = \theta_0 + \theta_1 \times mz_{i,l,p} + \theta_2 \times \gamma_{i,l,p}.$$

(13a)

The conditional posterior distribution of $\theta_0, \dots, \theta_2$ can be postulated as follows:

$$g_4(\theta_0, \theta_1, \theta_2 | \mathbf{mz}, \boldsymbol{\gamma}, \mathbf{miss}) = \prod_{\ell=0}^2 g_{4,\ell}(\theta_\ell | \mathbf{mz}, \boldsymbol{\gamma}, \mathbf{miss}) \propto f_2(\mathbf{miss} | \mathbf{mz}, \boldsymbol{\gamma}, \theta_0, \theta_1, \theta_2) \times \prod_{\ell=0}^2 g_{4,\ell}(\theta_\ell)$$

,

(13b)

where $\mathbf{mz} = \{mz_{i,l,p}\}$ represents the vector of m/z ratios, and $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}^{obs}, \boldsymbol{\gamma}^{miss}\} = \{\gamma_{i,l,p}\}$ represents the vector of intensity values including the missing ones. $g_{4,\ell}$ is the density function for the priors of θ_ℓ . $f_2(\mathbf{miss} | \mathbf{mz}, \boldsymbol{\gamma})$ is the logistic regression likelihood function of (13a)..

As the missing peptide intensity is not observable in the data, no information can be attained from the observations, θ_2 needs informative prior. Let $\boldsymbol{\theta} = (\theta_0, \dots, \theta_2)$, the joint posterior distribution for all unknown parameters including $\boldsymbol{\theta}$ with the factorized terms is as followed:

$$\begin{aligned}
g(\mathbf{B}, \boldsymbol{\theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma} | X, \boldsymbol{\gamma}, \mathbf{miss}) &= g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma} | X, \boldsymbol{\gamma}) \times g_4(\boldsymbol{\theta} | X, \boldsymbol{\gamma}, \mathbf{miss}) \\
&\propto f_1(\boldsymbol{\gamma}^{obs} | \boldsymbol{\gamma}^{miss}, \mathbf{X}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \times f_1(\boldsymbol{\gamma}^{miss} | \mathbf{X}^{miss}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \times g_3(\boldsymbol{\gamma}^{miss}) \times g_1(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \\
&\quad \times f_2(\mathbf{miss} | \mathbf{mz}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \times \prod_{\ell=0}^2 g_{4,\ell}(\theta_\ell)
\end{aligned}
\tag{14}$$

The joint posterior distribution $g(\mathbf{B}, \boldsymbol{\theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma} | X, \boldsymbol{\gamma}, \mathbf{miss})$ is not straightforward in this model and cannot be derived directly using the numerical integration, but it can be approximated by the conditional posterior distributions using an empirical Bayesian approach i.e. Gibb sampling. In the computation, $\gamma_{i,l,p}$ will not be available for the missing, but its value is computed conditional upon the other unknown parameters from the posterior sampler.

The doodle graph for the Bayesian model is shown as follows,

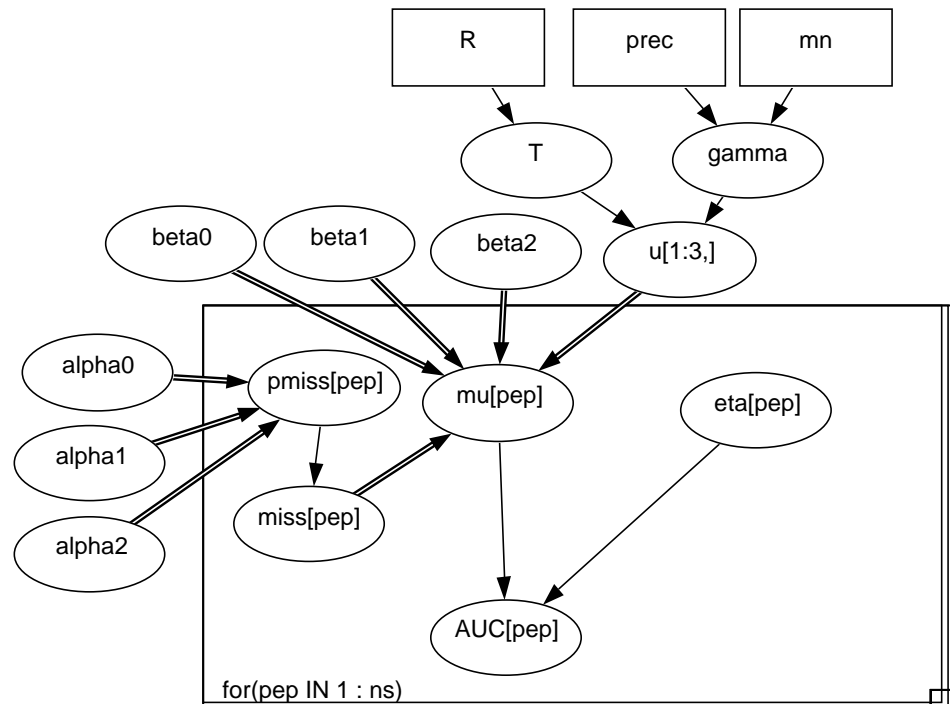


Figure legend: R is the inverse wishart hyper prior for T, prec is the inverse wishart distributed hyper-prior and mn is the multivariate normal distributed hyper prior for gamma. T and gamma are the precision and mean priors for the random parameter U of protein respectively. beta0, beta1 and beta2 are the coefficients for subject intercepts, runs and labels respectively. Pmiss is the probability of missing intensity data, and miss is the binary indicator for missing. alpha0, alpha1, and alpha2 are the regression coefficients for intercept, m/z and intensity value respectively in the logistic regression predicting pmiss. AUC is the intensity value for each peptide with mean mu and variance eta.

The derivations for the joint conditional posterior distribution of the unknown parameters are further explained in the following paragraphs. The derivations focus on the parameters at the protein level and assume the error variance σ^2 and the variance covariance matrix is either known or unknown.

In the following section, $\boldsymbol{\gamma} = \{\gamma_{i_1,l,p}^{obs}, \gamma_{i_2,l,p}^{miss}\}$ denotes the vector for completed peptide intensities including missing values that has defined in equations (9), and \mathbf{X} is the design matrix for explanatory variables. For the posterior sampling programming, the joint posterior of unknowns in (14) can also be factorized step-wisely as follows:

$$\begin{aligned}
g(\mathbf{B}, \boldsymbol{\theta}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma} | \mathbf{X}, \boldsymbol{\gamma}, \mathbf{miss}) &= g_4(\boldsymbol{\theta} | \mathbf{mz}, \boldsymbol{\gamma}, \mathbf{miss}) \times g_6(\mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma} | \mathbf{X}, \boldsymbol{\gamma}, \mathbf{miss}) \\
&= g_4(\boldsymbol{\theta} | \mathbf{mz}, \boldsymbol{\gamma}, \mathbf{miss}) \times g_7(\mathbf{B} | \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, \mathbf{X}, \boldsymbol{\gamma}, \mathbf{miss}) \times g_8(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma} | \mathbf{X}, \boldsymbol{\gamma}, \mathbf{miss}) \\
&= g_4(\boldsymbol{\theta} | \mathbf{mz}, \boldsymbol{\gamma}, \mathbf{miss}) \times g_7(\mathbf{B} | \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, \mathbf{X}, \boldsymbol{\gamma}, \mathbf{miss}) \times g_9(\boldsymbol{\alpha}, \boldsymbol{\Sigma} | \sigma^2, \mathbf{X}, \boldsymbol{\gamma}, \mathbf{miss}) \times g_{10}(\sigma^2 | \mathbf{X}, \boldsymbol{\gamma}, \mathbf{miss})
\end{aligned}
\tag{15}$$

where

$g_4(\boldsymbol{\theta} | \mathbf{mz}, \boldsymbol{\gamma}, \mathbf{miss})$, $g_7(\mathbf{B} | \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, \mathbf{X}, \boldsymbol{\gamma}, \mathbf{miss})$, and $g_9(\boldsymbol{\alpha}, \boldsymbol{\Sigma} | \sigma^2, \mathbf{X}, \boldsymbol{\gamma}, \mathbf{miss})$ are the conditional posteriors upon the likelihood and the posteriors of the other parameters.

The marginal posterior $g_{10}(\sigma^2 | \mathbf{X}, \boldsymbol{\gamma}, \mathbf{miss})$ is independent of \mathbf{miss} , $g_{10}(\sigma^2 | \mathbf{X}, \boldsymbol{\gamma}, \mathbf{miss})$ will be simplified as $g_{10}(\sigma^2 | \mathbf{X}, \boldsymbol{\gamma})$; it is proportional to the product of the prior for σ^2 and the likelihood of the observations,

$$\begin{aligned}
g_{10}(\sigma^2 | \mathbf{X}, \boldsymbol{\gamma}) &\propto g_{10}(\sigma^2) \times f_1(\mathbf{X}, \boldsymbol{\gamma} | \sigma^2) \\
&= g_{10}(\sigma^2) \times f_1(\mathbf{X}, \boldsymbol{\gamma}^{miss}, \boldsymbol{\gamma}^{obs} | \sigma^2) \\
&= g_{10}(\sigma^2) \times g_3(\boldsymbol{\gamma}^{miss}) \times f_1(\boldsymbol{\gamma}^{miss} | \mathbf{X}^{miss}, \sigma^2) \times f_1(\boldsymbol{\gamma}^{obs} | \mathbf{X}^{obs}, \sigma^2)
\end{aligned}
\tag{16}$$

where

\mathbf{B} in the multivariate setting can be expressed as a univariate regression vector $\boldsymbol{\beta}$ and can be derived through the conventional regression method via the likelihood function (Gelman et al., 1995);

And

$$f_1(\boldsymbol{\gamma}^{obs} | \boldsymbol{\beta X}^{obs}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n_{obs}} N(\gamma_{i,l,p} | \boldsymbol{\beta X}^{obs}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma});$$

$$f_1(\boldsymbol{\gamma}^{miss} | \boldsymbol{\beta X}^{miss}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n_{miss}} N(\gamma_{i+1,l,p} | \boldsymbol{\beta X}^{miss}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma});$$

Where, $\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}$ are used to construct the variance-covariance matrix for $\boldsymbol{\gamma}$.

A conjugate prior for the $\boldsymbol{\gamma}^{miss}$ is the normal distribution: $g_3(\boldsymbol{\gamma}^{miss}) = N(0, \eta)$.

The posterior of $\boldsymbol{\theta}$ is further defined as follows:

$$g_4(\boldsymbol{\theta} | \mathbf{mz}, \boldsymbol{\gamma}, \mathbf{miss}) \propto f_2(\mathbf{miss} | \mathbf{mz}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \times \prod_{\ell=0}^2 g_{4,\ell}(\theta_\ell) = \prod_{i=1}^n bin(pm_{i,l,p} | \mathbf{mz}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \times \prod_{\ell=0}^2 g_{4,\ell}(\theta_\ell).$$

If normal priors are assigned to $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, $(\theta_0, \theta_1, \theta_2)$ and multivariate normal prior is assigned to $(\mathbf{U}_{0,p}, \mathbf{U}_{1,p}, \mathbf{U}_{2,p})$, since $\boldsymbol{\gamma}$ is multivariate normal distributed, the joint conditional posteriors of all the unknowns will be multivariate normal distributed given known $\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}$.

When $\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}$ are unknown, a conjugate prior for σ^2 is the inversed gamma distribution, denoted as $g_{10}(\sigma^2) \sim InvGamma(a, b)$; a conjugate prior for $\boldsymbol{\Sigma}$ will be the inversed-Wishart distribution (Gelman et al., 1995). $(\boldsymbol{\alpha}, \boldsymbol{\Sigma}^{-1})$ use the four parameters of inversed-Wishart-normal density for the multivariate normal distribution parameters as defined by (Gelman et al., 1995),

$$\boldsymbol{\Sigma}^{-1} \sim Wish(\nu_0, d, \Lambda_0^{-1})$$

$$\boldsymbol{\alpha} | \boldsymbol{\Sigma} \sim N(\boldsymbol{\alpha}_0, \boldsymbol{\Sigma} / \kappa_0)$$

where ν_0, κ_0 are positive numbers, Λ_0 is a $d \times d$ positive definite matrix, where $\nu_0 = d - 1$.

The joint prior density for $(\boldsymbol{\alpha}, \boldsymbol{\Sigma}^{-1})$ is defined as

$$|\boldsymbol{\Sigma}|^{-\left(\frac{\nu_0+d+1}{2}\right)} \exp\left(-\frac{1}{2} \text{tr}\left(\Lambda_0 \boldsymbol{\Sigma}^{-1}\right)\right) \times |\boldsymbol{\Sigma}|^{-\left(\frac{1}{2}\right)} \exp\left(\frac{-\kappa_0}{2} (\mathbf{a}-\mathbf{a}_0)' \boldsymbol{\Sigma}^{-1} (\mathbf{a}-\mathbf{a}_0)\right). \quad (17)$$

The product of $g_9(\sigma^2)$ and $g_{10}(\mathbf{a}, \boldsymbol{\Sigma})$ will be shown as below:

$$\begin{aligned} & \frac{b^a}{\Gamma(a)} \delta^{-2(a+1)} \times \exp\left(\frac{-b}{\delta^2}\right) \times |\boldsymbol{\Sigma}|^{-\left(\frac{\nu_0+d+1}{2}\right)} \exp\left(-\frac{1}{2} \text{tr}\left(\Lambda_0 \boldsymbol{\Sigma}^{-1}\right)\right) \\ & \times |\boldsymbol{\Sigma}|^{-\left(\frac{1}{2}\right)} \exp\left(\frac{-\kappa_0}{2} (\mathbf{a}-\mathbf{a}_0)' \boldsymbol{\Sigma}^{-1} (\mathbf{a}-\mathbf{a}_0)\right) \end{aligned} \quad (18)$$

Let $\Psi_0 = (\mathbf{a}-\mathbf{a}_0)(\mathbf{a}-\mathbf{a}_0)'$, using the trace to replace the inner terms of the second exponential function of (17) gives us a gamma like the format for (18):

$$\begin{aligned} & \frac{b^a}{\Gamma(a)} \delta^{-2(a+1)} \times \exp\left(\frac{-b}{\delta^2}\right) \times |\boldsymbol{\Sigma}|^{-\left(\frac{\nu_0+d+1}{2}\right)} \times \exp\left(-\frac{1}{2} \text{tr}\left(\Lambda_0 \boldsymbol{\Sigma}^{-1}\right)\right) \times |\boldsymbol{\Sigma}|^{-\left(\frac{1}{2}\right)} \times \exp\left(-\frac{1}{2} \text{tr}\left(\kappa_0 \Psi_0\right) \boldsymbol{\Sigma}^{-1}\right) \\ & = \frac{b^a}{\Gamma(a)} \delta^{-2(a+1)} \times \exp\left(\frac{-b}{\delta^2}\right) \times |\boldsymbol{\Sigma}|^{-\left(\frac{\nu_0+d+2}{2}\right)} \times \exp\left(-\frac{1}{2} \text{tr}\left((\Lambda_0 + \kappa_0 \Psi_0) \boldsymbol{\Sigma}^{-1}\right)\right) \\ & = \frac{b^a}{\Gamma(a)} \left(\delta^{-2(a+1)} |\boldsymbol{\Sigma}|^{-\left(\frac{\nu_0+d+2}{2}\right)} \right) \times \exp\left(\frac{-b}{\delta^2} - \frac{1}{2} \text{tr}\left((\Lambda_0 + \kappa_0 \Psi_0) \boldsymbol{\Sigma}^{-1}\right)\right) \end{aligned}$$

The joint posterior distribution of unknown parameters can be computed from the conditional posterior distribution using the Gibb sampler, or using the Non U turn sampler of Hamiltonian Monte Carlo.

4.3.2 The missing mechanism for the iTRAQ data- a Bayesian approach using Hamiltonian Monte Carlo and Non U Turn posterior sampling

4.3.2.1 The mechanism of HMC and non U turn sampling

Hamiltonian Monte Carlo (HMC) originated from a physics phenomenon Hamiltonian dynamics. Hamiltonian dynamics describes the movement of a puck sliding from a random starting point to other points of a surface with various highs. This approach was introduced

by Alder and Wainwrigth (Alder and Wainwright, 1959) to simulate the molecular motions, six years after the Markov Chain Monte Carlo (MCMC) was introduced in the same area. The HMC method has an elegant formation comprising a pair of variables to describe the momentum and position of each state in the molecular movement; it has been coined as Hamiltonian dynamic since then. The statistical application of the HMC is attributed to Radford Neal who started to use it in his neural network model in 1993 (Neal, 2011). In the MCMC framework using HMC, the position variable p is set up to describe the unknown variables, the fictitious momentum q describes the kinetics of the molecules. The potential energy $U(p)$, which is a function of the position variable, is used to describe the proposed joint probability distribution that needs sampling from. In the statistical application of HMC, the potential energy $U(p)$ equals to minus log of the joint probability density. The kinetic energy $K(q)$ is equivalent to $|p|^2 / (2m)$, Where m is the mass of the puck. The pair of potential and kinetic energy is described by the Hamiltonian function $H(p, q) = U(p) + K(q)$ for a case using the HMC. $H(p, q)$ needs to be operated on its partial derivative space with relation to time t , $\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}$, $\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$. These paired partial equations define a mapping T_s from any state of s to $t+s$. When the paired differential equations of the $H(p, q)$ are in simple forms, the analytical format for the trajectory on t can be achieved. When they are complex, the computation of trajectories can be approximated on the discrete time interval due to the property of invariant volume of Hamiltonian dynamics. As suggested in Neal's first chapter of the MCMC hand book, the "leap frog" method is an optimal way to update the auxiliary momentum variable q and the position variable p with the metropolis acceptance probability. The leap frog method creates a set of discrete steps ε , 2ε , 3ε on the time interval, and the potential energy and kinetic energy take their trajectory through these steps. The updates from one step to another follow the scheme described in Neal (Neal, 2011):

$$\begin{aligned}
p_i(t + \frac{\varepsilon}{2}) &= p_i(t) - (\frac{\varepsilon}{2}) \frac{\partial U}{\partial q_i}(q(t)) \\
q_i(t + \varepsilon) &= q_i(t) + \varepsilon \frac{p_i(t + \varepsilon)}{m_i} \\
p_i(t + \varepsilon) &= p_i(t + \varepsilon) - (\frac{\varepsilon}{2}) \frac{\partial U}{\partial q_i}(q(t + \varepsilon))
\end{aligned}$$

where p_i and q_i represent the position and momentum at the i step respectively. The leap frog scheme, which gives a longer distance to the next state and has a higher acceptance probability compared to the random walk, provides a better proposal than a random walk in the simulation for a probability distribution in a high dimensional continuous space. It is reported by (Creutz, 1988; Neal, 2011) that the cost of HMC is about $D^{5/4}$ of which D is equivalent to the number of dimensions. This will give us an approximate number of iterations required for using HMC.

In the Bayesian data analysis, HMC provides a mean to sample the posterior joint conditional probability $U(p)$. Although the HMC is much more efficient than the random walk or Gibbs sampling, it comes at a price. HMC requires the derivation of the gradient of the log posterior density, and also needs manual tuning for the leap frog step and number of steps. If the step ε is too large, the simulation will not be accurate and the acceptance rate will be low; if the step ε is too small, the simulation will waste lots of computing steps. If the number of steps is set too large, the trajectory will start U turns and retrace back to original samples. If the number of steps is set to be too small, the HMC will be similar to a random walk MC (Hoffman and Gelman, 2011).

Rstan is a new software recently developed by Gelman's team to implement HMC modelling for Bayesian data analysis. In Rstan, the posterior sampling can choose the No-U-turn Sampling method (NUTS)-an extension of HMC. NUTS (Hoffman and Gelman, 2011) implements a recursive algorithm that will enable auto-tuning of the numbers of leap frog steps and ε . The main outstanding feature of the No-U-turn Sampling is that once the new updated state starts to double back and retrace, the sampling will automatically stop. NUTS uses the leap frog integrator to double the position-momentum states forwardly or backwardly at each step in the fictitious time. A binary tree with nodes of the position-momentum states is generated in this recursive manner. It stops whenever a sub tree from the

left most to the right most has a node that makes a U turn. As defined in Hoffman's paper, the simulation will stop whenever the following states occurred,

$(p^+ - p^-) \times q^- < 0$ or $(p^+ - p^-) \times q^+ < 0$, Where p^+ represents the positions generated forwardly and p^- represents the position state generated backwardly. No-U-turn sampling provides an alternative for sampling from a high dimensional correlated joint posterior distribution.

4.3.2.2 The HMC/NUTS Rstan computing method for our current proposed method

Based on the joint posterior function configured in 2.3, the potential energy function which represents minus log of the posterior function can be shown as:

$$U(p) = -\log \left(\begin{array}{l} f_1(\gamma_{i,l,p}^{obs} | \mathbf{X}, \boldsymbol{\theta}, \mathbf{B}, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \times g_4(\boldsymbol{\theta} | \mathbf{mz}, \boldsymbol{\gamma}, \mathbf{miss}) \\ \times g_7(\mathbf{B} | \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, X, \boldsymbol{\gamma}, \mathbf{miss}) \times g_9(\boldsymbol{\alpha}, \boldsymbol{\Sigma} | \sigma^2, X, \boldsymbol{\gamma}, \mathbf{miss}) \times \\ g_{10}(\sigma^2) \times g_3(\boldsymbol{\gamma}^{miss}) \times f_1(\boldsymbol{\gamma}^{miss} | X^{miss}, \sigma^2) \end{array} \right).$$

In Rstan, the missing values cannot be mixed with the non-missing values. In the Rstan program for the proposed model (15), the relationship of missing values of $\gamma_{i,l,p}$ and the explanatory variables are separately coded for censored and completely missing.

The Hamiltonian function $H(\mathbf{p}, \mathbf{q})$ used in HMC for the proposed model (15) according to suggestions from (Neal, 2011) is further described as follows:

$H(\mathbf{p}, \mathbf{q}) = U(\mathbf{p}) - K(\mathbf{q}) = U(\mathbf{p}) - \frac{1}{2} \mathbf{q} \times \mathbf{q}$, where $U(\mathbf{p})$ is the minus of log of the product of priors and likelihood given the data; and \mathbf{p} represents the vector of all the unknown parameters in the model $\mathbf{p} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \theta_0, \theta_1, \theta_2, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, \sigma^2, \boldsymbol{\gamma}^{miss})$. The fictitious function is given as $K(\mathbf{q}) = \frac{1}{2} \mathbf{q} \times \mathbf{q}$, where \mathbf{q} represents the momentum of the fictitious particle in the Hamiltonian dynamics.

A Rstan program for the model described in equation (15) is summarized in the following table 4.1 with a separate section for setting up censoring and completely missing values. In this program, priors and hyper priors are all re-parameterized in the transformed parameter section.

Table 4.1 The algorithm of the Rstan HMC/NUTS computing method

<p>data</p> <p>step1: Assign records of all the known peptide intensities γ^{obs} and their corresponding explanatory variables \mathbf{X}^{obs} to parameters in the data block, excluding records with censored and missing intensities γ^{miss} and their corresponding observations of explanatory variables \mathbf{X}^{miss} ;</p> <p>step2: Assign records of all the known explanatory variable \mathbf{X}^{miss} that link to the censored and missing peptide quantities γ^{miss} to parameters in the data block;</p> <p>step3: Assign priors and hyper priors for the covariance matrix Λ_0 and other known values (i.e. number of subjects) to parameters in the data block.</p> <p>transformed data</p> <p>Any data needing to be transformed is assigned in the transformed data block. For instance, to derive a prior for the covariance matrix if the prior is given as precision matrix will be defined in this block.</p> <p>parameters</p> <p>Any unknown variables including the missing and censored intensities and variables that will be used to re-parameterize the unknown variables are defined in the parameter block. Variables used for re-parameterizations are called latent variables. For example, in the NUTS program given in appendix 4.2, the latent parameters defined in this section are named by adding extension <code>_latent</code> to the unknown parameters. For example, <code>beta4_latent</code> is the latent variable for <code>beta4</code>.</p> <p>transformed parameters</p>

Any unknown variables with re-parameterization or comprising known values and unknown variables can be defined in the transformed parameters block. In the given NUTS program (appendix 4.2), the unknown parameters $\beta_0, \beta_1, \beta_2, \theta$, the missing and censored peptide quantities γ^{miss} and the probability of missing are defined in this block.

For example, the probability of missing are defined in the following statements:

```
pmiss[pep]<-inv_logit(alpha+alpha1*m_z_centered[pep]+alpha2*logofAUC[pep]);  
  
pmiss_m[pep]<-  
inv_logit(alpha+alpha1*m_z_centered_m[pep]+alpha2*logofAUC_m[pep]);
```

Where alpha, alpha1 and alpha2 denote the regression coefficients for intercept, m/z and abundance respectively, m_z_centered and m_z_centered_m denote the m/z ratio in the X^{obs} and X^{miss} respectively, and logofAUC and logofAUC_m denote the response γ^{obs} and γ^{miss} respectively.

Model

Step 1: Define priors and hyper priors: distributions of all priors or hyper priors for unknowns are defined firstly in the model block.

Step 2: Re-parameterization: The re-parameterization is equivalent to a two-steps sampling which firstly samples the paired latent location and scale parameters according to their hypothesized distributions in the model block and secondly derives the unknown parameters using the latent parameters in the transformed parameters block.

For example, the following statements are used to sample the latent parameters for re-parameterization of the regression coefficients for subject intercept β_0 ,

```
for (sub in 1:64)  
{beta2_latent[sub]~normal(0,1);  
beta2_mu[sub]~normal(0,1);}
```

, where beta2_latent is the latent location variable and beta_mu is the latent scale variable

of β_0 . This pair of variables is used to simulate variable β_0 (named as beta2 in the program) which is normal distributed with mean beta_latent and standard deviation beta2_mu defined in the transformed parameters block.

Another re-parameterization example in this model would be,

```
g~multi_normal(mn,T);  
  
pVAR~inv_wishart(3,invprec);  
  
for (prot in 1:nprotein)  
  
  U_latent[prot]~multi_normal(mn,R);    //standard multinormal distributed
```

where g denotes the latent location variable, $pVAR$ denotes the latent scale parameters for the protein level parameters U . U_{latent} is a unit multivariate normal distributed variable. In the transformed parameters section, the following statements are used to generate the protein level parameter U that has location variable g and scale parameters $pVAR$.

```
for (prot in 1:nprotein)  
  
  U[prot]<-g+pVAR*U_latent[prot];
```

Step 3: Missing and censored data parameters

The missing values are treated as unknown variable. Logistic regression is used to model the missing probability and is defined in the transformed parameter section. The distributions of the latent variables for the missing values logofAUC_m_latent[] and the distribution of the probability of missing are codes as follows:

```
for (pep in 1:nobs)  
  
  miss[pep]~bernoulli(pmiss[pep]);  
  
for (pep in 1:nmiss)  
  
  {miss_m[pep]~bernoulli(pmiss_m[pep]);  
  
  logofAUC_m_latent[pep]~normal(0,1);}
```

where `miss[pep]` and `miss_m[pep]` denotes the dummy variables indicating if the peptide intensity is missing for the observed peptides and completely missing peptides intensities respectively, `pmiss[pep]` and `pmiss_m[pep]` are the probabilities of missing that can be predicted by the `m/z` and peptide abundance for the observed and completely missing intensities.

The probability of censored values is estimated by integrating out its marginal probability under the censored limit:

```
for (pep in 1:ncensor)
  increment_log_prob(log1m(1-normal_cdf(censor_lim,mu_cen[pep],eta)));
```

where `increment_log_prob` is the system variable for user-defined joint probability likelihood, `censor_lim` defines the censored limit, `mu_cen` denotes the imputed censored data from the last iteration and `eta` denotes the standard deviation.

There are differences in the BUGS program compared to the NUTS program. BUGS allows missing values being included in the observations for the analysis, Rstan requires missing values being defined as the unknown parameters and their relationships with the other variables being separately coded. In the BUGS program, the missing probability for censoring is assigned to be 1. In the Hamiltonian MC/NUTS program, the observations with missing intensities are separately coded. The probability of missing for observations with censored intensities is estimated as the cumulative probability of a value below a known detectable limit from the normal distribution, which is derived from using the numerical integration. More details of the BUGS and HMC/NUTS computing approaches are demonstrated in the simulation study and the case studies in chapter 5 and chapter 6.

4.4. A simulation study

4.4.1 *The simulated experiment*

An iTRAQ proteomic experiment was simulated with a balanced row and column design for 8 runs (row), 8 labels (column) and two classes that comprised 32 subjects from the normal control population and 32 subjects from the diseased population. The first set of data assumes that peptides were clustered within proteins and proteins are all observed by the subjects. A Poisson distribution ($\lambda=5$) was used to generate the number of peptides of 200 proteins. The intensity of the peptide was simulated by a function of run, label, total amount of protein, subject class, protein intercept, and mass-to-charge ratio (m/z). The regression coefficients of the run, label, total amount of protein and subject are normal distributed; and the regression coefficient of the protein intercept, mass-to-charge ratio (m/z) are normal distributed with the same mean and variance for the 200 proteins, abundance differences between subject class are normal distributed with different means and variances across different proteins.

A second set of data is simulated to capture the non-random missingness pattern based on the first simulated completed dataset. The two case studies and the literature (Wells et al., 2011) suggest that the peptide intensity data is left censored at a threshold where the signal is too weak to be detected, and the value of the censored peptide intensity is reported as zero in the raw data from the LC-MS/MS. The simulated probability of missingness is a function of the peptide intensity and the mass-to-charge ratio (m/z). The censoring threshold is set to 0.1 which is the minima logarithmic intensity of the current immunology study, and any simulated intensity values below this value is set to zero. The simulated data of 78400 records has 928 (1.2%) observations with censored peptide intensities and 14287 (18.2%) observations with completely missing peptide intensities.

4.4.2 *The analytical methods*

Three sets of analytical methods are used for the evaluations. The first set is the single protein model in which the protein is analyzed one-by-one as defined in (3). The second sets of models are three multivariate multilevel models in which proteins are analyzed simultaneously and each of which is the special case as defined in (8). The third set is the

Bayesian version multivariate multilevel model as defined in (9), of which the missingness values are also estimated as the unknown parameters of the joint posterior distribution through Markov Chain Monte Carlo (MCMC). Gibbs posterior sampling and Hamiltonian Monte Carlo using No U turn sampling of the posterior distributions are compared for the Bayesian model.

Method 1 (single protein model): Each protein was analysed separately using a mixed model with a random intercept for subject. The R *lmer* function of R package lme4 (Pinheiro et al., 2014) used for this model is:

```
fit<-lmer(proteinsub$logofAUC~factor(proteinsub$run)+factor(proteinsub$tlabel)
+factor(proteinsub$class)+proteinsub$mzcentered+(1|proteinsub$subject),data=proteinsub),
```

where logofAUC is the response variable of the intensity, run, tlabel, class and mzcentered represent the run, label, physiological condition and the centralized m/z ratio respectively.

Method 2 (multiple proteins model): Using a multivariate multilevel model for all proteins that include random intercepts, random slopes of mass-to-charge ratio (m/z) and random coefficients for abundance differences between subject classes across different proteins, and random intercepts for different subjects.

The three R *lmer* functions used for this model are:

```
fit1<-lmer(mockdata$logofAUC~factor(run)+factor(tlabel)+(1+mzcentered+factor(class)|protein)+
(1|subject),data=mockdata)
```

```
fit2<-lmer(mockdata$logofAUC~factor(run)+factor(tlabel)+mzcentered+(1+mzcentered+factor(class)|protein)
+ (1|subject),data=mockdata)
```

```
fit3<-
lmer(mockdata$logofAUC~factor(run)+factor(tlabel)+mzcentered+factor(class)+(1+mzcentered+factor(class)|protein)
+ (1|subject), data=mockdata)
```

,

where

Fit 1 defines the model with three protein random effects (intercept, m/z, and physiological conditions) and a random intercept for subject, with two fixed effects (run and label);

Fit 2 defines the model with three protein random effects (intercept, m/z, and physiological conditions) and a random intercept for subject, with three fixed effects (run, label and m/z);

Fit 3 defines the model with three protein random effects (intercept, m/z, and physiological conditions) and a random intercept for subject, with four fixed effects (run, label, m/z and class).

Method three is the Bayesian version of multivariate multilevel model (multiple proteins model) with parameters for the missing components as defined in (9). Uninformative normal distributed priors were chosen for the regression coefficients of run, label, and subject intercepts noted as $\beta_{3_}$, $\beta_{4_}$ and β_2 respectively. An uninformative gamma distributed prior was chosen for the variance of the peptide intensity noted as $\text{auc}[\text{pep}]$. The protein level parameters $(\mathbf{U}_{0,p}, \mathbf{U}_{1,p}, \mathbf{U}_{2,p})$ as defined in equation (9) is multivariate normal distributed. A hyper inverse-Wishart distributed prior was chosen for the covariance matrix of the protein level parameters. Informative normal priors were chosen for the missing parameters *alpha0-alpha2*.

The BUGS program for using software win-BUGS (Spiegelhalter et al., 2003) is set up as below:

```

model

{
  for (pep in 1:78400)

    { auc[pep]~dnorm(mu[pep],eta[pep])

mu[pep]<-
beta2[subject[pep]]+U[proteinid[pep],1]+U[proteinid[pep],2]*m_z_centered[pep]+U[proteinid[pep],3]*class[pep]+
beta3_1*run1[pep]+beta3_2*run2[pep]+beta3_3*run3[pep]+beta3_4*run4[pep]+beta3_5*run5[pep]+beta3_6*run6[pep]
+beta3_7*run7[pep]+beta3_8*run8[pep]+beta4_113*flab113[pep]+beta4_114*flab114[pep]+beta4_115*flab115[pep]+be
ta4_116*flab116[pep]+beta4_117*flab117[pep]+beta4_118*flab118[pep]+beta4_119*flab119[pep]+beta4_121*flab121[
pep]

    eta[pep]~dgamma(0.1,0.1)
    miss[pep]~dbin(pmiss[pep],1)
    pmiss.lim[pep]<-alpha0+alpha1*m_z_centered[pep]+alpha2*auc[pep]
    pmiss[pep]<-(1-censor[pep])*(max(0.001,min(0.99,pmiss.lim[pep]))) +censor[pep]*0.99
    }
#prior for random coefficients
for (protein in 1:200)
{U[protein,1:3]~dmnorm(gamma[1:3],T[1:3,1:3])}
for (sub in 1:64)
{beta2[sub]~dnorm(0,1)}
#prior for fixed coefficient
#use informative prior
alpha0~dnorm(1,0.01)
alpha1~dnorm(0.0085,2.5E7)
alpha2~dnorm(-0.45,4)
beta3_1~dnorm(0,0.1)
beta3_2~dnorm(0,0.1)
beta3_3~dnorm(0,0.1)
beta3_4~dnorm(0,0.1)
beta3_5~dnorm(0,0.1)
beta3_6~dnorm(0,0.1)
beta3_7~dnorm(0,0.1)
beta3_8~dnorm(0,0.1)
beta4_113~dnorm(0,0.1)
beta4_114~dnorm(0,0.1)
beta4_115~dnorm(0,0.1)
beta4_116~dnorm(0,0.1)
beta4_117~dnorm(0,0.1)
beta4_118~dnorm(0,0.1)
beta4_119~dnorm(0,0.1)
beta4_121~dnorm(0,0.1)
#hyper prior
gamma[1:3]~dmnorm(mn[1:3],prec[1:3,1:3])
T[1:3,1:3]~dwish(R[1:3,1:3],3)
}

```

where,

The $\mu[\text{pep}] \leftarrow$ statement assigns the relation between the explanatory variables and the response-the mean peptide intensity values;

The $p_{\text{miss.lim}}[\text{pep}] \leftarrow \text{alph0} + \text{alph1} * m_z_centered[\text{pep}] + \text{alph2} * \text{auc}[\text{pep}]$ statement assigns the relation between the probability of missing and the explanatory variables including m/z ratio($m_z_centered[\text{pep}]$) and the abundance ($\text{auc}[\text{pep}]$);

The $p_{\text{miss}}[\text{pep}] \leftarrow (1 - \text{censor}[\text{pep}]) \times (\max(0.001, \min(0.99, p_{\text{miss.lim}}[\text{pep}])) + \text{censor}[\text{pep}] \times 0.99$ statement assigns the probability of missing incorporating the probability for censored values, where $\text{censor}[\text{pep}]$ is a dummy variable with value 0 or 1; if the value is censored, the missing probability is assigned to be 0.99;

The \sim assign the distributions for priors and hyper priors, and also the distribution for the intensity values.

The Rstan Hamiltonian/NUTS program for using software package Rstan (Stan Development Team, 2013) can be downloaded from:

<https://github.com/ireneslzeng/proteomics>.

In the Rstan program, the data is separated into three parts according to the completeness of the peptide intensity: observed, completely missing and censored. The regression likelihoods are also defined in three different statements.

The $\mu[\text{pep}] \leftarrow, \mu_m[\text{pep}] \leftarrow, \mu_cen[\text{pep}] \leftarrow$ statements in the transformed parameter block define the relations between the observed, completely missing and censored peptide intensity as the response and the explanatory variables respectively. The samplings for priors, known and unknown parameters (including the latent variables) are defined in the model block.

4.4.3 The results

The estimated differences in proteins' abundances between normal and diseased subjects are used to compare across different models. These differences are referred to as the class differences across the whole section below. Since the parameters used to simulate the two sets of data are known, they are included as the "gold standard" for the evaluation of the different methods described in 4.2 and are referred to as the actual values in the following section.

4.4.3.1 Simulated study with no missing values

Mean class differences:

As demonstrated in the scatter plot (figure 4.2.), estimates of the mean class differences from the multiple proteins model have better linear agreements with the actual values compared to estimates from the single protein model. In the linear regression model that includes actual values of the class differences as dependents and estimates from the *single protein* model as independent, the R^2 is 0.99 and the slope is 0.94 (standard error: 0.01) (figure 4.2 (a)). In a similar linear regression model that includes actual values of the class differences as dependents and estimates from the *multiple protein* model as independent, the R^2 is 0.99 and the slope is 0.99 (standard error: 0.01) (figure 4.2 (b)).

4.4.3.2 Simulated study with missing values

Among the three R *lmer* models, fit2 has the smallest REML convergence criteria value (REML deviance) 253015.7, fit1 and fit3 have the REML convergence criteria value of 253500.7 and 253029.7 respectively. The results of fit2 are demonstrated in table 4.2, the variances of protein random intercept, m/z ratio and class are 2.32, 1.89e-06, and 0.624 respectively; the variance of subject intercept is 0.73 and variance of residual error is 3.11. The significant fixed effects include coefficients for run 6, run 7, label 115, label 117, label 119 and m/z.

Estimates from the multiple proteins model using the simulated data with missing values also demonstrated a better agreement with the actual values than estimates from the single protein model, as shown in the scatter plots figure 4.3a and 4.3b. In two sets of similar linear regressions of which the actual values are included as dependants and estimates of single or multiple protein models as independents, the R^2 is 0.87 with a slope of 1.21 (standard error: 0.034) for the single protein model and R^2 is 0.91 with a slope of 1.27 (standard error: 0.028) for the multiple protein model. The R^2 is bigger and standard error of the slope is slightly smaller in the multiple proteins model when compared with the single protein model.

When the non-random missingness components are modelled using the Bayesian method, the estimates of the mean class difference are closer to the actual values, and this is demonstrated in the scatter plot (figure 4.2b). The R^2 is 0.95 and the slope is 0.996 (standard error: 0.016) when using the actual values to regress on estimates from the BUGS model. The R^2 is bigger and standard error of the slope is smaller than those of the multiple and single protein *lmre* models. The HMC/NUTS results achieve a similar better agreement with the actual values for the class difference. In the linear regression including the actual values as dependent and the HMC/NUTS estimates as independent, the R^2 is 0.95 and the slope is 1.07 (standard error: 0.018).

Apart from the estimates for class differences, protein intercept, slope for m/z ratio of each protein, as well as subject intercepts from the 60000 BUGS and 3000 NUTS posterior samples are also compared to the actual values used for the simulation, they are shown to be similar.

The missing components parameters from the 3000 posterior HMC/NUTS samples are -7.55 (95% credible interval: -8.49,-6.31), 0.0085 (95% credible interval: 0.0085-0.0085) and 0.86 (95% credible interval: 0.72-0.97) for alpha0, alpha1 and alpha 2 respectively. Comparing to the actual values of -5.2, 0.0084 and 0.60 for alpha0, alpha1 and alpha 2 respectively, the intercept and regression coefficient for abundance are bigger in their magnitudes.

The variances of protein intercept, protein m/z, and class difference are 3.03 (95% credible interval: 1.82, 5.08), 0.49 (95% credible interval: 0.28, 1.02) and 0.81 (95% credible interval: -0.40, 1.38) respectively. Comparing to the *lmer* fit2 variances for random

intercept, m/z and class difference which are 2.32, 1.89e-06, and 0.624 respectively, the median variance of protein intercept and class difference from NUTS are slightly higher. Nevertheless, their credible intervals contain those values from fit2. The residual variance of NUTS is 1.87 (95% credible interval: 1.81, 1.93) which is similar to the actual value of 2.0 and smaller than the fit2 result of 3.11.

The proposed multilevel multiple proteins model incorporating the non-missing method is workable and an improvement when compared to the protein ratio and single protein model method.

Table 4.2 The multivariate multilevel model using R (using data with missing values)

Multiple protein model FIT TWO: with protein level slope of mass/charge ratio as fixed effect				
Linear mixed model fit by REML ['lmerMod']				
Formula: mockdata\$logofAUC ~ factor(run) + factor(tlable) + mzcentered + (1 + mzcentered + factor(class) protein) + (1 subject)				
Data: mockdata				
REML criterion at convergence: 253015.7				
Random effects:				
Groups	Name	Variance	Std.Dev.	Corr
protein	(Intercept)	2.324e+00	1.524550	
	mzcentered	1.892e-06	0.001376	-0.99
	factor(class)1	6.238e-01	0.789779	-0.02 -0.13
subject	(Intercept)	7.265e-01	0.852372	
Residual		3.111e+00	1.763940	
Number of obs: 63185, groups: protein, 200; subject, 64				
Fixed effects:				
		Estimate	Std. Error	t value
(Intercept)		6.0760910	0.4281569	14.19
factor(run)2		-0.3662296	0.4271577	-0.86
factor(run)3		-0.1860670	0.4271329	-0.44
factor(run)4		-0.1455336	0.4271275	-0.34
factor(run)5		-0.1483655	0.4271364	-0.35
factor(run)6		-1.3837084	0.4270851	-3.24
factor(run)7		0.4142777	0.4271996	0.97
factor(run)8		-0.9391780	0.4271184	-2.20
factor(tlable)114		-0.4886118	0.4271843	-1.14
factor(tlable)115		-2.2501244	0.4271403	-5.27
factor(tlable)116		-0.7965655	0.4271765	-1.86
factor(tlable)117		-1.1618414	0.4271655	-2.72
factor(tlable)118		-0.8214152	0.4271788	-1.92
factor(tlable)119		-1.4473531	0.4271432	-3.39
factor(tlable)121		-0.2230853	0.4272183	-0.52
mzcentered		-0.0051850	0.0001114	-46.53

Figure 4.1 : The structure of the iTRAQ label.

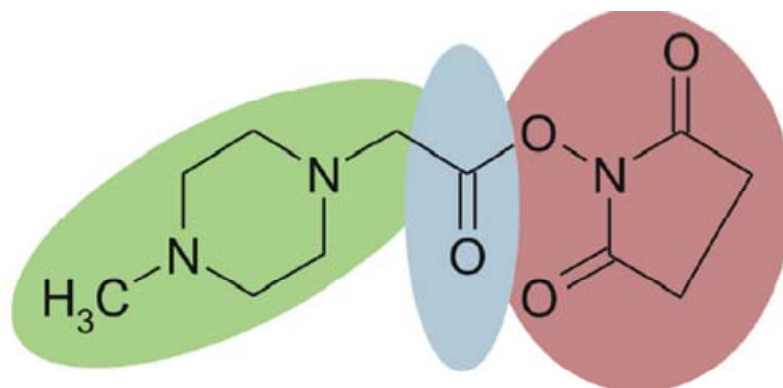
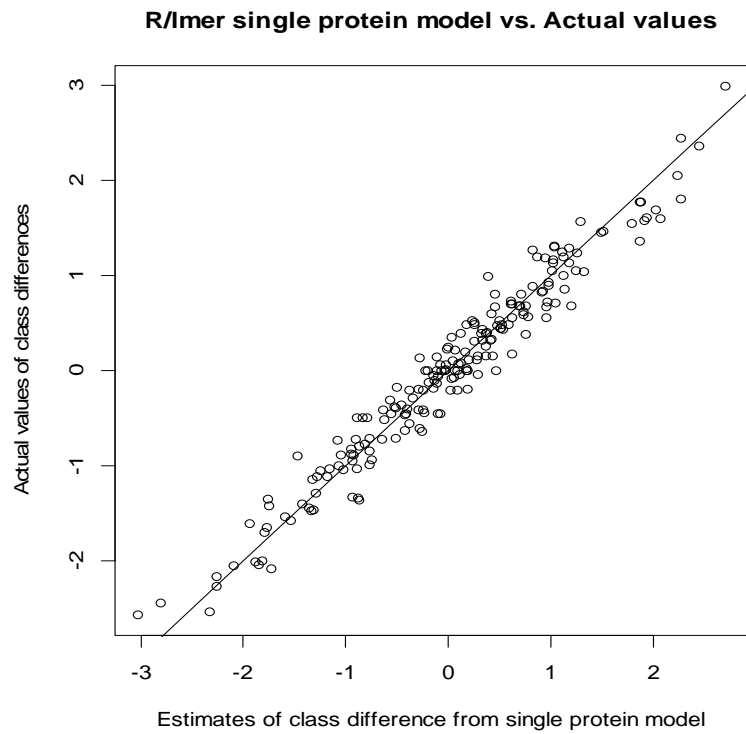


Figure 1

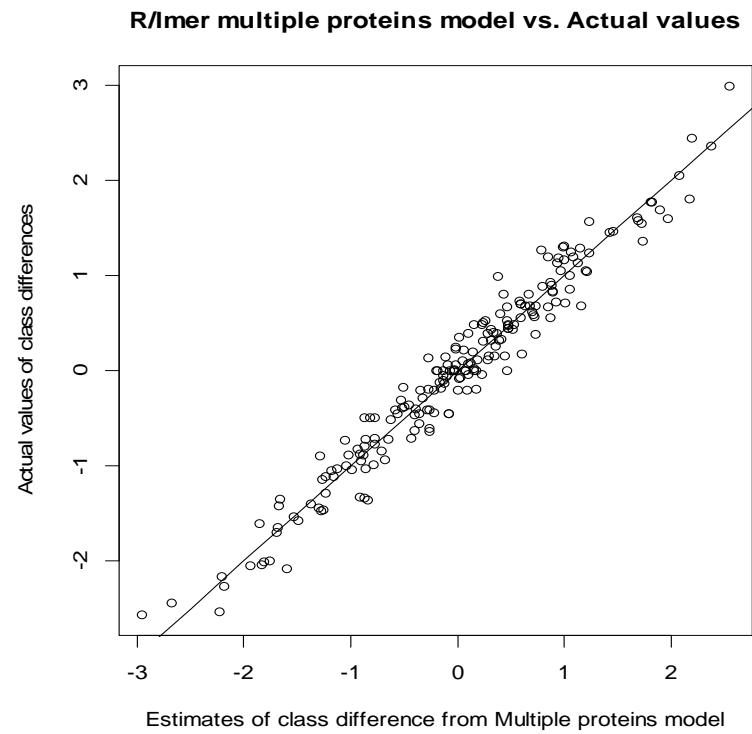
Chemical structure of the iTRAQ™ reagent. The label is composed of a peptide reactive group (red, NHS ester) and an isobaric tag of 145 Da, which consists of a balancer group (blue, carbonyl group) and a reporter group (green, N-methylpiperazine). The four available tags of identical overall mass vary in their stable isotope compositions such that the reporter group has a mass of 114–117 Da and the balancer of 28–31 Da. The fragmentation site between the balancer and the reporter group is responsible for the generation of the reporter ions in the region of 114–117 m/z.

Figure 4.2 Data without missing: the class differences from multiple protein models compared to single model and actual values

Figure Legend: The comparisons of the estimates for the class differences (a) between single protein model and the actual values, (b) between multiple protein model and the actual values, and (c) between the single protein and the multiple protein model.

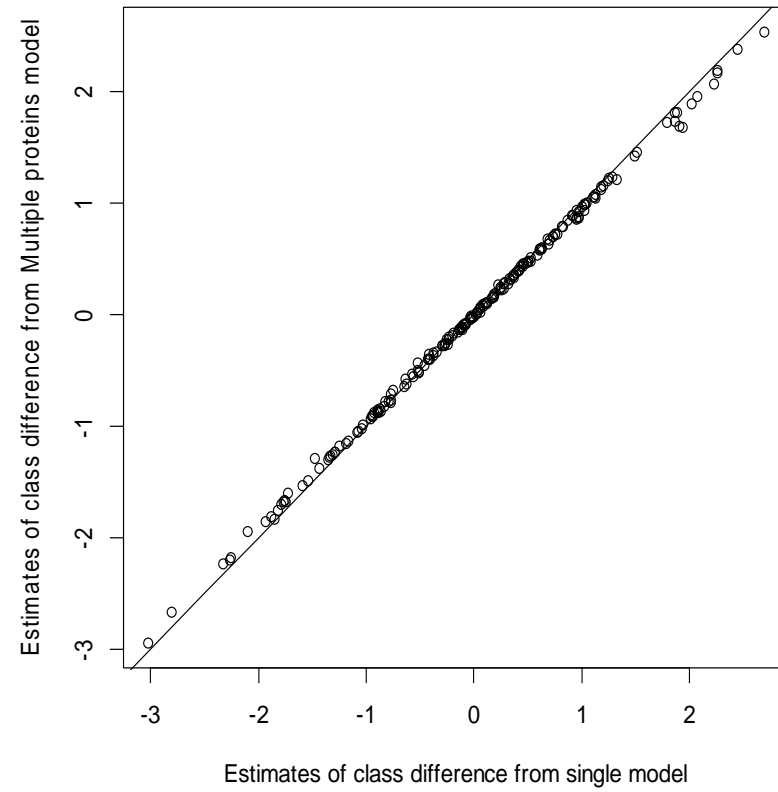


(a)



(b)

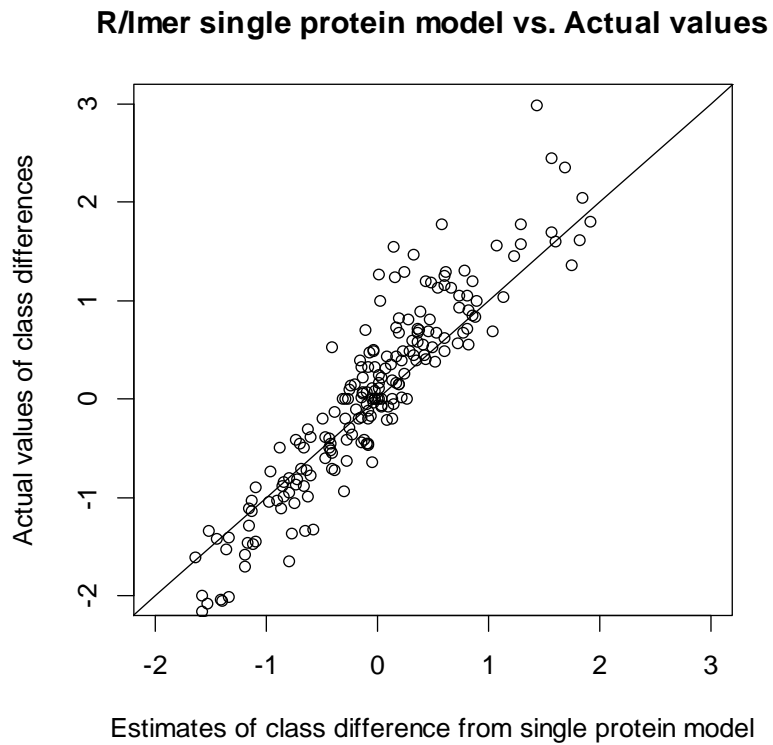
R/mer Single protein model vs. Multiple proteins model



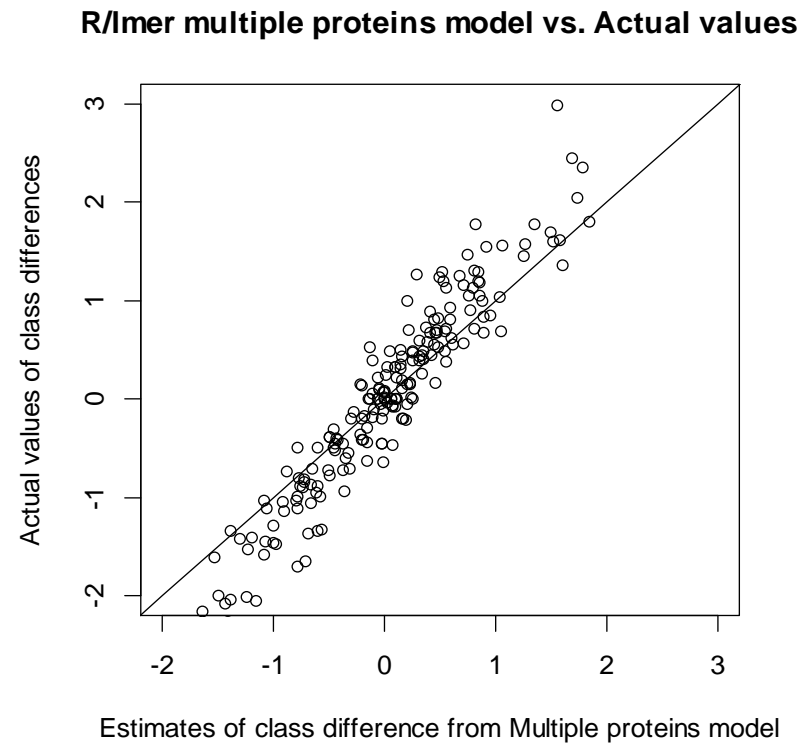
(c)

Figure 4.3a Data with missing values: the estimates of class differences from multiple protein models compared to single model and actual values.

Figure Legend: The comparisons of the estimates for the class differences from R/Imer (a) between single protein model and the actual values, (b) between multiple protein model and the actual values, and (c) between the single protein model and the multiple protein model.

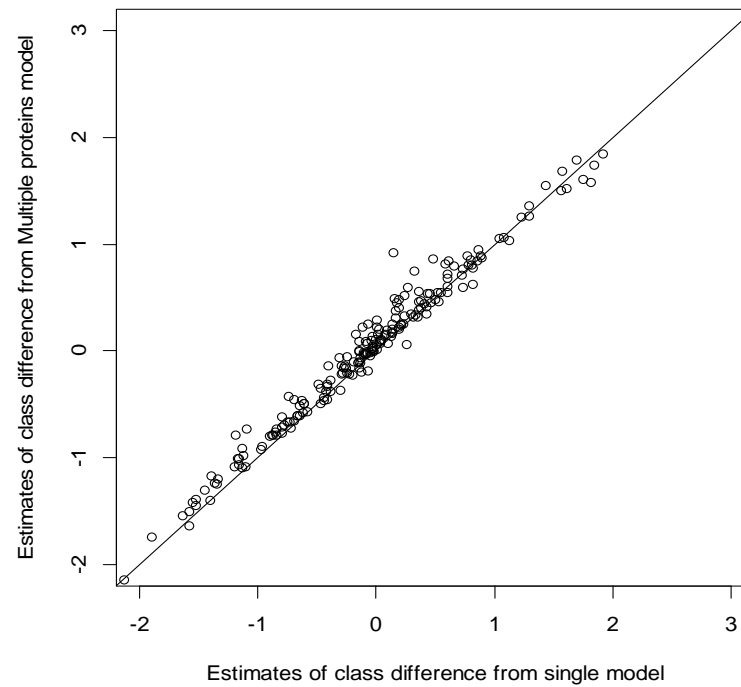


(a)



(b)

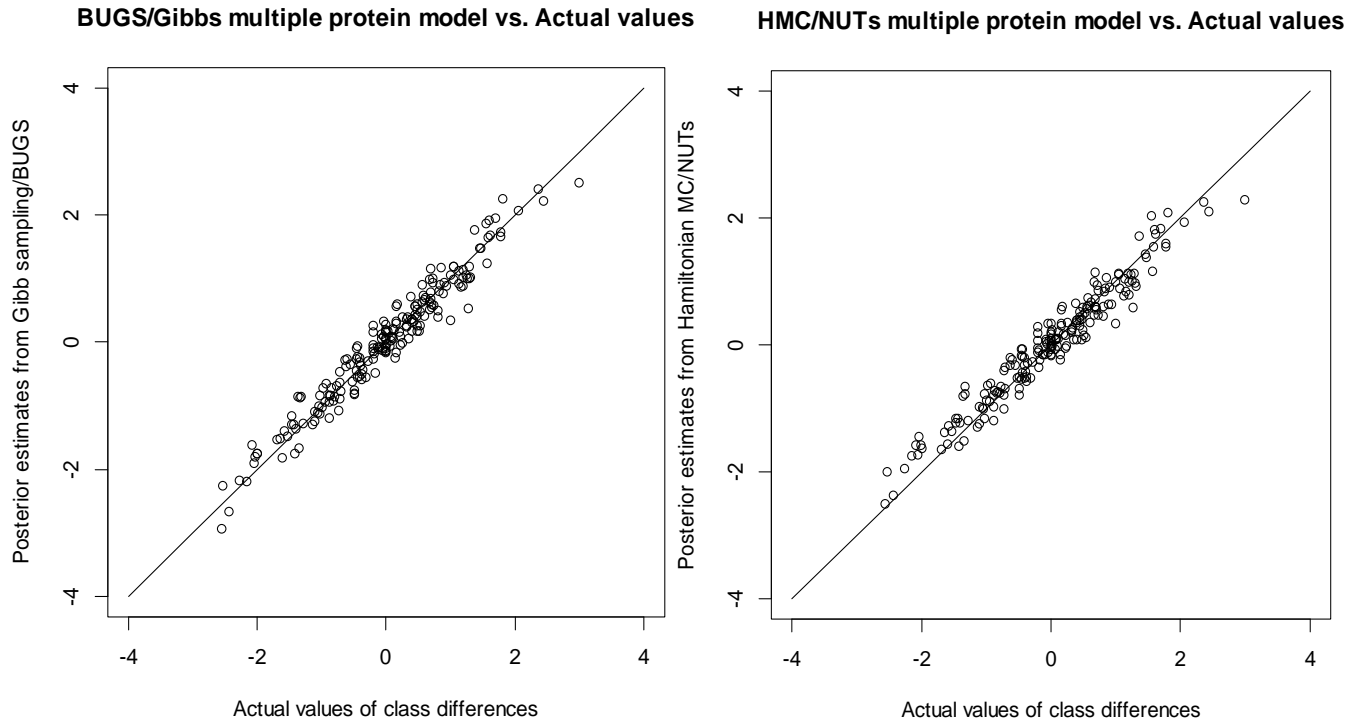
R/Imer Single protein model vs. Multiple proteins model



(c)

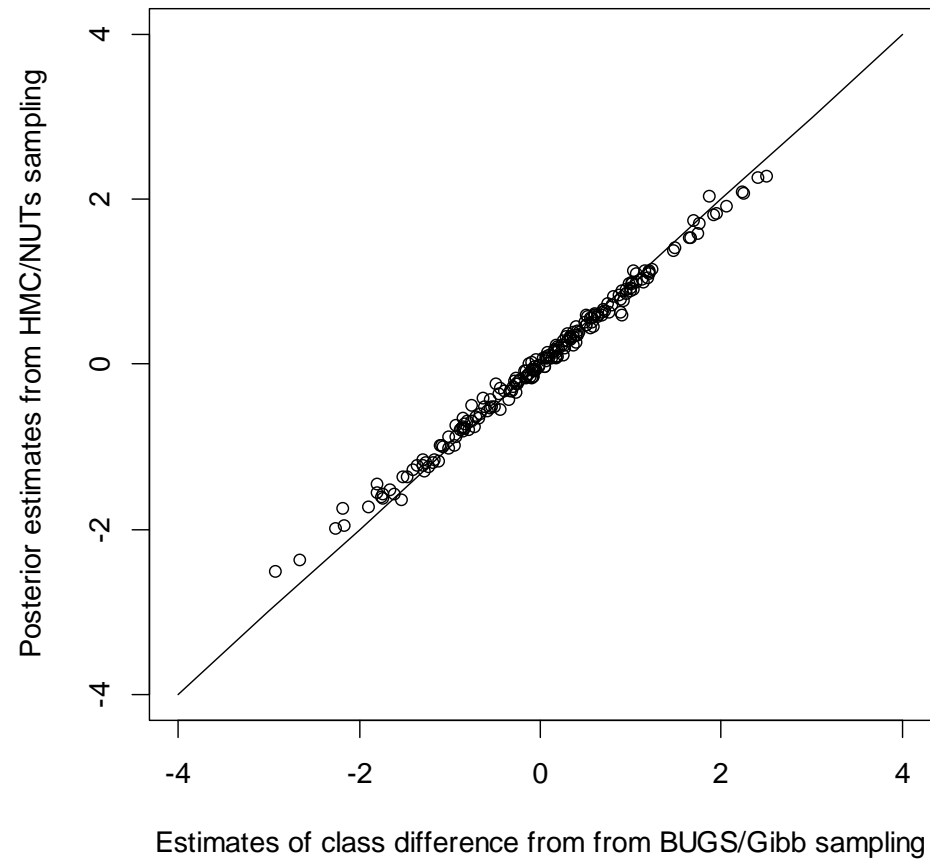
Figure 4.3b Data with missing: the class differences from BUGS and HMC/NUTS multiple protein models compared to the actual values

Figure Legend: The comparisons of the estimates for the class differences (a) between BUGS/Gibbs model and the actual values, (b) between HMC/NUTs model and the actual values.



**Figure 4.3c: data with missing: the estimates of class difference from the multiple protein BUGS
vs. the multiple protein NUTS algorithm**

HMC/NUTs multiple protein model vs. BUGS/Gibbs model



Appendix 4.1 The results from winbugs for the simulated study

Results of class differences from BUGS/Gibb sampling model

node	mean	sd	MC error	2.50%	median	97.50%	start	sample	Actual values
U[1,3]	-0.3563	0.3069	0.00856	-0.9588	-0.3556	0.243	10000	60002	-0.3999
U[2,3]	-0.08514	0.3001	0.00859	-0.6705	-0.08414	0.5023	10000	60002	-0.1001
U[3,3]	-1.187	0.4287	0.00867	-2.031	-1.186	-0.3532	10000	60002	-0.8907
U[4,3]	-0.1159	0.3547	0.00868	-0.8083	-0.117	0.5773	10000	60002	-0.1238
U[5,3]	-0.1742	0.3115	0.00861	-0.7852	-0.1745	0.4338	10000	60002	-0.1884
U[6,3]	-0.5338	0.379	0.00879	-1.279	-0.5307	0.2041	10000	60002	-0.4139
U[7,3]	-0.5056	0.2997	0.00859	-1.096	-0.5053	0.08082	10000	60002	-0.3858
U[8,3]	-0.5911	0.3104	0.00863	-1.202	-0.5896	0.01122	10000	60002	-0.3807
U[9,3]	-0.2628	0.3522	0.00863	-0.9516	-0.263	0.4247	10000	60002	-0.1976
U[10,3]	-0.1253	0.2945	0.00856	-0.7046	-0.1225	0.4499	10000	60002	-0.0492
U[11,3]	0.1764	0.3358	0.00868	-0.4858	0.1787	0.8287	10000	60002	0.5038
U[12,3]	0.3658	0.305	0.00859	-0.2302	0.3672	0.9583	10000	60002	0.4025
U[13,3]	0.3403	0.38	0.00860	-0.3981	0.3396	1.083	10000	60002	0.9923
U[14,3]	0.2646	0.3189	0.00868	-0.3597	0.2652	0.8846	10000	60002	0.3205
U[15,3]	0.602	0.3861	0.00877	-0.1584	0.6002	1.361	10000	60002	0.1709
U[16,3]	0.5071	0.3228	0.00863	-0.1285	0.5077	1.137	10000	60002	0.6712
U[17,3]	0.3695	0.3007	0.00860	-0.2224	0.3694	0.9576	10000	60002	0.3899
U[18,3]	0.09318	0.3147	0.00866	-0.5217	0.09103	0.7084	10000	60002	-0.0388
U[19,3]	0.191	0.3533	0.00862	-0.5079	0.1919	0.8854	10000	60002	0.1166
U[20,3]	-0.01149	0.3185	0.00857	-0.6363	-0.01076	0.6131	10000	60002	0.2436
U[21,3]	-1.663	0.3477	0.00857	-2.348	-1.661	-0.9874	10000	60002	-1.3456
U[22,3]	0.2539	0.3241	0.00873	-0.3823	0.2557	0.8922	10000	60002	-0.2069
U[23,3]	2.229	0.3045	0.00861	1.633	2.23	2.823	10000	60002	2.4402
U[24,3]	-1.475	0.3211	0.00858	-2.105	-1.477	-0.8473	10000	60002	-1.5304
U[25,3]	-0.8535	0.3197	0.00860	-1.483	-0.8543	-0.2264	10000	60002	-1.3245
U[26,3]	0.898	0.3183	0.00859	0.2737	0.8978	1.52	10000	60002	0.5599
U[27,3]	-0.06493	0.3205	0.00857	-0.6842	-0.06685	0.5626	10000	60002	0.0575
U[28,3]	-0.7983	0.3038	0.00863	-1.394	-0.7974	-0.2017	10000	60002	-0.4889
U[29,3]	-1.537	0.3588	0.00894	-2.243	-1.536	-0.8311	10000	60002	-1.6972
U[30,3]	1.196	0.3726	0.00860	0.47	1.198	1.922	10000	60002	1.0424
U[31,3]	-0.8398	0.3129	0.00859	-1.458	-0.8395	-0.2287	10000	60002	-0.8875
U[32,3]	1.474	0.3377	0.00864	0.8109	1.475	2.135	10000	60002	1.4641
U[33,3]	0.9895	0.3201	0.00859	0.3588	0.9889	1.617	10000	60002	1.0529
U[34,3]	-2.923	0.2995	0.00858	-3.517	-2.924	-2.335	10000	60002	-2.5599
U[35,3]	1.199	0.319	0.00853	0.5773	1.198	1.818	10000	60002	1.2860
U[36,3]	-0.09446	0.3486	0.00864	-0.7784	-0.09311	0.586	10000	60002	-0.0621
U[37,3]	2.246	0.3138	0.00863	1.628	2.248	2.86	10000	60002	1.8068
U[38,3]	0.121	0.3076	0.00860	-0.4902	0.1229	0.7196	10000	60002	-0.0787
U[39,3]	-0.7534	0.3198	0.00858	-1.385	-0.7522	-0.1305	10000	60002	-0.4890
U[40,3]	-1.238	0.3058	0.00859	-1.835	-1.239	-0.6409	10000	60002	-1.1132
U[41,3]	-1.61	0.3746	0.00864	-2.351	-1.607	-0.8812	10000	60002	-2.0841

node	mean	sd	MC error	2.50%	median	97.50%	start	sample	Actual values
U[43,3]	-0.2689	0.3295	0.00860	-0.915	-0.267	0.3732	10000	60002	-0.4643
U[44,3]	0.1575	0.3117	0.00859	-0.4513	0.1597	0.7638	10000	60002	-0.1981
U[45,3]	-2.657	0.3471	0.00859	-3.334	-2.655	-1.976	10000	60002	-2.4399
U[46,3]	1.185	0.3787	0.00869	0.444	1.183	1.928	10000	60002	0.8543
U[47,3]	-1.289	0.3484	0.00861	-1.969	-1.29	-0.6036	10000	60002	-1.4409
U[48,3]	-0.1438	0.3587	0.00860	-0.8536	-0.1433	0.556	10000	60002	-0.2016
U[49,3]	-1.286	0.3297	0.00872	-1.936	-1.285	-0.6379	10000	60002	-1.1397
U[50,3]	0.2539	0.422	0.00856	-0.5707	0.254	1.081	10000	60002	0.5303
U[51,3]	-0.4893	0.3704	0.00859	-1.216	-0.4899	0.238	10000	60002	-0.1762
U[52,3]	0.4877	0.3116	0.00859	-0.1241	0.4897	1.102	10000	60002	0.5272
U[53,3]	-0.6573	0.4166	0.00864	-1.473	-0.6574	0.1549	10000	60002	-0.9355
U[54,3]	-0.08731	0.3225	0.00860	-0.7164	-0.08673	0.5433	10000	60002	-0.2024
U[55,3]	-0.6134	0.3191	0.00858	-1.241	-0.6113	0.01037	10000	60002	-0.5191
U[56,3]	0.06914	0.3222	0.00859	-0.5673	0.07102	0.7019	10000	60002	0.0861
U[57,3]	0.2395	0.2894	0.00857	-0.33	0.2413	0.8041	10000	60002	0.3132
U[58,3]	-0.5553	0.312	0.00860	-1.16	-0.5545	0.05745	10000	60002	-0.4501
U[59,3]	0.3046	0.3071	0.00862	-0.2975	0.3032	0.9029	10000	60002	0.1566
U[60,3]	-1.07	0.321	0.00863	-1.695	-1.071	-0.4422	10000	60002	-0.7320
U[61,3]	1.025	0.3345	0.00873	0.3685	1.026	1.677	10000	60002	1.1896
U[62,3]	0.4074	0.3156	0.00859	-0.2109	0.4083	1.025	10000	60002	0.4717
U[63,3]	-0.2497	0.3119	0.00857	-0.8638	-0.2491	0.358	10000	60002	0.1348
U[64,3]	-0.2644	0.316	0.00859	-0.8823	-0.2656	0.3607	10000	60002	-0.6039
U[65,3]	-0.3185	0.3063	0.00865	-0.918	-0.318	0.2809	10000	60002	-0.4141
U[66,3]	-1.019	0.3485	0.00853	-1.705	-1.019	-0.3375	10000	60002	-0.9945
U[67,3]	2.062	0.3478	0.00858	1.381	2.063	2.744	10000	60002	2.0498
U[68,3]	-1.3	0.3205	0.00860	-1.929	-1.301	-0.6736	10000	60002	-1.4650
U[69,3]	0.6726	0.3224	0.00854	0.04381	0.6738	1.302	10000	60002	0.5855
U[70,3]	0.1716	0.2948	0.00859	-0.407	0.1719	0.7452	10000	60002	0.0183
U[71,3]	2.406	0.3107	0.00862	1.796	2.407	3.012	10000	60002	2.3590
U[72,3]	0.7609	0.3235	0.00856	0.129	0.7614	1.394	10000	60002	0.8880
U[73,3]	-1.75	0.3272	0.00860	-2.387	-1.751	-1.106	10000	60002	-1.4180
U[74,3]	0.1757	0.3505	0.00862	-0.5084	0.1753	0.8646	10000	60002	0.4820
U[75,3]	-0.4358	0.3461	0.00859	-1.111	-0.4358	0.2403	10000	60002	-0.3600
U[76,3]	1.946	0.3558	0.00861	1.248	1.945	2.649	10000	60002	1.6929
U[77,3]	1.73	0.3029	0.00861	1.139	1.733	2.317	10000	60002	1.7721
U[78,3]	0.3072	0.3318	0.00855	-0.3464	0.3084	0.951	10000	60002	0.3943
U[79,3]	0.8985	0.2988	0.00857	0.3134	0.8992	1.483	10000	60002	0.8308
U[80,3]	0.915	0.3228	0.00858	0.2812	0.9155	1.547	10000	60002	1.1303
U[81,3]	-0.6303	0.3035	0.00858	-1.224	-0.6304	-0.02838	10000	60002	-0.7234
U[82,3]	1.001	0.3348	0.00872	0.3426	1.001	1.658	10000	60002	0.7147
U[83,3]	-0.8603	0.3238	0.00865	-1.492	-0.8616	-0.2231	10000	60002	-1.3631
U[84,3]	1.643	0.4229	0.00865	0.8218	1.643	2.472	10000	60002	1.5779
U[85,3]	-0.04467	0.3049	0.00860	-0.6432	-0.04488	0.5507	10000	60002	0.2232
U[86,3]	0.08859	0.3013	0.00861	-0.5036	0.08996	0.673	10000	60002	0.0755
U[87,3]	0.5075	0.3126	0.00855	-0.1052	0.5093	1.118	10000	60002	0.4859
U[88,3]	-0.227	0.3209	0.00857	-0.8584	-0.2281	0.399	10000	60002	-0.4449

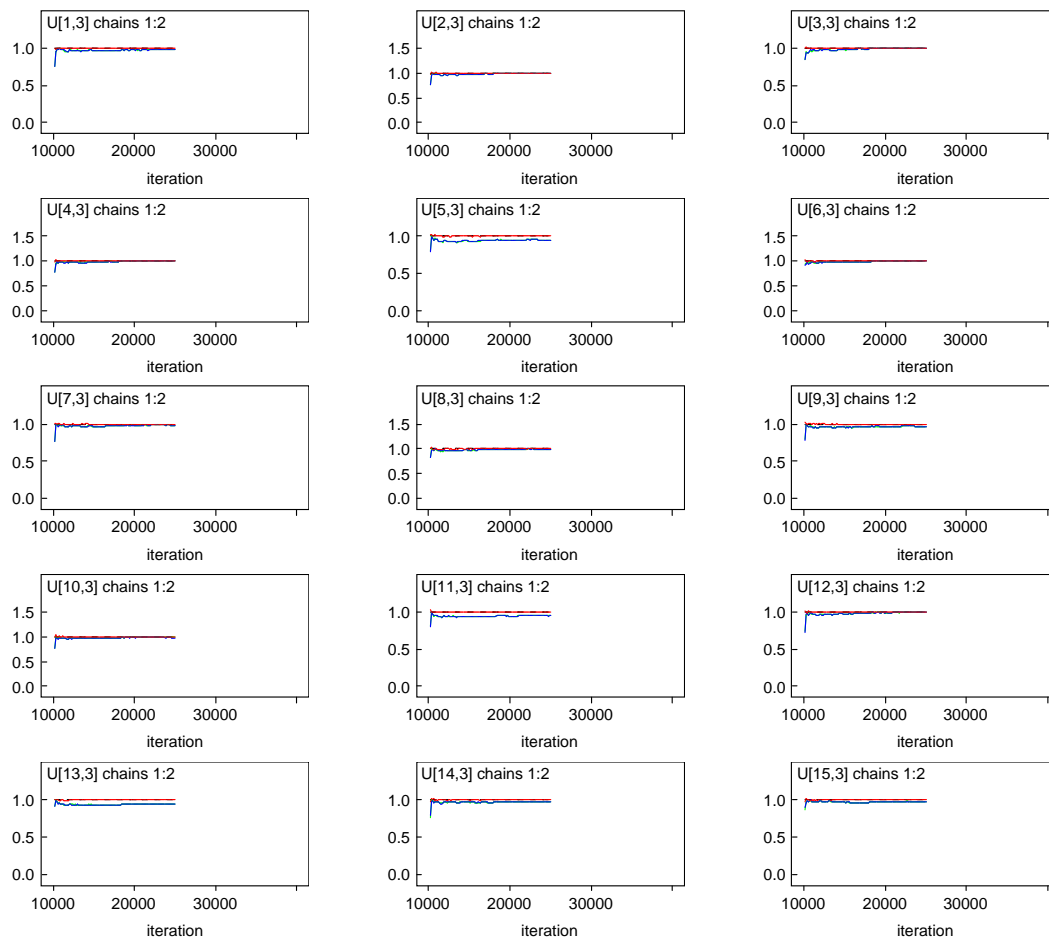
node	mean	sd	MC error	2.50%	median	97.50%	start	sample	Actual values
U[90,3]	0.03107	0.3241	0.00865	-0.6055	0.03308	0.6681	10000	60002	0.0664
U[91,3]	0.06002	0.3311	0.00856	-0.5854	0.0584	0.712	10000	60002	0.3512
U[92,3]	-1.742	0.3193	0.00858	-2.364	-1.741	-1.12	10000	60002	-1.9964
U[93,3]	1.008	0.3365	0.00855	0.3458	1.01	1.668	10000	60002	1.2492
U[94,3]	1.005	0.3463	0.00875	0.328	1.006	1.682	10000	60002	1.2948
U[95,3]	-0.8937	0.3242	0.00860	-1.533	-0.8922	-0.2643	10000	60002	-0.7213
U[96,3]	0.4998	0.406	0.00870	-0.3013	0.5003	1.298	10000	60002	0.8072
U[97,3]	0.8011	0.3295	0.00870	0.1527	0.8021	1.447	10000	60002	0.8228
U[98,3]	0.3959	0.301	0.00855	-0.1937	0.3961	0.9888	10000	60002	0.2562
U[99,3]	1.14	0.3128	0.00857	0.5275	1.14	1.754	10000	60002	1.1956
U[100,3]	-1.358	0.3322	0.00858	-2.008	-1.355	-0.7119	10000	60002	-1.4010
U[101,3]	1.861	0.4836	0.01043	0.9223	1.861	2.812	10000	60002	1.5428
U[102,3]	-1.902	0.3578	0.00866	-2.599	-1.903	-1.199	10000	60002	-2.0535
U[103,3]	-0.765	0.2885	0.00859	-1.328	-0.765	-0.1998	10000	60002	-0.7086
U[104,3]	-1.799	0.2997	0.00859	-2.384	-1.798	-1.214	10000	60002	-2.0401
U[105,3]	-0.7171	0.3048	0.00859	-1.308	-0.7187	-0.1181	10000	60002	-0.9866
U[106,3]	0.3936	0.3289	0.00866	-0.2482	0.3928	1.043	10000	60002	0.3256
U[107,3]	0.3953	0.3345	0.00863	-0.2654	0.396	1.048	10000	60002	0.8028
U[108,3]	-0.8683	0.3324	0.00860	-1.516	-0.8688	-0.2163	10000	60002	-1.3338
U[109,3]	0.6742	0.3097	0.00856	0.0642	0.6761	1.277	10000	60002	0.6806
U[110,3]	1.479	0.3394	0.00865	0.8112	1.48	2.146	10000	60002	1.4519
U[111,3]	-0.07207	0.3629	0.00857	-0.7825	-0.07315	0.6421	10000	60002	-0.1312
U[112,3]	0.8818	0.3025	0.00862	0.2902	0.8823	1.476	10000	60002	1.1946
U[113,3]	-0.9431	0.3002	0.00860	-1.532	-0.943	-0.3584	10000	60002	-0.8810
U[114,3]	0.5877	0.3539	0.00855	-0.1101	0.5876	1.283	10000	60002	0.7319
U[115,3]	0.9821	0.3401	0.00857	0.3184	0.9834	1.643	10000	60002	0.6719
U[116,3]	-0.05857	0.3575	0.00869	-0.7641	-0.05925	0.6435	10000	60002	-0.4504
U[117,3]	1.051	0.3084	0.00864	0.4504	1.052	1.659	10000	60002	1.2376
U[118,3]	0.5985	0.3702	0.00853	-0.1177	0.5995	1.323	10000	60002	0.4839
U[119,3]	0.5479	0.3144	0.00858	-0.06697	0.55	1.159	10000	60002	0.7038
U[120,3]	-0.1591	0.3082	0.00858	-0.7592	-0.159	0.4419	10000	60002	0.1484
U[121,3]	-1.746	0.3361	0.00856	-2.408	-1.744	-1.086	10000	60002	-2.0053
U[122,3]	0.3349	0.3151	0.00860	-0.2814	0.3341	0.956	10000	60002	0.1560
U[123,3]	-0.02037	0.3186	0.00865	-0.6451	-0.01756	0.6008	10000	60002	0.0157
U[124,3]	-0.04692	0.3633	0.00885	-0.7643	-0.04637	0.6602	10000	60002	-0.2022
U[125,3]	0.8883	0.3474	0.00860	0.209	0.8892	1.564	10000	60002	0.9316
U[126,3]	0.3268	0.3503	0.00860	-0.3627	0.3263	1.01	10000	60002	-0.0451
U[127,3]	-0.2429	0.3189	0.00857	-0.8688	-0.2426	0.3822	10000	60002	-0.4150
U[128,3]	1.028	0.3128	0.00861	0.4121	1.029	1.636	10000	60002	1.3100
U[129,3]	-2.191	0.305	0.00859	-2.791	-2.19	-1.59	10000	60002	-2.1579
U[130,3]	-0.6776	0.326	0.00856	-1.319	-0.6776	-0.03479	10000	60002	-0.7702
U[131,3]	0.08465	0.3213	0.00860	-0.5444	0.08594	0.7153	10000	60002	0.2215
U[132,3]	0.06586	0.3112	0.00856	-0.5393	0.06518	0.6763	10000	60002	-0.0762
U[133,3]	0.942	0.3108	0.00861	0.3322	0.9415	1.552	10000	60002	0.7201
U[134,3]	1.161	0.3353	0.00860	0.4995	1.161	1.816	10000	60002	0.6829
U[135,3]	-1.166	0.3129	0.00858	-1.783	-1.165	-0.5588	10000	60002	-1.4675
U[136,3]	0.3326	0.2901	0.00858	-0.2381	0.3329	0.8968	10000	60002	0.4364

node	mean	sd	MC error	2.50%	median	97.50%	start	sample	Actual values
U[137,3]	1.664	0.3359	0.00864	1.001	1.662	2.327	10000	60002	1.7755
U[138,3]	-0.8133	0.3106	0.00856	-1.424	-0.8124	-0.2071	10000	60002	-0.4908
U[139,3]	0.5963	0.408	0.00877	-0.211	0.5991	1.395	10000	60002	0.6981
U[140,3]	1.239	0.3118	0.00859	0.6259	1.241	1.842	10000	60002	1.5655
U[141,3]	-2.172	0.3148	0.00866	-2.792	-2.17	-1.557	10000	60002	-2.2697
U[142,3]	-0.3042	0.3234	0.00863	-0.938	-0.3036	0.3303	10000	60002	-0.2860
U[143,3]	-0.458	0.3487	0.00860	-1.143	-0.4578	0.2293	10000	60002	-0.7136
U[144,3]	1.922	0.3347	0.00859	1.264	1.92	2.582	10000	60002	1.6020
U[145,3]	-0.937	0.3352	0.00860	-1.593	-0.9381	-0.2795	10000	60002	-0.9479
U[146,3]	-0.388	0.3488	0.00861	-1.063	-0.3877	0.2995	10000	60002	-0.6285
U[147,3]	-1.087	0.3044	0.00865	-1.683	-1.086	-0.4941	10000	60002	-1.1145
U[148,3]	0.1641	0.3315	0.00861	-0.4869	0.1633	0.813	10000	60002	0.1950
U[149,3]	-0.7269	0.3357	0.00861	-1.379	-0.7264	-0.06781	10000	60002	-0.8448
U[150,3]	0.5596	0.307	0.00862	-0.0448	0.5593	1.16	10000	60002	0.4434
U[151,3]	0.7939	0.3088	0.00854	0.1868	0.7938	1.399	10000	60002	0.6832
U[152,3]	-1.013	0.3326	0.00860	-1.659	-1.013	-0.3589	10000	60002	-1.0407
U[153,3]	0.7141	0.292	0.00855	0.141	0.7152	1.284	10000	60002	0.3856
U[154,3]	0.7426	0.3204	0.00860	0.1103	0.7425	1.369	10000	60002	0.5698
U[155,3]	1.766	0.347	0.00856	1.087	1.763	2.442	10000	60002	1.3641
U[156,3]	-0.09278	0.3485	0.00860	-0.7747	-0.09247	0.5922	10000	60002	-0.4570
U[157,3]	-1.519	0.333	0.00868	-2.177	-1.519	-0.8727	10000	60002	-1.6470
U[158,3]	0.2044	0.2976	0.00862	-0.3825	0.2059	0.7828	10000	60002	0.1129
U[159,3]	1.687	0.4269	0.00880	0.8595	1.689	2.521	10000	60002	1.6094
U[160,3]	0.5652	0.3049	0.00859	-0.03138	0.5654	1.165	10000	60002	0.4380
U[161,3]	0.3629	0.3229	0.00858	-0.2711	0.3643	0.9959	10000	60002	0.3205
U[162,3]	-1.275	0.3231	0.00860	-1.911	-1.273	-0.6413	10000	60002	-1.2842
U[163,3]	0.05481	0.326	0.00866	-0.5836	0.05476	0.6938	10000	60002	0.1020
U[164,3]	-0.4461	0.3353	0.00859	-1.105	-0.4459	0.2089	10000	60002	-0.4548
U[165,3]	-0.8538	0.2992	0.00858	-1.436	-0.854	-0.2676	10000	60002	-0.7972
U[166,3]	-1.114	0.3278	0.00855	-1.755	-1.114	-0.4721	10000	60002	-1.0535
U[167,3]	-0.9286	0.2987	0.00858	-1.515	-0.9283	-0.3411	10000	60002	-0.8223
U[168,3]	1.056	0.3495	0.00863	0.37	1.055	1.741	10000	60002	1.0021
U[169,3]	0.8766	0.3219	0.00869	0.2438	0.8775	1.508	10000	60002	1.1623
U[170,3]	-0.8343	0.3346	0.00866	-1.492	-0.835	-0.178	10000	60002	-1.0329
U[171,3]	-0.3383	0.3468	0.00856	-1.022	-0.3383	0.3379	10000	60002	-0.5538
U[172,3]	0.941	0.3739	0.00861	0.206	0.9421	1.671	10000	60002	0.8997
U[173,3]	0.407	0.3127	0.00860	-0.2067	0.4081	1.022	10000	60002	0.5937
U[174,3]	-1.401	0.3124	0.00856	-2.014	-1.401	-0.7888	10000	60002	-1.5762
U[175,3]	-1.81	0.3309	0.00857	-2.455	-1.811	-1.162	10000	60002	-1.6044
U[176,3]	1.124	0.3338	0.00856	0.4744	1.124	1.781	10000	60002	1.1324
U[177,3]	0.2677	0.3331	0.00863	-0.3856	0.2677	0.9236	10000	60002	0.4848
U[178,3]	0.5283	0.435	0.00880	-0.3235	0.5284	1.385	10000	60002	1.2640
U[179,3]	0.4271	0.3204	0.00854	-0.1999	0.4282	1.051	10000	60002	0.4474
U[180,3]	0.6268	0.32	0.00858	-0.00146	0.6273	1.249	10000	60002	0.5572
U[181,3]	0.6374	0.3492	0.00857	-0.04196	0.6358	1.321	10000	60002	0.6805
U[182,3]	0.1783	0.3834	0.00865	-0.5763	0.1792	0.9299	10000	60002	0.3966

node	mean	sd	MC error	2.50%	median	97.50%	start	sample	Actual values
U[183,3]	-0.2847	0.3777	0.00866	-1.027	-0.2835	0.4535	10000	60002	-0.6381
U[184,3]	-1.124	0.3067	0.00861	-1.731	-1.124	-0.5273	10000	60002	-1.0281
U[185,3]	-0.8432	0.3468	0.00853	-1.527	-0.8419	-0.1612	10000	60002	-0.8735
U[186,3]	0.5686	0.3644	0.00896	-0.1412	0.5684	1.284	10000	60002	0.1583
U[187,3]	0.7006	0.3206	0.00857	0.06725	0.7016	1.328	10000	60002	0.6154
U[188,3]	-2.258	0.3205	0.00858	-2.889	-2.256	-1.631	10000	60002	-2.5297
U[189,3]	1.196	0.3017	0.00856	0.6071	1.197	1.787	10000	60002	1.0499
U[190,3]	-0.1481	0.3162	0.00856	-0.7654	-0.151	0.4745	10000	60002	0.0000
U[191,3]	0.19	0.3512	0.00872	-0.4949	0.1887	0.8833	10000	60002	0.0000
U[192,3]	0.288	0.3424	0.00868	-0.3788	0.2872	0.963	10000	60002	0.0000
U[193,3]	-0.1345	0.3316	0.00860	-0.7891	-0.1335	0.5149	10000	60002	0.0000
U[194,3]	0.04397	0.3348	0.00859	-0.6149	0.04408	0.7041	10000	60002	0.0000
U[195,3]	0.1714	0.3326	0.00857	-0.4738	0.172	0.819	10000	60002	0.0000
U[196,3]	-0.0949	0.339	0.00869	-0.7584	-0.0957	0.5696	10000	60002	0.0000
U[197,3]	0.1637	0.3157	0.00862	-0.4549	0.163	0.7849	10000	60002	0.0000
U[198,3]	-0.05965	0.2993	0.00858	-0.6475	-0.0598	0.5202	10000	60002	0.0000
U[199,3]	0.06	0.3463	0.00862	-0.6159	0.06056	0.7398	10000	60002	0.0000
U[200,3]	-0.1525	0.3786	0.00863	-0.8988	-0.1535	0.5875	10000	60002	0.0000

Appendix 4.2

Gelman Rubin statistics for Protein level class differences (Winbugs results for the first 1-15 proteins, the rest of the proteins have similar ranges in the Gelman Rubin statistics (< 1.2) as these first 15 proteins. Due to the limitation of pages, they are not all included.)



CHAPTER 5

A cardiac proteomics study-case study I

5.1 Description of the study

The cardiac proteomic study is one part of Dr. Ralph Stewart's (Department of medicine, University of Auckland, and Green Lane Cardiovascular Service, Auckland City Hospital) double blinded randomized control trial that investigated how the intra-coronary metoprolol (beta blocker) changed the myocardial metabolism of peptides, proteins and metabolites profiles in patients admitted to hospital with a first myocardial infarction, who had serum troponin T > 0.1mmol/l and needed coronary angioplasty/stenting. In the main study, only several candidate markers were selected in the biochemical analysis. The selected peptides, proteins and metabolites included Brain Natriuretic Peptide, free fatty acids, glucose and lactic acid. Brain Natriuretic Peptide (BNP) is a peptide secreted from the myocardial tissue cells which is proved to be a useful marker for the diagnosis of heart failure and other heart diseases (Beck et al., 2011; Lin and R.A.H.Stewart, 2011). Free fatty acids is one of the fatty acids molecules which can be transported into blood stream without aids of other carriers and it can be used in any part of the body where needed ("Free Fatty Acids in the Blood,"). Glucose is a metabolite for energy. Lactic acid is produced in the muscle tissue during hard exercise. These candidate molecules are hypothesized to change during the coronary intervention.

In the proteomic section of the study, we investigated how the whole plasma proteome profile changed after coronary angioplasty intervention; it involved eight patients from the placebo arm. The main study collected five plasma samples at different time points of each patient during their percutaneous coronary intervention, two of these plasma samples were selected in the proteomic study.

5.2 The laboratory methods (a brief summary of the clinical laboratory sample preparation and the iTRAQ experiment in the University lab)

5.2.1 *The clinical laboratory processing for biological samples*

Blood for proteomic analysis was collected in EDTA tubes spiked with 8µM Pepstatin and 16µM Bestatin (Protease inhibitors). Plasma was extracted after centrifugation at

3,350 xg for 5 mins. The resulting plasma was snap frozen with dry ice/methanol slurry, before storage at minus 70 degrees centigrade. This process was performed in the cardiac catheterization laboratory within 60 seconds of collection by Dr. Patrick Gladding (*North Shore Hospital*).

5.2.2 The MS/MS iTRAQ experiments for coronary sinus plasma tissue samples

The processed plasma tissue samples were transformed to the Center for Proteomic and Genomic, University of Auckland for proteomic analysis. Each tissue sample was independently dissolved and digested by trypsin and mixed with iTRAQ reagents for the preparation of the 4-plex MS/MS assays in the Proteomic lab. The top 12 abundant plasma proteins (Table 5.1) were depleted with IgY-12 SC Spin Column (Beckman) with 10 salt-steps for the fractionations.

The depletion is a process to exclude the high abundant proteins from the final mass spectrometric analysis. The high sensitive mass spectrometer can detect proteins with low or very low abundance when excluding those high abundant ones.

Protein database (The Swiss-Prot human protein sequence database) was used in the Protein Pilot Software to match observed peptides with their corresponding proteins for the identification. In an effort to screen out low scoring protein-hits from a large scale analysis, a reversed database search was applied to assess how accurate the protein-hits are. Any protein-hits with a confidence score below a criterion as an unsatisfactory (false) reversed database matches have been excluded. In this study, according to the recommendation from the biochemist in the Center for Proteomic and Genomic, only those peptides with confidence score >10 were included for protein quantitation.

5.2.3 Clinical study design and experimental design

The main study is a randomization control trial with parallel groups of placebo and intra-coronary metoprolol. In the placebo arm, a pair of blood samples (one taken before and one taken 20 minutes after the percutaneous coronary intervention (PCI)) was selected from the series of measurements for each patient. A total of six right coronary artery lesion patients, and two left anterior descending lesion patients were included in this sub-study.

Sixteen blood tissues in total were allocated across 4 runs and 4 iTRAQ™ isobaric tags. In each run, a 4-plex MS/MS assay was used to randomly allocate paired samples from 2 patients. The randomization occurred separately for each run. Four 4-plex assays were used to allocate the 16 samples from 8 patients. This design, Randomized Complete Block Design with paired samples, achieves the orthogonality between label and time of sample collection (before or after PCI).

Tryptic peptides were labeled by the iTRAQ™ isobaric tags, followed by fractionation and separation of 2d of LC, and analyzed by tandem mass spectrometry (MS/MS). Numbers of 535, 254, 242, and 237 proteins were discovered from the 4 ITRAQ runs respectively.

Table 5.1 List of the top 12 high abundant proteins in plasma

- Albumin
- IgG
- α 1-Acid Glycoprotein
- α 1-Antitrypsin
- Apolipoprotein A-I
- Apolipoprotein A-II
- Fibrinogen
- Haptoglobin
- IgA
- IgM
- Transferrin
- α 1-Acid Glycoprotein
- α 1-Antitrypsin
- α 2-Macroglobulin

5.3. The analytical methods

5.3.1 A single protein multilevel model

When proteins are analyzed one by one in a multilevel mixed model, the intensities of the peptides for a given protein will be modeled as the response variable; the variations of the intensities of peptides are expected to be composed of the variations from experiments, from the instrumental features of the mass spectrometer, and from subject's physiological conditions.

In this case study, the experimental factors that potentially contribute to the variances of the peptide intensities includes the iTRAQ labels and runs effects. The value of the mass-to-charge ratio (m/z) is one type of the instrumental features. The subject's physiological conditions include different subject's baseline peptide intensity level which can be defined as subject intercepts and the sampling time of the peptide intensity level. The sampling time records if the sample is collected either before or after the coronary intervention. A single protein model similar as equation (3) defined in chapter 4 is constructed below,

$$y_{i,l} = b_{0,l} + \beta_1 mz_{i,l} + \sum_{h=1}^4 \beta_{2,h} label_{h,i,l} + \sum_{r=1}^8 \beta_{3,r} run_{r,i,l} + e_{i,l}, \quad (1)$$

$$b_{0,l} = \gamma_{0,0} + \gamma_1 \times sampling_time_{i,l} + u_{0,l}, \quad (2)$$

where,

$y_{i,l}$ represents the intensity on the natural log scale for the i th observed peptide and l th subject, i ranges between 1 and the total number of peptides included in the analysis, $l = 1, \dots, 8$;

The symbol $b_{0,l}$ represents the intercept for the l th subject;

The symbol mz_i represents the centralized m/z ratio for the i th peptide observation; and β_1 is its regression coefficient;

The symbol $label_{h,i,l}$ and $run_{r,i,l}$ defines the label and run that $y_{i,l}$ are observed, $\beta_{2,h}$ and $\beta_{3,r}$ are regression coefficients for label h and run r respectively;

$\gamma_{0,0}$ represents the subject level regression coefficients for intercept and sampling time;

$sampling_time_{i,l}$ is a binary variable that indicates if the blood sample was collected before or after the coronary intervention; its coefficient γ_1 represents the abundance difference between the two sampling time which is the effect caused by the coronary intervention;

$u_{0,l}$ represents the random residual term at the subject level.

Equation (1) of level 1 includes mz_i , $label_{h,i,l}$ and $run_{r,i,l}$ as fixed effects and a random residual error term $e_{i,l}$. Equation (2) of level 2 includes the different sampling time $sampling_time_{i,l}$ as a fixed effect. Since every biological subject has multiple peptide records, and we hypothesized that each subject may have a different level of abundance for a protein, equation (2) includes a random residual term $u_{0,l}$ for the subjects.

Substituting (2) to (1) gives us

$$y_{i,l} = \left[\gamma_{0,0} + \gamma_1 \times sampling_time_{i,l} + \beta_1 mz_{i,l} + \sum_{h=1}^4 \beta_{2,h} label_{h,i,l} + \sum_{r=1}^8 \beta_{3,r} run_{r,i,l} \right] + [u_{0,l} + e_{i,l}] \quad (3)$$

, where $\beta_1, \dots, \beta_{3,r}$ are the fixed effects regression coefficients for the experimental factors and $\gamma_{0,0}, \gamma_1$ are the fixed effect coefficients for intercepts and sampling time respectively; $u_{0,l}, e_{i,l}$ are the random effects for subjects and unexplained errors respectively.

The results of model described in (3) showed that, Eighty-four of the 151 proteins have sufficient peptide information to achieve the convergence in the algorithm of the mixed models for estimating both random and fixed effects. The variance of subjects' intercepts range from 0.07 to 8.88 among these 84 single protein models. The differences in the relative quantities between two sampling times on the log-scale ranged between -3.52 - 0.57, which is equivalent to fold changes ranging between 0.03-1.77.

5.3.2 Multiple protein multivariate model with random effects at protein and subject levels ignoring missing values

For the data structure described in chapter 4, peptides are nested within proteins and proteins are crossed over subjects. A hierarchical multivariate approach will enable us to analyze all proteins using one model. This model estimates the intervention and the slope of the m/z ratio as random effects at the protein level and includes a random intercept representing the observed abundance of various proteins. iTRAQ Label and run are included as fixed effects, assuming that their effects are the same across different proteins. The slopes of m/z within different combinations of run and label were not shown to vary (figure 5.1) and the interactions of m/z and run or label were therefore not included in the model. Subject is included as a random effect.

The two-level model is defined as follows:

$$\begin{aligned} \gamma_{i,l,p} = & \phi_{0,l} + \beta_{0,p} protein_{i,l,p} + \beta_{1,p} m/z_{i,l,p} + \sum_{h=1}^4 \beta_{2,h} label_{h,i,l,p} + \sum_{r=1}^8 \beta_{3,r} run_{r,i,l,p} \\ & + \beta_{4,p} sampling_time_{i,l,p} + e_{i,l,p} \end{aligned} \quad (4)$$

$$\beta_{0,p} = \alpha_0 + b_{0,p}$$

$$\beta_{1,p} = \alpha_1 + b_{1,p}$$

$$\beta_{4,p} = \alpha_2 + b_{2,p}$$

(5)

$$\phi_{0,l} = \gamma_{0,0} + c_{0,l}, \quad (6)$$

Equation (4) defines the level 1 relation between the peptide intensity $\gamma_{i,l,p}$ of protein p subject l and independents m/z , label, run, and sampling time; $\beta_{2,h}, \beta_{3,r}$ are the regression coefficients for label h and run r , they are the same across all proteins. $\beta_{0,p}$, $\beta_{1,p}$ and $\beta_{4,p}$ represents the protein intercept, coefficients for m/z and sampling time respectively. $\beta_{0,p}, \beta_{1,p}$ and $\beta_{4,p}$ vary across different proteins.

Equation (5) assigns the relation between protein level random coefficients and their explanatory variables. In this case study, only an intercept $\alpha_0, \dots, \alpha_2$ and a random residual $b_{0,p}, \dots, b_{2,p}$ for each protein are included as the explanatory variables for each random coefficient.

Equation (6) assigns the relation between subject level random coefficients and their explanatory variables. Only an intercept $\gamma_{0,0}$ and a random residual $c_{0,l}$ for each subject are included as the explanatory variables.

Substituting (5)-(6) into (4) gives us the following mixed effected model:

$$\gamma_{i,l,p} = \left[\begin{array}{l} \gamma_{0,0} + \alpha_0 \times protein_{i,l,p} + \alpha_1 mz_{i,l,p} + \sum_{h=1}^4 \beta_{2,h} label_{h,i,l,p} + \sum_{r=1}^8 \beta_{3,r} run_{r,i,l,p} \\ + \alpha_2 sampling_time_{i,l,p} \end{array} \right] + \left[b_{1,p} \times mz_{i,l,p} + b_{2,p} \times sampling_time_{i,l,p} + b_{0,p} \times protein_{i,l,p} + c_{0,l} + e_{i,l,p} \right] \quad (7)$$

where,

$$(b_{0,p}, b_{1,p}, b_{2,p}) \sim MVN(\mathbf{0}, \Phi); c_{0,l} \sim N(0, \tau_0^2); e_{i,l,p} \sim N(0, \sigma^2);$$

$$\Phi = \begin{bmatrix} \sigma_{b,0}^2 & \sigma_{b,0}\sigma_{b,1} & \sigma_{b,0}\sigma_{b,2} \\ \sigma_{b,0}\sigma_{b,1} & \sigma_{b,1}^2 & \sigma_{b,1}\sigma_{b,2} \\ \sigma_{b,0}\sigma_{b,2} & \sigma_{b,1}\sigma_{b,2} & \sigma_{b,2}^2 \end{bmatrix}.$$

Based on the model defined in equation (7), three multivariate multilevel proteins models were used for the cardiac proteomic study and their results are shown in Table 5.2.

As shown in Table 5.2, Model 1 and Model 2 do not include a fixed effect for the intervention as Model 3 does. In model 3, the fixed effect for the intervention is shown to be significant as are the centralized mass-to-charge ratio (m/z), label 116 and 117. Model 3 has the smallest AIC, BIC, deviance and REML deviance, and is chosen as the best model for this dataset.

In model 3, the variance of random effects for proteins parameters is 0.09 (std: 0.29), $7.0e-07$ (std: 0.0009), and 0.52 (std: 0.72) for $\sigma^2_{b,0}$ (intercept), $\sigma^2_{b,1}$ (m/z) and $\sigma^2_{b,2}$ intervention respectively. The variance of subject intercept τ_0^2 is 1.5 (std: 1.23). The variance of residual errors σ^2 is 2.57. The variance of protein intercepts is relatively smaller compared to the variance of subject intercepts. The intra-class correlation coefficient (ICC) is a measure of variance for the estimated random effect compared to the unexplained variance. ICC for intervention effect at the protein level $\frac{\sigma^2_{b,2}}{\sigma^2_{b,2} + \sigma^2}$ is 0.20, and the ICC for the subjects $\frac{\tau_0^2}{\tau_0^2 + \sigma^2}$ is 0.58.

The fixed effects of centralized m/z and intervention are -0.0015 and -1.42, respectively, and both are shown to be significant ($t = -10.25$ and -19.25 respectively). Label 116 and label 117 are also significantly different from the reference label 113. The significant fixed effect of intervention reveals a systematic change of -1.42 on the log-scale in all protein expressions, which can be explained by a possible dilution impact on the molecules in the bloodstream caused at the time when they flow through the coronary sinus from the aorta. The predicted random effects of intervention for different proteins indicate the magnitudes of fold changes in the protein abundances introduced by the intervention. The predicted random effects of intervention for each protein can be equivalently treated as the adjusted protein ratios derived from the model.

Table 5.2 The multivariate multilevel models

Model 1: intensity~ factor(run)+factor(label)+ (1+m/z+ factor(intervention) | protein) + (1 | subject)

Random effects:				
Groups	Name	Variance	Std. Dev.	Corr
protein	(Intercept)	0.29	0.54	
	centralized m/z	2.98e-06	0.0017	-0.74
	intervention)	2.54	1.60	-0.90 0.82
subject	(Intercept)	1.50	1.23	
	Residual	2.57	1.60	
Number of obs: 15895, groups: protein, 151; subject, 8				
Fixed effects:				
		Estimate	Std. Error	t value
	(Intercept)	3.46	0.87	3.99
	run2	-0.0080	1.22	-0.006
	run3	-0.0067	1.22	-0.005
	run4	-1.50	1.22	-1.22
	label115	-0.085	0.043	-1.99
	label116	0.43	0.052	8.24
	label117	-0.23	0.047	-4.90
AIC	BIC	log Likelihood	deviance	REMLdev
60812	60927	-30391	60775	60782

Model 2: intensity~ factor (run)+ factor(label)+ m/z +(1+m/z+ factor(intervention) | protein) + (1 | subject)

Random effects:				
Groups	Name	Variance	Std.Dev.	Corr
protein	(Intercept)	0.09	0.29	
	centralized m/z	7.0e-07	0.0009	-0.053
	intervention	0.52	0.72	-0.640 0.205
subject	(Intercept)	1.50	1.23	
	Residual	2.57	1.60	
Number of obs: 15895, groups: protein, 151; subject, 8				
Fixed effects:				
		Estimate	Std. Error	t value
run2		-0.01	1.22	-0.008
run3		-0.0079	1.22	-0.006
run4		-1.50	1.22	-1.23
label115		-0.088	0.0423	-2.07
label116		0.4296	0.0521	8.24
label117		-0.2334	0.0471	-4.96
centralized m/z		-0.0012	0.0001	-8.65
AIC	BIC	log likelihood	deviance	REMLdev
60813	60935	-30390	60757	60781

Model 3: intensity~ factor (run)+ factor(label)+ m/z + factor(intervention)+(1+m/z+ factor(intervention) | protein) + (1 | subject)

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
protein	(Intercept)	0.09	0.29	
	centralized m/z	7.0e-07	0.0009	-0.053
	intervention	0.52	0.72	-0.640 0.205
subject	(Intercept)	1.50	1.23	
	Residual	2.57	1.60	

Number of obs: 15895, groups: protein, 151; subject, 8

Fixed effects:

	Estimate	Std. Error	t value
run2	-0.012	1.22	0.010
run3	-0.0087	1.22	-0.007
run4	-1.50	1.23	-1.23
label115	-0.07	0.04	-1.66
label116	0.43	0.05	8.22
label117	-0.22	0.047	-4.68
centralized m/z	-0.0015	0.0001	-10.15
intervention	-1.42	0.074	-19.25

AIC	BIC	logLik	deviance	REMLdev
60625	60756	-30296	60564	60591

5.3.3 Multiple protein multivariate model with random effects at protein and subject levels and with the missing mechanisms modeled by a Bayesian approach.

This case study has 17780 peptide observations of which 1667 (9.4%) had censored and 218 (1.2%) had completely missing values in the peptide intensities.

In the Bayesian model, the same linear multilevel mixed regression model as defined in (7) were used to define the relation between the peptide intensities and explanatory variables. The explanatory variables include protein level parameters (intercept, m/z ratio and sampling time), peptide level parameters (m/z ratio, run and label), and subject level intercepts. Logistic regression was used to model the missing probability as a function of m/z ratio and quantities (observed or unobserved).

The Bayesian model:

$$\begin{aligned} \mu_{i,l,p} &= \phi_{0,l} subject_{i,l,p} + \beta_{0,p} protein_{i,l,p} + \beta_{1,p} protein_{i,l,p} \times mZ_{i,l,p} \\ &+ \beta_{4,p} protein_{i,l,p} \times sampling_time_{i,l,p} + \sum_{h=1}^4 \beta_{2,h} label_{h,i,l,p} + \sum_{r=1}^8 \beta_{3,r} run_{r,i,l,p}, \\ \text{logit}(pm = \Pr(\text{missed } \gamma_{i,l,p})) &= \alpha_0 + \alpha_1 mZ_{i,l,p} + \alpha_2 \gamma_{i,l,p} \end{aligned} \quad (8)$$

where

$\mu_{i,l,p}$ defines the mean for the peptide intensity including completed, completely missing and censored values, $\gamma_{i,l,p} \sim N(\mu_{i,l,p}, \sigma^2)$;

$subject_{i,l,p}$ defines the subject identity and $\phi_{0,l}$ defines the intercepts for each subject;

$\beta_{0,p}, \beta_{1,p}, \beta_{4,p}$ define the protein level parameters: intercept, m/z ratio and sampling time and $protein_{i,l,p}$ is a variable recording the identity of protein;

$\beta_{2,h}$ and $\beta_{3,r}$ are regression coefficients for label h and run r respectively;

pm defines the probability of having a missing intensity value, $\alpha_0, \dots, \alpha_1$ define the regression coefficients in the logistic regression for pm .

Priors:

As described in chapter 4, the missingness is modeled for the censoring and completely missing using logistic regression. The coefficients of logistic regression α_1 and α_2 are part of the joint unknown parameters with a pair of chosen normally distributed informative priors $N(0.0085, 4e-8)$, $N(-0.45, 0.25)$ for m/z ratio and peptide abundance respectively (Hrydziuszko and Viant, 2012). The protein level parameters $\beta_{0,p}, \beta_{1,p}, \beta_{4,p}$ used a multivariate normal distributed prior (γ, T) where,

$$\begin{aligned} \gamma &\sim MVN(\mu, \Omega^{-1}), T \sim invWISHART(\Phi^{-1}, 3) \\ \mu &= (0.5, 0.5, 0.5), \\ \Omega &= \begin{pmatrix} 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 \end{pmatrix}, \Phi = \begin{pmatrix} 0.01 & 0.001 & 0.001 \\ 0.001 & 0.01 & 0.001 \\ 0.001 & 0.001 & 0.01 \end{pmatrix} \end{aligned}$$

The non-informative priors for the peptide level parameters $\beta_{2,h}$ and $\beta_{3,r}$ are normally distributed and denoted as $N(0,10)$ with mean 0 and precision 0.1, and the non-informative prior for the subject level parameters $\phi_{0,l}$ is normal distributed and denoted as $N(0,1)$ with mean 0 and precision 1.

Computing programs:

In the program for Gibbs sampling, the missing probability for censored peptide intensity is fixed to be 1; In the HMC/NUTs program, the missing probability for the censoring is the cumulative probability of an intensity value lower than the detectable limit and it is derived by integrating a standard normal density function between the negative infinite and the standardized detectable limit $c \int_{-\infty}^c N(x|0,1) dx$, where c is standardized by the mean intensity values $\mu_{i,l,p}$ and the variance σ^2 .

The Rstan HMC/NUTS program uses the same priors as the BUGS program; except for the scaled matrix Φ of the inversed Wishart distribution that 0.1 is set as the diagonal

value and 0.05 as the off-diagonal value. Both the BUGS and the Rstan program utilize the m/z ratio information from the censored and missing peptide observations for deriving the posterior estimates of the unknown parameters.

Comparison of the results of posteriors of the unknown parameters across R/lmer, BUGs and HMC/NUTs program

Posterior median, 2.5% and 97.5% for the interventional effect for different proteins from both BUGs and HMC/NUTS program are listed in the appendix. Nineteen proteins are elevated and 17 proteins are suppressed post the coronary intervention shown in both the BUGs and NUTs results. Amongst the proteins with elevated abundance, the biggest fold change 4.5 is from Apolipoprotein A1 (*APOA1 Apoli*) protein, and 19 proteins have folder changes > 1.5. Among proteins with suppressed abundances, the biggest fold reduction of 0.70 is from A2M Alpha-2 protein, and 5 proteins have folder reduction > 0.50.

The Monte Carlo sampling error is smaller in the NUTs results. The point estimates of the *R/lmer* single model that are in the range of -2 and -4 are shown to shrink towards the ranges of 0 and -2, between which the grand mean lies (figure 5.2). There are more disagreements in the point estimates of the interventional effects from the results of NUTS than results from BUGS when compared to results of *R/lmer* (figure 5.3). The disagreements in the estimated interventional effects for proteins are plausibly due to the large proportion of censoring values in this case study. Figure 5.4(b) demonstrated the comparison between NUTS and BUGS result when the censored values were excluded, we observed a better agreement under this circumstance.

As shown in table 5.3, the cross proteins variances for intercept, m/z ratio and intervention from NUTS program are 0.34, 0.034 and 0.70 respectively; and these values are 0.05, 8.3e-5 and 0.03 from BUGS respectively. Compared to the variance components in model 3 of *R/lmer*, which are 0.09, 7.0e-07 and 0.52 respectively, NUTS has a similar intervention variance but a much bigger m/z ratio variance. BUGS has a similar protein intercept variance, a bigger m/z variance but smaller intervention variance.

The unexplained residuals variance in the NUTS and BUGS models are 1.62 and 0.37 respectively; both are smaller than 2.57 in the *R/lmer* multiple protein models. The

missing parameters for peptide abundance of BUGS are different from the NUTS estimates, but the missing parameters for m/z of these two different models are similar.

5.4. Discussion

The multiple proteins model which utilized the information across runs and labels for different proteins improved the accuracies for the estimates of intervention effect and slope of m/z ratio predicting the peptide intensity for different proteins, in particular for proteins with small number of observations. The multiple protein model also enable us to identify the systematic post-coronary intervention reduction in the intensities for all proteins; while gives us the prediction of the change for every protein. Adding the unknown parameters of the missing components using the Bayesian approach reduced the variance in the unexplained random errors and the across protein variance in the interventional effects, but it also introduces more variances in the coefficients of m/z ratio across proteins.

Table 5.3 The posterior parameters derived from *T* in the multivariate multilevel model from Gibbs sampling and *HMC/Non U Turn Sampling*

Gibbs:			NUTS:		
Posterior of the protein level variance components			Posterior of the protein level variance components		
Groups	Name	Variance(median(IQR))	Groups	Name	Variance(median(IQR))
Protein	(Intercept)	0.053 (0.081,0.034)	Protein	(Intercept)	0.34 (0.30, 0.35)
	centralized m/z	8.3e-5 (6.6e-5,1.1e-4)		centralized m/z	0.034 (0.031,0.037)
	intervention	0.032 (0.021,0.049)		intervention	0.70 (0.66, 0.75)
Residual		0.37 (0.36, 0.38)	Residual		1.62 (1.609,1.63)
Posterior distribution of the logistic regression coefficients for the missing model			Posterior distribution of the logistic regression coefficients for the missing model		
Groups	Name	Coefficient (median (2.5%-97.5%))	Groups	Name	Coefficient (median(IQR))
Missing	Intercept	-8.20 (-10.14,-7.25)	Missing	Intercept	-4.18 (-4.20, -4.07)
	m/z	0.0084 (0.0084, 0.0088)		m/z	0.0085 (0.00849, 0.0085)
	peptide abundance	-2.89 (-2.29,-1.97)		peptide abundance	-0.77 (-0.90,-0.75)

Figure 5.1 The associations between mass-to-charge ratios (m/z) and the relative intensities (on log scale) by different combinations of runs and labels

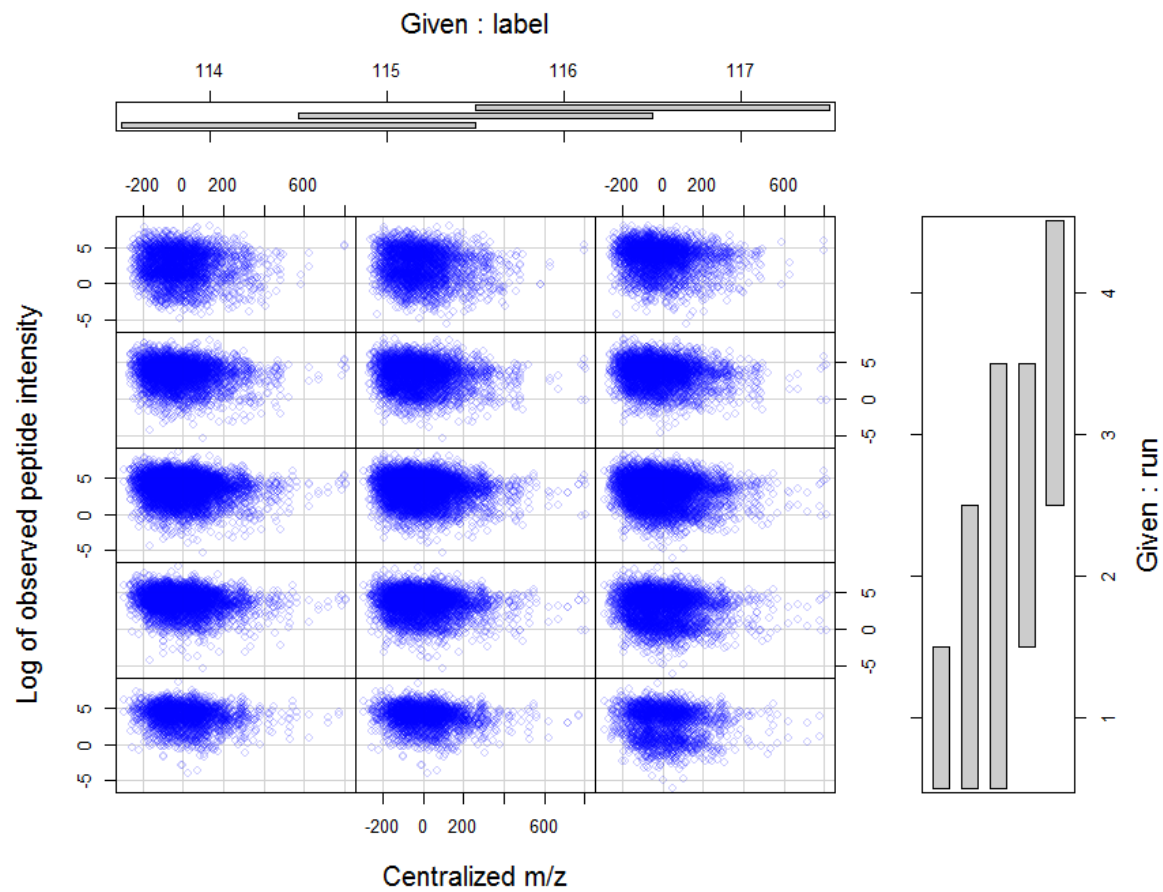


Figure 5.2 The comparison in the estimates of the interventional effect between single protein model and multiple proteins model

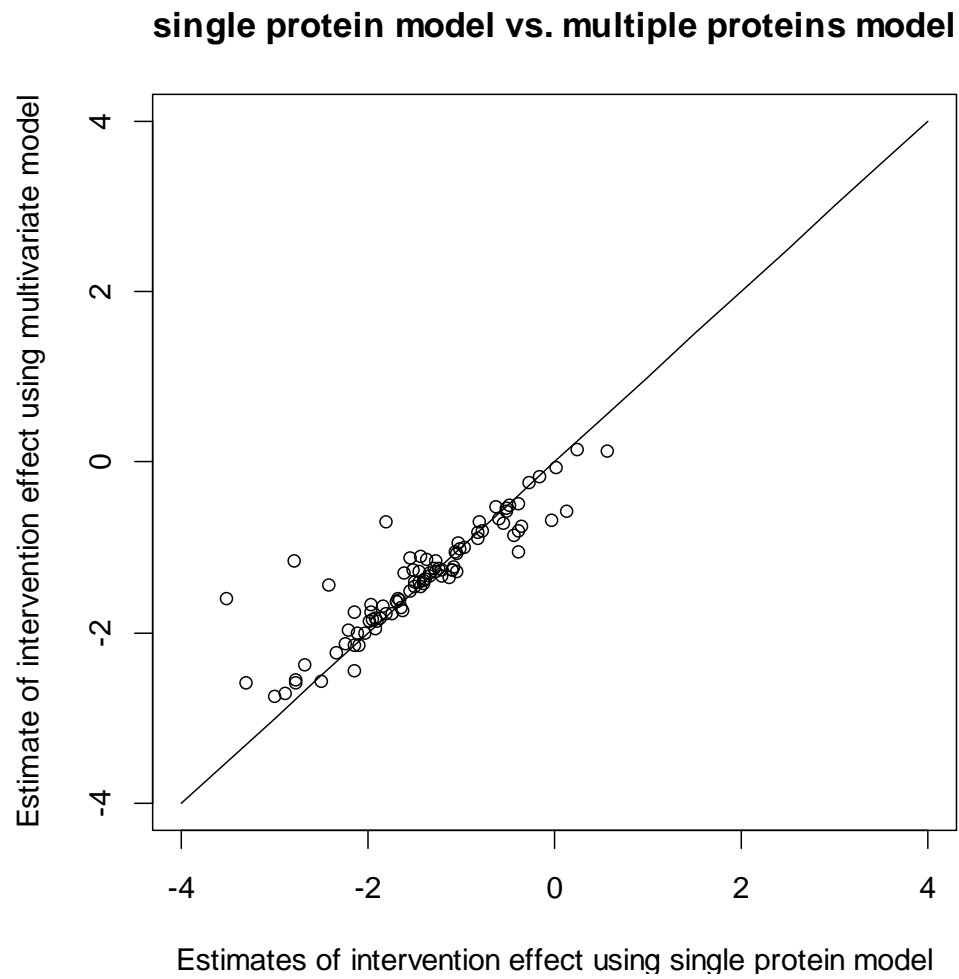


Figure 5.3 The comparison in the estimates of intervention effect between R/Imer, Gibbs and HMC/NUTs methods

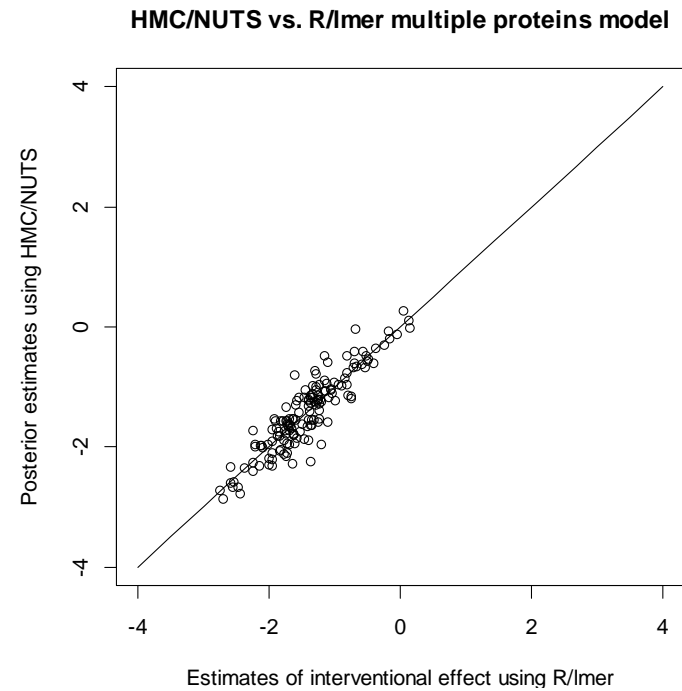
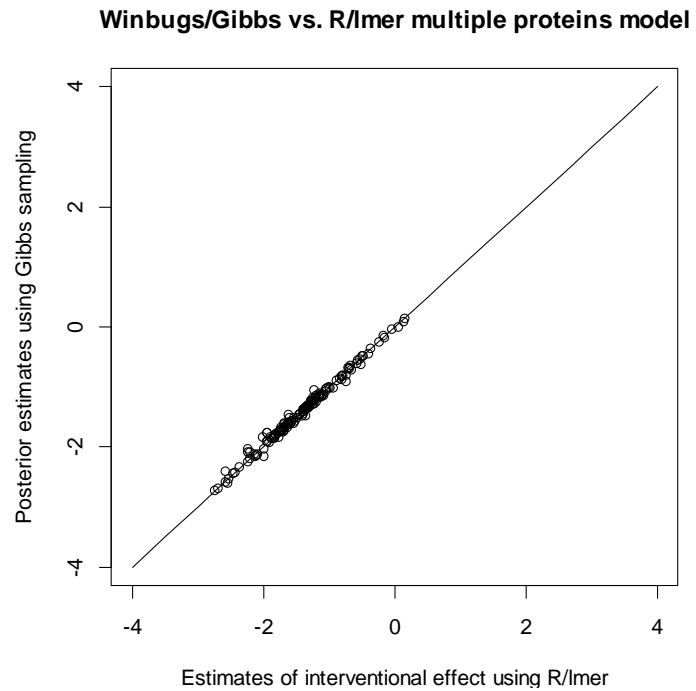
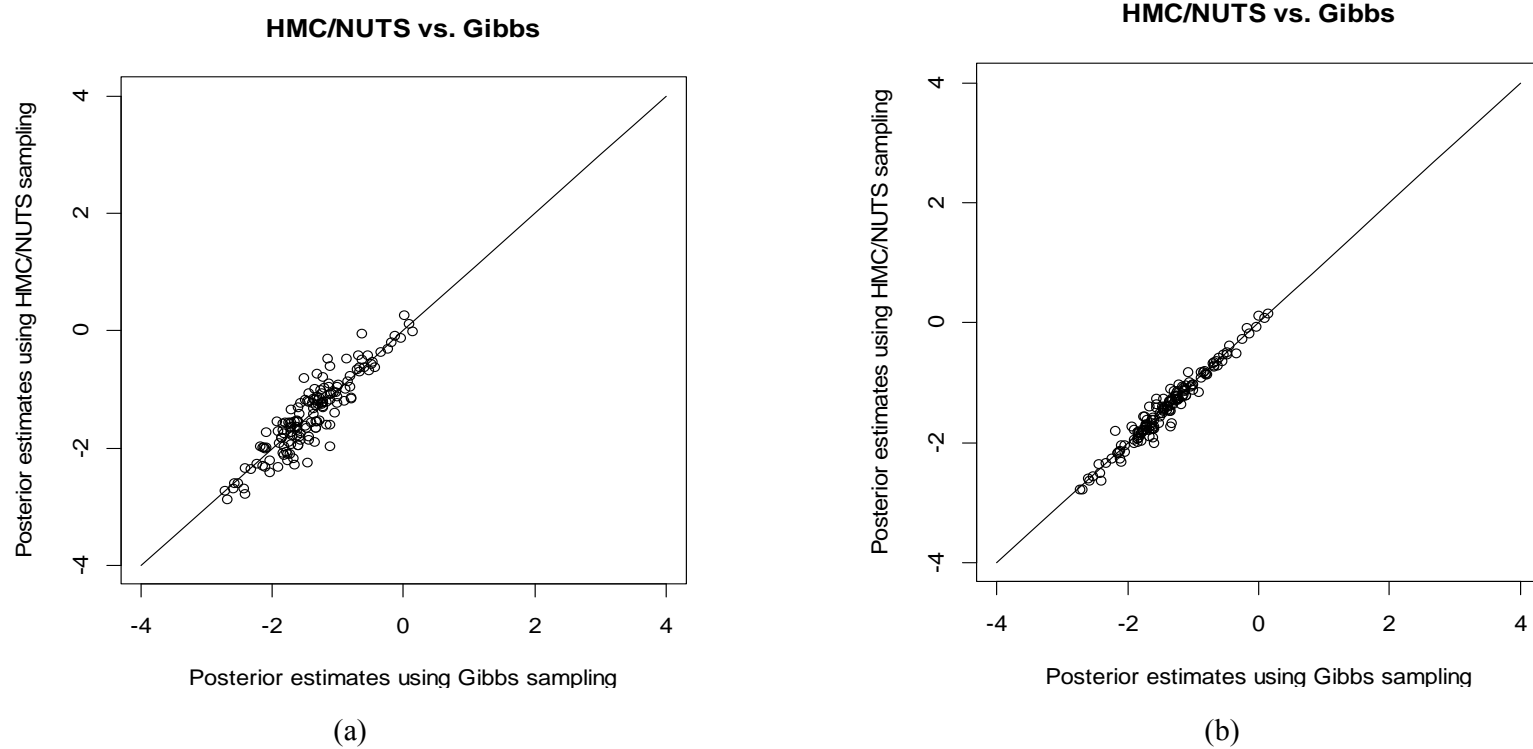


Figure 5.4 Comparison of intervention effect between HMC/NUTS and Gibbs methods
using all data and data without censored intensities

Figure Legends: (a) Comparison of estimates of the changes in intensities after the intervention from NUTS vs. BUGS using all data;
(b) Comparison of estimates of the changes in intensities after the intervention from NUTS vs. BUGS using data excluding the observations with censored intensity.



Protein	WinBUGS/Gibb Sampling Results				Rstan/N NUT Sampling results		
	12.5 percent ile	median	97.5 percentile	Direction of changes	12.5 percentile	median	97.5 percentile
- 104 kDa p	-0.875	-0.477	-0.085	downward	-0.422	-0.407	-0.352
- Hypothesi	0.127	0.605	1.081	upward	0.085	0.197	0.406
- Ig kappa	-1.611	-0.502	0.575		-0.467	0.629	0.677
- Lambda-ch	-0.199	0.426	1.049		-0.451	0.187	0.248
- Rheumatoi	-1.455	-0.293	0.855		-0.132	0.386	0.539
A1BG alpha	0.283	0.74	1.195	upward	0.404	0.475	0.749
A2M Alpha-2	-1.47	-1.177	-0.884	downward	-1.219	-1.154	-1.108
ACTG1 Actin	-1.308	-0.127	1.047		-0.562	-0.284	1.048
AFM Afamin	0.038	0.411	0.785	upward	0.312	0.328	0.462
AGT Angiote	-0.657	0.046	0.75		-0.064	0.528	0.685
AHSG Alpha-	0.534	0.941	1.349	upward	0.747	0.796	1.159
ALB Isoform	1.112	1.378	1.647	upward	0.893	1.013	1.219
AMBP AMBP p	0.175	0.747	1.326	upward	0.265	0.282	0.569
APOA1 Apoli	0.965	1.504	2.046	upward	0.646	0.709	1.327
APOA2 Apoli	-1.077	-0.298	0.473	/upward	0.223	0.303	0.572
APOA4 Apoli	0.286	0.609	0.933	upward	0.381	0.804	1.001
APOB Apolip	-0.701	-0.321	0.064	/downward	-0.451	-0.431	-0.403
APOC2;APOC4	-0.995	0.085	1.159		-0.364	0.429	0.795
APOC3 13 kD	-0.509	0.312	1.135		-0.434	-0.343	0.305
APOE Apolip	-2.114	-1.026	0.033	/downward	-1.517	-0.547	-0.033
APOH Beta-2	0.792	1.176	1.556	upward	0.957	1.24	1.37
APOM Apolip	-1.058	-0.023	1.023	/upward	0.02	0.287	0.973
ATRN 152 kD	-1.061	0.115	1.266		-0.303	0.132	0.949
AZGP1 alpha	-0.631	-0.038	0.554		-0.72	-0.048	0.812
B2M B2M pro	-0.507	0.344	1.211		-0.841	-0.179	0.093
BCHE Cholin	-1.291	-0.246	0.803		-0.546	-0.144	0.366
BTD biotini	-1.032	0.071	1.178		-0.855	-0.407	0.718
C1QB comple	-0.773	0.157	1.086		-0.484	-0.307	-0.155
C1QC Comple	-1.146	0.053	1.266		-1.302	-0.024	1.163
C1R Complem	-1.085	-0.41	0.255		-0.342	-0.15	0.228
C1RL Comple	-0.893	0.058	1.017		-0.065	-0.023	0.144
C1S Complem	-0.114	0.409	0.934		-0.009	0.192	0.315
C2 Compleme	-0.332	0.173	0.68	/upward	0.148	0.32	0.464
C3 Compleme	-0.008	0.188	0.385	/upward	0.174	0.291	0.299
C4A Complem	-1.914	-0.715	0.455		-0.521	-0.287	0.416
C4B complem	0.496	0.815	1.138	upward	0.012	0.609	0.667
C4BPA C4b-b	-0.783	-0.305	0.178		-0.626	-0.222	0.018
C4BPB Isofo	-1.019	-0.172	0.673		-0.97	-0.541	0.602
C5 Compleme	-1.277	-0.73	-0.177	downward	-0.882	-0.675	-0.618

Protein	WinBUGS/Gibb Sampling Results			Direction of changes	Rstan/NUT Sampling results		
	12.5 percentile	median	97.5 percentile		12.5 percentile	median	97.5 percentile
<i>C6 Compleme</i>	-0.231	0.186	0.606		-0.108	0.203	0.665
<i>C7 Compleme</i>	-0.793	-0.204	0.374		-0.399	-0.345	-0.063
<i>C8A 65 kDa</i>	-0.701	0.08	0.855	/upward	0.111	0.23	0.267
<i>C8B Complem</i>	-1.841	-1.017	-0.195	downward	-0.97	-0.945	-0.062
<i>C8G Complem</i>	-0.711	0.197	1.117	/upward	0.09	0.473	1.06
<i>C9 Compleme</i>	-0.981	-0.419	0.136	/downward	-0.298	-0.098	-0.028
<i>CA1 Carboni</i>	-1.173	0.065	1.324	/upward	0.489	0.575	0.713
<i>CD14 Monocy</i>	-1.22	-0.191	0.835		-0.56	0.005	0.392
<i>CD44 Isofor</i>	-0.308	0.77	1.87		-0.915	0.108	0.752
<i>CD5L CD5 an</i>	-0.603	0.168	0.934	/downward	-0.795	-0.465	-0.048
<i>CFB Isoform</i>	0.586	0.886	1.19	upward	0.413	0.661	0.942
<i>CFD complem</i>	-1.22	-0.147	0.923		-0.442	0.593	0.958
<i>CFH Isoform</i>	-0.096	0.194	0.482	/upward	0.04	0.131	0.763
<i>CFHR2 Isofo</i>	-1.247	-0.289	0.664		-0.324	-0.178	0.108
<i>CFI Complem</i>	0.06	0.567	1.081	upward	0.484	0.511	0.664
<i>CFP 50 kDa</i>	-0.566	0.245	1.067		-0.66	-0.637	0.175
<i>CLEC3B Hypo</i>	-1.31	-0.208	0.885		-0.016	0.14	1.23
<i>CLU 52 kDa</i>	-0.966	-0.148	0.67		-0.213	0.023	0.404
<i>CLU 54 kDa</i>	-0.417	0.412	1.248		-0.579	-0.23	0.317
<i>CNDP1 Beta-</i>	-1.321	0.005	1.339	/downward	-0.961	-0.525	-0.011
<i>CP Cerulopl</i>	-0.743	-0.437	-0.13	downward	-0.522	-0.366	-0.047
<i>CPB2 Isofor</i>	-1.001	-0.031	0.945		-0.197	0.631	0.902
<i>CPN1 Carbox</i>	-1.425	-0.19	1.003		-1.017	-0.803	0.538
<i>CPN2 simila</i>	-1.097	-0.301	0.488		-1.072	0.344	0.514
<i>CRP Isoform</i>	-1.469	-0.265	0.93		-0.246	0.005	0.798
<i>DBH dopamin</i>	-1.448	-0.191	1.063		-0.481	0.202	0.468
<i>ECM1 Extrac</i>	-0.3	0.549	1.403		0.14	0.163	0.194
<i>F10 Coagula</i>	-0.961	0.072	1.104		-0.842	0.158	0.476
<i>F11 Isoform</i>	-1.462	-0.445	0.556		-0.059	0.355	0.661
<i>F12 Coagula</i>	-0.047	0.544	1.14		-0.077	0.485	0.542
<i>F2 Prothrom</i>	-0.451	-0.089	0.269		-0.738	-0.357	0.123
<i>F5 Coagulat</i>	-1.157	-0.212	0.741		-1.002	0.183	0.301
<i>F9 Coagulat</i>	-1.621	-0.437	0.723	/upward	0.051	0.109	0.278
<i>FCN3 Isofor</i>	-1.375	-0.487	0.38	/upward	0.295	0.321	0.528
<i>FETUB 42 kD</i>	-1.323	-0.301	0.721	/downward	-0.761	-0.411	-0.302
<i>FGA Isoform</i>	0.65	0.931	1.21	upward	0.536	0.804	0.816
<i>FGB Fibrino</i>	-0.904	-0.618	-0.326	downward	-0.968	-0.626	-0.252

<i>Protein</i>	WinBUGS/Gibb Sampling Results				Rstan/NUT Sampling results		
	12.5 percentile	median	97.5 percentile	Direction of changes	12.5 percentile	median	97.5 percentile
<i>FGG Isoform</i>	-1.473	-1.109	-0.748	downward	-1.065	-0.869	-0.429
<i>FN1 Isoform</i>	-0.572	0.139	0.854		-0.187	-0.047	0.017
<i>FN1 fibrone</i>	-0.587	-0.183	0.211		-0.231	-0.049	0.055
<i>GC vitamin</i>	-0.293	0.311	0.918	/upward	0.264	0.33	0.361
<i>GPLD1 Isofo</i>	-1.634	-0.397	0.813		-0.664	0.622	0.886
<i>GPX3 Glutat</i>	-1.859	-0.714	0.413	/downward	-0.498	-0.456	-0.228
<i>GSN Isoform</i>	0.321	0.714	1.101	upward	0.023	0.667	0.821
<i>HABP2 Hyalu</i>	-1.715	-0.69	0.316	/downward	-0.599	-0.529	-0.012
<i>HBA1;HBA2 H</i>	-0.407	0.267	0.949		-0.328	0.867	1.008
<i>HBB Hemoglo</i>	-0.707	-0.063	0.577		-0.601	-0.274	0.661
<i>HGFAC Hepat</i>	-1.963	-0.758	0.389	/upward	0.21	0.288	0.331
<i>HP Haptoglo</i>	0.934	1.28	1.624	upward	0.122	1.048	1.215
<i>HPR Isoform</i>	-1.518	-0.251	1.007		-0.313	-0.016	0.374
<i>HPX Hemopex</i>	-0.53	-0.208	0.114		-0.288	-0.052	-0.04
<i>HRG Histidi</i>	-0.272	0.183	0.635		-0.066	-0.001	0.16
<i>IGF2 Isofor</i>	-0.712	0.301	1.315		-0.368	0.029	1
<i>IGFALS Insu</i>	-1.708	-1.006	-0.31	downward	-1.088	-0.53	-0.317
<i>IGFBP3 CDNA</i>	0.019	0.869	1.714	upward	-0.622	-0.18	0.276
<i>IGHA1 CDNA</i>	0.239	1.232	2.272	upward	-0.419	0.031	0.319
<i>IGHA1 Hypot</i>	-0.792	0.355	1.523		-0.169	-0.07	0.296
<i>IGHA1 IGHAI</i>	-0.08	0.739	1.58		-0.037	0.03	0.092
<i>IGHG2 Hypot</i>	-0.065	0.622	1.317	/upward	0.097	0.387	0.52
<i>IGHG3 IGHG3</i>	-0.63	0.291	1.217	/downward	-0.409	-0.363	-0.102
<i>IGHG4 IGHG4</i>	-0.404	0.616	1.649	/upward	0.522	0.53	0.715
<i>IGHM IGHM p</i>	-0.328	0.093	0.525		-0.423	-0.111	-0.013
<i>IGHV4-31 Hy</i>	-1.597	-0.392	0.782		-0.51	-0.214	0.053
<i>IGJ immunog</i>	-0.429	0.536	1.52		-0.319	0.272	0.423
<i>IGKC IGKC p</i>	0.188	1.072	1.965	upward/	-0.034	0.247	1.392
<i>IGKV3D-11 S</i>	-0.96	0.203	1.383		-1.144	0.557	0.65
<i>IGL@ IGL@ p</i>	-1.373	-0.348	0.677		-1.029	0.247	0.322
<i>ITIH1 Inter</i>	-1.186	-0.714	-0.252	downward	-0.7	-0.615	-0.337
<i>ITIH2 Inter</i>	-0.864	-0.438	-0.01	downward	-0.613	-0.401	0.205
<i>ITIH3 Inter</i>	-0.401	0.246	0.901	/downward	-0.303	-0.243	-0.084
<i>KLKB1 Plasm</i>	-1.031	-0.286	0.442	/upward	0.32	0.414	0.429
<i>KNG1 Isofor</i>	0.018	0.791	1.577	upward	-0.132	0.456	2.131
<i>KNG1 Kinino</i>	-0.445	0.057	0.555		-0.29	-0.101	0.325
<i>KRT1 Kerati</i>	-1.474	-0.145	1.153		-0.608	-0.24	0.913
<i>LCAT Phosph</i>	-1.402	-0.142	1.103	/upward	0.055	0.065	0.643
<i>LGALS3BP Ga</i>	-1.048	0.021	1.089		-0.748	-0.687	0.378
<i>LPA Lipopro</i>	0.095	0.971	1.875	upward	0.245	0.52	0.817
<i>LRG1 Leucin</i>	-0.92	-0.333	0.253	/downward	-0.54	-0.389	-0.352
<i>LUM Lumican</i>	-1.39	-0.699	-0.012	downward	-1.003	-0.071	-0.018
<i>MBL2 Mannos</i>	-1.403	-0.306	0.767		-0.883	-0.344	0.575
<i>MST1 Hepato</i>	-0.854	0.12	1.101		-0.252	0.063	0.287
<i>ORM1 Alpha-</i>	-0.489	0.306	1.112		0.149	0.26	0.5
<i>ORM2 Alpha-</i>	-0.572	0.279	1.131		-0.195	0.734	0.887

<i>Protein</i>	WinBUGS/Gibb Sampling Results			Direction of changes	Rstan/NUT Sampling results		
	12.5 percentile	median	97.5 percentile		12.5 percentile	median	97.5 percentile
<i>PGLYRP2 Iso</i>	-0.54	0.05	0.647		-0.431	0.367	0.778
<i>PI16 protea</i>	-0.871	0.175	1.219		-0.435	0.104	0.14
<i>PLG Plasmin</i>	-0.3	0.021	0.342		-0.035	0.073	0.472
<i>PON1 Serum</i>	-1.51	-0.422	0.644		-0.517	0.165	0.202
<i>PRG4 Isofor</i>	-1.315	-0.243	0.818		-0.85	-0.444	-0.258
<i>PROS1 Vitam</i>	-0.806	-0.217	0.37	downward	-0.49	-0.478	-0.182
<i>PTGDS Prost</i>	-1.254	-0.094	1.055		-0.248	-0.031	0.309
<i>RBP4 Retino</i>	-1.491	-0.74	0.004	/downward	-0.654	-0.108	-0.035
<i>SELL L-sele</i>	-0.817	0.273	1.367		-0.526	-0.322	0.394
<i>SEPP1 Selen</i>	-0.98	0.101	1.174		-0.21	-0.048	0.367
<i>SERPINA1 AI</i>	-1.002	-0.42	0.156	/downward	-0.682	-0.389	-0.083
<i>SERPINA10 P</i>	-1.297	-0.306	0.672		-0.642	-0.398	0.192
<i>SERPINA3 Is</i>	-1.796	-1.301	-0.804	downward	-1.344	-0.776	-0.355
<i>SERPINA4 Ka</i>	-1.664	-0.913	-0.17	Downward/	-0.716	0.269	1.096
<i>SERPINA5 PI</i>	-1.51	-0.388	0.719	/downward	-0.695	-0.59	-0.365
<i>SERPINA6 Co</i>	-1.978	-0.73	0.47		-0.112	0.133	0.461
<i>SERPINA7 Th</i>	-1.945	-1.169	-0.4	downward	-0.931	-0.634	0.126
<i>SERPINC1 An</i>	-0.002	0.387	0.779	/upward	0.231	0.256	0.3
<i>SERPIND1 He</i>	-2.065	-1.269	-0.488	downward	-0.8	-0.018	0.018
<i>SERPINF1 Pi</i>	-0.418	0.127	0.674		-0.328	0.114	0.257
<i>SERPINF2 AI</i>	-0.877	-0.37	0.135		-0.27	0.233	0.261
<i>SERPING1 PI</i>	-1.247	-0.819	-0.392	downward	-0.892	-0.731	-0.153
<i>SOD3 Extrac</i>	-1.182	-0.084	1.031		-1.021	-0.433	0.762
<i>TF Transfer</i>	-0.216	0.118	0.452		-0.011	0.339	0.486
<i>TTR Transth</i>	-0.754	-0.034	0.687		-0.324	-0.024	0.099
<i>VTN Vitrone</i>	-0.316	0.246	0.798		-0.01	0.196	0.215
<i>VWF 309 kDa</i>	-1.489	-0.498	0.47		-0.818	-0.321	0.208

*The mean difference on log scale has adjusted by adding the dilution factor of 1.4

CHAPTER 6

An immunology proteomic study-Case study II

6.1 Description of the study

Common Variable Immunodeficiency Disorder (CVID) is the most common primary immunodeficiency disorder encountered in clinical practice. CVID is also known as acquired hypogammaglobulinemia, where patients have low levels of immunoglobulin G, A and M. CVID patients are susceptible to recurrent infections. Currently, there is no cure for this disease. Patients are given frequent immunotherapy which consists of transfusing human antibodies harvested from donated plasma, to maintain a normal level of immunity (M. A. Park et al., 2008; J. H. Park et al., 2012).

According to the prevalence estimation from J. H. Park et al. (2012), 1:25,000 of the population suffer from this disorder. These patients have low titer of immunoglobulin and are usually prone to frequent infection. In 2008, an immunology proteomic CVID study was set up at LabPLUS by Drs Rohan Ameratunga and See-Tarn Woon. This project aims to identify potential cellular protein markers for differentiating CVID subgroups and predicting clinical phenotypes. CVID patients and healthy normal controls were planned to be recruited in the study so that the comparison of their protein profile of the lymphocyte cells can be made.

Participants had discussed participation in this study with their immunologist (Dr. Rohan Ameratunga, LabPLUS) prior to giving their consents. Once the consent was given, blood samples were obtained and transported to LabPLUS for processing and analysis.

6.2. The laboratory method (short summary of the clinical laboratory sample preparation and the iTRAQ experiment in the University lab)

6.2.1 *The clinical laboratory processing for biological samples*

Peripheral blood mononuclear cells (PBMC) from 17 CVID patients and 42 normal donors were isolated from whole blood and the whole cell lysate from PBMC were prepared. The PMBC were then incubated in lysis buffer (0.2% SurfactAmps, 50 mM

phosphate buffer, pH 7.0, 100 mM NaCl, 0.5 mM EDTA, protease inhibitor) on ice for 30 min, followed by pelleting the cell debris at 15,000 x g for 30 min. The haemoglobin presenting in the clarified cell lysate were removed by adding 50% Ni-NTA agarose (Life Technologies Invitrogen, Carlsbad, CA, USA) and the agarose was removed by low speed centrifugation. The sample preparation was performed by Dr. See-tarn Woon at LabPLUS.

6.2.2 The MS/MS iTRAQ experiments for blood lymphocytes tissue samples

The processed cell lysate tissue samples were transformed to the Center for Proteomics and Genomics, University of Auckland for proteomic analysis. The proteins were dissolved and digested by trypsin and mixed with iTRAQ reagents for the preparation of the 8-plex (CVID case) MS/MS assays. The most abundant serum proteins were depleted with IgY-12 SC Spin Column (Beckman) with one salt-step for the fractionation.

The labeled samples were combined and fractionated by nanoLC and analyzed by tandem mass spectrometry. The observed labeled peptide results were used for matching with the protein database for the identification of their corresponding proteins. In the reverse database search, unused scores greater than 0 to 0.47 were used for protein quantitation.

6.2.3 Clinical study design and experimental design

This is the first time that lymphocyte tissues are analyzed systematically through LC-MS/MS at the university lab. A reproducibility assessment is considered to be essential before conducting the mass spectrometry analysis for the discovery. The clinical proteomic study thus has two sections: 1) reproducibility assessment; 2) protein marker discovery. There were patient recruitment difficulties in the study, as a result, the total number of patients and controls deviated slightly from the original plan; 17 CVID patients and 42 normal controls from Auckland clinical centers participated in the study. A frequency matching scheme was assigned to make sure the normal controls were well matched with CVID patients in the proportion of gender and ethnicity and were restricted within the same age band. The ratio of patient-to-control was set to 2:1. The gender and ethnicity distribution of the cases were updated periodically for the matching.

In the reproducibility section, blood samples of the first 4 patients and 4 normal controls were analyzed four times within a week for the reproducibility evaluation in the proteomics analysis.

Row and column design is used for the LC-MS/MS reproducibility evaluation of which two Latin squares were used for the 4 runs (row) x 8 plex (column) assay layouts for the reproducibility assessment (Table 6.1). This design achieved the orthogonality between label and the participant's class (patient vs. control). Four patients and four matched controls were selected and four samples of their blood proteins were analyzed in the four mass spectrometry runs. Each run contains one blood proteins sample for the replicated biological samples (P4, P6, P9, P17, N10, N25, N34 and N41).

Table 6.1 The experimental design layout for the reproducibility assessment

Run	113	114	115	116	117	118	119	121
1	P4	N10	P9	N25	P6	N34	P17	N41
2	N25	P4	N10	P9	N41	P6	N34	P17
3	P9	N25	P4	N10	P17	N41	P6	N34
4	N10	P9	N25	P4	N34	P17	N41	P6

- Patients P4, P6, P9, and P17 were selected as the cases for the evaluations.
- Normal controls N10, N25, N34, and N41 were selected as the matched control for the evaluations.

The discovery section is a case-control study. For the LC/MS-MS analysis, 13 CVID patients and 37 normal controls were allocated in a 7 runs (row) x 8 plex (column) assay layout. An unbalanced row and column layout was used instead of a balanced one because of the unbalanced recruiting; the layout of the experimental design (Table 6.2) aimed to achieve orthogonality for the label, run in respect to the participant's class (i.e. patients vs. controls).

Table 6.2 The experimental design layout for the discovery section

Run	113	114	115	116	117	118	119	121
1	N1	P1	N2	N3	P2	N4	N5	P3
2	N6	N7	P5	N8	N9	P7	N11	N12
3	P13	N13	N14	P8	N15	N16	P10	N17
4	N18	P14	N19	N20	P11	N21	N22	P12
5	N23	N41	P15	N26	N27	P16	N28	N29
6	P8	N30	N31	P5	N32	N33	P16	N35
7	N36	N37	N38	N39	N40	P2	P11	P10

- Patients : P1-P17 were selected as the cases for the discovery analysis, excluding patient cases P4,P6,P9, and P17 who were included in the reproducibility evaluations;
- Normal controls: N1-42 were selected as the normal controls for the discovery analysis, excluding normal controls N10, N25, N34, and N41 who were included in the reproducibility and N24 who is missing.

6.3. The reproducibility assessment

Permutation based assessment for the reproducibility as described in Chapter two was used for the COVID study. Since 8 biological samples are allocated by 8-plex in each run and repeatedly analyzed for 4 runs, the permutation test assesses the results from each run against the results from the averages of all 4 runs. The assessment identified two plasma contaminated lymphocytes tissues (Figure 6.1 A). Excluding these two samples (P6 and N43) and run 2 which had a bad yield, the final reproducibility assessment used the three biological replicates from the 3 runs. The permutation tests results for $Tmax$ statistic and rank test statistic are demonstrated in Table 6.3. As described in Chapter 2, the $Tmax$ statistic is the parametric statistics proposed for the reproducibility assessment, it is the maximum t statistic ($\text{Max}_{1 \leq i < m} T_i$) of the m paired differences of the principal components (PC). Set $T_i = \frac{(\mu_i - 0)}{std_i}$, where μ_i is the mean difference and std_i is the standard deviation of the difference in the i^{th} PC. The non-parametric statistic is a two-

dimensional sign score equivalent to $\log\left(\frac{P_+}{P_-}\right)$, where P_+ is the total number of positive differences and P_- is the total number of negative differences in m principal components of n samples. For three runs of the experiments that achieved a good discovery, the permutation tests did not detect significant disagreements between the repeated tests from all proteins. The first principal component and second component graphs also show that there are no systematic patterns between the repeated analyses for the same biological subject (FPC plot in figure 6.1B and SPC plot 6.1C); and there is no linear or polynomial trends in the differences between replicates and the averaged results (y axis) against the averaged principal component scores (x axis). These results indicate that the analysis results are reproducible so long as the lymphocyte samples are without contamination from the plasmas cells and the experiment has optimal discovery.

Table 6.3 The reproducibility permutation test for the CVID proteomic studies: results from each replicated run vs. the average of three runs

	Permutation test p values for Tmax	Permutation test p values for Rank test statistic
Run 1:	0.46	0.58
Run 3:	0.15	0.39
Run 4:	0.21	>0.90

6.4. The analytical methods for the discovery section

6.4.1 A single protein multilevel model

A single protein multilevel model with subject as a random effect has set up similarly as defined in equation (3) in Chapter 4. In this single protein model, the peptide intensity is the response. The explanatory variables include peptide level variables (label, m/z, and run) and subject level variables (total amount of protein, subject class [normal vs. control], age and gender).

$$y_{i,l} = b_{0,l} + \beta_1 mz_{i,l} + \sum_{h=1}^8 \beta_{h,2} label_{h,i,l} + \sum_{r=1}^7 \beta_{r,3} run_{r,i,l} + e_{i,l}, \quad (1)$$

Equation (1) defines the level 1 of the model where,

$y_{i,l}$ denotes the peptide intensity for peptide i and subject l , i ranges between 1 and the total number of peptides observed for the protein being analyzed, and l ranges between 1 and the total number of subjects;

$mz_{i,l}$ denotes the centralized m/z ratio for the peptide i , and β_1 is the regression coefficient for m/z ratio;

$label_{h,i,l}$ and $run_{r,i,l}$ denote the label h and run r respectively, $\beta_{h,2}$ and $\beta_{r,3}$ are their regression coefficients respectively;

$b_{0,l}$ denotes the intercept for the subject l , it varies across different subjects.

Level 2 of the model defines the relation between the subject intercept $b_{0,l}$ and the explanatory variables

$$b_{0,l} = \gamma_{0,0} + \gamma_1 \times class_{i,l} + \gamma_2 \times total_protein_{i,l} + \gamma_3 \times age_{i,l} + \gamma_4 \times gender_{i,l} + u_{0,l}, \quad (2)$$

where

$\gamma_{0,0}$ denotes the grand intercept;

$class_{i,l}$ is a binary variable indicates if the subject belongs to the patient group or the normal control group, γ_1 denotes the regression coefficient for class;

$total_protein_{i,l}$ denotes the total amount of protein from the sample of subject l , γ_2 denotes the regression coefficient for the total amount of protein;

$age_{i,l}$ denotes the age of subject l , γ_3 is the regression coefficient of the age;

$gender_{i,l}$ denotes the gender of subject l , γ_4 denotes the regression coefficient of the gender;

$u_{0,l}$ denotes the subject level random residual term.

Substituting (2) to (1) gives us

$$y_{i,l} = \left[\begin{aligned} &\gamma_{0,0} + \gamma_1 \times class_{i,l} + \gamma_2 \times total_protein_{i,l} + \gamma_3 \times age_{i,l} \\ &+ \gamma_4 \times gender_{i,l} + \beta_1 mz_{i,l} + \sum_{h=1}^8 \beta_{h,2} label_{h,i,l} + \sum_{r=1}^7 \beta_{r,3} run_{r,i,l} \end{aligned} \right] + [u_{0,l} + e_{i,l}]$$

, where β_1, \dots, β_3 are the fixed effects coefficients for the experimental factors run and label, and $\gamma_1, \dots, \gamma_3$ are the fixed effect for the total amount of proteins, age and gender respectively; $u_{0,l}, e_{i,l}$ are the random effects for subjects and unexplained errors respectively.

6.4.2 Multiple proteins multivariate model with random effects at protein and subject levels ignoring missing

Under the multivariate multilevel framework as defined in (8) of Chapter 4, hierarchical multivariate models were set up to analyze the functional group of proteins, i.e. immunity group. This model includes fixed effects at the peptide level such as run, label, and m/z; and fixed effects at the subject level such as age and gender. Protein level factors, which include the intercept, slopes of m/z and the class effect, are modeled as random effects assumed to vary across proteins. A random intercept is also included at the subject level assumed that there are potential variations in the intensities introduced by subjects. The multilevel multivariate model is defined as follows,

$$\begin{aligned} \gamma_{i,l,p} = & \phi_{0,l} + \beta_{0,p} \text{protein}_{i,l,p} + \sum_{h=1}^8 \beta_{h,2} \text{label}_{h,i,l,p} + \sum_{r=1}^7 \beta_{r,3} \text{run}_{r,i,l,p} \\ & + \beta_{3,p} \text{class}_{i,l,p} + \beta_{4,p} \text{mz}_{i,l,p} + e_{i,l,p} \end{aligned} \quad (3)$$

$$\begin{aligned} \beta_{0,p} &= \alpha_0 + b_{0,p} \\ \beta_{3,p} &= \alpha_1 + b_{1,p} , \\ \beta_{4,p} &= \alpha_2 + b_{2,p} \end{aligned} \quad (4)$$

$$\phi_{0,l} = \gamma_{0,0} + \gamma_2 \times \text{total_protein}_{i,l,p} + \gamma_3 \times \text{age}_{i,l,p} + \gamma_4 \times \text{gender}_{i,l,p} + c_{0,l}, \quad (5)$$

where $\beta_{h,2}$ and $\beta_{r,3}$ are the fixed effects coefficients, same for all proteins; $\beta_{0,p}$, $\beta_{3,p}$ and $\beta_{4,p}$ are the random effect coefficients for proteins; $\gamma_{0,0}$, $\gamma_2, \dots, \gamma_4$ are the fixed effects coefficients for the grand intercept, total proteins, age and gender respectively. The random residuals terms include the random error residual term $e_{i,l,p}$, the subject level residual $c_{0,l}$, and the three protein level residual terms $b_{0,p}$, $b_{1,p}$ and $b_{2,p}$ for protein intercept, m/z and class differences respectively.

Substituting (4)-(5) into (3) results in the following mixed effect model

$$\begin{aligned} \gamma_{i,l,p} = & \left[\gamma_{0,0} + \alpha_0 + \alpha_2 \text{mz}_{i,l,p} + \sum_{h=1}^8 \beta_{h,2} \text{label}_{h,i,l,p} + \sum_{r=1}^7 \beta_{r,3} \text{run}_{r,i,l,p} + \alpha_1 \text{class}_{i,l,p} + \right. \\ & \left. \gamma_2 \times \text{total_protein}_{i,l,p} + \gamma_3 \times \text{age}_{i,l,p} + \gamma_4 \times \text{gender}_{i,l,p} \right] \\ & + \left[b_{2,p} \times \text{mz}_{i,l,p} + b_{1,p} \times \text{class}_{i,l,p} + b_{0,p} \times \text{protein}_{i,l,p} + c_{0,l} + e_{i,l,p} \right] \end{aligned} \quad (6)$$

where

$$(b_{0,p}, b_{1,p}, b_{2,p}) \sim MVN(\mathbf{0}, \mathbf{\Phi}); c_{0,l} \sim N(0, \tau_0^2); e_{i,l,p} \sim N(0, \sigma^2);$$

$$\mathbf{\Phi} = \begin{bmatrix} \sigma_{b,0}^2 & \sigma_{b,0}\sigma_{b,1} & \sigma_{b,0}\sigma_{b,2} \\ \sigma_{b,0}\sigma_{b,1} & \sigma_{b,1}^2 & \sigma_{b,1}\sigma_{b,2} \\ \sigma_{b,0}\sigma_{b,2} & \sigma_{b,1}\sigma_{b,2} & \sigma_{b,2}^2 \end{bmatrix}.$$

$(b_{0,p}, b_{1,p}, b_{2,p})$ is multivariate normal distributed with variance-covariance matrix $\mathbf{\Phi}$, subject level residual $c_{0,l}$ is normal distributed with variance τ_0^2 , error residual $e_{i,l,p}$ is normal distributed with variance σ^2 .

Equation (6) defines a generic model for this case study; five special cases of equation (6) including different combinations of fixed effects are compared and summarized in Table 6.4.

6.4.3 A multiple protein multivariate model with random effects at protein and subject levels and with the missing mechanisms modeled by Bayesian approach.

In the Bayesian model, the same linear multilevel mixed regression model as defined in (6), which includes independents of protein level parameters, peptide level parameters and subject level parameters, are used to predict the response-the logarithmic peptide intensities. Logistic regression was used to model the probability of missing as a function of m/z and peptide intensity (predicted and observed).

The Bayesian model

$$\begin{aligned} \mu_{i,l,p} = & \phi_{0,l} subject_{i,l,p} + \beta_{0,p} protein_{i,l,p} + \beta_{1,p} protein_{i,l,p} \times mz_{i,l,p} \\ & + \beta_{4,p} protein_{i,l,p} \times class_{i,l,p} + \sum_{h=1}^8 \beta_{h,2} label_{h,i,l,p} + \sum_{r=1}^7 \beta_{r,3} run_{r,i,l,p} + \\ & \gamma_2 \times total_protein_{i,l,p} + \gamma_3 \times age_{i,l,p} + \gamma_4 \times gender_{i,l,p} + e_{i,l,p} \end{aligned} \quad (7)$$

$$\text{logit}(pm_{i,l,p} = \Pr(\text{missed } \gamma_{i,l,p})) = \alpha_0 + \alpha_1 m z_{i,l,p} + \alpha_2 \gamma_{i,l,p}, \quad (8)$$

where

the equation (7) defines the relation between the mean intensities $\mu_{i,l,p}$ and the independents; $\phi_{0,l}$ denotes the subject intercept varied across different subjects;

$\beta_{0,p}, \beta_{1,p}, \beta_{4,p}$ are the protein level regression coefficients for the intercept, m/z slope and class respectively that vary across different proteins;

The label and run regression coefficients $\beta_{h,2}$ and $\beta_{r,3}$ are constant across all the proteins, and the grand intercept, total protein, age and gender coefficients $\gamma_{0,0}, \gamma_2, \dots, \gamma_4$ are constant across all the proteins;

$e_{i,l,p}$ denotes the residual error term.

Equation (8) defines the relation between the probability of missing $pm_{i,l,p}$ and the independents; $\alpha_0, \alpha_1, \alpha_2$ are the logistic regression coefficients for the intercept, m/z and abundance respectively.

Definition for the prior distributions:

The observational data fitted by the model described in (6) showed that most of the proteins do not demonstrate any differences in their abundances between normal and COVID patients. That is, the $\beta_{4,p}$ estimated from the observations are mostly zero.

Two different distributions were thus assigned to the priors of the class coefficients $\beta_{4,p}$ for the evaluation. One is the normal distribution; the other one is the double exponential distribution.

1) Normal prior for $\beta_{4,p}$ and multivariate normal prior for $(\beta_{0,p}, \beta_{1,p})$

$$\begin{aligned}\beta_{4,p} &\sim N(0, \eta^2), \eta \sim \text{invGamma}(1, 1); \\ (\beta_{0,p}, \beta_{1,p}) &\sim \text{MVN}(\gamma, T); \\ \gamma &\sim \text{MVN}(\mu, \Omega^{-1}), T \sim \text{invWISHART}(\Phi^{-1}, 2); \\ \mu &= (0, 0), \\ \Omega &= \begin{bmatrix} 0.01 & 0.001 \\ 0.001 & 0.01 \end{bmatrix}, \Phi = \begin{bmatrix} 0.01 & 0.01 \\ 0.01 & 0.01 \end{bmatrix};\end{aligned}$$

$$\begin{aligned}\alpha_0 &\sim N(0, 1), \alpha_1 \sim N(0.0085, 4e-8), \alpha_2 \sim N(-0.45, 0.25); \\ e_{i,l,p} &\sim N(0, \sigma^2), \sigma \sim \text{invGamma}(1, 1),\end{aligned}$$

where $\beta_{4,p}$ uses a non-informative normal distributed prior with inversed Gamma distributed hyper prior η , the couple protein level parameters of intercept and m/z $(\beta_{0,p}, \beta_{1,p})$ use non-informative multivariate distributed prior with a pair of hyper priors (γ, T) that is multivariate and inverse Wishart distributed respectively.

As described in Chapter 4, the probability of missing is modeled for the censoring and completely missing using logistic regression. The coefficients of logistic regression $\alpha_0, \dots, \alpha_2$ are part of the joint unknown parameters, and they are given normal distributed informative priors learning from the cardiac case. The missing values of the peptide quantities are also treated as unknown parameters in the Bayesian models of which are sampled from the conditional posterior distribution.

The scale parameter for the residual term $e_{i,l,p}$ uses an inversed gamma distribution.

2) Double exponential (DE) priors for $\beta_{4,p}$ and multivariate normal prior for $(\beta_{0,p}, \beta_{1,p})$

$$\beta_{4,p} \sim \text{doubleExp}(0, \eta), \eta \sim \text{invGamma}(1, 1);$$

The second prior for $\beta_{4,p}$ is a double exponential distribution with location parameter 0 and an inversed gamma distributed scale parameter η , other priors are kept the same as defined in 1).

Computing programs:

In the Gibbs sampling program, the probability for having censored intensity is fixed to be 1; In the HMC/NUTs program, the probability for having censored intensity are estimated from the normal cumulative probability density between the negative infinity and the known lowest detectable limit.

The Rstan HMC/NUTS program uses the same priors as the BUGS program; except for the scaled matrix Φ of the inversed Wishart distribution where 0.1 is set as the diagonal value and 0.05 as the off-diagonal value. Both the BUGS and the Rstan programs utilize the m/z ratio information from the censored and missing peptide observations for deriving the posterior estimates of the unknown parameters.

The codes for the BUGs and HMC models are listed in the appendix.

6.5. Results

6.5.1 Analysis of proteins in the immunity group

The immunity function group classified by Dr. See-Tarn Woon has 35 proteins consisting of 6000 peptide observations of which 56 (0.9%) had censored and 174 (2.9%) had completely missing values in the peptide intensities. Results from multiple protein models and the Bayesian models with missing data parameters included are discussed in the following sections 5.1.1-5.1.3.

6.5.1.1 Results from the multiple protein models (Table 6.4)

Comparing the AIC, BIC, and REML deviance across different models as demonstrated in Table 6.4, model 2 is a simple model with the smallest AIC, the second smallest BIC and REML deviance. Model 2 includes fixed effects of the run, label, gender, age, total amount of protein and m/z; it includes random effects of protein intercept, m/z, class, and subject intercept. Model 3 is similar to model 2 but includes class as an additional fixed effect; the class effect -0.28 (std error: 0.23) is not shown to be significant. The non-significant result for the fixed class effect indicates that there is no systematic difference in the protein abundances between normal and controls among all discovered proteins. Model 2 is selected as the final model for the interpretation in the following sections.

In model 2, centralized m/z is the only fixed effect significant with an estimate of -3.6e-03 (std error: 0.001; t statistic:-3.4). The other fixed effects such as label, run, m/z, age and gender are not significant. The Best Linear Unbiased Predicted values of random class effects of the 35 proteins are ranged between -0.12 and 0.28, compared to the range of -0.64 and 1.11 from single protein models. The multiple protein model demonstrated that eight proteins are irregularly regulated; they are PSME1, RPS6 40S, SAMHD1, MPO, S100A9, CAP1 ADERYL, PD1A3, and PSME2 with class differences in the logarithmic intensities -0.85, 0.15, -0.12, -0.07, -0.07, -0.09, 0.28, and -0.12 respectively (Figure 6.4); These are equivalent to a folder change of 0.43, 1.16, 0.89, 0.93, 0.93, 0.91, 1.32, and 0.89 respectively.

The variance for the protein intercept is 5.26 and the variance for the residual is 1.17. These values are used to derive the ICC for the protein intercept of 0.82. The variance for the class difference is 0.02, and, similarly, the ICC for protein class difference can be derived as 0.017.

The variance of m/z slope across proteins is 3.38e-05. Including m/z ratio as a random effect achieves a better fit as shown in models 2 and 3. ICC for m/z slope across proteins is 2.9e-05. The variance for the subject intercept is 2.21 and leads to an ICC of 0.65.

In the model 2 version excluding fixed effect age and gender, the variance components for protein intercept, m/z slope and class are 1.26, 2.37e-5 and 0.01 respectively. The variance for subject intercept is 0.35 and the residual variance is 1.19.

6.5.1.2 Results from the Bayesian model -Mean effects $\beta_{4,p}$

The NUTS program gives the medians of the posterior class difference $\beta_{4,p}$ between -0.02 and 0.04 when a normal distributed prior is used; and gives the medians of the posterior $\beta_{4,p}$ between -0.06 and 0.03 when a DE prior is used. None of the proteins was shown to be irregularly regulated in the NUTS results.

The Bugs program gives the median of the posterior $\beta_{4,p}$ between -0.27 and 0.14 when a normal distributed prior is used; and gives the medians of the posterior $\beta_{4,p}$ between -0.10 and 0.08 when a DE prior is used. None of the protein was shown to be irregularly regulated in the BUGS result too, except that DEFA1 and MYH9 have a plausible trend of down regulation when a narrower credible interval is used.

6.5.1.3 Result comparison in variance components

Inter proteins variances – η^2 of $\beta_{4,p}$ the class difference

As defined in 2) of the prior section, when the double exponential (DE) distributed prior is given to $\beta_{4,p}$ and the hyper prior η is inverse Gamma distributed, the posterior of the variances η^2 has median 0.13 (95% credible interval: 0.08, 0.22) from the BUGS program. It has median 0.64 (95% credible interval: 6.8e-04, 1.90) from the HMC/NUTS program (Figure 6.4a). When the normal prior is given to $\beta_{4,p}$ as defined in 1) of the prior section, the posterior of variance has median 0.12 (95% credible interval: 0.07, 0.21) from the BUGS program, and has median 0.41 (95% credible interval: 0.07, 1.63) from the HMC/NUTS program (Figure 6.4b). Using these two different priors of $\beta_{4,p}$ produced similar results in the variance components from either the BUGS or the NUTS program. NUTS produced greater variance of the class difference across proteins.

The inter proteins variance of $\beta_{4,p}$ is 0.02 from the R/lmer model and are much smaller compared to the results from the Bayesian models.

Inter proteins variances-slope of mass to charge ratio (m/z)

Figure 6.2 demonstrated that the m/z has a negative association with the peptide intensity, and the associations represented by the slopes in Figure 6.3b are similar across **runs**. Although the ICC indicated that the variation in slope across **proteins** is very small compared to the unexplained error variance, Figure 6.3a demonstrates that the slopes between intensity and m/z vary across different **proteins**. In particular, those proteins with small numbers of peptides will not have a reliable estimate for the slope of m/z if they are analyzed separately; it will be an advantage to include groups of proteins in one model.

When double exponential prior for $\beta_{4,p}$ is used, BUGS results in an inter protein variance 0.003 (95% credible interval: 0.002, 0.005) for the slope of m/z; NUTS results in a variance of 0.85 (95% credible interval: 0.49, 1.48) for the slope of m/z. Using BUGS has a larger variation in the m/z slope across proteins.

Inter proteins variances-intercepts for proteins

In the BUGS program with double exponential prior for $\beta_{4,p}$, the variance for the protein intercept is 0.52 (95% credible interval: 0.32, 0.92), with normal prior results in a similar variance of 0.52 (95% credible interval: 0.32, 0.91). In the NUTS program, the variance of the protein intercept term is 26.5 (95% credible interval: 13.0, 48.6) and 24.9 (95%

credible interval: 13.2, 44.8) for using the double exponential and normal prior of $\beta_{4,p}$ respectively. Compared to the R/lmer REML(MODE) method, NUTS has a greater variance. Both BUGS and NUTS show that, the posterior estimates of the inter protein variances of intercepts are not dissimilar when using two different priors for $\beta_{4,p}$.

Table 6.4 The comparison of different R: *lmre* Models including 35 immunity proteins

<i>Models:</i> <i>Fixed effect</i>	<i>Models:</i> <i>Random effects</i>	<i>AIC</i>	<i>BIC</i>	<i>REML deviance</i>
<i>Run, label, gender, age, total amount of protein</i>	<i>Protein intercept, m/z, class; Subject intercept</i>	<i>17834</i>	<i>18001</i>	<i>17784</i>
<i>Run, label, gender, age, total amount of protein, m/z</i>	<i>Protein intercept, m/z, class; Subject intercept</i>	<i>17833</i>	<i>18006</i>	<i>17781</i>
<i>Run, label, gender, age, total amount of protein, m/z, class</i>	<i>Protein intercept, m/z, class; Subject intercept</i>	<i>17834</i>	<i>18014</i>	<i>17780</i>
<i>Run, label, gender, age, total amount of protein, m/z</i>	<i>Protein intercept, class; Subject intercept</i>	<i>18180</i>	<i>18333</i>	<i>18134</i>
<i>Run, label, m/z, gender, age, total amount of protein, m/z, protein</i>	<i>Protein level intercept with with class; subject intercept</i>	<i>18174</i>	<i>18554</i>	<i>18060</i>

**Table 6.5 The selected final model for proteins in immunity group
(listed as the second model in Table 6.4- nlme version 3.1-108)**

Linear mixed model fit by REML ['lmerMod']				
Random effects:				
Groups	Name	Variance	Std. Dev.	Corr.
Subject	(Intercept)	2.21	1.49	
Protein	(Intercept)	5.26	2.29	
	centralized m/z	3.38e-05	0.0058	0.66
	factor(class)P	0.02	0.0063	0.88 0.76
	Residual	1.17	1.08	
Number of obs: 5770, groups: subject, 49; protein, 35				
Fixed effects:				
	Estimate	Std. Error	t value	
(Intercept)	3.98	1.86	2.14	
Log of total protein	-1.74e-02	0.51	-0.03	
factor(run)2	0.19	0.64	0.30	
factor(run)3	0.39	0.78	0.50	
factor(run)4	9.60e-02	0.72	0.13	
factor(run)5	-0.10	0.66	-0.14	
factor(run)6	-0.35	0.63	-0.56	
factor(run)7	0.16	0.65	0.25	
factor(tlable)114	-0.24	0.77	-0.32	
factor(tlable)115	-0.08	0.71	-0.11	
factor(tlable)116	0.25	0.48	0.52	
factor(tlable)117	-0.24	0.72	-0.34	
factor(tlable)118	-0.11	0.74	-0.15	
factor(tlable)119	0.32	0.68	0.47	
factor(tlable)121	-0.40	0.69	-0.58	
factor(gender)M	0.31	0.54	0.59	
age	-1.78e-04	0.02	-0.01	
centralized m/z	-3.64e-03	0.001	-3.43	
Random components in model 2 excluding age and gender				
Random effects:				
Groups	Name	Variance	Std.Dev.	Corr
subject	(Intercept)	3.48e-01	0.59	
protein	(Intercept)	1.26e+00	1.12	
	prec_m_z_center	2.37e-05	0.005	0.34
	factor(class)P	9.27e-03	0.096	0.02 0.88
	Residual	1.19	1.088	
Number of obs: 5770, groups: subject, 49; protein, 35				

6.5.1.3 Missing data modeling parameters

The posterior estimates of missing data parameters have overlapping 95% credible intervals from BUGS compared to HMC/NUTS program. Using BUGS, the regression coefficients for intercept α_0 , m/z α_1 and abundances α_2 are -4.6, 0.0084, and -2.5 respectively. Using NUTS, the posterior estimates for regression coefficients $\alpha_0, \alpha_1, \alpha_2$ are -6.5, 0.0085, and -1.5 respectively (Table 6.6a). In both programs, the different prior for $\beta_{4,p}$ does not result in different posteriors for $\alpha_0, \alpha_1, \alpha_2$.

**Table 6.6a The posterior estimates of the variance components
(double exponential prior for $\beta_{4,p}$)**

Gibbs:			NUTS:		
Posterior of the protein level variance components			Posterior of the protein level variance components		
Groups	Name	Variance	Groups	Name	Variance
		(median(95% credible interval))			(median(95% credible interval))
Protein	(Intercept)	0.52 (0.32, 0.92)	Protein	(Intercept)	26.5 (13.0,48.6)
	centralized m/z	0.003 (0.002,0.005)		centralized m/z	0.85 (0.49,1.48)
	factor(intervention)	0.13 (0.08,0.22)		factor(intervention)	0.64 (6.8e-04,19.0)
	Residual	1.19 (1.14, 1.23)		Residual	1.09 (1.07, 1.11)
Posterior distribution of the logistic regression coefficients for the missing model			Posterior distribution of the logistic regression coefficients for the missing model		
Groups	Name	Coefficient	Groups	Name	Coefficient
		(median(95% credible interval))			(median(95% credible interval))
Missing	Intercept	-4.6 (-6.2, -3.3)	Missing	Intercept	-6.5 (-8.7, -5.2)
	m/z	0.0084 (0.0080,0.0088)		m/z	0.0085 (0.0085,0.0085)
	peptide abundance	-2.5 (-3.1, -2.0)		peptide abundance	-1.8 (-2.3,-1.4)

**Table 6.6b The posterior estimates of the variance components
(normal prior for $\beta_{4,p}$)**

Gibbs: Posterior of the protein level variance components			NUTS: Posterior of the protein level variance components		
Groups	Name	Variance	Groups	Name	Variance
		median(95% credible interval)			(median(95% credible interval))
Protein	(Intercept)	0.52 (0.32,0.91)	Protein	(Intercept)	24.9 (13.2,44.8)
	centralized m/z	0.003 (0.002,0.005)		centralized m/z	0.78 (0.48,1.32)
	factor(class)	0.12 (0.07, 0.21)		factor(intervention)	0.64 (0.02,3.71)
	Residual	1.19 (1.14,1.23)		Residual	1.09 (1.07, 1.11)
Posterior distribution of the logistic regression coefficients for the missing model			Posterior distribution of the logistic regression coefficients for the missing model		
Groups	Name	Coefficient	Groups	Name	Coefficient
		median(95% credible interval)			median(95% credible interval)
Missing	Intercept	-4.7 (-6.4, -3.3)	Missing	Intercept	-6.5 (-8.4, -5.2)
	m/z	0.0084 (0.0080,0.0088)		m/z	0.0085 (0.0085,0.0085)
	peptide abundance	-2.5 (-3.2, -2.0)		peptide abundance	-1.8 (-2.2,-1.4)

6.5.2 Analysis of 76 proteins in one model

Among the discovered proteins, 76 of them had > 5 peptides observations. Attempts to including these 76 proteins in one model were made using both BUGS and HMC/NUTS programs.

All proteins' 95% credible intervals from the NUTS program contained zero. Narrower intervals produced four candidates with trends in irregular abundance in CVID patients. The HMC/NUTS algorithm achieved convergence in the protein level estimates, while the Gibbs algorithm did not achieve convergence within 120000 iterations.

6.6 Discussion

In this case study, both the estimates for central tendency and variance of the protein level class difference are shrunk towards the grand estimates in the multiple protein models. Compared to the simulated study and the cardiac case study, the CVID proteomic discovered more proteins but many more proteins had smaller numbers of observations. The resultant unbalanced design, and having sparse numbers of observations in the

combined cells of protein by subjects, adds challenges in the multiple protein model analysis. The posterior inter-protein variance in the class difference has mode closing to zero in the Immunity proteins (shown in Figure 6.4). The posterior inter-protein variance is similar to those shown in Figure 6.4 in the model including 76 proteins. The violation of multivariate normal distribution in the protein level estimates makes it difficult to analyse all proteins in one model, but analysing them by different functional groups has been shown to be a solution from this case study. Intensities of proteins belonging to a same functional group would be more likely to be multivariate normal distributed and correlated; when correlated responses are analysed in a multiple responses model, the efficiency will be higher than analysing them via single response models (Goldstain, 1999).

In all the trials, we found that NUTS achieved better convergence than Gibbs, and need fewer numbers of iterations for convergence. Analysing a group of proteins made it easier to achieve convergence in the program than analysing a larger number of proteins. We also found that, the BUGS and NUTS programs are not sensitive to the two different prior distributions for class difference. NUTS produced larger variance components compared to BUGS due to the censoring missing mechanism. Both BUGS and NUTS indicated that when missing data are taken into account, proteins of the immunology function group may not be irregulated in COVID patients, although a couple proteins need further investigation. The *R/lmer* model 2 gives different candidates, perhaps caused by the uncertainty introduced by the non-random missingness. A larger number of missing observations, treated as unknown parameters in the Bayesian model, reduced the statistical power of the data. Since no informative priors for the class difference are available, the advantage of using the Bayesian method is to provide estimates for the non-random missingness, thus preventing the potential bias caused by missing values and to reducing the false discoveries.

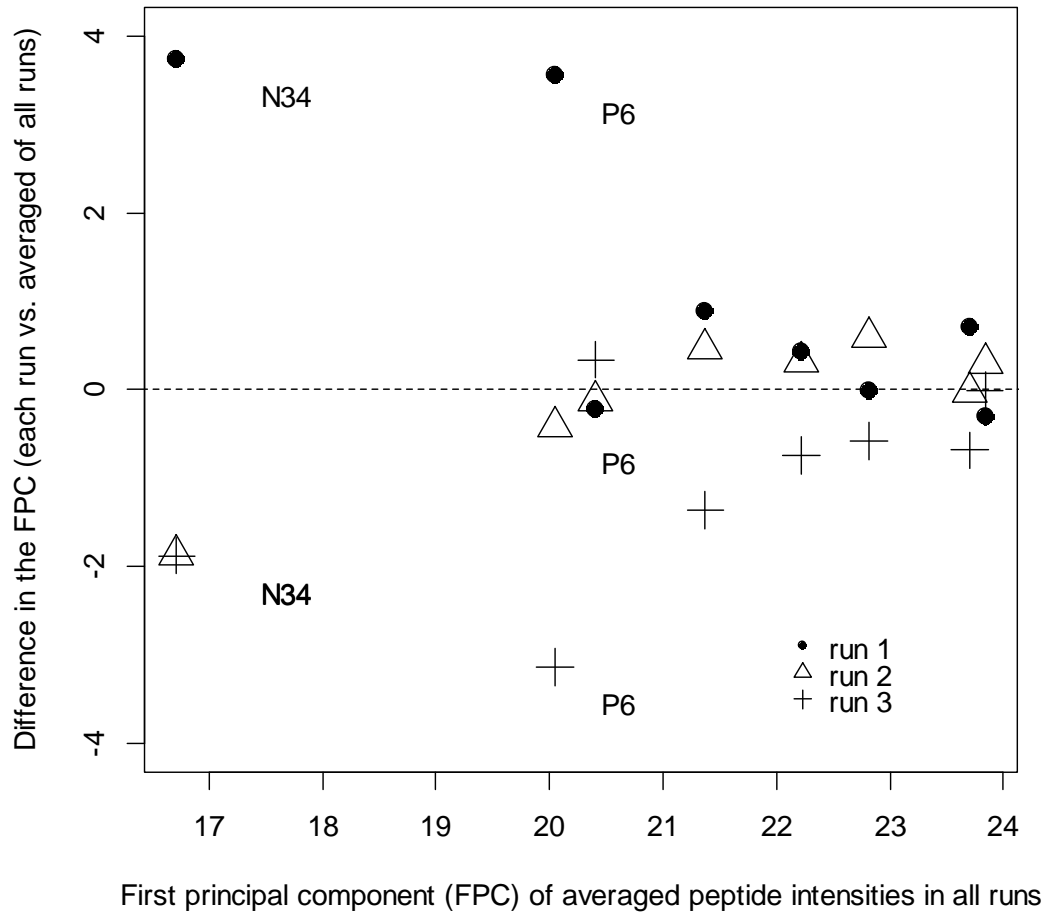
In the computation for the multivariate multilevel model, the variance-covariance matrix \mathbf{V} of the response $\gamma_{i,l,p}$ is derived from equation of the variance-covariance matrix of the protein level estimates $\mathbf{\Phi}$ and the variance-covariance matrix of the subject level estimates \mathbf{G} as defined in section 2.3 of Chapter 4. When the data comprises large numbers of proteins and subjects, the computation for \mathbf{V} will become a challenging task. The sparse and unbalanced numbers of peptide observations in the combination of

subjects and proteins also slows down the MCMC computing and makes convergence hard.

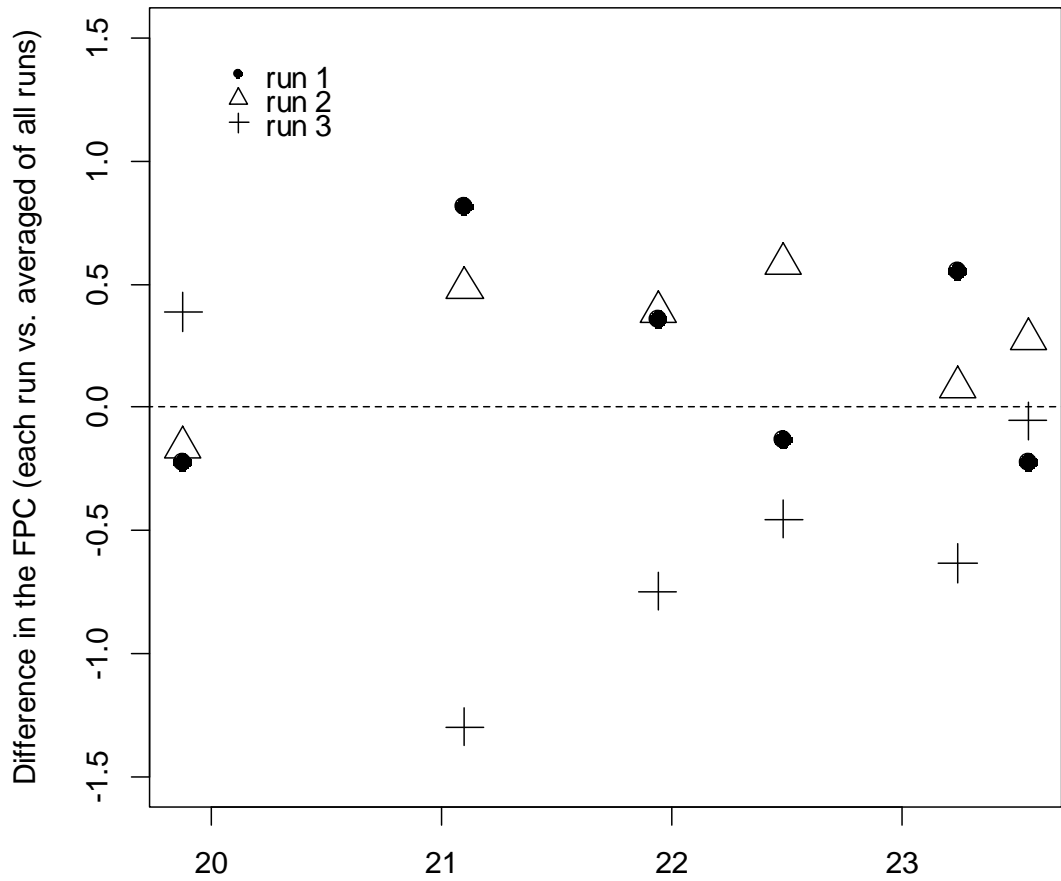
In the future work, the variance component estimates can be considered to be different for different protein groups. The correlations information between different proteins in the same functional group from the current case can also be useful as the prior information for the future study.

Figure 6.1 First principal component (FPC) plots for the reproducibility assessment

Legend: A) FPC with all 8 samples. B) FPC excluding the two contaminated samples. C) Second principal component plot excluding the two contaminated samples.

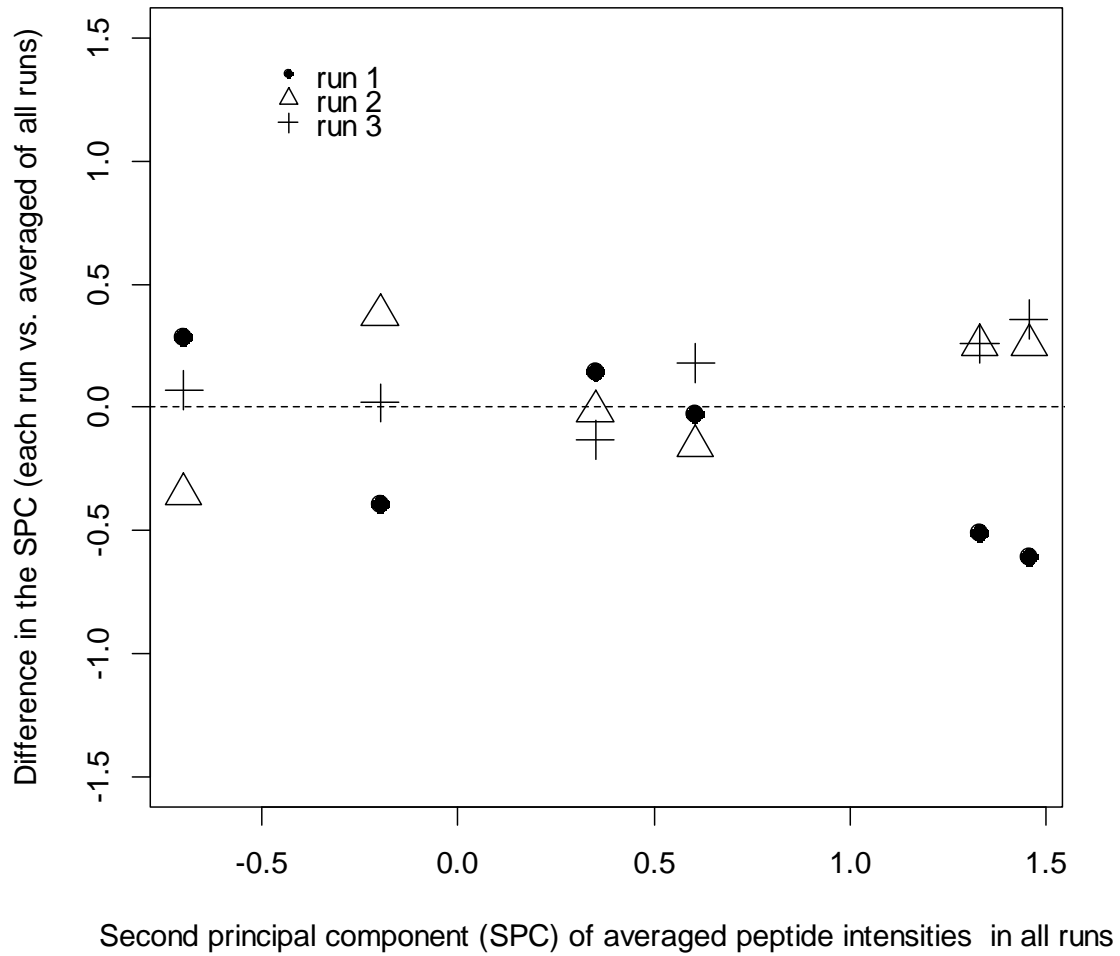


(A)



First principal component (FPC) of averaged peptide intensities in all runs

(B)



(C)

Figure 6.2 The mass to charge ratio and relative intensity by runs

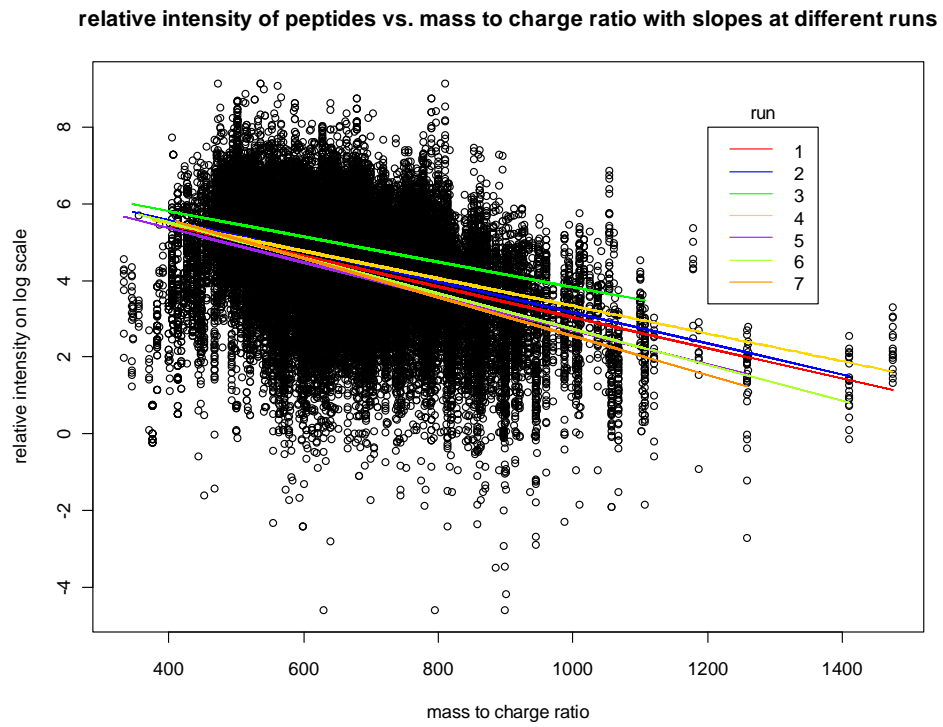


Figure 6.3 The mass to charge ratio and relative intensity by proteins

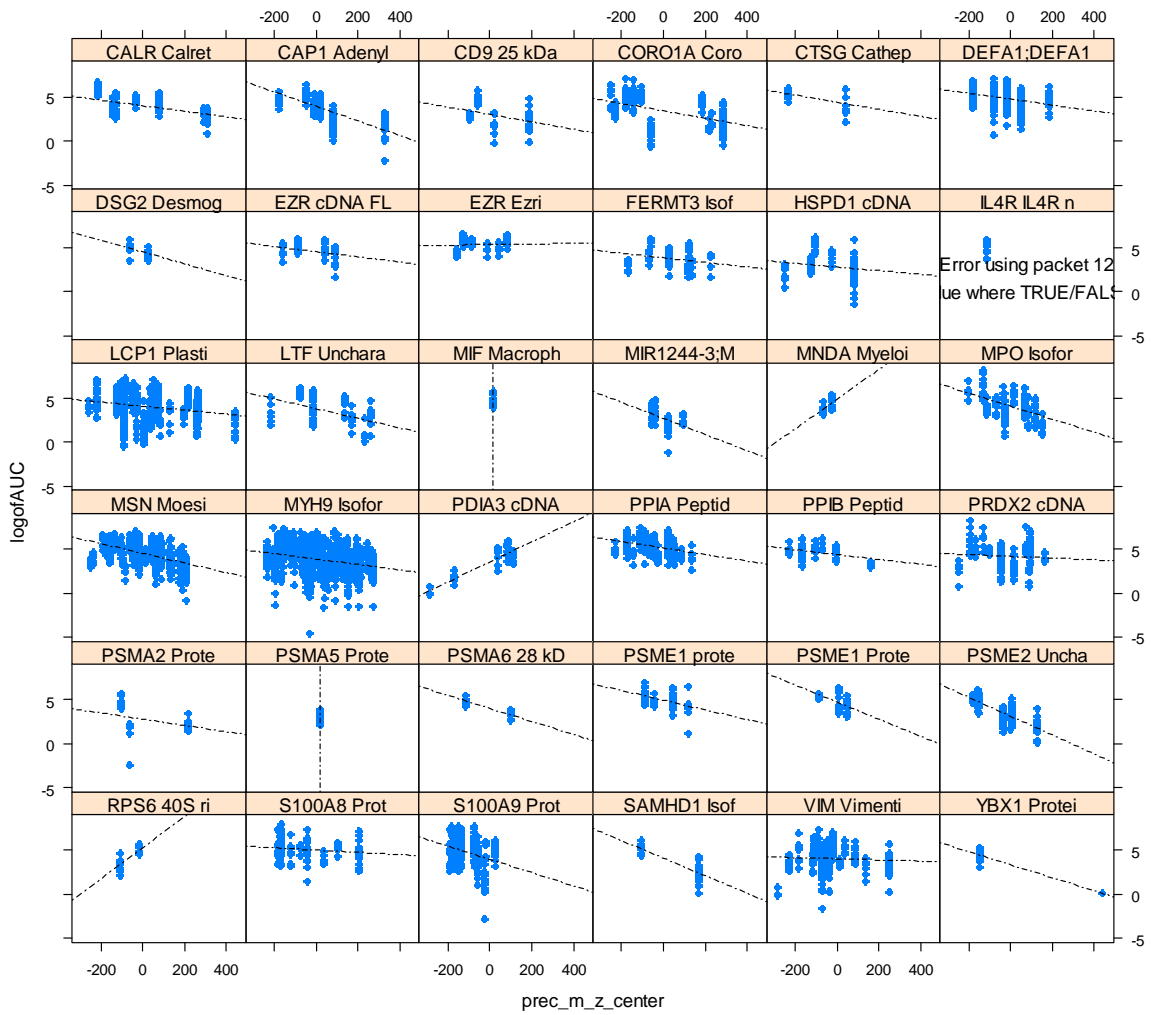


Figure 6.3b The mass to charge ratio and relative intensity by label and runs

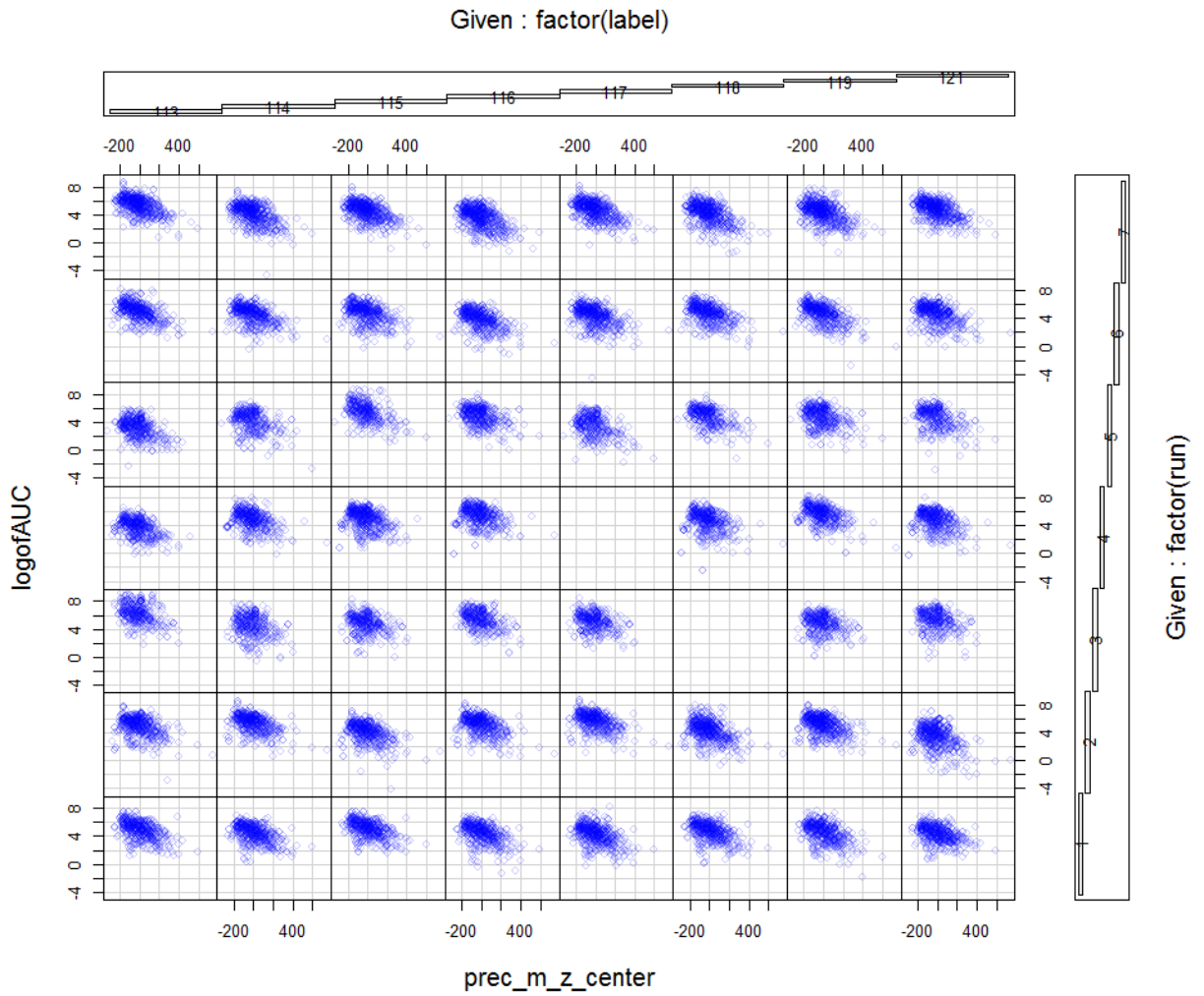


Figure 6.4 The caterpillar plots for multiple protein model (model 2)

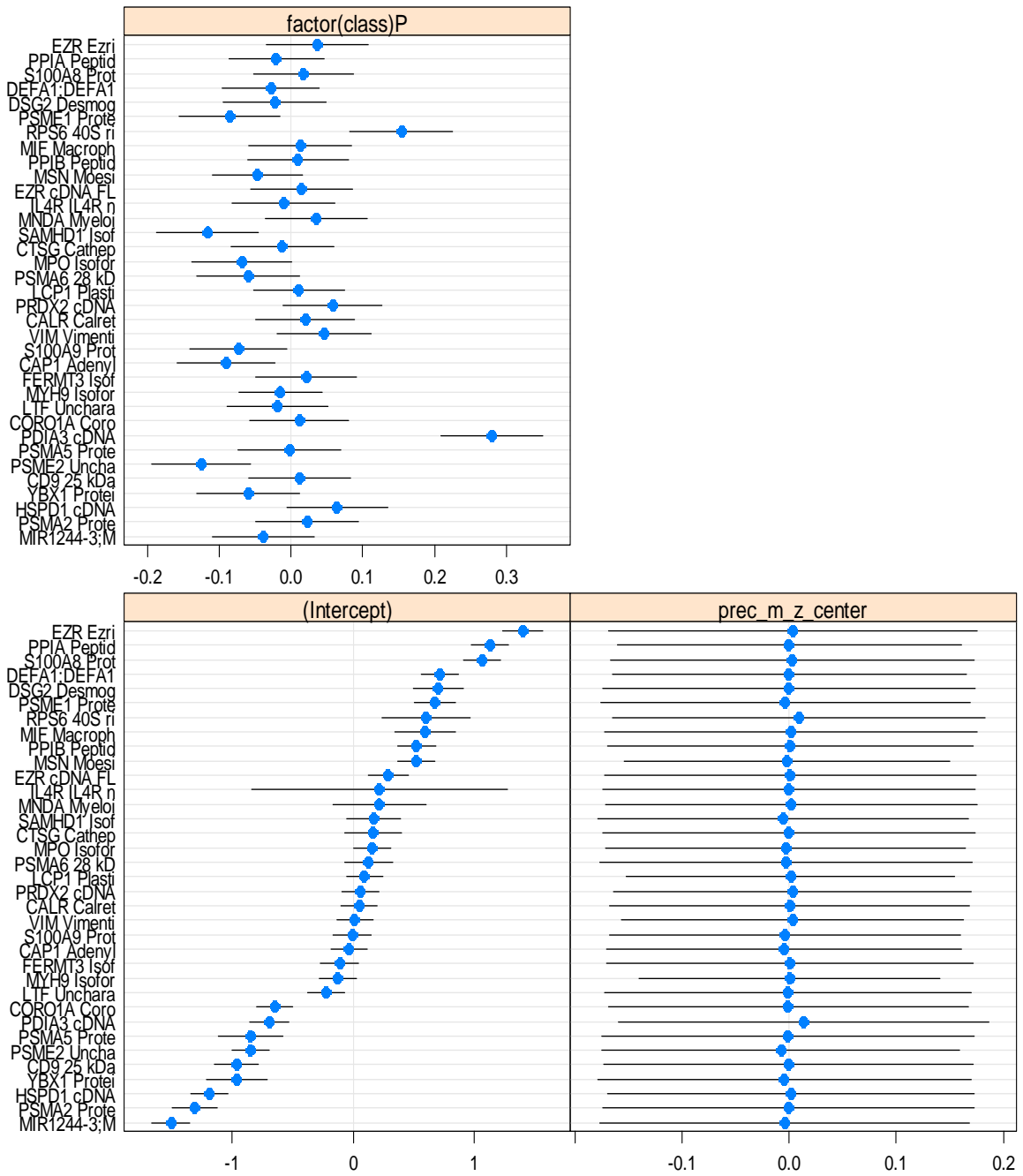
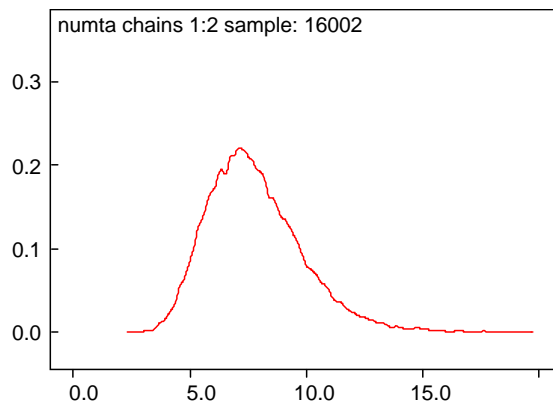
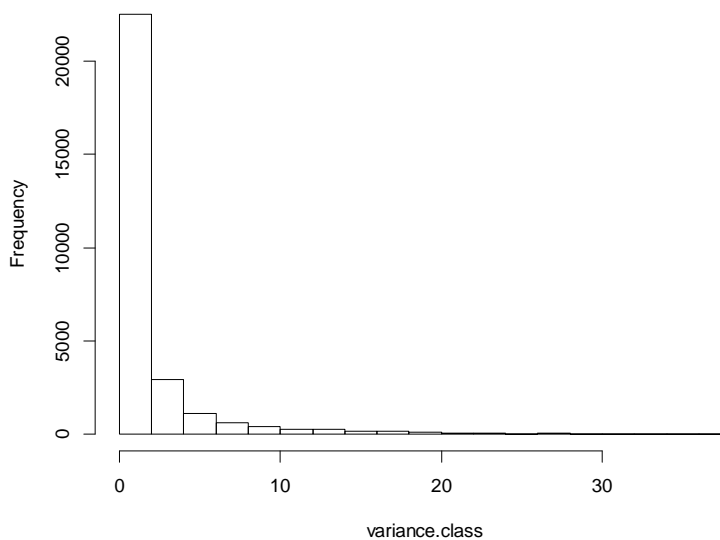


Figure 6.5a The posterior distribution of the protein level variance for class mean –
DE prior for $\beta_{4,p}$



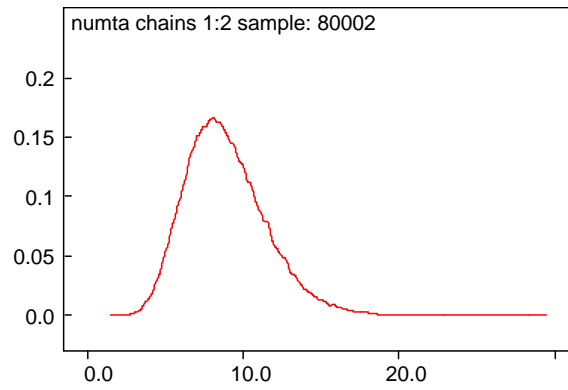
(A)- From BUGS, Of note: numta is the precision: 1/variance.

Histogram of variance.class

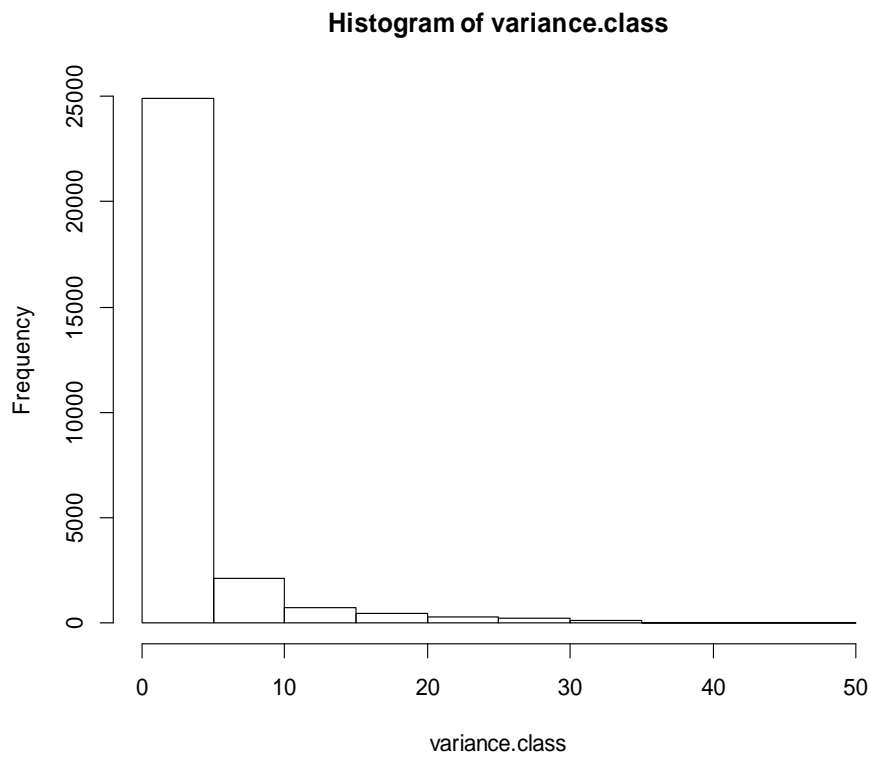


(B)-From NUTS.

Figure 6.5b The posterior distribution of the protein level variance for class mean-normal prior $\beta_{4,p}$



(A)- From BUGS, Of note: numta is the precision: 1/variance.



(B)-From NUTS.

Appendix 6.1: The RSTAND and Bugs programs

The RSTAN program

```
proteincv_code <-'
data {
  int nobs;
  int ncensor;
  int nmiss;
  int nprotein;
  real censor_lim; //number of records
  int<lower=0> subject_obs[nobs];
  int<lower=0> subject_m[nmiss];
  int<lower=0> subject_cen[ncensor];

  int<lower=0> proteinid_obs[nobs];
  int<lower=0> proteinid_m[nmiss];
  int<lower=0> proteinid_cen[ncensor];
  int<lower=0,upper=1> classno_obs[nobs];
  int<lower=0,upper=1> classno_m[nmiss];
  int<lower=0,upper=1> classno_cen[ncensor];

  int<lower=0,upper=1> run1[nobs];
  int<lower=0,upper=1> run2[nobs];
  int<lower=0,upper=1> run3[nobs];
  int<lower=0,upper=1> run4[nobs];
  int<lower=0,upper=1> run5[nobs];
  int<lower=0,upper=1> run6[nobs];

  int<lower=0,upper=1> tlab113[nobs];
  int<lower=0,upper=1> tlab114[nobs];
  int<lower=0,upper=1> tlab115[nobs];
  int<lower=0,upper=1> tlab116[nobs];
  int<lower=0,upper=1> tlab117[nobs];
  int<lower=0,upper=1> tlab118[nobs];
  int<lower=0,upper=1> tlab119[nobs];

  int<lower=0,upper=1> miss_obs[nobs];
  int<lower=0,upper=1> miss_m[nmiss];

  real logofAUC[nobs];
  real m_z_centered[nobs];
  real m_z_centered_cen[ncensor];
  real m_z_centered_m[nmiss];

  real totalprotein[nobs];

  cov_matrix[2] prec;
  cov_matrix[2] R2;
  vector[2] mn2;
}

transformed data
{
  cov_matrix[2] invprec;
  invprec<-inverse(prec);
}

parameters {
  vector[2] U_latent[nprotein];
  vector[nprotein] s;
  vector[2] gsub[nprotein];

  cov_matrix[2] pVAR;
  real<lower=0> ita;
  real numta;

  real alpha_latent;
```

```

real alpha_mu;
real alpha_theta;

real alpha1_latent;
real alpha2_latent;

real beta2_latent[50];
real beta2_theta;
real beta2_mu[50];

real beta3_latent[6];
real beta3_theta;
real beta3_mu[6];

real beta4_latent[7];
real beta4_theta;
real beta4_mu[7];

real beta5_latent;
real beta5_theta;
real beta5_mu;

real logofAUC_m_latent[nmiss];
}

transformed parameters
{
  vector[2] U[nprotein];
  vector[nprotein] g3sub;
  real beta3[6];
  real beta4[7];
  real beta5;
  real beta2[50];
  real alpha;
  real alpha1;
  real alpha2;
  matrix[2,2] L;

  for ( run in 1:6)
  beta3[run]<-beta3_mu[run]+beta3_theta*beta3_latent[run];

  for (label in 1:7)
  beta4[label]<-beta4_mu[label]+beta4_theta*beta4_latent[label];

  for (sub in 1:50)
  beta2[sub]<-beta2_mu[sub]+beta2_theta*beta2_latent[sub];

  beta5<-beta5_mu+beta5_theta*beta5_latent;
  L<-cholesky_decompose(pVAR);

  g3sub<-0+numta*s;

```

```

for (prot in 1:nprotein)
{ U[prot]<-gsub[prot]+L*U_latent[prot];
alpha<-alpha_mu+alpha_theta*alpha_latent;
alpha1<-0.0085+2.5e-7*alpha1_latent;//mass ratio dependant,the standard error can change
alpha2<-0.45+0.4*alpha2_latent;//abundant dependant
}

model
{
real mu[nobs];
real mu_m[nmiss];
real mu_cen[ncensor];
real logofAUC_m[nmiss];
real pmiss[nobs];
real pmiss_m[nmiss];

for (pep in 1:nobs)
{mu[pep]<-beta2[subject_obs[pep]]+U[proteinid_obs[pep]][1]+U[proteinid_obs[pep]][2]*m_z_centered[pep]
+g3sub[proteinid_obs[pep]]*classno_obs[pep]+beta3[1]*run1[pep]+beta3[2]*run2[pep]+beta3[3]*run3[pep]+beta3[4]*run4[pep]
+beta3[5]*run5[pep]+beta3[6]*run6[pep]+beta4[1]*tlab113[pep]+beta4[2]*tlab114[pep]+beta4[3]*tlab115[pep]+beta4[4]*tlab
116[pep]+beta4[5]*tlab117[pep]+beta4[6]*tlab118[pep]+beta4[7]*tlab119[pep]+beta5*totalprotein[pep];
pmiss[pep]<-inv_logit(alpha+alpha1*m_z_centered[pep]+alpha2*logofAUC[pep]);
}

for (pep in 1:nmiss)
{mu_m[pep]<-
beta2[subject_m[pep]]+U[proteinid_m[pep],1]+U[proteinid_m[pep],2]*m_z_centered_m[pep]+g3sub[proteinid_obs[pep]]*class
no_m[pep];
logofAUC_m[pep]<-mu_m[pep]+logofAUC_m_latent[pep]*ita;
pmiss_m[pep]<-inv_logit(alpha+alpha1*m_z_centered_m[pep]+alpha2*logofAUC_m[pep]);
}

for (pep in 1:ncensor)
{mu_cen[pep]<-
beta2[subject_cen[pep]]+U[proteinid_cen[pep],1]+U[proteinid_cen[pep],2]*m_z_centered_cen[pep]+g3sub[proteinid_obs[pep]]
*classno_cen[pep];
if (mu_cen[pep]> censor_lim) mu_cen[pep]<-censor_lim;
}

for (sub in 1:50)
{beta2_latent[sub]~normal(0,1);
beta2_mu[sub]~normal(0,1);
}
beta2_theta~gamma(1,1);

for (run in 1:6)
{beta3_latent[run]~normal(0,1);
beta3_mu[run]~normal(0,1);
}
beta3_theta~gamma(1,1);

for (label in 1:7)
{beta4_latent[label]~normal(0,1);
beta4_mu[label]~normal(0,1);
}
beta4_theta~gamma(1,1);

beta5_latent~normal(0,1);
beta5_mu~normal(0,1);

for (prot in 1:nprotein)

```

```

gsub[prot]~multi_normal(mn2,R2);

s~double_exponential(0,1);
numta~gamma(1,1);

pVAR~inv_wishart(2,invprec);
for (prot in 1:nprotein)
  U_latent[prot]~multi_normal(mn2,R2); //standard multinormal distributed

ita~gamma(1,1);
logofAUC~normal(mu,ita);

alpha_latent~normal(0,1);
alpha_mu~normal(0,1);
alpha_theta~gamma(1,1);

alpha1_latent~normal(0,1);
alpha2_latent~normal(0,1);

for (pep in 1:nobs)
  miss_obs[pep]~bernoulli(pmiss[pep]);

for (pep in 1:nmiss)
  {miss_m[pep]~bernoulli(pmiss_m[pep]);
   logofAUC_m_latent[pep]~normal(0,1);}

for (pep in 1:ncensor)
  lp__ <- lp__ + log(Phi((censor_lim-mu_cen[pep])/ita)+0.001);//the difference between censored limit and mu_cen can ne negative
}

```


The BUGS program

```

model
{
  for (pep in 1: ns)
  {
    auc[pep]~dnorm(mu[pep],ita)
    mu[pep]<-beta2[subject[pep]]+U[proteinid[pep],1]+U[proteinid[pep],2]*m_z_centered[pep]+U[proteinid[pep],3]*class[pep]
    +beta3_1*run1[pep]+beta3_2*run2[pep]+beta3_3*run3[pep]+beta3_4*run4[pep]+beta3_5*run5[pep]+beta3_6*run6[pep]
    +beta4_113*tlab113[pep]+beta4_114*tlab114[pep]+beta4_115*tlab115[pep]+beta4_116*tlab116[pep]+beta4_117*tlab117[pep]
    +beta4_118*tlab118[pep]+beta4_119*tlab119[pep]+beta5*totalprotein[pep]
    pmiss.lim[pep]<-alph0+alph1*m_z_centered[pep]+alph2*auc[pep]
    pmiss[pep]<-(1-censor[pep])*(max(0.001,min(0.99,pmiss.lim[pep]))) +censor[pep]*1
    missp[pep]~dbin(pmiss[pep],1)
    #the missing included censor in bug program, but not in NUTS program
  }
  ita~dgamma(1,1)

  #prior for random coefficients
  for (protein in 1:35)
  {U[protein,1:2]~dmnorm(gamma[1:2],T[1:2,1:2])}

  for (protein in 1:35)
  {U[protein,3]~ddexp(0,numta)} # need to assign the prior assume the class effect is independant from the slope,proteins abundance.

  for (sub in 1:50)
  {beta2[sub]~dnorm(0,1)}
  #prior for fixed coefficient
  #use informative prior
  alph0~dnorm(1,0.01)
  alph1~dnorm(0.0085,2.5E7)
  alph2~dnorm(-0.45,4) #use informative prior
  beta3_1~dnorm(0,0.1)
  beta3_2~dnorm(0,0.1)
  beta3_3~dnorm(0,0.1)
  beta3_4~dnorm(0,0.1)
  beta3_5~dnorm(0,0.1)
  beta3_6~dnorm(0,0.1)

  beta4_113~dnorm(0,0.1)
  beta4_114~dnorm(0,0.1)
  beta4_115~dnorm(0,0.1)
  beta4_116~dnorm(0,0.1)
  beta4_117~dnorm(0,0.1)
  beta4_118~dnorm(0,0.1)
  beta4_119~dnorm(0,0.1)
  beta5~dnorm(0,0.1)
  #hyper prior

  #hyper prior
  gamma[1:2]~dmnorm(mn[1:2],prec[1:2,1:2])
  T[1:2,1:2]~dwish(R[1:2,1:2],2)
  numta~dgamma(1,1)
}

Data
list(ns=6000,R=structure(.Data=c(0.01,0.001,0.001,0.1),.Dim=c(2,2)),prec=structure(.Data=c(0.01,0.01,0.01,0.01),.Dim=c(2,2)),mn=c(0,0))

```

**Appendix 6.2: BUGS and NUTS results for protein level parameter of class-immunity
group proteins**

protein	Using normal prior			Using Double EXP prior		
	2.5% percentile	median	97.5% percentile	2.5% percentile	median	97.5% percentile
CALR Calret	-0.13	0.08	0.38	-0.20	0.14	0.47
CAP1 Adenyl	-0.21	0.00	0.22	-0.35	-0.04	0.26
CD9 25 kDa	-0.40	-0.02	0.26	-0.63	-0.13	0.35
CORO1A Coro	-0.13	0.06	0.37	-0.23	0.11	0.44
CTSG Cathep	-0.37	0.00	0.37	-0.60	-0.02	0.56
DEFA1;DEFA1	-0.38	-0.10	0.09	-0.54	-0.24	0.05
DSG2 Desmog	-0.30	0.02	0.44	-0.47	0.10	0.69
EZR cDNA FL	-0.34	0.00	0.35	-0.54	-0.02	0.50
EZR Ezri	-0.21	0.04	0.42	-0.33	0.12	0.58
FERMT3 Isof	-0.25	0.01	0.31	-0.41	0.01	0.42
HSPD1 cDNA	-0.24	0.01	0.32	-0.40	0.01	0.41
IL4R IL4R n	-0.41	0.00	0.37	-0.67	-0.03	0.60
LCP1 Plasti	-0.23	-0.05	0.10	-0.36	-0.14	0.08
LTF Unchara	-0.37	-0.03	0.21	-0.57	-0.15	0.26
MIF Macroph	-0.30	0.02	0.43	-0.50	0.07	0.66
MIR1244-3;M	-0.31	0.00	0.32	-0.50	-0.02	0.45
MNDA Myeloi	-0.34	0.01	0.40	-0.55	0.03	0.61
MPO Isofor	-0.32	-0.03	0.17	-0.49	-0.14	0.20
MSN Moesi	-0.17	-0.01	0.14	-0.30	-0.08	0.14
MYH9 Isofor	-0.23	-0.07	0.05	-0.35	-0.16	0.03
PDIA3 cDNA	-0.22	0.05	0.47	-0.36	0.14	0.65
PPIA Peptid	-0.17	0.01	0.20	-0.30	-0.03	0.22
PPIB Peptid	-0.21	0.03	0.36	-0.34	0.07	0.49
PRDX2 cDNA	-0.20	0.01	0.25	-0.34	-0.02	0.30
PSMA2 Prote	-0.60	-0.06	0.23	-0.87	-0.27	0.29
PSMA5 Prote	-0.34	0.01	0.37	-0.56	0.01	0.59
PSMA6 28 kD	-0.35	0.00	0.38	-0.57	0.01	0.58
PSME1 Prote	-0.25	0.03	0.39	-0.39	0.08	0.56
PSME2 Uncha	-0.24	-0.01	0.22	-0.39	-0.07	0.26
RPS6 40S ri	-0.34	0.01	0.39	-0.56	0.02	0.59
S100A8 Prot	-0.36	-0.04	0.17	-0.56	-0.18	0.20
S100A9 Prot	-0.30	-0.05	0.12	-0.46	-0.17	0.11
SAMHD1 Isof	-0.37	-0.01	0.30	-0.59	-0.07	0.44
VIM Viment	-0.26	-0.05	0.12	-0.40	-0.15	0.11
YBX1 Protei	-0.48	-0.02	0.31	-0.76	-0.14	0.45

HMC/NUTS results for protein level parameter of class-immunity group proteins

protein	Using normal prior			Using Double Exp prior		
	2.5% percentile	median	97.5% percentile	2.5% percentile	median	97.5% percentile
CALR Calret	-3.68	0.01	2.78	-6.07	-0.01	5.49
CAP1 Adenyl	-2.97	-0.01	2.33	-5.94	-0.01	5.01
CD9 25 kDa	-3.02	0.01	3.78	-3.43	-0.01	4.64
CORO1A Coro	-2.98	-0.01	2.50	-4.00	0.00	4.92
CTSG Cathep	-3.42	-0.01	2.59	-4.21	0.00	5.29
DEFA1;DEFA1	-3.25	-0.01	2.65	-6.02	-0.06	3.36
DSG2 Desmog	-2.19	0.03	3.80	-7.42	-0.01	7.02
EZR cDNA FL	-2.84	-0.01	2.81	-5.46	-0.01	4.76
EZR Ezri	-2.90	0.00	3.06	-3.85	0.01	4.53
FERMT3 Isof	-3.16	-0.01	2.34	-4.05	-0.03	4.71
HSPD1 cDNA	-2.65	0.01	2.90	-4.40	0.02	5.14
IL4R IL4R n	-3.29	-0.01	2.53	-4.61	0.00	5.86
LCP1 Plasti	-2.84	0.00	2.31	-3.32	-0.02	4.39
LTF Unchara	-2.49	0.04	2.88	-5.15	0.00	3.86
MIF Macroph	-2.77	0.01	3.03	-2.85	0.02	4.70
MIR1244-3;M	-4.15	-0.01	2.59	-5.38	0.00	3.67
MNDA Myeloi	-2.08	0.00	3.60	-5.09	0.00	5.12
MPO Isofor	-2.79	0.01	2.98	-4.38	-0.01	5.04
MSN Moesi	-2.92	0.03	3.12	-6.18	-0.02	3.19
MYH9 Isofor	-2.54	0.02	2.95	-3.32	-0.01	4.88
PDIA3 cDNA	-2.86	0.01	3.03	-3.82	0.00	5.23
PPIA Peptid	-3.23	-0.01	2.21	-4.49	0.02	4.83
PPIB Peptid	-2.81	0.00	2.31	-3.48	0.00	4.60
PRDX2 cDNA	-2.70	0.00	2.73	-4.99	-0.03	2.80
PSMA2 Prote	-2.77	0.00	3.30	-4.98	-0.01	5.41
PSMA5 Prote	-3.30	0.02	2.93	-3.31	0.01	4.57
PSMA6 28 kD	-2.73	0.00	2.81	-4.94	0.00	5.08
PSME1 Prote	-2.65	0.00	2.86	-3.31	0.01	3.81
PSME2 Uncha	-2.95	0.02	2.71	-3.53	0.03	5.32
RPS6 40S ri	-3.54	0.00	2.75	-5.66	-0.01	2.75
S100A8 Prot	-2.93	-0.01	2.76	-3.86	0.00	4.63
S100A9 Prot	-2.83	0.00	3.15	-3.18	0.00	5.61
SAMHD1 Isof	-2.91	0.00	3.13	-3.92	-0.01	4.58
VIM Viment	-3.15	-0.02	2.69	-3.53	0.01	4.14
YBX1 Protei	-2.89	0.01	2.87	-4.82	-0.03	5.11

Bugs using Normal priors has a better convergence in the posterior samples of class difference.

Appendix 6.3: NUTs results for protein level parameter of class difference -76

proteins

protein name	class.de.p2.5	class.de.median	class.de.p97.5	class.de.p12.5	class.de.p87.5
ACTB cDNA F	-0.005	-0.004	-0.003	-0.005	-0.003
ACTBL2 Beta	-0.911	1.529	1.581	-0.055	1.571
ACTG1 cDNA	-1.626	0.368	1.089	-0.244	0.377
ACTG1 Uncha	-1.734	-1.551	1.779	-1.624	0.034
ACTN1 Actin	-1.845	-0.429	1.454	-0.440	0.045
ACTN1 cDNA	-0.941	0.213	1.421	-0.041	0.224
ALB Isofor	-1.288	0.989	1.477	-0.095	1.016
ANXA1 Uncha	-1.758	0.043	0.893	-0.282	0.047
ANXA6 annex	-1.027	2.298	2.381	0.012	2.367
CALR Calret	-2.319	6.494	6.720	-0.399	6.684
CAP1 Adenyl	-1.555	2.003	2.078	-0.060	2.063
cDNA FLJ5	-1.246	-1.206	1.242	-1.239	0.016
CFL1 Unchar	-0.838	-0.425	7.810	-0.440	0.148
CLIC1 Chlor	-1.755	-1.575	1.444	-1.619	0.040
CORO1A Coro	-2.711	-0.369	4.687	-0.407	0.314
DEFA1;DEFA1	-1.267	8.628	8.928	-0.225	8.878
ENO1 Isofor	-1.374	0.048	1.547	-0.047	0.121
EZR Ezri	-2.050	3.537	3.644	-0.179	3.628
F13A1 Coagu	-1.750	-0.368	0.407	-0.393	-0.015
FGA Isofor	-4.091	2.373	2.459	-0.104	2.444
FGG Unchara	-2.550	-2.463	1.879	-2.537	0.223
FLNA Isofor	-2.115	2.252	2.362	-0.137	2.335
FLNA Unchar	-2.364	-0.823	0.688	-0.847	0.036
GSN Unchara	-3.090	-2.990	0.956	-3.071	0.045
HBA2;HBA1 H	-0.610	0.260	5.333	-0.020	0.265
HBB Hemoglo	-0.730	-0.472	2.227	-0.479	0.121
HBD					
Hemoglo	-1.450	6.806	7.040	-0.181	7.000
HNRNPK cDN	-1.117	0.217	1.067	-0.080	0.220
HSP90AA1 Is	-1.017	-0.982	0.417	-1.010	0.034
HSPA1B;HSPA	-0.962	4.227	4.375	-0.028	4.350
HSPA8 Uncha	-1.097	-0.020	1.580	-0.050	0.184
ITGA2B Isof	-1.734	-0.874	1.469	-0.906	0.012
KRT1 Kerati	-1.331	0.383	1.493	-0.048	0.393
LCP1 Plasti	-0.636	-0.092	2.329	-0.095	0.213
LDHA L-lact	-1.845	1.796	1.864	-0.046	1.853
LDHB L-lact	-1.434	0.471	1.690	-0.041	0.484
MPO Isofor	-0.373	-0.134	3.566	-0.138	0.338
MSN Moesi	-2.355	3.743	3.874	-0.083	3.854
MYH9 Isofor	-2.148	-2.076	2.353	-2.135	0.092
MYL6 17 kD	-1.264	1.530	1.581	-0.106	1.571
MYL6 cDNA F	-0.830	0.579	1.129	-0.092	0.595
PARK7 Prote	-0.671	3.800	4.877	0.000	3.901
PF4 Platele	-2.024	2.179	2.255	-0.508	2.242
PFN1 Profil	-1.268	2.371	2.453	-0.255	2.439
PGK1 Phosph	-5.919	3.727	3.859	-0.088	3.841
PKM2 pyruva	-1.184	0.341	2.577	-0.113	0.348
PKM2 Pyruva	-1.516	2.006	2.075	-0.027	2.064
PLEK Plecks	-2.323	-2.180	1.154	-2.298	0.085
POTEE Isofo	-4.562	-4.415	1.757	-4.537	0.356

PPBP Platel	-0.958	2.049	2.126	-0.094	2.114
PPIA Peptid	-0.847	0.622	1.852	-0.014	0.638
PRDX2 cDNA	-1.916	2.425	2.873	-0.033	2.492
S100A8 Prot	-2.206	1.665	1.721	-0.065	1.708
S100A9 Prot	-0.780	3.886	4.020	-0.029	3.998
SOD1 Supero	-1.604	4.692	4.854	-0.114	4.827
TAGLN2 24 k	-1.600	-0.096	1.020	-0.176	0.018
TALDO1 Tran	-1.513	1.335	1.455	-0.117	1.371
THBS1					
Throm	-1.619	1.258	1.301	-0.143	1.292
TLN1 Talin-	-0.486	1.313	1.539	-0.015	1.347
TMSB4X					
TMSB	-2.105	-2.036	1.111	-2.091	0.083
TMSL3 8 kD	-1.292	1.394	1.439	-0.108	1.433
TPI1;TPI1P	-1.352	0.311	1.339	-0.150	0.323
TPM1 Isofor	-1.007	0.195	1.576	-0.023	0.214
TPM1					
Unchar	-1.468	-0.001	0.983	-0.090	0.052
TPM3 Isofor	-1.277	-1.232	1.707	-1.257	0.166
TPM4 Isofor	-3.517	1.240	1.272	-0.489	1.264
TUBA1B					
Tubu	-1.072	0.170	1.022	-0.103	0.175
TUBA4A cDN	-1.224	-1.060	1.490	-1.086	0.102
TXN Unchara	-1.871	-0.064	0.367	-0.170	-0.011
VCL Isofor	-1.202	1.253	1.458	-0.060	1.288
VIM Vimenti	-0.658	1.120	1.161	-0.033	1.154
YWHAB Isofo	-0.711	1.486	2.084	-0.042	1.525
YWHAQ 14-					
3-	-0.697	2.263	2.343	-0.034	2.330
YWHAQ					
Uncha	-1.800	0.443	1.743	-0.083	0.452
YWHAZ					
Uncha	-6.623	-2.510	0.425	-2.580	-0.096
ZYX Zyxi	-2.102	0.955	0.996	-0.270	0.992

*Data presented are 2.5 percentile, median, 97.5 percentile, 12.5 percentile and 87.5 percentile of the posterior estimates of class differences, from left to right respectively.

CHAPTER 7

Discussion

7.1 Overall review

By the time this PhD research was close to its finish, several notable changes had occurred in the proteomics world. The first was the announcement of a free data portal of clinical proteomic studies supplied by the National Cancer Institution. The release of the free data portal was the results of the 2008 international summit on proteomics data released and sharing policy (aka, the Amsterdam principles) (Rodriguez et al., 2009), and the 2010 follow-on workshop that the National Cancer Institute convened to address quality metrics for proteomics with an emphasis on mass spectrometry. The second exciting change was the most recently released landmark editorial article “Assays Without Borders” in the December 2013 issue of the journal *Nature Methods* (Kennedy et al., 2014). It reported research of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) partnered with 3 international labs to demonstrate the reproducibly quantifiable Human proteins using targeted mass spectrometry-based assay across three labs, two countries, and two continents. This research developed a multiplex of 645 assays representing 319 proteins expressed in human breast cancer. The third remarkable change is the increasing amount of integrated research in “omics” including proteomics and genomics. Using the SCOPUS database and searching the terms “integrating proteomics and genomics” and limiting to the published years between 2007 to present, 145 research articles were identified, of which 18 are published in computing and mathematical journals (SCOPUS classified statistical journals as mathematical journals). Using the same search terms and limiting publication years to 2000 to 2006 identified 83 research articles, of which 6 are from computing and mathematical journals. An approximately 75% increase of published research was seen in the studies integrating “omics” in the same period as this PhD research when compared to that in the years 2000 to 2006, and there was a three-fold increase for studies with computing and mathematical methodologies.

These changes support the relevance of this PhD research which was initiated in 2006-2007. The following sections summarize the learning, improvements and future areas of expansions for each section of this PhD study.

7.2. Reproducibility assessment for high throughput devices

7.2.1 Another literature review for the methodology of reproducibility assessment in the same period as the current PhD research

A recent search of articles from the SCOPUS database using the term “reproducibility in proteomics” identified 8 publications published between 2007 and the present. Of these, two are related to the statistical methodology. One of these studies (Merciera et al., 2009) was published at the same time as the multi-feature reproducibility paper (I.S.L. Zeng et al., 2009). They applied the standard exploratory principal component analysis and a single level mixed model analysis for the reproducibility measurement. Both analyses were applied to the peptide (peak) intensity level. The second study (Dazard et al., 2012) introduced a reproducibility index and confidence scores (ROCS) for a protein interaction proteomics study using Affinity-Purification mass spectrometry. This ROCS reproducibility assessment was invented for use in the protein identification phase. Until today, there is still little research on multi-feature reproducibility assessments, as proposed in chapter 2.

7.2.2 The theoretical review and interpretation of the methodologies related to the proposed method

7.2.2.1 Principal component, big random matrix and Tracey-Widom distribution for the largest Eigenvalues

Principal component analysis has been used as an exploratory data analysis (EDA) and a multivariate analysis tool to reduce the number of dimensions of a high dimensional data matrix (Kshirsagar, 1972). The principal components are the best linear combinations of the original variables, that explain the maximal proportion of the variances from the original

$n \times p$ data matrix, where n is the number of observations and p is the number of variables. They are derived by minimizing the sum of squared perpendicular distances between the original coordinates of data points and the proposed new axes of the principal component subspaces. Computationally, this can be realized by minimizing the trace of the product of residual matrix E and its transpose E' from the equation $Y = XB + E$, where Y represents the centralized data matrix, X represents the principal component scores and B represents the eigenvectors of the correlation or covariance matrix of the original data (Rao, 1964; SAS Institute Inc., 2010). The properties of the principal components are:

- 1) The resultant eigenvalues from the principal components analysis are in descending order. The first j th eigenvalue is equal to the proportion of variance of the centralized data explained by the j th principal component, if eigenvalues are normalized to sum to 1 (scaled eigenvalues). The first eigenvalue is the largest value of the eigenvalues and the first principal component explains most of the variances of the centralized data.
- 2) The eigenvectors are orthogonal vectors and the principal components are the perpendicular projection from the original data matrix.
- 3) The resultant principal components scores are uncorrelated.

When the number of variables p exceeds the number of samples n , PCA utilize the sample covariance matrix and enable the reduction in the dimension of the data.

PCA used the sample covariance matrix often referred to as the Wishart matrix, to derive the sample ordered eigenvalues. As for any other sample statistics, the sample ordered eigenvalues are estimates of the population eigenvalues. Among the sample ordered eigenvalues, the largest sample eigenvalue followed the Tracey-Widom distribution of order one under the random matrix theory (RMT) framework.

Random matrix theory has emerged as a mathematical framework applied in multivariate statistical analysis in the last decade. Its most common application is in the well-known method PCA and factor analysis, which is popularly used for high dimensional data produced from “omics”, imaging field, or the macroeconomic data from the stock market. Among these high dimensional data, the underlying structures are usually believed to be buried by

the noise (Berry et al., 2011). However, the earliest application of RMT in multivariate data analysis can be traced back to (Pearson, 1901) who introduced the reduction of data dimension through PCA (SAS Institute Inc., 2010). Studying the property of the sample covariance matrix of a high dimensional rectangular data matrix can be replaced by the study of the distribution of its eigenvalues. This is because the empirical distribution of the eigenvalues of the population covariance matrix Σ and its estimate from the sample variance \mathbf{S} can be both decomposed. For the population covariance, it can be decomposed as $\Sigma v_k = \lambda_k \times v_k$, and analogously the sample variance \mathbf{S} can be used to derive the sample Eigenvalues $\hat{\lambda}_k$ as in $\mathbf{S} v_k = \hat{\lambda}_k \times v_k$ through PCA.

Tracy and Widom (1996) proved that the largest eigenvalue of a Wishart matrix which has standard Gaussian distributed complex values as the matrix entries, asymptotically converged to the Tracy-Widom law. Johnstone (2001) proved that the largest eigenvalue of the Wishart matrix $X = (X_{jk})_{n \times p}$, with i.i.d standardized Gaussian entries also converged to the Tracy-Widom law of order one asymptotically. According to the theory 1.1 in his paper, let n be the sample size and p be the number of variables (dimension), if $\frac{n}{p} \rightarrow \gamma^{-1} \geq 1, n \rightarrow \infty, \text{ or } \frac{p}{n} \rightarrow \gamma \in (0, 1]$, then

$\frac{l_1 - \mu_{np}}{\sigma_{np}} \Rightarrow \mathbf{W}_1 \sim F_1$, where l_1 denotes the largest eigenvalue for X , μ_{np} and σ_{np} define the centre and scaling constants respectively, and

$$\mu_{np} = (\sqrt{n-1} + \sqrt{p})^2; \sigma_{np} = (\sqrt{n-1} + \sqrt{p}) \left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}} \right)^{1/3}.$$

F_1 is the Tracy-Widom distribution for order one defined as

$F_1(s) = \exp\left(-\frac{1}{2} \int_s^\infty q(x) + (x-s)q^2(x) dx\right), s \in \mathbb{R}$, where q solve the nonlinear Painlevé II differential equation, $q''(x) = xq(x) + 2q^3(x), q(x) \sim Ai(x)$ as $x \rightarrow +\infty$ and $Ai(x)$ denotes the airy function. Johnstone's theory works well even when n and p are as small as 5. It can

be applied to data when $n < p$ and both are large, simply by reversing the role in the centre and scale definition. Karoui (2007) extended Johnstone's theorem 1.1 to $p/n \rightarrow 0$, as $n, p \rightarrow \infty$. Onatski (2008) further extending the work of Karoui (2007) to establish the joint distribution of the several largest eigenvalues of a singular complex Wishart matrix (when $n < p$). Onatski's work provides a framework to test the first m largest eigenvalues which has a similar functionality to the scree plot in PCA. It provides inferential information such as the confidence limit for the test statistics - a function of the ordered eigenvalue for a complex singular matrix. This test not only provides the information for the underlying population structure, it also takes into account the uncertainty from the sample.

7.2.2.2 Multivariate permutation

In 1934, Fisher (Berry et al., 2011) proposed the exact probability of guessing right in the "lady tasting tea" experiment, by permuting all possible arrangements of the tea (tea with milk added first vs. tea with milk added second) to calculate the probability of a right guess in his design of experiment text book (Fisher, 1935).

Since 1980, the permutation method has gained attention in many areas after the booming in computing techniques. It is widely used to simulate the empirical distributions of test statistics for comparing quantities between two groups or pair groups. In addition, the multiple comparison consequence can be addressed using permutation method when the sample size is less than the number of variables in the observed data.

In the last decade, many works have also achieved by using permutation method in multivariate data analysis (M. J. Anderson and Legendre, 1999). It has been used when the assumption for the asymptotic test is violated, such as the multivariate normality assumption, and the sphericity assumption of the sample variance. It has also been commonly used in genome-wide association studies, in which the sample sizes are far less than the number of variables, to adjust for the false positive rate.

7.2.3 The improvement in the current reproducibility method and how this method works

As the traditional clinical reproducibility assessment only works for a single measure, a multidimensional approach will be necessary for the assessment of multiple measures from the proteomic platform. In particular, when the high dimensional data are formed by low dimensional signals embedded in isotropic noise, such as the data from mass spectrometer, reduction of the dimension will improve the efficiency in the reproducibility assessment.

The proposed method in this study provides a new approach to assess the repeatability at the proteome-wide scale. It uses principal component analysis to identify the underlying correlated pattern of the multiple features. The features which vary across different types of mass spectrometry experiments (i.e. SELDI-MS, iTRAQ MALDI-MS) are defined as either the discovered peaks or the summarized proteins expression. Principal component analysis is applied to the averaged quantities from those common features discovered from the replicated MS analysis. The first principal component scores derived from the PCA, which take up the most variance from the original data, provide a global quantity for the reproducibility assessment.

The by-product of this analysis, the First Principal Component (FPC) plot, provides a simple tool to visualize the agreement between replicated samples at the projected principle component subspace. Since the first principal component explains the largest variance of the original data, the FPC plot preserves the proteome-wise information. It has an analogue idea as the MA plot in a micro array study. The advantage of this method is that, it does not include the noise data, only operating on the common features.

The eigenvalues analysis is based on the property of the m largest eigenvalues of a sample covariance Wishart matrix. This property describes that the largest eigenvalue followed the Tracey-Widom limit law asymptotically. The Tracey-Widom distributed statistics in the eigenvalue analysis provide an inferential approach to identify the underlying dimensions of the data matrix.

In order to test the hypothesis that there is no difference across the replicates on the principal component space, the four classical multivariate test statistics (Wilks's Lambda, Pillai's trace, Hotelling–Lawley trace, and Roy's maximum root) have been considered. However,

with data $n < p$, the asymptotic multivariate test statistics are not appropriate. Two alternative multivariate test statistics are proposed for the inferential analysis. The permutation method is applied to these two test statistics to assess the hypothesis of no significant difference between the technical replicates. It operates on the global scale but in the projected principal component space. For these two test statistics, Max t stat uses the actual coordinates of the transformed data point, and the sign rank statistics only uses the sign of the difference between the paired data. The theoretical distribution of the test statistics is not available; using an alternative multivariate permutation test becomes necessary.

The current research provides a permutation method for testing the reproducibility for tests with multivariate responses. It can be extended in medical statistics to imaging, proteomics, metabolomics, and other areas.

7.2.4 The limitation of this method and what next to consider

The limitation of this method is that, it operates on the post-identification stage, which assumes that the peak has been well identified and it is only applied to the common features across the replications.

The eigenvalue test relies on the asymptotic assumption where the n and p approach to infinity, but this may not work well for data with a small sample size.

The two proposed multivariate test statistics are a parametric and a non-parametric version for the testing, both of them work well when sample size is larger than 15.

With the further development of random matrix theory, the asymptotic assumption for analysis take the infinity for both sample size and dimensionality (Paul and Aue, 2013). The increased computer efficiency in the 20th Century would make it common to perform permutation even used the enumeration method, instead of the Monte Carlo method.

The $n < p$ problem will always exist, especially when the discovery technology improve and p is getting bigger. The current method needs to expand to the condition when p is much bigger and $p/n \gg 1$, and the empirical distribution of m largest Eigenvalues test statistics needs to

extend to when $m > 100$. The new development of RMT, such as joint distribution of the eigenvalue, and the convergence rate of the empirical spectral distribution of the **big random matrix** will be the new mathematical theory for us to look at (Tao and Vu, 2010b; O'Rourke et al., 2011; Tao and Vu, 2011).

7.3. Multi-stage design is the necessary strategy in clinical proteomic study

7.3.1 Review of what has happened since this PhD research started, and focus on any new research that has emerged

A recent literature search using the term “multi-stage design in proteomics studies” identified two relevant studies. One is the paper (I.S.L Zeng et al., 2013) presented as chapter 3 in this thesis, while the other paper is a study utilizing multi-stage experimental design to validate a systematic antibody-screening tool of protein antibodies and immunohistochemistry (Williams et al., 2010). The latter paper introduced a systematic approach to validate an antibody-screening tool and provide a new strategy in exploration of human proteomes. This study showed a high successful rate of 93% of the discovered antibody protein array. Multi-stage design is shown to be a promising strategy in discovering next generation clinical biomarkers for the diagnosis, and prognosis of a disease and for predicting the response of a therapy for the disease.

Using the term “false discovery rate, family wise error rate, number of type I errors”, 10 articles were found in the SCOPUS related to the multiple hypothesis testing problem, and three of them proposed new methods for this statistical problem using numbers of type I errors.

7.3.2 The theoretical review and interpretation of the methodologies related to the proposed method

In multiple hypothesis testing literature, many works focused on how to control type I error rates when a single hypothesis test was performed many times and the conventional type I

error rate of 0.05 was not appropriate. The multiple tests phenomenon is especially ubiquitous in genomic studies when millions of tests are required for million genes. Several measures had been proposed for the overall type I error rate across multiple comparisons, including the widely used False discovery rate (FDA) (Soric, 1989; Benjamini and Hochberg, 1995) and Family wise error rate (FWER) (Shaffer, 1995). False discovery rate (FDR) is the expected proportion of the false positives from the significance findings, while Family wise error rate is the expected probability of having at least one false positive from the significance findings. Correspondingly, many control procedures had been invented to minimize these global type I error rates (Holm, 1979; Hochberg, 1988; Benjamini and Hochberg, 1995; Hwang et al.).

While a few works in the past focused on maximizing the power of multiple test, Storey (2007)'s development of a new theory using an optimal discovery procedure (ODP) to maximize number-of-true-positives with a fixed number-of-false-positives is gaining attention. His works are also of the most relevance to our study, and the following section will discuss our method under his proposed ODP framework.

Storey's multiple test procedure theory was proposed as a comparison to Neyman-Pearson's optimality lemma for a single test. An optimal discovery procedure (ODP) is defined as a multiple test procedure that maximizes the Expected number of True Positives (ETP) at a fixed Expected number of False Positives (EFP) for all Single Threshold Procedures (STP). A Single Threshold Procedure (STP) is a multiple test procedure that uses a single threshold as test significance for all tests. Storey (2007) proved that every multiple testing procedure that is invariant to the labelling of the test is a STP in the lemma 1 of his paper. In his ODP, firstly the tests were ranked in an order by the threshold function and secondly a threshold would be selected as the cut-off to define a test being significant. His proposed ODP borrowed the information from other tests of the multiple tests to determine the **relative** significance of each single one.

In a multi-stage clinical proteomic study, the multiple hypothesis testing error rate controls is less important than the statistical power. This is because the number of proteins being tested is much less than the number of genes tested in a genome association study. The sample size is expected to be small in the first discovery stage. As such, power is more of concerns than

controlling for overall type I errors because the false discoveries will be ruled out in the following verification and validation stages. However, if any false negative occur in the first stage, it will be more difficult to identify the missed protein candidates which could be of real value in comparison with making a false positive discovery. On the other hand, costs of proteomic assay and studies impose large constraints on the number of proteins to be selected at stages I and II. To maximize the statistical power under constraints of cost and a fixed overall type I error thus becomes the most appropriate optimal procedure for the current clinical proteomic studies.

7.3.3 The improvement in the current multi-stage design method and how this method works

In the development of a rigorous optimal procedure in multiple-testing, three components should be involved. The first is the optimal goal, which is the final result to be achieved; the second is the constraints under which the optimality to be found; the third is the procedure, which is the objective function and the optimization method, that achieves this optimality (Storey, 2007). These three components for our proposed method are described in detail below.

The optimality proposed in our study includes two different algorithms SA-a and SA-b. Both of the algorithms assumed that we have known information for stage I sample size, a cost function with fixed costs for assay, recruitment and other items. Algorithm SA-a selects proteins based on the t test of each protein. Algorithm SA-b selects proteins based on the t test of each protein and the F test for the biological group. The optimal goal is to find the design solution to maximize the expected number of true discoveries. The design solution is a vector of design parameters for a multi-stage proteomic study. The design parameters include significance thresholds for the p values based on the statistical tests used at stages I and II, and the sample size used at stages II and III. For SA-a, the constraints are the overall cost. For SA-b, the constraints are the expected number of false positives (EFP) and cost. Since we know that the false discovery rate (FDR) can be approximated by EFP and expected number of true positives (ETP),

$$FDR \approx \frac{EFP}{EFP + ETP} \text{ (Storey, 2007)}$$

Using EFP as a constraint will be approximated as the constraint of FDR. The expected number-of-true-positives (ETP) can be approximated by p which is the total number of discovered proteins. The procedure for the optimality of this multiple-test problem is the characteristics function built by the simulation or the analytical approximation for a three stage design proteomic study. The objective functions use the nominal type I error and the sample size at each stage to derive the expected number of true discoveries. Hybrid simulated annealing is used as the optimization method in this multiple procedure. In comparison to the conventional simulated annealing, which does not have a structure, the nested simulated annealing builds a structure for the searching space. Subsets of the solution spaces are constructed by using a beta-distributed jumping length and a uniformly distributed radius. The jumping length determines the size of the jump from one centre to another, and the radius provides the size of each searching sub-space.

Our proposed algorithms are also equivalent to Single Threshold Procedures (STP) defined by Storey. The single protein selection algorithm SA-a is a STP. At each stage, a single threshold (i.e. c_1 , c_2 and c_3) is used for p values of all the single protein tests. The group protein selection algorithm SA-b is also a STP, as it used single threshold for tests at each stage. At stage I, the single threshold is a combination of p value (significance levels) of the F test for the group a protein belonging to and p value of t test for each protein. These combinations of p values thresholds are denoted as α_{t_i} and α_{f_i} for the protein test and group test at stage I respectively. These paired p value thresholds vary in different solutions given by the proposal function in the simulated annealing search according to the following criterion,

$$\left(T^2 > F^{-1}_{df(p_i), df(n_i-p_i)}(1-\alpha_{f_i}) \right) \cup \left(T > Pt^{-1}_{df(n_i)}(1-\alpha_{t_i}/2) \cap T^2 > F^{-1}_{df(p_i), df(n_i-p_i)}(0.95) \right),$$

where T^2 is the F - distributed Hotelling's T -squared statistics with degrees of freedom determined by the number of proteins and the sample size at each stage; T is the Student t -statistics; and F^{-1} is the quantile function for the F -statistics. The criterion is set to select groups with a changeable significance level and proteins with a changeable significance level

from a group significant at the fixed 0.05 level during the optimization. This setting will pick all proteins of a group if this group is of high significance at stage I.

Similarly, at stage II, the single threshold is also a pair of p values (significance levels) of the F test for the group and the T test for the single protein $(\alpha_{t_2}, \alpha_{f_2})$:

$$\left(T^2 > F_{df(p_2), df(n_2 - p_2)}^{-1}(1 - \alpha_{f_2}) \cap Pt_{df(n_2)}^{-1}(0.975) \right) \cup \left(T > Pt_{df(n_2)}^{-1}(1 - \alpha_{t_2} / 2) \right)$$

The second-stage criterion is set to select proteins with a changeable significance level, and proteins significant at 0.05 levels but belonging to groups with a changeable significance level. This setting will pick a protein if it is of high significance, or if it is significant at the 0.05 significance level and in a highly significant group at stage II.

The proposed methods of multi-stage design for a clinical proteomic study have the completed components of a rigorous optimal multiple test procedure as defined by Storey. SA-b utilizes the biological group function which may ultimately improve the biological and clinical relevant discovery, when the grouping information is informative. The SA-a only using a single protein test also provide a tool for the sample size estimations for a multistage design when grouping information is not required, or is not informative.

Both the SA-a and SA-b utilize the nice property of the expected values for a summed probability. We know that it is also equivalent to the sum of the expected value of a probability. Since the probability of each protein being selected at the final stage is estimable, the sum of this probability across all proteins will be an estimate of the expected total number of true discoveries. This initiates the essential idea of the SA-a and SA-b.

The analytical approximation procedure for estimating the expected number of true discoveries of a three stage design in SA-b is proved to yield a similar solution as using the simulation function. The approximation decomposed the components of the probability of a protein being discovered in the three stage study. This approximation approach is an improvement for an optimization problem when the exact analytical function is not easily obtained and the simulation function takes a longer time to run. The nested simulated annealing method is also shown to speed up the optimization process.

The proposed method for a multi-stage clinical proteomic study can be generalized and used for any multi-stage studies that involve screening hundreds and thousands of candidates at early stages, and verification and validation at later stages.

7.3.4 The limitation of this method and future development

The proposed method for the multiple stage clinical proteomic design is suitable to use when the discovery study of the first stage has already been conducted. In the future, when the discovery technology becomes more stable, the multiple-stage design can include stage I sample size as one of the unknown design parameters with an approximated total number of proteins discovered at the first stage. Another future development will be to consider the correlation and hierarchical structure of the protein groups, and incorporate these structures in the objective functions.

Recently, Nomaa and Matsuib (2011) developed the ODP using Bayesian estimate under the same conceptual framework as Storey. Future optimality for the similar problem can also consider using the threshold function proposed by Storey (2007) in lemma 1, and adding known prior information for the mean of each protein and known weights according to their biological functions under the Bayesian framework.

7.4. Using Bayesian methods will be an advance for analysing proteomics studies

7.4.1 Review of what has happened since this PhD research started, and focusing on any new research that has emerged

A recent search using the term " 'statistical method' AND 'proteomics' ", and " 'statistical method' AND 'Mass spectrometry' " in SCOPUS produced 4 papers and two books. Oberg and Mahoney (2012) described a model for explaining the variations of the abundance data from protein, peptide, and experimental factors. Cox and Mann (2012) proposed an annotation enrichment method to integrate the proteomic data and other complementary high throughput devices. Two other papers are about reviewing the method for protein identification. Two books provided a general framework for quantitative proteomics and

basic statistical analytical methods for proteomic data. Utilizing Bayesian method to incorporate the missingness remains an advancing method for analysing proteomics studies.

7.4.2 The theoretical review and interpretation of the methodologies related to the proposed method

7.4.2.1 Multivariate multilevel mixed model

Multilevel models have been used in many disciplines and mostly in social science and public health studies in recent decades. An early advocate of multivariate multilevel model was Goldstain (Goldstain, 1995) who described a simple two-level multivariate multilevel model for students' examination results of writing and coursework. This formulation allows the multivariate normal distributed response even in unbalanced design for repeated measures and with random missing data. Since then some extensions of multivariate Gaussian responses to other mixed types of order, unordered distributed responses were also emerged under the multilevel framework. A number of different approaches have been explored for the estimations of parameters in these mixed type responses multilevel models, such as maximal likelihood estimate (ML), Expectation-Maximization (EM), and Markov Chain Monte Carlo(MCMC)(Goldstein et al., 2009).

Accompanying with these works, Goldstein et al. (2009) defined a unified framework which extended the univariate multilevel model to a multivariate model with mixed type responses of continuous Gaussian distributed, ordered or un-ordered categorical variables. This model is named as GCKL model in their study and is aimed to generalize to roughen data including missing values. Under the GCKL model framework, MCMC was the main approach for the derivations of the posterior distribution of unknown parameters. The MCMC algorithm is shown to be computationally efficient when there are a large number of unknown parameters involved, and feasible when there are non-random missing data with informative prior information (such as the probabilistic data linkage problem described in Goldstein's paper). Coincidentally, our approach described in chapter 4 is a special case in the GCKL framework, but has used different components for handling non-random missing data. Our multivariate multilevel model has peptide abundance as the response at level one and protein

abundance as the latent response at level two. The responses at both levels are multivariate Gaussian distributed. The GCKL framework used standard Gibbs sampling to derive the posterior distribution for the unknown parameters, including fixed coefficients, random coefficients, and the covariance matrix at both levels. Fixed and random parameters were sampled from a multivariate Gaussian distribution with uniform priors and the covariance matrices at both levels are sampled from Wishart distributions. When the response has missing values, the missing values were sampled from a multivariate Gaussian distribution based on the model parameters at each sampling iteration.

7.4.2.2 MCMC using Gibbs sampling and Hamiltonian using Non U Turn Sampling

Our proposed method uses MCMC and utilizing Gibbs sampling and Hamiltonian MC/Non U Turn Sampling. Compared to the standard Gibbs sampling, Hamiltonian MC avoids the random walk by introducing the leap frog function. It provides an alternative to approximating the solution on the continuous time scale from the solutions on the discrete time scale with a specified step size. The logarithmic posterior probability function was simulated by one of the paired partial differentiated equations, namely the Hamiltonian. Larger moving steps are generated from the leap frog scheme and this helps to improve the convergence compared to the random walk. It has been shown to have higher efficiency in sampling high dimensional correlated multivariate distributions.

7.4.3 The improvement in the current proposed method and how this method works

The multilevel model described in chapter 4 provided a new approach to analysing a group or all groups of proteins when the data has non-random missing values. The multivariate responses are the abundance for the multiple proteins and treated as the second level in the hierarchy, and are nested within individual subjects. The multivariate multi-level model allows the unbalanced design structure among the responses. This is an advantage in a proteomic study when the numbers of peptides are unequal across the protein. The experimental factors can be utilized altogether in the multiple proteins model for deriving the

results of a single protein. This improves the estimation for the variations across proteins and results in shrunk estimates for a single protein when it does not have many abundance observations. This multiple protein approach is considered to be more reliable than a single protein model when there is a lack of information for the single protein. It also enables us to identify the proteome wide difference, such as a systematic difference in the protein abundance between the two sampling times as described in the cardiac case of Chapter 5, while providing a separated predicted estimate for each protein at the mean time. In these specified models in the case studies, utilizing the relationship between the mass-to-charge ratio (m/z) and the intensity values has improved the discovery. A further relationship of the m/z ratio and the missingness also improved the reliability of the analysis.

When the non-random missingness needs to be considered and modelled, using numerical integration or EM will not be feasible. MCMC will still be considered as an efficient approach to derive the solutions for the unknown parameters in this proposed multivariate multilevel model. Using Hamiltonian NUTS is an improvement, especially in the case of utilizing non-standard distribution for handling the missing data. Compared to BUGS, RSTAN has more flexibility for handling missing data via the censored or truncated distribution function. In the “advanced use of BUGS language” section in the BUGS manual, it briefly introduces how the censoring and truncation being handled. The truncated distribution is dealt with by using the “zeros” and “ones” trick, but it is less efficient (as stated in the manual), as it produces high auto-correlation, poor convergence and high MC errors. The NUTS algorithm is more efficient in dealing with the truncated distribution and non-standard distribution. The user can define the censored distribution by integration for a normal distributed variable; the user can also define a non-standard distribution by specifying the probability density function.

The philosophy of the proposed multilevel multivariate method can be generalized to other proteomics or “omic” study with different experimental structures. Through separately defining the factors and covariates related to the experimental subjects and the experimental outputs such as the intensity of the molecule, the variations of the experimental structure can be clearly defined. The current method has been assessed in model with hundreds of proteins. It can be generalized to larger number of proteins but requires larger computing capacity.

When the protein number is large, analysing proteins in functional groups will reduce the total numbers of proteins.

7.4.4 The limitation of this method and future development

Using the proposed model, the protein level covariance matrix is assumed to be the same across proteins. The parameters for estimating protein to protein association also have not been incorporated yet. But the multilevel structure newly defined in Chapter 4 will enable us to make future improvements, such as 1) considering different covariance matrices for the protein level variables, and 2) adding known factors as the explanatory variables to estimate the components of the covariance matrices. When the number of proteins is large, the multiplicity adjustment will also need to be considered under the Bayesian framework.

Using the current algorithm, censoring data is different from missing; the way NUTS deals with missing will cause some difference in the final results when compared to the results of BUGS. The future version RHMC of RSTAN may improve the computing efficiency for high dimensional correlated proteomic data.

Sparse observations for some proteins and unbalanced design are shown to have their adverse impacts in the final estimation for the variance components in one of the case study. Currently, how these unfavourable factors influence the results is still unknown. A model with flexibility to cope with sparseness in the estimation will also be desirable.

7.5 Conclusion

This PhD research firstly proposed a new method to assess the reproducibility in clinical proteomic studies when a new device or new tissue is being used for a proteomic experiment. The reproducibility assessment utilizes a dimensional reduction technique and permutation approach to make the assessment extend to a proteome-wise scale. It secondly proposed two optimal design algorithms and realized them via a R package to assist the multiple stage study design through biomarker discovery to clinical utility. Finally, a multivariate multilevel

model has been proposed for the analysis of proteomic data when non-missing data is presented, and the method was tested and used in two real life clinical proteomic studies. The analytical method is shown to have large improvements and to gain statistical powers for assisting the discovery in a clinical proteomic study.

Bibliography

- Alder, B. J., & Wainwright, T. E. (1959). Studies in molecular dynamics. I. General method. *Journal of Chemical Physics*, 31, 459-466.
- Anderson, L. (2005). Candidate-based proteomics in the search for biomarkers of cardiovascular disease. *Journal of Physiology*, 563.1, 23-60.
- Anderson, M. J., & Legendre, P. (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of statistical computation and simulation*, 62, 271-303.
- Anderson, T. W. (1984). *An introduction to Multivariate Statistical Analysis* (2nd ed.). New York: Wiley.
- Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., & Adbersold, R. (2011). The quantitative proteome of a human cell line. *Molecular System Biology*, 7.
- BeJan, A. I. (2005). largest Eigenvalues and Sample covariance Matrices. Tracy-Widom and Painleve II: Computational Aspects and realization in S-Plus with applications. *Mathematics Subject Classification.*, 1991.
- Belisle, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithm on R^d *Journal of Applied Probability*, 29, 885-895.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistics Society*, 57, 289-300.
- Berry, K. J., Johnston, J. E., & Mielke, P. E. J. (2011). Permutation methods. *WIREs Computational Statistics*(3), 527-542. doi:10.1002/wics.177
- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1, 307-310.
- Boehm, A. M., Putz, S., Altenhofer, D., Sickmann, A., & Falk, M. (2007). Precise protein quantification based on peptide quantification using iTRAQ™. *BMC Bioinformatics*(8). doi:10.1186/1471-2105-8-214
- Breitwieser, F. P., Muller, A., Dayon, L., Kocher, T., Hainard, A., Pichler, P., Schmidt-Erfurth, U., Superti-Furga, G., Sanchez, J.-C., Mechtler, K., Bennett, K. L., & Colinge, J. (2011). General Statistical Modeling of Data from Protein Relative Expression Isobaric Tags. *Journal of Proteome Research*, 10, 2758-2766.
- Chapman, J. R. (Ed.). (1996). *Protein and peptide analysis by Mass spectrometry*. Totowa, New Jersey: Humana Press Inc. .
- Chen, J. J., Hsueh, H. M., Delongchamp, R. R., Lin, C. J., & Tsai, C. A. (2007). Reproducibility of microarray data: A further analysis of microarray quality control (MAQC) data *BMC Bioinformatics*, 8(SUPPL. 9) art. no. S20
- Chong, P. K., Gan, C. S., Pham, T. K., & Wright, P. C. (2006). Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: Implication of multiple injections *Journal of Proteome Research*, 5(5), 1232-1240.
- Chornoguz, O., Grmai, L., Sinha, P., Artemenko, K. A., Zubarev, R. A., & Ostrand-Rosenberg, S. (2010). Proteomic Pathway Analysis Reveals Inflammation Increases Myeloid-Derived Suppressor Cell Resistance to Apoptosis. *Molecular & Cellular Proteomics*, 10.3. doi:10.1074/mcp.M110.002980
- Corthals, G. L., & Rose, K. (2007). Quantitation in Proteomics. In *Proteome Research: Concepts, Technology and Application*. Berlin: Springer.
- Cox, J., & Mann, M. (2012). 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with

- complementary high-throughput data. *BMC Bioinformatics*, 13(Suppl 16:S12), S12.
- Creutz, M. (1988). Global Monte Carlo algorithms for many-fermion systems. *Physical Review D*, 38, 1228-1238.
- D'Ascenzo, M., Choe, I., & Lee, K. H. (2008). iTRAQPak: an R based analysis and visualization package for 8-plex isobaric protein expression data. *Briefings in Functional Genomics and Proteomics*, 7(2)(No 2.), 127-135.
- Dazard, J.-E., Saha, S., & Ewing, R. M. (2012). ROCS: a Reproducibility Index and Confidence Score for Interaction Proteomics Studies. *BMC Bioinformatics*, 13, 128-148.
- Fisher, R. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- . Free Fatty Acids in the Blood. *World of Sports Sciences*. from <http://www.faqs.org/sports-science/Fo-Ha/Free-Fatty-Acids-in-the-Blood.html>
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). Multivariate models. In *Bayesian data analysis* (pp. 407-419). U.S.A: Chapman & Hall.
- Goldstain, H. (1995). *Multilevel Statistical Models*. London;E. Arnold; New York: Halsted Press
- Goldstain, H. (1999). Multivariate Multilevel model. In *Multilevel statistics models* (pp. 5-6). London: Institute of Education.
- Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3), 173-197.
- Good, P. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses: A Practical Guide to Resampling Methods for Testing Hypotheses*.
- Greef, J. V. D., Martin, S., Juhasz, P., Adourian, A., Plasterer, T., Verheij, E. R., & McBurney, R. N. (2007). The art and practice of systems biology in Medicine: Mapping patterns of relationship. *Journal of proteome research*, 6, 1540-1558.
- Greenbaum, D., Colangelo, C., Williams, K., & Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology*, 4(117).
- Hajek, B. (1988). Cooling schedules for optimal annealing. *Math. Opera. Research*, 13, 311-329.
- Hale, J. E., Gelfanova, V., Ludwig, J. R., & Knierman, M. D. (2003). Application of proteomics for discovery of protein biomarkers. *Briefings in Functional Genomics and Proteomics*, 2(3)(3), 185-193.
- Hamdan, M., & Righetti, P. G. (2002). Modern strategies for protein quantification in proteome analysis: advantages and limitations. *Mass Spectrom Review*, 21(4), 287-302.
- Hill, E. G., Schwacke, J. H., Walters, S. C., Slate, E. H., Oberg, A. L., Eckel-Passow, J. E., Therneau, T. M., & Schey, K. L. (2008). A Statistical Model for iTRAQ Data Analysis. *Journal of Proteome Research*, 7, 3091-3101.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800-803.
- Hoffman, M. D., & Gelman, A. (2011). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of machine learning research*, 12.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70.
- Hoorn, E. J., Hoffert, J. D., & Knepper, M. A. (2005). Combined proteomics and pathways analysis of collecting duct reveals a protein regulatory network activated in vasopressin escape. *J Am Soc Nephrol.*, 16(10), 2852-2863.
- Hrydziuszko, O., & Viant, M. R. (2012). Missing values in mass spectrometry based metabolomics:an undervalued step in the data processing pipeline. *Metabolomics*, 8, S161–S174.

- Hwang, Y.-T., Kuo, H.-C., Wang, C.-C., & Lee, M. F. (2013). Estimating the number of true null hypotheses in multiple hypothesis testing. *Statistical Computing*. doi:<http://dx.doi.org/10.1007/s11222-013-9377-5>
- Issaq, H. J., Conrads, T. P., Prieto, D. A., Tirumalai, R., & Veenstra, T. D. (2003). SELDI-TOF MS for diagnostic proteomics. *Analytical Chemistry*, *April (1)*, 149-153.
- Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal component analysis. *The Annals of Statistics*, *29(2)(2)*, 295-327.
- Karoui, N. E. (2007). Tracy-Widom Limit for the largest Eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability*, *35(2)*, 663-714.
- Karp, N. A., Huber, W., Sadowski, P. G., Charles, P. D., Hester, S. V., & Lilley, K. S. (2010). Addressing Accuracy and Precision Issues in iTRAQ Quantitation. *Molecular & Cellular Proteomics*, *9*, 1885-1897.
- Kennedy, J., Abbatiello, S. E., Whiteaker, J. R., Lin, C., Kim, J., Zhang, Y., Wang, X., Ivey, R. G., Zhao, L., Min, H., Lee, Y., Yu, M.-H., Yang, E. G., Lee, C., Wang, P., Rodriguez, H., Kim, Y., & Carr, S. A. (2014). Demonstrating the feasibility of large-scale development of standardized assays to quantify human proteins. *Nature Methods*, *11*, 149-155.
- Kitamura, N., Akazawa, K., Miyashita, A., Kuwano, R., Toyabe, S.-i., Nakamura, J., Nakamura, N., Sato, T., & Hoque, M. A. (2009). Programs for calculating the statistical powers of detecting susceptibility genes in case-control studies based on multistage designs. *Bioinformatics*, *25*, no 2, 272-273.
- Kiyonami, R., Schlabach, T., & Miller, K. (2005). Identification and quantification of iTRAQ labeled peptides on the Finnigan LTQ using MS/MS and MS: Thermo electron coporation: application notes
- Kline, K. G., & Sussman, M. R. (2010). Protein quantitation using isotope-assisted mass spectrometry. *Annual Review of Biophysics*, *39*, 291-308.
- Kshirsagar, A. M. (1972). *Multivariate Analysis*. New York: Marcel Dekker.
- Leitner, A., & Lindner, W. (2004). Current chemical tagging strategies for proteome analysis by mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci*, *25(813)*, 1-26.
- Lin, A., & R.A.H.Stewart. (2011). Natriuretic Peptides in Severe Aortic Stenosis - Role in Predicting Outcomes and Assessment for Early Aortic Valve Replacement. In *Aortic Stenosis - Etiology, Pathophysiology and Treatment* (pp. 203-220): InTech.
- Luo, R., Colangelo, C. M., Sessa, W. C., & Zhao, H. (2009). Bayesian Analysis of iTRAQ Data with Nonrandom Missingness: Identification of Differentially Expressed Proteins. *Stat Biosci*, *1*, 228-245.
- Lyne, R., Burns, G., Mata, J., Penkett, C. J., Rustici, G., Chen, D., Langford, C., Vetrie, D., & Bähler, J. (2003). Whole-genome microarrays of fission yeast: Characteristics, accuracy, reproducibility, and processing of array data. *BMC Genomics*, *4*, art. no. 27
- MacCoss, M. J., & Matthews, D. E. (2005). Quantitative MS for proteomics: teaching a new dog old tricks. *Analytical Chemistry*, *77(15)*, 294A-302A.
- Maddalo, G., Petrucci, F., Iezzi, M., Pannellini, T., Del Boccio, P., Ciavardelli, D., Biroccio, A., Forli, F., Di Ilio, C., Ballone, E., Urbani, A., & Federici, G. (2005). Analytical assessment of MALDI-TOF Imaging Mass Spectrometry on thin histological samples. An insight in proteome investigation. *Clinica Chimica Acta*, *357(2)*, 210-218.
- Marida, K. V., Kent, J. T., & Bibby, J. M. (1980). *Multivariate Analysis*: Academic Press.
- Mcguire, N. J., Overgaard, J., & Pociot, F. (2008). Mass spectrometry is only one piece of the puzzle in clinical proteomics. *Briefings in Functional Genomics and Proteomics*, *7(1)(1)*, 74-83.

- McShane, L. M., Radmacher, M. D., Freidlin, B., Yu, R., Li, M. C., & Simon, R. (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, *18*(11)(11), 1462-1469.
- Meani, F., Pecorelli, S., Liotta, L., & Petricoin, E. F. (2009). Clinical application of proteomics in ovarian cancer prevention and treatment. *Molecular diagnosis therapy*, *13*(5), 297-311.
- Merciera, C., Truntzere, C., Pecqueur, D., Gimeno, J.-P., Belz, G., & Roy, P. (2009). Mixed-model of ANOVA for measurement reproducibility in proteomics. *Journal of proteomics*, *72*, 974-981.
- Moerkerke, B., & Goetghebuer, E. (2008). Optimal screening for promising genes in 2-stage designs. *Biostatistics*, *9*(4), 700-714.
- Muirhead, R. J. (1982). *Aspect of Multivariate Statistical Theory*. New York: Wiley.
- National Cancer Institute. (2007). *Building the Foundation for Clinical Cancer Proteomics Clinical proteomic technologies for cancer 2007 Annual Report*. Retrieved from <http://proteomics.cancer.gov/>.
- Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. U.S.A: Chapman & Hall/CRC.
- Neubert, K., & Brunner, E. (2007). A studentized permutation test for the non-parametric Behrens-Fisher problem [statistics]. *Computational Statistics and Data Analysis*, *51*(10), 5192-5204.
- Nikolaev, A. G., & Jacobson, S. H. (2010). Simulated Annealing. In J.-Y. P. M. Gendreau (Ed.), *Handbook of Metaheuristics*: Springer
- Nocedal, J., & Wright, S. J. (1999). *Numerical Optimization*. New York: Springer.
- Nomaa, H., & Matsuura, S. (2011). The optimal discovery procedure in multiple significance testing: an empirical Bayes approach. *Statistics in medicine*, *31*, 165-176.
- O'Rourke, S., Renfrew, D., & Soshnikov, A. (2011). *Fluctuations of Matrix Entries of Regular Functions of Sample Covariance Random Matrices*.
- Oberg, A. L., & Mahoney, D. W. (2012). Statistical methods for quantitative mass spectrometry proteomic experiments with labeling. *BMC Bioinformatics*, *13*(Suppl 16:S7).
- Onatski, A. (2008). The Tracy-Widom Limit for the largest Eigenvalues of Singular Complex Wishart Matrices. *The Annals of Applied Probability*, *18*(2)(2), 470-490.
- Palagi, P. M., Walther, D., Zimmermann-Ivol, C. G., & Appel, R. D. (2007). Proteome imaging. In *Proteomic research: concepts, technology and application* (pp. 123-144). Berlin: Springer.
- Palmblad, M., Tiss, A., Cramer, R. . (2009). Mass spectrometry in clinical proteomics - From the present to the future *Proteomics - Clinical Applications*, *3*(1), 6-17
- Park, J. H., Resnick, E. S., & Charlette, C.-R. (2012). Perspectives on common variable immune deficiency. *Ann N Y Acad Sci.*(Aug).
- Park, M. A., Li, L. T., Hagan, J. B., Maddox, D. E., & Abraham, R. S. (2008). Common Variable Immunodeficiency: a new look at an old disease. *Lancet*, *372*, 489-502.
- Patterson, S. D., Eyk, J. E. V., & Banks, R. E. (2010). Report from the Wellcome Trust/EBI "Perspectives in Clinical Proteomics" retreat- A strategy to implement Next-Generation Proteomic Analyses to the clinic for patient benefit: Pathway translation. *Proteomics Clin. Appl.*, *4*, 883-887.
- Paul, D., & Aue, A. (2013). Random matrix theory in statistics: A review. *Journal of statistical planning and inference*. doi:<http://dx.doi.org/10.1016/j.jspi.2013.09.005>
- Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, *6*, 559-572.
- Pelikan, R., & Bigbee, W. L. (2007). Intersession reproducibility of mass spectrometry profiles and its effect on accuracy of multivariate classification models. *Bioinformatics*, *23*(22)(22), 3065.

- Pelzing, M., & Neuss, C. (2005). Separation techniques hyphenated to electrospray-tandem mass spectrometry in proteomics: capillary electrophoresis versus nanoliquid chromatography. [Comparative Study]. *Electrophoresis*, 26(14), 2717-2728.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & Team, R. C. (2014). nlme: Linear and Nonlinear Mixed Effects Models.
- Rao, C. R. (1964) The Use and Interpretation of Principal Component Analysis in Applied Research. *The Indian Journal Of Statistics(SANKHYA)* (pp. 329–358).
- Rencher, A. C. (2002). *Methods of multivariate analysis* A JOHN WILEY & SONS, INC.
- Rodriguez, H., Snyder, M., Uhlén, M., Andrews, P., Beavis, R., Borchers, C., Chalkley, R., Cho, S., Cottingham, K., Dunn, M., Dylag, T., Edgar, R., Hare, P., Heck, A., Hirsch, R., Kennedy, K., Kolar, P., Kraus, H., Mallick, P., Nesvizhskii, A., Ping, P., Pontén, F., Yang, L., Yates, J., Stein, S., Hermjakob, H., Kinsinger, C., & Apweiler, R. (2009). Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: the Amsterdam principals. *Journal of Proteome Research*, 8(7), 3689-3692.
- Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S. P., Bartlet-Jones, M., He, F., Jacobson, A., & Pappin, D. J. (2004). Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, 3(12), 1154-1169.
- SAS Institute Inc. (2010). SAS Help Manual. . Cary, NC: SAS Institute Inc. .
- Satagopan, J. M., & Elston, R. C. (2003). Optimal two-stage Genotyping in population-based association studies. *Genetic Epidemiology*, 25(2), 149-157.
- Satagopan, J. M., Venkatraman, E. S., & Begg, C. B. (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics*, 60, 589-597.
- Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E., & Begg, C. B. (2002). Two-stage designs for gene-disease association studies. *Biometrics*, 58, 163-170.
- Sechi, S., & Oda, Y. (2003). Quantitative proteomics using mass spectrometry. *Current Opinion in Chemical Biology*, 7(1), 70-77.
- Semmes, O., Feng Z, Adam BL, Banez LL, Bigbee WL, Campos D, Cazares LH, Chan DW, Grizzle WE, Izbicka E, Kagan J, Malik G, McLerran D, Moul JW, Partin A, Prasanna P, Rosenzweig J, Sokoll LJ, Srivastava S, Srivastava S, Thompson I, Welsh MJ, White N, Winget M, Yasui Y, Zhang Z, Zhu L. (2005). Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. *Clinical Chemistry*, 51(1), 102-112.
- Seshi, B. (2006). An integrated approach to mapping the proteome of the human bone marrow stromal cell. *Proteomics*, 6(19), 5169-5182.
- Shadforth, I. P., Dunkley, T. P., Lilley, K. S., & Bessant, C. (2005). i-Tracker: for quantitative proteomics using iTRAQ. [Research Support, Non-U.S. Gov't]. *BMC Genomics*, 6, 145.
- Shaffer, J. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46, 561-584.
- Shen, Y., & Smith, R. D. (2005). Advanced nanoscale separations and mass spectrometry for sensitive high-throughput proteomics. *Expert Review of Proteomics*, 2(3), 431-447.
- Silva, J. C., Denny, R., Dorschel, C. A., Gorenstein, M., Kass, I. J., Li, G. Z., McKenna, T., Nold, M. J., Richardson, K., Young, P., & Geromanos, S. (2005). Quantitative proteomic analysis by accurate mass retention time pairs. *Analytical Chemistry*, 77(7), 2187-2200.
- Skol, A. D., Scott, L. J., Abecasis, G. R., & Boehnke, M. (2007). Optimal designs for two-stage genome-wide association studies. *Genetic Epidemiology*, 31(7), 776-788.
- Soric, B. (1989). Statistical discoveries and effect-size estimation. *Journal of American Statistical Association*(84), 608-610.

- Spencer, C. C. A., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLoS Genetics*, 5(5).
- Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS user manual.
- Stan Development Team. (2013). RStan: the R interface to Stan, Version 1.3.
- Steffens, B. (2010). Feasible and successful: Genome-wide interaction analysis involving all 1.9×10^{11} pair-wise interaction tests. *Human Heredity*, 69(4), 268-284.
- Stephens, A. N., Quach, P., & Harry, E. J. (2005). A streamlined approach to high-throughput proteomics. *Expert Review of Proteomics*, 2(2), 173-185.
- Storey, J. D. (2007). The Optimal Discovery Procedure: A New Approach to Simultaneous Significance Testing. *Journal of the Royal Statistical Society.*, 69, 347-368.
- Tao, T., & Vu, V. (2010b). Random matrices: universality of local eigenvalue statistics up to the edge. *Communications in Mathematical Physics*, 298, 549-572.
- Tao, T., & Vu, V. (2011). Random matrices: universality of local eigenvalue statistics. *Acta Mathematica*, 206, 127-204.
- Tracy, C., & Widom, H. (1996). On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177, 727-754.
- Vitzthum, F., Behrens, F., Anderson, N. L., & Shaw, J. H. (2005). Proteomics: from basic research to diagnostic application. A review of requirements & needs. [Review]. *Journal of Proteome Research*, 4(4), 1086-1097.
- Wang, H., Thomas, D. C., Pe'er, I., & Stram, D. O. (2006). Optimal two-stage genotyping designs for genome-wide association scans. *Genetic Epidemiology*, 30, 356-368.
- Wells, G., Prest, H., & IV, C. W. R. (2011). In I. Agilent Technologies (Ed.). U.S.A: Agilent Technologies, Inc. .
- Wheldon, M. C., Anderson, M. J., & Johnson, B. W. (2007). Identifying treatment effects in multi-channel measurements in electroencephalographic studies: multivariate permutations tests and multiple comparisons. *Australian and New Zealand Journal of statistics*, 49(4)(4), 397-413
- Whitford, D. (2005). An introduction to protein structure and function. In *Proteins structure and function*. USA: John Wiley & Sons, Ltd.
- Wiese, S., Reidegeld, K. A., Meyer, H., & Warscheid, B. (2007). Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics*(7), 340-350. doi:10.1002/pmic.200600422
- Williams, R., Chung, J., Ylaya, K., Whiteley, G., & Hewitt, S. M. (2010). Characterizations and validations of novel antibodies toward translational research. *Proteomics Alinical Application*, 4, 618-625. doi:10.1002/prca.200900186
- Zehetmayer, S., Bauer, P., & Posch, M. (2008). Optimized multi-stage designs controlling the false discovery or the family-wise error rate. *Statistics in medicine*, 27(21), 4145-4160.
- Zeng, I. S. L., Browning, S., Gladding, P., Jullig, M., Middleditch, M., & Stewart, R. A. H. (2009). A multi-feature reproducibility assessment of mass spectral data in clinical proteomic studies. *Clinical Proteomics*(5), 170-177.
- Zeng, I. S. L., Lumley, T., Ruggiero, K., Middleditch, M., Woon, S.-T., & Stewart, R. A. H. (2013). Two optimization strategies of multi-stage design in clinical proteomic studies. *Statistical Applications in Genetics and Molecular Biology*, 12(2), 263-283.
- Zhanhua, C., Gan, J. G., Lei, L., Mathura, V. S., Sakharkar, M. K., & Kanguane, P. (2005). Identification of critical heterodimer protein interface parameters by multi-dimensional scaling in euclidian space. *Frontiers in Bioscience*, 10, 844-852.

- Zotenko, E., Guimarães, K. S., Jothi, R., & Przytycka, T. M. (2006). Decomposition of overlapping protein complexes: A graph theoretical method for analyzing static and dynamic protein associations. *Algorithms for Molecular Biology*, 1(7).
- Zou, Y., Zou, G., & Zhao, H. (2006). Two-stage designs in case-control association analysis. *Genetics*, 173(1747-1760).
- Zuo, Y., Zou, G., Wang, J., Zhao, H., & Liang, H. (2008). Optimal two-stage design for case-control association analysis incorporating genotyping errors. *Annals of Human Genetics*, 72 (3)(3), 375-387