Statistical Analysis of Genomic Assays in Complex Study Designs

by

Elham Khodayari Moez

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Epidemiology

School of Public Health
University of Alberta

Abstract

Human genomic data are being generated at an increasing rate owing to the advancement of high-throughput technology. Wider availability of genomics, transcriptomics, proteomics and metabolomics data motivated complex study questions with the intention to gain higher degree of understanding of system biology. These study questions inspired novel study designs and demanded compelling statistical analysis. Although beneficial for understanding the disease progression, recently-proposed directions of integrative and longitudinal analysis of multiple omics call for advanced statistical methods.

Phenotypes are not determined by merely presence of single or few genes, but by the interconnection of many genes and their downstream pathways. The regulation of human genome at multiple levels may be revealed by integrative analysis of omics and helps the establishment of personalized clinical practices. In our study of prostate cancer, tumor and healthy samples manifested the differential interdependency of oncogene expressions (MYC and AKT1) and metabolite pathways. We showed the inability of classic statistical analysis approaches to deal with this complex design and offered Linear Combination Test (LCT) as a solution for linking genomics and metabolomics, working directly with multiple continuous and correlated measurements.

Despite promoting an insight into the temporal progression of the disease and providing more accurate data, the longitudinal design of genetic studies is out of reach for scientists, due to lack of adequate statistical methods that accounts for the within-subject correlation. In this thesis, a Longitudinal Linear Combination Test (LLCT), a self-contained gene set analysis method, is proposed to detect the genes which are differentially expressed in association with different

trajectories of one or multiple phenotypes. LLCT is a high-dimensional data analysis method applicable to a wide range of longitudinal omics data. It allows adjusting for potentially time-dependent covariates and works well with unbalanced and incomplete data. An extension of LLCT is applicable to family-based data with an additional layer of correlation between subjects. The reasonable performance of LLCT for different sample sizes, gene set sizes, number of follow-up visits, within-gene-set correlation and within-subject correlation and the outperformance of LLCT compared to other methods were demonstrated in simulation studies. The application study illustrated the adequacy of LLCT to detect genes whose differential expression significantly alters the dynamic of blood pressure in related and unrelated datasets. We also proposed a generalization of LLCT that can handle time-course omics datasets

Efforts to investigate the genomic network may be wasted by poorly designed studies and inappropriate analytical tools. The success of genetic investigations depends on the development of comprehensive analysis methods appropriate for complex studies, designed to minimize the potential error and biases in the hope of achieving a greater level of consistency among the study findings.

Preface

This thesis is the original final work of my PhD study with supervision of Dr. Irina Dinu and co-supervision of Dr. Jeffrey L. Andrews, conducted at School of Public Health of University of Alberta. The literature review in chapter 1 and concluding chapter 4 are my original works.

The chapter 2 of this dissertation has been submitted as a manuscript and it is at the second round of revisions: "Khodayari Moez E., Pyne S., Dinu I., Association Between Bivariate Expression of Key Oncogenes and The Metabolic Phenotypes of Patients with Prostate Cancer, Computers in Biology and Medicine". I was responsible for literature review, data analysis, interpretations of the results and preparation of the first draft of manuscript. Dr. Saumyadipta Pyne identified the need for a better method to analyze the complex data, contributed in acquisition of the data and interpretation of findings. Dr. Irina Dinu contributed in the concept formation and design of the study as well as analysis and interpretation of the findings. Dr. Saumyadipta Pyne and Dr. Irina Dinu reviewed the manuscript and provided critical comments.

Chapter 3 of this thesis is being submitted as Khodayari Moez E., Andrews JL., Dinu I., Longitudinal Linear Combination Test for Gene set Analysis of Longitudinal Data. I was responsible for implementation of the project, analysis of application data, design and analysis of simulation data, interpretation of results and preparing the first draft of manuscript. Dr. Irina Dinu identified the need for a new method to analyze longitudinal data measured by biotechnologies, provided the project outline and she made substantial contribution in revision of manuscript. Dr. Jeffrey Andrews provided key advices about the conduct of project and manuscript preparation. The data we used to illustrate the method is part of the Genetic Analysis Workshop 19. GAWs are supported by NIH grant R01 GM031575. The GAW19

exome and whole genome sequence data were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. Additional genetic and phenotypic data for GAW19 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants R01 HL0113323, P01 HL045222, R01 DK047482, and R01 DK053889. Additional Starr County genotype and phenotype data were supported by NIH grants R01 DK073541 and R01 HL102830. The VAGES study was supported by a Veterans Administration Epidemiologic grant. The FIND-SA study was supported by NIH grant U01 DK57295.

To God for walking beside me on this journey

&

To my parents who guided me to where I am today

&

To my supportive husband, Navid

Acknowledgement

This work has not been possible without the help and support of many kind people around me and I would like to express my appreciation to them here.

First of all, I owe my deepest gratitude to my supervisor, Dr. Irina Dinu. Her overwhelming supports and understanding, her patient guidance and dedication was always stimulating for me. I have learnt a lot from her genius and invaluable advices in both research and teaching. Without her persistent support and encouragement, this thesis would not have been possible.

I would like to sincerely thank my co-supervisor, Dr. Jeffrey Andrews for his support, guidance and constructive feedbacks throughout the program.

Also, I would like to express my great appreciation to Prof. Saumyadipta Pyne for sharing his expertise and his clever and creative insights. His mentorship has been a significant contribution to this thesis and my program.

I am deeply grateful to Prof. Anita Kozyrskyj for providing me the opportunity to work with her expert team. Her constructive comments, thoughtful questions and clever advice, as well as her constant support were so inspiring for me. What I have learnt from her will light up my future career path.

My thanks and appreciations also go to my colleagues and friends at School of Public Health who have willingly helped me out in this journey.

I am deeply indebted to my parents; my brother and my sister whose sacrifice and unconditional support have helped me achieve my current level of success. I have no word to

thank enough my husband for his support and understanding and his endless supplies of love and encouragement.

Thank you all for your unwavering support.

TABLE OF CONTENTS

Chapter 1 **Introduction** 1

1.1 Review of High Dimensional Data Analysis 1

1.2 Review of Microarray Data Analysis 2

1.3 The Demand for Linkage Analysis of Omics 6

1.4 Omics Data Analysis with Longitudinal Phenotype 7

1.5 Application on the Data of Genetic Analysis Workshop 19 8

1.6 Time-Course Omics Data Analysis 9

1.7 Application on Microbiome 12

Chapter 2 **Association between Bivariate Expression of Key Oncogenes and the Metabolic Phenotypes of Patients with Prostate Cancer** 13

2.1 Abstract 13

2.2 Background 14

2.3 Methods 19

2.3.1 Collecting Data 19

2.3.2 Statistical Method 20

2.3.3 Simulation Study Design 22

2.4 Results 23

2.4.1 Application 23

2.4.2 Simulation 27

2.5 Discussion 31

2.6 Conclusions 35

2.7 Summary 35

Chapter 3 **Longitudinal Linear Combination Test for Gene set Analysis** 36

3.1 Abstract 36

3.2 Introduction 37

3.3 Method 41

3.3.1 Longitudinal Linear Combination Test (LLCT) 41

3.3.2 Generalization 1: LLCT for Family-Based Data 45

3.3.3 Generalization 2: Time-Course Microarray Data Analysis 48

3.3.4 Design of Simulation Study 50

3.4 Results....52
3.4.1 Simulation Study....52
3.4.2 Application....59
3.5 Discussion....88
3.6 Conclusion....91
Chapter 4 **Discussion**....93
4.1 The demand for Novel Statistical Methods....93
4.2 Strengths....97
4.3 Limitations....99
4.4 Conclusions and Public Health Implications....100
4.5 Future Directions....101
4.6 Software Package....102
References....103
Appendices....117

LIST OF TABLES

Table 2.1 The comparison of normal and cancer cells about the association between oncogenes (bivariate or univariate MYC and AKT1) and different metabolite sets. The data is analyzed using LCT. 25
Table 3.1 Summary information (mean (standard deviation)) of covariates and outcomes at different time points: GAW19 application, studies of related and unrelated subjects 61
Table 3.2 The number of significant gene sets found by LLCT at different levels of confidence, testing a variety of outcomes and datasets 62
Table 3.3 Results of LLCT of association between the expressions of different gene sets and various measures of blood pressure for UNRELATED subjects in GAW19 database 66
Table 3.4 Results of LLCT of association between the expressions of different gene sets and various measures of blood pressure for RELATED subjects in GAW19 database 69

LIST OF FIGURES

Figure 2.1 An illustration of scenarios derived from real data where LCT can improve analysis of metabolites and oncogenes associations. 19
Figure 2.2 Cluster analysis of subjects based on significant (A and B), and insignificant (C) metabolite sets signatures: dendrogram of subjects clustered based on their metabolite set signature, and scatterplots of AKT1 vs. MYC with cluster-specified observations. 27
Figure 2.3 Power comparison between GSEA and LCT analyses in the presence of different metabolite set sizes 28
Figure 2.4 Power comparison between GSEA and LCT analyses in the presence of different number of metabolite sets 28
Figure 2.5 Power comparison between GSEA and LCT analyses in the presence of different sample size 29
Figure 2.6 Power comparison between GSEA and LCT analyses in the presence of different between-metabolite-set correlation 29
Figure 2.7 Power comparison between GSEA and LCT analyses in the presence of proportion of metabolite sets consisting significant metabolites (significant metabolite sets) 30
Figure 2.8 Power comparison between GSEA and LCT analyses in the presence of proportion of the metabolites with moderate to strong association with phenotype within significant metabolite sets 30
Figure 3.1 Calculation of the power of LLCT using simulated data generated with different within-gene set correlation. Type I error is set at 5%. For each plot, the simulation variables except the one mentioned on the title varies but remains comparable among the curves. 52
Figure 3.2 Calculation of the power of LLCT using simulated data generated with different sample size. Type I error is set at 5%. For each plot, the simulation variables except the one mentioned on the title varies but remains comparable among the curves. 53
Figure 3.3 Calculation of the power of LLCT using simulated data generated with different gene set size. Type I error is set at 5%. For each plot, the simulation variables except the one mentioned on the title varies but remains comparable among the curves. 53
Figure 3.4 Calculation of the power of LLCT using simulated data generated with different number of repeated. Type I error is set at 5%. For each plot, the simulation variables except the one mentioned on the title varies but remains comparable among the curves. 54
Figure 3.5 Calculation of the power of LLCT using simulated data generated with different within-subject correlation. Type I error is set at 5%. For each plot, the simulation variables except the one mentioned on the title varies but remains comparable among the curves. 54
Figure 3.6 Comparison of powers of LLCT method and the method of pathway analysis via regression (PAVR) proposed by Adewale et a1, using simulated data generated with different within-gene-set correlation. **B1** denotes the gene effect and **B3** denotes the gene effect over time referring to equation 3.14. 56
Figure 3.7 Comparison of powers of LLCT method and the method of pathway analysis via regression (PAVR) proposed by Adewale et a1, using simulated data generated with different sample size. **B1** denotes the gene effect and **B3** denotes the gene effect over time referring to equation 3.14. 57
Figure 3.8 Comparison of powers of LLCT method and the method of pathway analysis via regression (PAVR) proposed by Adewale et a1, using simulated data generated with different gene set size. **B1** denotes the gene effect and **B3** denotes the gene effect over time referring to equation 3.14. 58

Figure 3.9 Comparison of powers of LLCT method and the method of pathway analysis via regression (PAVR) proposed by Adewale et a1, using simulated data generated with different number of repeated measurements. **B1** denotes the gene effect and **B3** denotes the gene effect over time referring to equation 3.14..... 59

Chapter 1
Introduction

1.1 Review of High Dimensional Data Analysis

Advances in high-throughput technology which allows for the exact and simultaneous measurements of thousands of gene expressions, proteins and metabolites are calling for novel statistical methods, specifically tools for analysis of high-dimensional diverse biological data. In case of high-dimensionality where the number of variables recorded for each sample(p) exceeds the sample size(n): p>>n, the classical statistical methods are unable to analyze the data. With this setting, the deviation from fundamental assumptions of statistical methods, e.g. the Law of Large Numbers and the Central Limit Theorem[1], the singularity and ill-condition matrices, unidentifiable distributions and low computational efficiency are important challenges facing the data analysis.

Statisticians primarily proposed dimension reduction methods in dealing with high-dimensional data. These methods vary from explanatory approaches, visualizing the possible associations and relationships, to hypothesis-testing approaches, testing the validity of observations[2]. Among the explanatory methods, clustering analysis proceeded with Principle Component Analysis (PCA) or Principle Coordinate Analysis (PCoA). These approaches have received much attention due to their graphical features[3]. PCA and PCoA de-noise the dataset and decrease the dimension. Clustering analysis such as k-means clustering with Euclidean distance and hierarchical clustering are frequently used data analysis approaches in biological studies[4]. Penalized likelihood regression[5] is a popular regression-based method focused on dimension reduction. Later, Least Absolute Shrinkage and Selection Operator (LASSO) [6] was

developed to force models to be sparse by setting some effect sizes to be zero and, therefore, reduce the dimension. When the number of predictors far exceeds the sample size, it is not unrealistic to assume that some predictors do not contribute to describe the response, and therefore, the vector of effect sizes is likely to be sparse. Once the direct estimation of the effect sizes for high-dimensional data through the classical methods is impossible, these methods impose some restrictions to make the effect sizes identifiable. Despite the popularity and efficiency in sparse estimation, LASSO is not appropriate for inference due to difficulty in characterization of the estimators' distributions[7]. Among the solutions proposed [8–10], one instructs splitting the sample to two sub-samples with equal sizes, applying LASSO to the first half and using the selected variables for ordinary least square analysis of the second half and, thereby, providing inferential analysis[8]. This method, is sensitive to the sub-sample selection and the p-values may vary substantially from one sub-sample to the other. To solve this problem, an iterative approach was presented by Fan and Lv[11]. Alternative method to find the estimations of coefficients is Empirical Bayesian Method which considers a priori for the distribution of coefficients with a variance depending on an unknown variance. Alternatively, the method of Global Test [12] employed a hierarchical modelling approach to transform high-dimensional data to a low-dimensional data. In this method, the units of analysis are considered to be genes which are nested within subjects. As such, the p-dimensional data is turned to an n-dimensional data and becomes more computationally efficient.

1.2 Review of Microarray Data Analysis

Correlation learning methods play a very important role in hypothesis testing of high-dimensional data. These methods are very popular in DNA Microarray data analysis, helping

researchers detect differential gene expressions across two or more than two biological states, e.g. cases and controls. Phenotype-genotypes correlations are ranked and the most influential genotypes are detected[11]. In the case of binary phenotype, this method relies on the t-test statistics. Tibshirani et al. [13] used this approach to identify the subset of genes that optimally describe each state. Similarly, the commonly-used method, Significance Analysis of Microarrays (SAM) proposed by Tusher et al. [14], ranked the genes using a score similar to a t-statistic and determined the significant genes with the scores larger than a specific threshold. There are many more of these methods in the literature, some of them are discussed in the next paragraph.

Correlation-based approaches constitute a large body of developed methods for analysis of DNA Microarray data. The initial attempts to analyze DNA microarray data were to investigate the effect of each gene, separately. In this so-called Individual Gene Analysis (IGA), the genes are assumed to express independently. This assumption is against the biological concept of genetic linkage and may lead to faulty results. While IGA always detects the genes with strongest effect, accumulation of the mild or moderate effects of multiple genes may in fact determine the phenotypic condition. IGA, usually, ends up in a long list of significant genes which may be difficult to interpret biologically. Quite often, a disease does not involve a single gene, but multiple genes, sharing a common biological function. Such collections of genes are called biological pathways [15]. This concept motivated a shift from analysis at a univariate gene level, or IGA to analysis at a multivariate level, or Gene set analysis (GSA). GSA wasinitially suggested by Mootha et al. in 2003 [15] and Subramanian et al.in 2005 [16]. GSA brought up the possibility to account for within-gene-set correlations and produced results with better replication across studies. GSA focuses on analysis of biological pathways. The Cancer

Genome Atlas (TCGA) [17], Gene Expression Omnibus (GEO) [18], Kyoto Encyclopedia of Genes and Genomes (KEGG)[19], BioCarta [20], Molecular Signature Database of the Broad Institute [20] provide an archive for these a-priori defined gene sets. Gene Set Enrichment Analysis (GSEA)[16], the most popular GSA method, relies on a correlation-based approach. It utilizes the genotype-phenotype correlation to rank the genes and examines the significance of a given gene set by locating its members in the ranked list. The degree to which a gene set is overrepresented in the extremes is calculated by Kolmogorov-Smirnov-like Enrichment Score and used to indicate if the gene set is differentially expressed in association with phenotype. Later, Efron and Tibshirani[21] proposed maxmean statistic as a substitute for GSEA Enrichment Score with superior power characteristics. Significance Analysis of Function and Expression (SAFE)[22], as the other commonly used method, compares the measures of the association between the genes inside and outside of a given gene set. A considerable difference determines a significant effect. SAM became a popular IGA method, soon after it was published. SAM was generalized to SAM-GS [24] to analyze multiple genes, organized in biological pathways. The SAM-GS statistic is derived from the summation of SAM statistic for all the genes within a gene set.

Due to the difficulty in parametric determination of the test statistic distributions, these methods rely on permutation-based inferential analysis. Under the null hypothesis of no association, the labels of the study units are interchangeable. The distribution of the test statistic is approximated by random permutations of the labels. This approach assumes independency of the study units. GSA methods, such as GSEA, are termed “competitive” methods[23] if they employs *gene* permutation to test whether the association between a gene set and the outcome is equal to those of the other genes. Competitive methods have been criticized about their

untenable statistical independence assumption across genes. Ignorance of the correlation causes overstating of statistical significance[24]. Obviously, the results of competitive methods could be severely subjective regarding the choice of gene sets to enrol in the study. On the other side, there is the category of so-called "self-contained" methods[23] where a *subject* permutation method is employed to test if there is no gene within the gene set associated with the phenotype. Global test[12], SAFE[22] and SAM-GS[25] belong to this category.

Analyzing the associations between multiple genes-sets and phenotypes requires an adjustment for multiple testing. False Discovery Rate (FDR) [26] is one of the most known tools for estimating the multiple test error and provides a good alternative to the traditional Bonferroni approach. In this thesis, we used FDR adjusted p-value, also known as q-value, which is the proportion of false significant tests.

While there is a sufficient literature on development of methods for association of microarray data and binary or categorical phenotype, researchers believe that many biological variables (e.g. Blood Pressure and Cholesterol Level) are associated with the disease under study, in a continuous fashion[27]. As such, assigning any threshold to meet the requirement of an inappropriate analytical method may be very misleading in the genomic studies. The recently-developed method of Linear Combination Test (LCT), which is described in detail throughout the thesis, is unique in its ability to handle one or multiple continuous phenotypes despite its low computational cost. As this method brings lots of flexibility in terms of sample size, correlation structure, gene set size, proportion of missing values and number of phenotypes, it is a very good candidate for methodology developments for analysis of complex data. In this thesis, we took advantage of favorable features of LCT to link different omics measurements and investigate the temporal patterns of phenotype or omics data.

1.3 The Demand for Linkage Analysis of Omics

As discussed before, the complexity of the research questions is directly related to the technological advancements. With increased availability of microarrays and high throughput sequencing of biological data, system biology emerged as a new and promising research area. System biology is a term assigned to mathematically understanding the interactions between biological networks. At the molecular level, understanding collaborative functioning of the DNA or genome, the RNA or transcriptome, the proteins or proteome and the metabolite profiles or metabolome in developing a phenotype is the focus of biological system studies. In general, the connection between genotype and phenotype is known to depend on how genes alone, or in combination with other genes, are expressed to messenger RNA and act through the proteins and metabolites. The importance of the study of complex biological system mediating the effects of DNA diversities on the phenotype has been emphasized by many authors [28] and should rely on a sophisticated analysis method. Inadequate statistical methods are frequently applied to handle challenging properties of this data, specifically the complex correlation structure. The second chapter of this thesis proposes a statistical method for investigation of system biology, which imposes minimal limitations on the analysis, compared to the commonly used methods. Previously, researchers dealt with the analytical limitations by employing clustering methods to dichotomize the continuous measurements of omics data, or simply ignore possible correlation structures. Beside the high dimensionality of data, we think that the correlation or even interaction structure within each genomic assay is the main challenge of combining several different omics to discover meaningful biological signatures. LCT brings great flexibility to accommodate these complexities.

In chapter 2, we linked metabolome and transcriptome in healthy and prostate cancer subjects, separately. Using LCT, we addressed the within-metabolome correlation present in different biological pathways, the within-transcriptome correlation and the within-transcriptome interaction, high dimensionality of the metabolome data and small sample size.

1.4 Omics Data Analysis with Longitudinal Phenotype

The differential gene expression profiles may not only indicate the phenotype value at a specific time point, but the variation of phenotype over time. Motivated by the efficiency of longitudinal designs in enhancing the understanding of temporal progression of a phenotype and considering the feasibility of such studies by the recent advances in technology, longitudinal designs are becoming increasingly popular in genetic studies. Despite a rich literature on analysis of binary or categorical phenotype variables, there are few methods developed to handle a continuous phenotype and fewer to handle repeatedly measured phenotypes. When measuring a single subject repeatedly over time, the correlation between these measurements imposes a new layer of complexity to the data analysis. This analytical challenge was addressed by a correlation learning method developed by Adewale et al. [29]. In this method, the summation of the gene-specific regression coefficients estimated by a longitudinal model formed the enrichment score of a given gene set. This method assumed that the effect of gene expression is constant over time which may not always hold in the presence of complex molecular systems. Thus, novel approaches for providing longitudinal data analyses for omics studies are needed. We developed Longitudinal Linear Combination Test (LLCT) to detect differentially expressed gene sets associated with temporal trajectory of one or multiple phenotypes. LLCT is a two-step approach for explaining both within and between subject

variations. We believe and will show in chapter 3 that our method accounts for all the above-mentioned complexities in correlation structure, at very low computational cost.

1.5 Application on the Data of Genetic Analysis Workshop 19

Starting from 1982, Genetic Analysis Workshops provided a forum for developing novel statistical methods and comparing existing methods for identification of genetic factors on complex diseases. GAW13, GAW16, GAW18 and GAW19 focused on analysis of longitudinal phenotypes. In GAW13, the joint and two-step models were applied on a genome data including 2885 subjects with the phenotypes measured as frequently as 21 times. Since the sample was large enough, classic analysis methods such as random effect modelling was applied[30]. The sample size for GAW16 was also large enough to practice more advanced classic methods of linear mixed models, generalized estimating equations, growth modeling and multivariate adaptive splines to analyze longitudinal data. GAW16 acknowledged the additional information provided by longitudinal design[31]. More recently developed methods such as Bayesian methods or LASSO were employed in GAW18. These studies found it difficult to tackle the challenges of missing data and high-dimensionality[32]. In all GAW18 studies (except one[33]), they failed to analyze the whole dataset because of the inability of the methods to handle high-dimensionality. However, they admitted the increased power gained by repeated measurements of the phenotype[34]. Similar to the previous GAWs, GAW19 experienced inconsistent and incomparable findings due to heterogeneity in methodology and data [35]. In chapter 3, we present results of our proposed method on GAW19 data.

Since many genetic studies are family-based, GAWs also attempted to discuss the applicability of their proposed methods on family-based data. With this data structure,

additional statistical consideration is required to account for the within-family correlations. The pedigree-based genetic studies benefit from decreased heterogeneity among subjects and increased power and control of Type I error[34]. However, this study design comes with a complex data correlation structure. In chapter 3, we will show the applicability of LLCT on the data with pedigree-based setting.

1.6 Time-Course Omics Data Analysis

Gene expression is a time-dependent process. Although examination of the gene expression at a single time point may reveal some genes contributing to the development of a health condition, a time-course evaluation is required to recognise all the genes involved[36]. Moreover, time-course omics studies increase our understanding of the dynamic of biological processes. As such, these studies are becoming very popular in recent microarray literature. Large number of genes, small number of replicates, small sample size, and high rate of missing values are common analytical challenges of these studies. While small number of replicates and missing values prevent classical time-series analysis methods from functioning, large number of genes and small sample sizes present important challenges for classical longitudinal methods. In contrast to longitudinal phenotype methods, there is a large body of literature developing and reviewing the time-course microarray methods[36–38]. The proposed methods can be divided into two main categories, according to their objectives: 1. Identification of group of genes with a similar temporal expression fashion, and 2. Examination of differential time-course expression patterns corresponding to different biological conditions. Besides the primary goal of the first class which is learning from grouping the genes, these methods help future studies select an optimal subset from the entire gene set. Clustering approaches are very popular in this

category. Distance-based clustering[39,40], splines and Hidden Markov Models (HMM) are some attempts for temporal pattern recognition. HMM results in high quality clustering. In contrast to popular clustering approaches, HMM considers the temporal association of gene expression measurements[41]. Orthogonal polynomials and splines may also help explaining the non-linear variations over time[36].

In the hypothesis testing category, there has been a rapid advance in developing the methods for identification of the genes differentially expressed over time in association with a biological condition. A review by Ruan and Yuan[37] classified these methods into three groups corresponding to "static experiments methodology", "smoothing methods" or "time-series analysis approaches". As an example, an extension of ANOVA for testing the differential expression over time was developed and took advantage of permutation method for tackling the deviation of independent observation assumption[42]. This method suffers from low sensitivity due to treating the time variable as a qualitative variable and therefore disregarding the temporal order of measurements. In the second group, there are various methods proposed based on linear combinations of B-spline basis function which help compare the temporal patterns of gene expression among different classes of subjects[43]. A method developed by Bar-Joseph et al.[44] calculated the difference between two times series related to two biological conditions, previously smoothed by B-splines. This method fails to detect the similarity of the curves in case one curve is a noisy realisation of the other one. This method was also criticized for improper handling of short time-course data[38]. Storey et al. estimated the spline curves under the null and alternative hypotheses and defined an F-test to compare their goodness of fit [45]. In the third category, gene expressions are analyzed as outputs of autoregressive process [46].

Besides the categories defined by the review of Ruan and Yuan[37], the regression-based approaches may constitute a separate category. The two-step method of maSigPro[47] fits each gene separately against the experimental group and time variables and identifies the significant genes at the first step. This procedure is followed by a stepwise selection for finding the condition for which significant genes are expressing differentially. In a recently-proposed method [48], the linear mixed models were fitted under null and alternative hypotheses and LR test was employed to compare them.

The methods developed for temporal pattern discovery of the genes may also be applicable for hypothesis testing after some modification. Researchers utilized PCA to compare the fundamental patterns between two or multiple biological conditions[49,50]. Likewise, Hidden Markov Models helped detecting the genes with differential expressions. Assuming that the expression profile of each gene can be described by a Markov Chain, the most likely configuration of the states can be estimated by EM algorithm. These gene-specific probabilities are then compared to find the differential expressions[51].

As described above, most existing methods analyze binary or categorical biological conditions. Hence, there is a need for development of a method to detect differentially expressed genes over time in association with a continuous biological condition. LLCT, which is primarily developed to analyze longitudinal phenotypes, is generalized to address this analytical gap in chapter 3. LLCT is derived from LLC which was developed for static experiments and thus classified into the first group reviewed by Ruan and Yuan[37].

1.7 Application on Microbiome

Outside the genomics, there are many other scientific fields which can benefit from advances in high-throughput analytical tools. Microbiome, which carries distinct DNA signatures, is sequenced via high-throughput technology and is measured in huge number and diversity. This is one of the potential areas of application for the methods reviewed in this section. Human microbiome is characterised by Operational Taxonomic Units (OTU) which are nearly identical 16S rRNA-tagged sequences. OTUs, like genes, can be clustered to different taxonomic classes, based on their shared characteristics. The effect of human microbiome, specifically gut microbiota, on development of a variety of diseases is under active investigation. The temporal pattern of gut microbiota acquisition, which starts at birth, is introduced as a possible risk factor, and should be studied longitudinally. The application of novel time-course microarray analysis methods, such as LLCT, may benefit microbiota studies.

Chapter 2

Association between Bivariate Expression of Key Oncogenes and the Metabolic Phenotypes of Patients with Prostate Cancer

2.1 Abstract

Background: AKT and MYC are two of the most prevalent oncogenes associated with prostate cancer. The precise effects of these two key oncogenes' overexpression on the regulation of metabolic pathways in Prostate Cancer are under active investigation. However, few studies have looked into the joint effects on the Prostate Cancer patients' metabolic phenotypes in terms of their bivariate oncogene-pair expressions. This is primarily due to the lack of a suitable statistical method to analyse the data in the presence of the interaction between oncogenes and within-metabolite set correlation.

Methods: We analysed data on the expressions of phosphorylated AKT1 and MYC and concentrations of 228 metabolites from 60 human prostate tumor samples and 16 normal tissue samples. The metabolomics data allowed us to study not only the measurement of individual metabolites, which can exhibit a dynamic range, but the enriched phenotypes in terms of "metabolite sets" that come from known metabolic pathways. We studied 71 metabolite sets defined by KEGG annotation. We used generalized Linear Combination Test (LCT) for multiple continuous outcomes to find associations between metabolite sets and oncogenic expressions, after accounting for the correlation between AKT1 and MYC expressions and

correlation between the metabolites in a metabolite set. LCT performance was evaluated using a simulation study.

Results: Through a comprehensive analytical method, our study linked onco-genomics and metabolomics data from the patients, to gain better understanding of the inter-connected mechanisms underlying prostate cancer. This study showed that dysregulations of AKT1 and MYC significantly alter the metabolic pathways activated by non-glucose nutrient sources and their downstream. Our findings highlighted the role of MYC as the leading, but not the only, oncogene in prostate oncogenesis. In our simulation study, LCT performed better than the known alternative method, Gene-Set Enrichment Analysis (GSEA).

Conclusions: Overall, our study offers a solution for linking genomics and metabolomics, working directly with multiple continuous and correlated measurements.

2.2 Background

Prostate cancer (PCa) remains a leading cause of death in men [52], accounting for 6.6% (307,000 deaths) of male deaths and 14.8% (1.1 million cases) of male cancer incidence worldwide in 2012 [53]. The prostate cancer incidence rate increases by age and the risk of death caused by prostate cancer is higher in less developed countries and the highest in predominantly black population [54,55]. There is a high demand for biomarkers to detect prostate cancer in early stages[56]. The use of Prostate-Specific Antigen (PSA) as a prostate cancer biomarker has been criticized for low sensitivity and specificity[57–60]. Early diagnosis of prostate cancer is critical because prostate cancer is prevalently asymptomatic, unless it is in an advanced stage, and a proper therapy cannot be successfully administered unless the disease

is at an early stage (early detection increases the 5-year survival rate from 1% to about 50%) [61,62].

Encouraged by the recent advances in metabolomics technologies and motivated by findings that cancer alters cellular metabolism [63], metabolomics, which is the study of chemical processes involving metabolites, begins to serve a critical role in better understanding the complex nature of cancers [64,65]. Over the last few years, there has been an increasing number of publications aiming to introduce metabolomics as a means for early detection of cancer, determination of the cancer progression, identification of surrogate biomarkers for screening prostate cancer [66–69] and, more importantly, targets for therapeutic and preventive interventions [70–76].

Compared to other "omes", metabolome analysis can reveal ideal biomarkers. Metabolites are known to be a sensitive indicator of any alteration in a biochemical system, as they are the downstream of any changes in genes, transcripts and proteins [67,77] . As metabolomics profiling becomes cost and time efficient, high throughput and fully automated [68,70], discoveries about the metabolites' role in cancer screening, detection, progression, therapy or prevention can significantly impact the clinical health practices. While the upstream gene or gene products may vary across different species, the corresponding downstream metabolites are more comparable and also robust to analytical approaches [77]. Furthermore, metabolomics data benefit from lower dimensionality in statistical analysis, as they are measured in fewer number than genes or proteins. However, the fact that the biochemical reactions within a cell are not independent and could be linked via metabolic pathways imposes complexity into the analysis of metabolite data. The way we should analytically address this challenge will be discussed later in this section.

Over the past decade, there has been a growing number of studies looking at metabolite profiling of prostate cancer [78–80], although the search for ideal biomarkers is yet unresolved. This has led researchers to adopt more approaches such as investigating metabolic markers of PCa in oncogene-specific contexts [81]. Future direction of metabolomics research tends to involve integration of other omics data, say, expressions of oncogenes and oncoproteins, to create a molecular-level understanding of the context for the development of a metabolic phenotype of a prostate tumor [67,82].

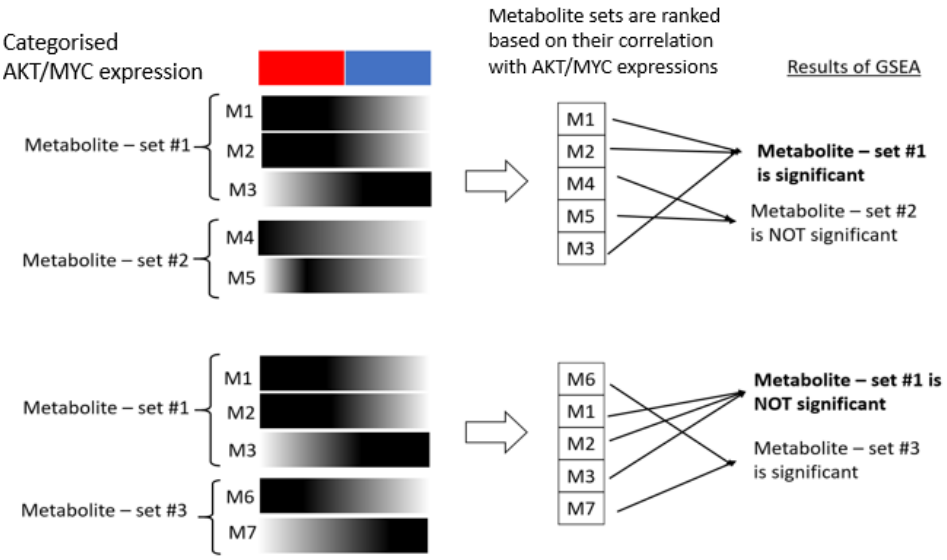
As upstream regulators of metabolites, there are multiple genetic and epigenetic alterations in tumor cells required to activate the growth factor receptors and signal protein, kinases and transcription factors. Among these genetic alterations is the loss of Phosphatase and Tensin Homolog(PTEN) and subsequent activation of Phosphoinositide 3-Kinase (PI3K/AKT), the pathway controlling cell growth, migration, differentiation and survival; and among the earliest alterations is proto-gene MYC overexpression that codes the transcription factor. Amplification of MYC was detected in about 75% of the advanced prostate cancers [83]. In a study conducted by Clegg et al. [84], the possibility that these two oncogenes may cooperate in prostate tumorigenesis was evaluated and a significant association was observed.

In the present study, our goal is to investigate the association between bivariate expression of the oncogene-pair (MYC and AKT1) and metabolite sets that are enriched in prostate tumors in patients. Significant associations can provide insights about the potential of oncogenic classification of PCa patients based on the observed tumor metabolic phenotypes. This hypothesis was previously tested in the study of Priolo et al., which was among the first attempts to explore this particular oncogene-pair associated metabolomics in PCa. The current study aims to introduce a more reliable analytical tool to examine this hypothesis.

The widely-used approach for analysis of metabolite sets, also employed by Priolo et al., is Gene Set Enrichment Analysis (GSEA) [16] where metabolites involved in the same pathway are ranked and compared with the metabolites outside the pathway. The over-representation of a pathway at the extremes of the ranked list determines the significance of the pathway. GSEA suffers from ignoring the correlation structure within-metabolite-sets, which may result in overstating the statistical significance[24]. Also, the competitive methods are subjective as their result varies according to the researcher's decision about the list of metabolite sets under study. Figure 2.1.a. illustrates GSEA and highlights its subjectivity and inability to accommodate within-pathway correlations. The other challenging feature of this data is to analyze the simultaneous effects of two oncogenes with continuous measures of expressions considering that they may interact. Since a method, like GSEA, is limited to comparing the effects of two (or more) variables, a frequent practice, as employed in Priolo et al.[81], is to classify the subjects according to their combined states and determine the metabolite signature in each subgroup separately. More specifically, Priolo et al.[81] handled the bivariate oncogenes outcome by using clustering algorithm to categorize the data into four groups: low AKT1 and low MYC, low AKT1 and high MYC, high AKT1 and low MYC, high AKT1 and high MYC. Then they ran separate GSEA analyses to identify metabolite sets different among paired groups. Exclusion of data that does not classify well into discrete categories is the main drawback of this method. This can be avoided by taking a modeling approach to examine the linear combination of the continuous measurements. As illustrated in Figure 2.1.b, our analysis enables the oncogenes to suppress or complement each other's effects. We reason that an alternative method such as generalized Linear Combination Test (LCT) improves our

understanding of the associations between metabolite sets and multiple oncogene expressions. This method is described in detail in the next section.

A.



B.

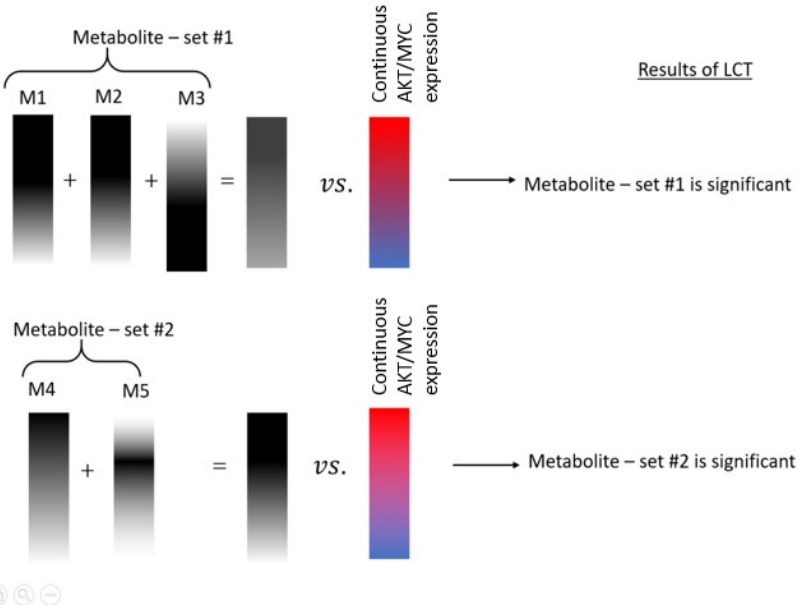


Figure 2.1 An illustration of scenarios derived from real data where LCT can improve analysis of metabolites and oncogenes associations.

A. According to GSEA analysis, metabolite set #1 is significant in the presence of metabolite set #2, while it is not significant in the presence of metabolite set #3. Assuming that M4 and M5 have a complementary effect, GSEA failed to detect this significant complementary effect. Another important limitation is that oncogenes expression values must be categorized and combined into a univariate value to comply with GSEA method. B. In contrast, LCT can handle multiple continuous oncogenes expression values: metabolite set #1 is significant and the result of LCT remains unchanged in the presence of other metabolite sets screened in the study. The complementary effect of M4 and M5 has been detected by LCT.

2.3 Methods

2.3.1 Collecting Data

The radical prostatectomy samples were obtained from fresh frozen tissues stored by Institutional Tissue Repository of Dana-Farber Cancer Institute/ Brigham and Women's Hospital (60 tumors and 16 normal). The non-metastatic tumor tissues with more than 80% purity were cut to two or three 8-μm sections. Deoxyribonucleic Acid (DNA), Ribonucleic Acid (RNA), protein and metabolite were isolated from these tissues. Phosphorylated AKT1 and MYC expressions of tumor cells were detected by immunoblotting. A detailed description of data collecting methods is given in Priolo et al. [81].

2.3.2 Statistical Method

Many methods have been developed and introduced to analyze microarray data. These methods have been widely borrowed to analyze metabolite datasets. An example is a web-based tool called Metabolite Sets Enrichment Analysis (MSEA) which uses the "global test", a gene set analysis (GSA) method for identifying the changing pattern of metabolite sets in a biologically meaningful context. Using the same strategy, we employed LCT, a method primarily defined for gene sets, to analyze the metabolite set data. LCT has many attractive features that will be described later.

In early microarray data analysis, the differentially expressed genes were identified using statistical methods such as t-test, principle components analysis and discrimination analysis. Then, the most significant genes were selected based on a predefined threshold and inspected for biological patterns. However, biological interpretation of the results was sensitive to the choice of the threshold, and this subjectivity became an important concern in analysis of individual gene sets. In order to overcome this problem, Gene Set Analysis (GSA) uses existing biological knowledge of genes and their pathways and tests pre-defined gene sets, instead of individual genes. For the same reason, we borrowed a GSA method to analyze the metabolite data. After grouping metabolite data based on the biological pathways they activate, the method we used takes into account the correlations within each metabolite set. Liu et al.[85] performed simulation studies and showed how other GSA methods treating the members of the set as independent measurements exhibit larger type II errors, and therefore smaller power.

While many GSA methods are designed for binary outcomes, such as cancer or control, LCT[86] is designed to work with continuous outcomes, and therefore allows us to look at an intermediate continuous phenotype, such as the oncogenes expressions.

LCT tests whether there is a significant linear relationship between the metabolite set $X = \{x_1, \dots, x_p\}$ consisting of p metabolites and the two oncogenes expressions $= \{Y_{Myc}, Y_{Akt}\}$. The multivariate null hypothesis can be expressed linearly and univariately as

H0 : There is no association between any of the linear combinations of metabolites in a set and any linear combinations of oncogenes.

If $Z(X, A)$ is a linear combination of metabolites within a metabolite set x_is with coefficient vector of A and $Z(Y, B)$ is a linear combination of oncogenes expressions y_is with coefficient vector of B, then we calculated the following statistic to test the null hypothesis:

$$T^2 = max|\rho((Z(X, A), Z(Y, B))^2| \quad (2.1)$$

In this method, the coefficient vectors A and B are estimated in a way that maximizes the Pearson correlation between $Z(X, A)$ and $Z(Y, B)$. T^2can also be rewritten as:

$$T^2 = max\frac{(A^T Cov(X,Y)B)^2}{(A^T Cov(X,X)A).(B^T Cov(Y,Y)B)} = \max\frac{(A^T \Sigma_{XY} B)^2}{(A^T \Sigma_{XX} A)\ .(B^T \Sigma_{YY} B)} \quad (2.2)$$

In the procedure of estimation of coefficient vectors, two problems arise: singularity caused by the high dimensionality of data (solved by shrinkage methods) and computational efficiency (solved by eigenvalue decomposition). Then, the p-value is calculated using sample permutations. Sample permutation method preserves the correlation structure within-metabolite-set and the correlation structure within oncogenes expressions[86].

We analyzed the data of this study using R 3.2.1 and calculated p-values based on 10,000 permutations.

2.3.3 Simulation Study Design

A simulation study was designed to evaluate the performance of LCT and compare it with GSEA to test for enrichment of different metabolite sets using the expression of a gene as a continuous phenotype. We varied correlation structures among the metabolite sets, metabolite set sizes, number of metabolite sets, sample sizes, proportions of significant metabolites within-metabolite-set and proportions of significant metabolite sets.

Metabolites were simulated for N (N=10, 20, 30) metabolite sets of sizes n (n=10, 20, 30) from a multivariate normal distribution with mean of 1.1 and correlation matrix of R. The correlation matrix was designed to maintain the within-metabolite-set correlation of ρ=0.5 and the correlation of r=0, 0.1 or 0.2 between all pairs of metabolite sets. The simulated continuous phenotype, gene expression here, satisfied the correlation of at least 0.6 with p×100% (p=0.2, 0.4, 0.6, 0.8) of the n metabolites belonging to P×100% (P=0.2, 0.4, 0.6, 0.8) of the metabolite sets. We used a correlation of less than 0.2 with the remaining metabolites. We performed k-means clustering method to cluster the continuous phenotype into two groups before running GSEA since GSEA uses discrete phenotype classes. A significant metabolite set was detected by a p-value smaller than 0.05 and a False Discovery Rate (FDR) adjusted p-value (q-value) smaller than 0.3, calculated by GSEA or LCT methods. The simulation was designed and executed in R.3.2.1. The power was calculated based on 100 iterations.

2.4 Results

2.4.1 Application

We employed the LCT method to test 71 nominated metabolite sets (KEGG annotation - Dataset S1). We first evaluated if metabolite sets predict both oncogenes alterations at the same time and then assessed if they predict any of them independently. We then included the interaction of the oncogenes in the model of both oncogenes to examine the possible multiplicative effect. We repeated the analysis for the normal sample. The full results related to tumor and normal samples are shown in Appendices 1 and 2, respectively. A comparison of the analysis results of tumor and normal samples was summarized in Table 2.1. We restricted the size of the sets to two or more metabolites.

Table 2.1 showed the metabolite sets significantly associated with AKT1 and MYC in tumor samples: fructose and mannose metabolism (p-value=0.02; q-value=0.35), purine (p-value=0.01; q-value=0.29) and pyrimidine metabolism (p-value<0.01; q-value=0.23). When the model considered the multiplicative effect of MYC and AKT1, in addition to all those pathways mentioned above, D-Glutamine and D-glutamate metabolism (p-value=0.01; q-value=0.31), fatty acid biosynthesis (p-value=0.01; q-value=0.31) and nitrogen metabolism (p-value=0.037; q-value=0.36) emerged as significant pathways. The contributions of oncogenes were described by estimating the coefficients presented in the last column. Based on the estimated coefficients, the metabolite set indicated alterations of oncogenes in similar or opposite directions. We note that the corresponding univariate level associations did not reach the significance level of 5% for most of the significant associations at the multivariate level in the tumor samples. This observation highlighted the advantage of the multivariate LCT method over the univariate analysis, and it was consistent to previous applications of LCT to other datasets [86].

Comparison of the results of tumor and cancer cells was quite revealing in several ways. First, none of the significant metabolite pathways in cancer samples was significant for the normal samples. In the analysis of tumor samples, we found no metabolite sets significantly associated with AKT1 alterations alone. In normal samples, the joint and separated effects of the oncogenes appeared insignificant in association with all the metabolite sets as the q-values were very high.

Hierarchical clustering was used to describe the difference between the signature of significant and insignificant metabolite pathways among tumor and normal subjects. The subjects were first clustered by their similar metabolite set signature and then the distribution of clusters for different values of oncogene expressions was examined. Figure 2.2 revealed the association between the metabolite signature and PCa. As shown in dendrograms, most of the normal subjects belong to the same cluster of significant metabolite signature. It also showed the association between oncogene expressions and PCa as almost all the normal subjects were scattered around the low values of MYC. As an evidence for the association between oncogene expressions and metabolite clusters, different significant-metabolite-driven clusters were not uniformly scattered over all the expression values. While some of them concentrated in AKT1-low area, some others appeared more frequently in MYC-low area. However, Figure 2.2-C, based on an insignificant metabolite set, illustrated a uniform distribution of different clusters over vertical and horizontal axes, indicating the lack of association between metabolite set and oncogenes expressions. Appendix A- Figure 1 provided the graphs related to other significant metabolite sets: nitrogen metabolism, fatty acid biosynthesis, D-glutamate metabolism, purine metabolism.

Table 2.1 The comparison of normal and cancer cells about the association between oncogenes (bivariate or univariate MYC and AKT1) and different metabolite sets. The data is analyzed using LCT.

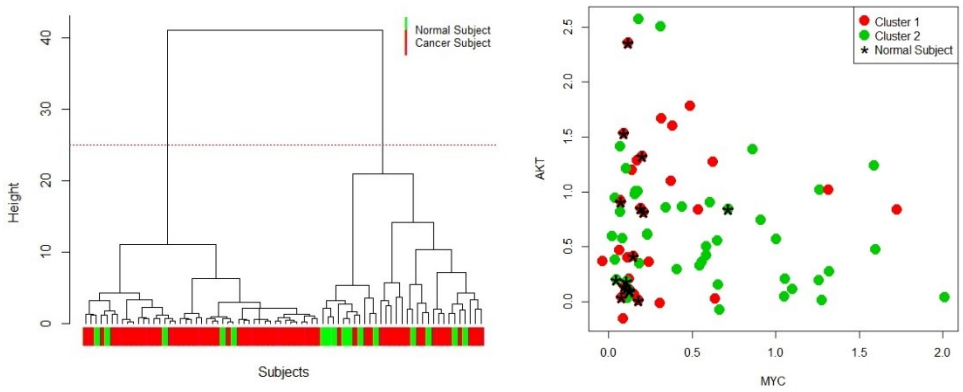
Metabolite Set†	Size of Metabolite Set	Type of Cell	p-value for (MYC, AKT1)	q-value for (MYC, AKT1)	P-value for (MYC, AKT1) Interaction	q-value for (MYC, AKT1) Interaction	p-value for MYC only	q-value for MYC only	p-value for AKT1 only	q-value for AKT1 only
D-Glutamine and D-glutamate metabolism	3	Cancer	0.035*	0.472	0.01*	0.309**	0.070	0.799	0.619	0.887
		Normal	0.286	0.979	0.255	0.964	0.467	0.911	0.283	0.988
Fatty acid biosynthesis	5	Cancer	0.043*	0.472	0.014*	0.309**	0.062	0.799	0.462	0.887
		Normal	0.601	0.979	0.540	0.964	0.585	0.911	0.662	0.988
Fructose and mannose metabolism	6	Cancer	0.018*	0.348**	0.022*	0.319**	0.125	0.799	0.404	0.887
		Normal	0.421	0.979	0.481	0.964	0.769	0.911	0.238	0.988
Nitrogen metabolism	5	Cancer	0.057	0.472	0.037*	0.358	0.115	0.799	0.585	0.887
		Normal	0.320	0.979	0.246	0.964	0.510	0.911	0.214	0.988
Purine metabolism	18	Cancer	0.01*	0.29**	0.03*	0.348**	0.021*	0.799	0.929	0.945
		Normal	0.629	0.979	0.649	0.964	0.862	0.943	0.361	0.988
Pyrimidine metabolism	12	Cancer	0.004*	0.232**	0.016*	0.309**	0.837	0.885	0.780	0.887
		Normal	0.815	0.979	0.707	0.964	0.693	0.911	0.760	0.988

† The metabolite sets significantly associated with oncogenes in tumor samples are listed here. Appendix A Tables 1 and 2 include the full analysis result for all metabolite sets.

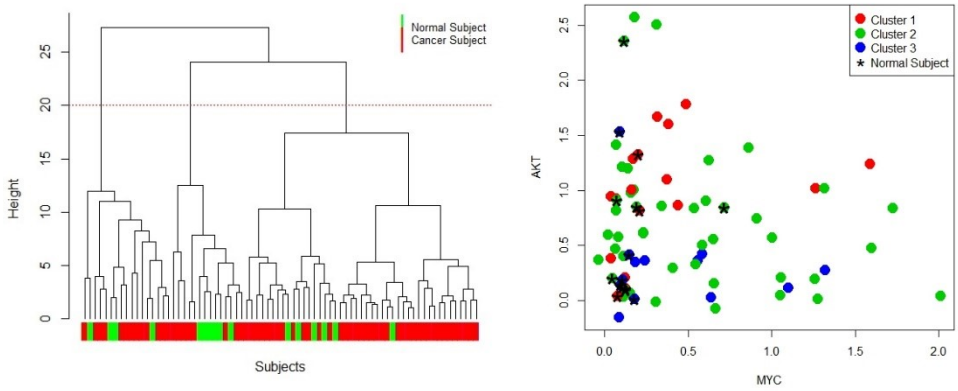
*Associations significant at p-value<0.05.

**Associations significant at q-value<0.35.

A. Fructose and mannose metabolism



B. Pyrimidine metabolism



C. Alanine, aspartate and glutamate metabolism

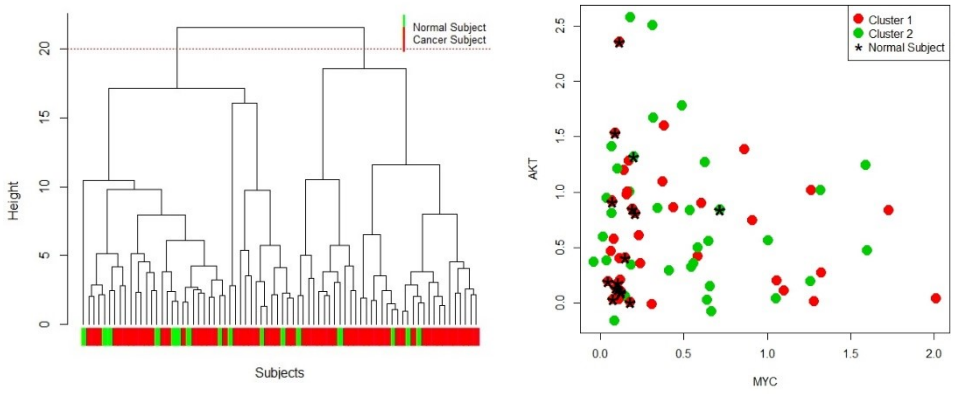


Figure 2.2 Cluster analysis of subjects based on significant (A and B), and insignificant (C) metabolite sets signatures: dendrogram of subjects clustered based on their metabolite set signature, and scatterplots of AKT1 vs. MYC with cluster-specified observations.

2.4.2 Simulation

The results of simulation study were shown in Figure 2.3. A general observation was LCT outperformed GSEA for various data structures. LCT power was always well above 80%, while GSEA power hardly reached values above 60%. Referring to Figure 2.3.A, metabolite set size had a substantial effect on the performance of GSEA, but not on LCT. A similar finding was reported in the simulation study of Dinu et al. [25]. Figure 2.3.B showed the consistent performance of both methods in handling different number of metabolite sets. Power was improved by enrolling larger number of subjects into the study (Figure 2.3.C). A powerful GSEA required significantly larger sample size in comparison with LCT analysis. Figure 2.3.D depicted the insufficiency of GSEA when the metabolite sets were not independent. In the presence of very low between metabolite set correlations (0.1 and 0.2), the power of GSEA declined sharply by about 30%. The performance of LCT did not vary with different correlation

structure of metabolite sets. GSEA failed to detect the significant metabolite sets properly when they were prevalent in the data (Figure 2.3.E). The metabolite sets consisting of higher proportion of metabolites with moderate to strong correlations with the phenotype were more likely to be detected by both LCT and GSEA methods (Figure 2.3.F).

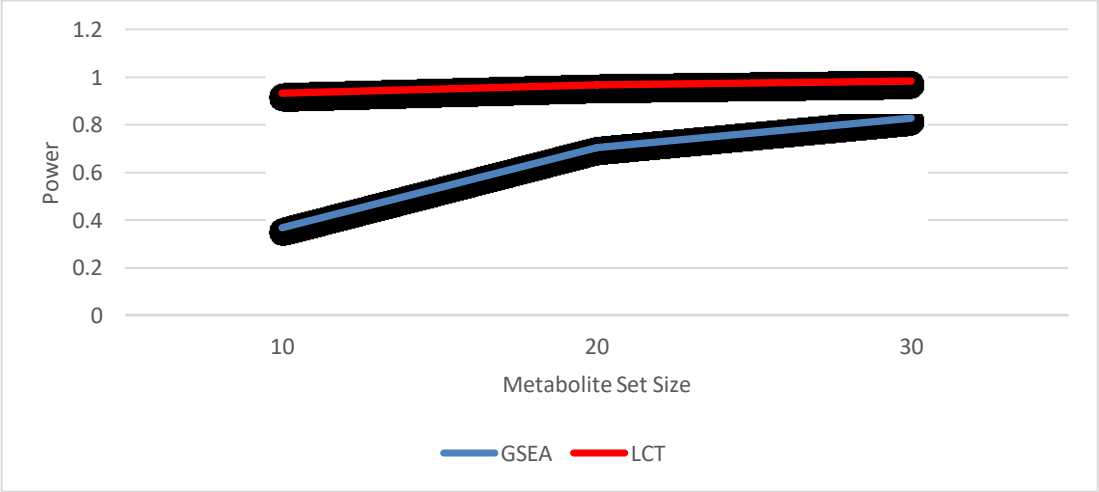


Figure 2.3 Power comparison between GSEA and LCT analyses in the presence of different metabolite set sizes

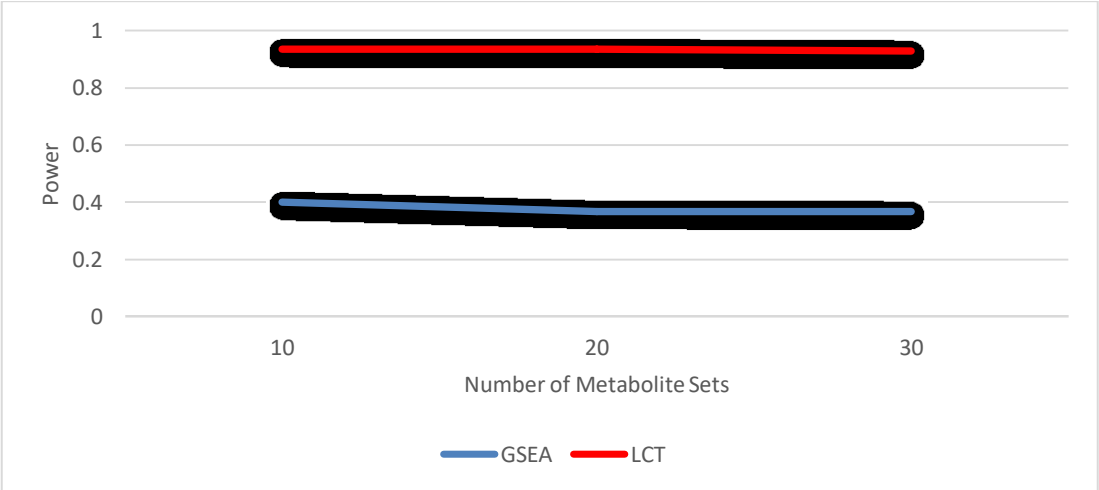


Figure 2.4 Power comparison between GSEA and LCT analyses in the presence of different number of metabolite sets

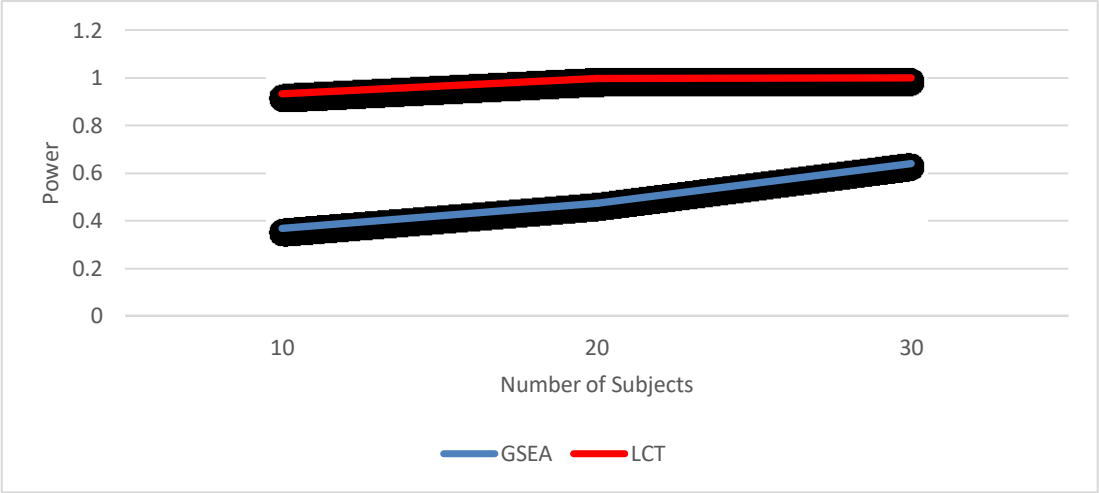


Figure 2.5 Power comparison between GSEA and LCT analyses in the presence of different sample size

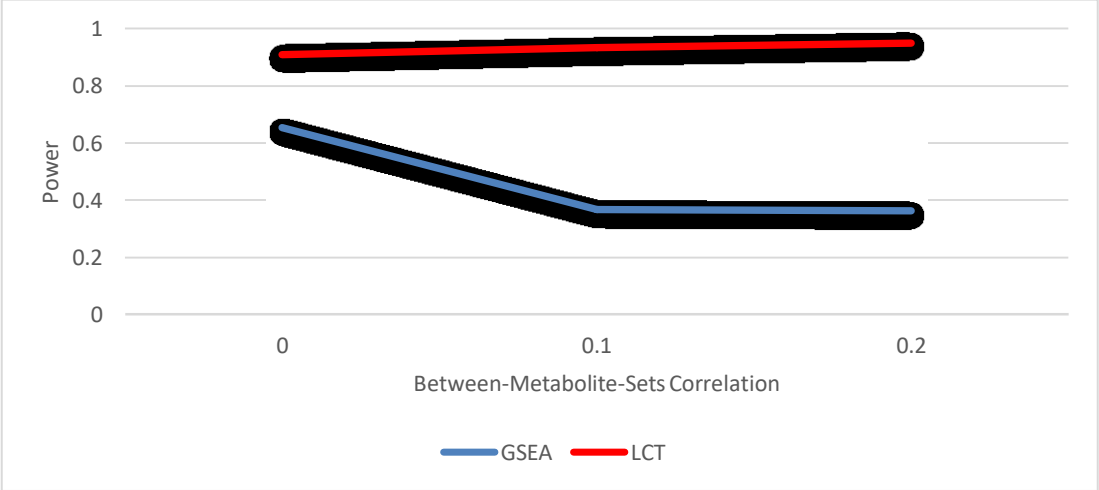


Figure 2.6 Power comparison between GSEA and LCT analyses in the presence of different between-metabolite-set correlation

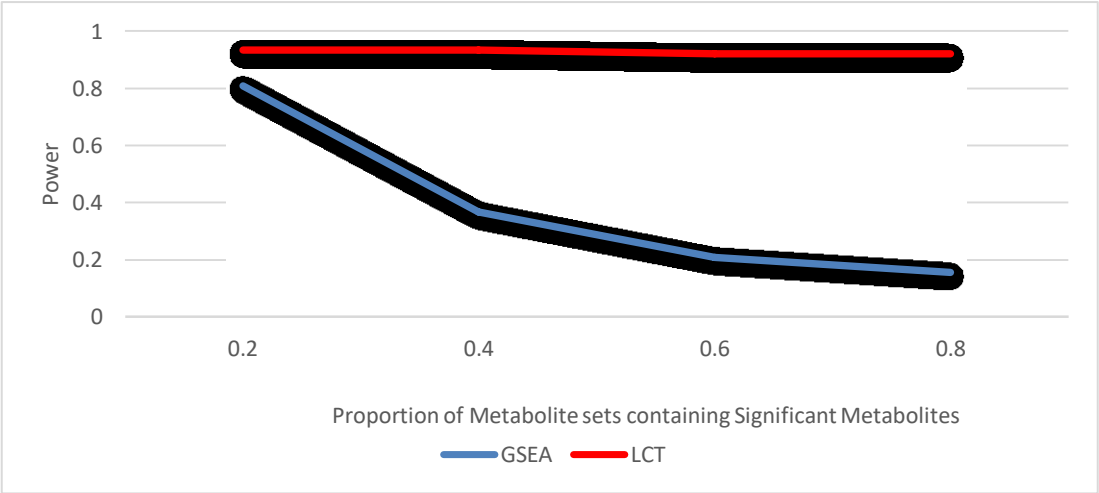


Figure 2.7 Power comparison between GSEA and LCT analyses in the presence of proportion of metabolite sets consisting significant metabolites (significant metabolite sets)

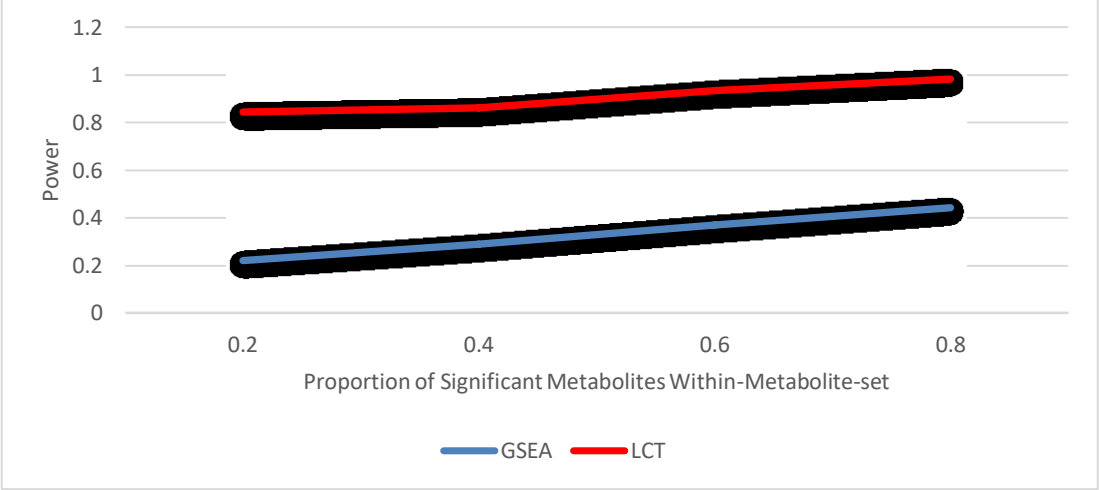


Figure 2.8 Power comparison between GSEA and LCT analyses in the presence of proportion of the metabolites with moderate to strong association with phenotype within significant metabolite sets

2.5 Discussion

More than half a century passed since Otto Warburg's theory of cancer cells was proposed. However, the characterization of metabolic alterations in tumors is still under investigation, and the involvement of pathways beyond glycolysis is being studied. Many recent studies supported the significant role of metabolite discoveries in the future direction of cancer research [74]. While many such studies have focused primarily on the differential concentrations of specific metabolites, the contexts of such alterations have rarely been associated with oncogenic expressions, much less by bivariate oncogene-pair expressions. The complex structure of the data collected to investigate this hypothesis calls for utilization of advanced statistical methods which account for the correlation within-metabolite-set, interaction of oncogenes and continuous nature of the measurements. Our simulation study and real application showed that GSEA, one of the most popular methods for gene set analyses, may not be the best choice to handle complex data structure. We note that GSEA findings vary with the size of the pathway, even if the same level of linkage between the omics is maintained[87]. This observation can be explained by GSEA assigning higher enrichment scores to larger sets[25]. GSEA assigns significance to a set, in the context of other sets tested[87]. Therefore, researchers may come up with different findings regarding the same set, if they consider testing different collections of metabolite sets. In this simulation study, we focused on analysis of a single gene expression and LCT is expected to perform even better with multiple phenotypes because GSEA requires combined dichotomizing of multiple phenotypes and therefore, lose more information. The clustering of multiple phenotypes, in contrast, may lead to larger loss of information, and lower power of GSEA.

Our study takes advantage of the relatively new statistical method (LCT) by testing the association between alterations in the metabolite sets and those in oncogenes AKT1 and MYC in human prostate tumor samples. Although the excess use of glucose is well regarded as a hallmark of cancer cells, other metabolic pathways may also be affected by the myriad oncogenic processes. Glutamine metabolism is one of the most important dysregulated pathways during oncogenesis [88] and we found it significantly altered in prostate tumors. This pathway dysregulation is known to be MYC-induced as MYC contributes in glutaminolysis by regulating glutaminase and was frequently observed to be overexpressed in tumors [89].

The other altered pathway is nucleotide synthesis pathway. Glutamine, which supplies carbon and nitrogen for proliferative active cells, acts as a source of nitrogen [88] and catalysts (e.g. thymidylate synthesis and inosine synthesis [90] for nucleotide synthesis, specifically purines and pyrimidines. Our findings suggested significant alterations in nitrogen, purine and pyrimidine metabolism in prostate cancer, believed to relate to upregulation of MYC for up taking of glutamine.

Fatty acid synthesis pathway was also found to be significant in our analysis. The cancer cells synthesize glucose and glutamine for *de novo* fatty acid synthesis in order to meet their energy needs, survive longer and increase proliferation. Elevated level of multi-enzyme complex fatty acid synthesis and overexpression of many enzymes of fatty acid synthesis, such as acetyl-CoA carboxylase and ATP citrate lyase, are common events during oncogenesis, whose inhibition leads to tumor cell apoptosis[72]. Several attempts have been made to investigate the association between altered fatty acid biosynthesis pathway and AKT and MYC dysregulation. Although some studies highlighted the link between high levels of MYC and

increased fatty acid synthesis [91–93], there are some others who believe that PI3K-AKT signaling pathway controls this alteration [94].

While AKT regulates the glycolysis pathway and has no control over glutamine metabolism [95], MYC is able to switch to non-glucose nutrient source like glutamine and fatty acids [96]. These previous knowledges along with our observation about significant glutamine, fatty acid and mannose and fructose pathways underline the leading role of MYC in prostate cancer. This critical effect is also shown in Figure 2.2, where all the normal subjects have low levels of MYC expression. This finding may explain the inefficiency of mTOR inhibitors in treatment of prostate cancer [97]. Priolo et al. also linked the dysregulation of downstream of glutamine and lipid metabolism to MYC overexpression.

In order to collect more evidence supporting our hypothesis that MYC is the master oncogene, we tested MYC and AKT1 as the sole oncogene altering the metabolite pathway. Although we found no pathway significantly altered by their separated effects, we noted an overall tendency of MYC p-values to be smaller than AKT1 p-values in the tumor samples wherever the association of pathway and multivariate outcome (MYC, AKT1) is significant. These findings are supporting our hypothesis that although MYC drives the mitochondrial function of the cells, its dysregulation may not be sufficient to complete a metabolite pathway alteration. This is where the main effect of AKT1, or moreover the interaction between AKT1 and MYC should be taken into account.

Several studies suggest that MYC and PI3K-AKT pathways indirectly interact [97,98]. If we accept that MYC is the leading, but not necessarily the only oncogene involved in developing prostate cancer, there could be two possible explanations of AKT action - either disjointly, e.g.,

activated AKT guarantees survival of the tumor cell [99], or interactively, e.g., activated AKT inhibits the antagonistic effect of FOXO on MYC through phosphorylation [98,100,101].

Consistently, many recent studies suggested different mechanisms for the inhibition of Myc-related apoptosis by upregulation of AKT[102–104]. The second scenario is more likely based on our observation, as our model showed a significant interaction between the oncoproteins in tumor cells.

Notably, we found that MYC and AKT1 did not appear to interact in a normal cell. This finding was in contrast with our observation in the tumor samples that the significantly enriched metabolic pathways were associated with oncogene-pair interactions.

Our findings confirmed the presence of complex structure in the data and highlighted the necessity of utilizing an appropriate analytical method. Failure to take care of the correlated effects of multiple oncogenes as continuous measurement, may prevent us from revealing the underlying mechanisms of metabolic programming.

Despite the strengths of LCT in analysis of system biology data, there are few limitations that should be discussed. Firstly, not all the metabolites within a significant metabolite set are significantly associated with the expression phenotype. Identification of the "core metabolite set" which may contribute to the significance of the whole set can help biologists develop novel preventive or therapeutic strategies targeting the core. The SAM-GSR method proposed by Vatanpour et al.[105] can be applied to reduce the metabolite sets to their cores. LCT, a self-contained method, assumes independence of the metabolite sets, an assumption not always valid. This weakness of self-contained methods needs to be addressed by future studies.

2.6 Conclusions

The present study aims to take advantage of compelling analytical methods to obtain a better understanding of the molecular and biochemical alterations in prostate tumors by connecting data from genomics and metabolomics. Using LCT method, we uncovered the role of MYC as the leading, but not the only oncogene associated with human prostate cancer metabolic phenotypes. In particular, we showed how the multiplicative (interaction) effect of MYC and (phosphorylated) AKT1 expressions determines the context for differential metabolite sets in prostate tumors, but not the healthy samples.

Our analytical approach can be applied to studies of other complex diseases where such contextual distinction among multivariate and correlated phenotypes would be useful.

2.7 Summary

The analysis of the multivariate effects of gene expressions on metabolite phenotypes is challenging due to the presence of high-dimensional continuous datasets, within-metabolite-set correlations and the possible interactions between the genes. The present study suggests an analytical solution for linking genomics and metabolomics. The association between bivariate expressions of AKT1 and MYC and metabolite signature of the prostate cancer patients was examined. MYC and significantly-altered metabolite sets were found to play an important role in the development of prostate cancer.

Chapter 3

Longitudinal Linear Combination Test for Gene set Analysis

3.1 Abstract

Background: Although microarray studies have greatly contributed in the recent genetic advances, lack of replication has always been a continuing concern in this area. The complex study designs have good potential to address this concern. However, they appeared to be unwelcomed by genetic investigators due to lack of proper analysis method. A primary challenge of analysis of complex microarray study data is handling the correlation structure within data, while dealing with a large number of genetic measurements and a small number of subjects. Motivated by the lack of available methods for analysis of repeatedly measured phenotypic or genomic data, we developed longitudinal linear combination test (LLCT).

Results: LLCT is a two-step method to analyze multiple longitudinal phenotypes when there is high dimensionality in response and/or explanatory variables. Dealing with within-subjects and between-subjects variations in two steps, LLCT examines if the maximum possible correlation between a linear combination of the time trends and a linear combination of the predictors given by the gene expressions is statistically significant. A generalization of this method can handle family-based study designs when the subjects are not independent. This method is also applicable to time-course microarray and identifies gene sets with significantly different expression patterns over time. Based on the results from a simulation study, LLCT

outperformed its alternative, the pathway analysis via regression method. LLCT was shown to be very powerful in analysis of large gene sets with low heterogeneity.

Conclusions: This method offers many interesting flexibilities to the analysis. This self-contained pathway analysis method is applicable to a wide range of longitudinal omics data, allows adjusting for potentially time-dependent covariates and works well with unbalanced and incomplete data. An important application of this method can be time-course linkage of omics, an attractive perspective of future genetics.

3.2 Introduction

Longitudinal designs are fast becoming a key instrument in genetics studies as they advance understanding of disease progression and underlying biological mechanism. Longitudinal studies provide information about age of onset and time-varying covariates that helps investigate a complex disease more precisely. A primary concern of these study designs is to find a proper analysis method which deals best with the correlation structure imposed by longitudinal data. Within-subject correlation cannot be addressed by traditional statistical analysis methods.

In recent years, there has been an increasing interest in microarray studies which has triggered rapid advances in microarray data analysis methods. From 2001, a considerable amount of literature has been published on methods of Individual Gene Analysis (IGA)[106] and Gene Set Analysis(GSA)[107–110]. Majority of these studies have proposed enrichment methods for binary and categorical phenotypes. Little attention has been paid on developing the

methods for other phenotypes, especially longitudinal. The current thesis contributes to fill this gap by proposing longitudinal linear combination test (LLCT).

A frequent practice to deal with longitudinal phenotypes in genetics studies is to simply average the multiple measurements. With this approach, the temporal variation of the phenotype is discarded and part of the information is lost[111]. To the best of our knowledge, the only GSA method developed to analyze longitudinal phenotype is the Pathway Analysis via Regression (PAVR) method proposed by Adewale et al. [29]. This method utilizes regression modelling to analyze binary, multi-class, continuous, count, rate, survival and longitudinal data and adjusts the results for potential covariates. In this method, the measure of association of a specific gene set with the phenotype is a sum of squares of Wald statistics from regression models fitted on the phenotype against the individual genes in the pathway of interest. We will compare this method with LLCT and discuss its limitations later in this thesis.

Our goal is to develop a statistical method which not only tackles the limitations of available methods but addresses challenges of complex designs in recent microarray studies. The main function of this method is to recognize differentially expressed gene sets associated with a phenotype trajectory over time. It is also applicable to family-based study designs when the subjects are not independent. A generalization of this method can handle time-course microarray studies and identifies gene sets with significantly different expression patterns over time.

Longitudinal microarray studies do not only consider the trajectories of phenotypes, but gene expression trajectories may also be the concern of many genetic studies. In time-course microarray studies, arrays are collected repeatedly over time, allowing one to examine the

dynamic behavior of gene expressions. GSA methods for time-course gene expressions received more attention than GSA methods for repeated measurements of phenotypes. These methods are exploratory in nature by clustering genes to co-expressed groups[47]. However, this development was not sufficient to address biologists' concerns about the association of gene expressions trajectories with one or more specific covariate(s). Many procedures have been proposed for time-course microarray experiments, to test if specific genes exhibit different expression profiles significantly associated with covariates. ANOVA-based methods [112,113] and regression-based approaches are very popular in this field. Linear Mixed Models (LMM) or Generalized Estimating Equations (GEE) are more mature statistical models accommodating the correlations between repeated measurements. However, they are not directly applicable, as the time-course expression data is often collected for a large number of genes, but only for few subjects. To deal with the high dimensionality of the data, Turner et al.[114] modeled the genes separately and then rescaled the data using Variance Inflation Factor (VIF) estimates to accommodate the correlation between the genes within gene sets. LMMs were also used in the methods developed by Hejblum et al. [48], Zhang et al.[115], and Conesa et al. (maSigPro method)[47], but they only work with categorical predictor variables. Our proposed method, LLCT can handle both categorical and continuous predictors.

Family-based data is another type of complex designs in microarray studies. Family-based study designs are advantageous compared to studies of unrelated subjects in terms of lower genomic or phenotypic heterogeneity. Also, we are more likely to detect any significant effect when we observe multiple copies of the significant effects in a family [116]. Over about the past 35 years, study designs incorporating information from related subjects have resulted in better scientific interpretations[117].

LLCT is a GSA method. Incorporating information about the group of genes which are linked via biological pathways, LLCT aims to discover gene sets associated with the phenotype trajectories. These biological pathway, or a-priori defined gene sets are archived in online databases, available for download: The Cancer Genome Atlas (TCGA) [17], Gene Expression Omnibus (GEO) [18], Keyoto Encyclopedia of Genes and Genomes (KEGG)[19], BioCarta[19], Molecular Signature Database of the Broad Institute[20] . Although imposing additional complexity into the analysis, this feature of LLCT is biologically very appealing. In contrast to IGA, GSA works based on a biologically realistic assumption that the genes are not independent and a cell's function can be accomplished by differential expression of a group of genes, even if all of them show only weak to moderate changes [118].

LLCT is a self-contained method. Reviews on GSA have attempted to draw distinction between self-contained and competitive GSA. A competitive method employs gene permutation to test whether the association between a gene set and the outcome is equal to those of the other gene sets (so-called "Q1 hypothesis"[23]). A self-contained method employs subject permutation to test the equality of the two mean vectors of gene set expressions corresponding to the two groups (so-called "Q2 hypothesis"[23]). Since competitive methods have been widely criticized for their inability to take care of the correlation within gene sets, we focus here on developing a self-contained method testing the Q2 hypothesis.

The remaining part of the chapter proceeds as follow. Section 3.3 is concerned with an overview of LLCT method. Section 3.4 presents the results of simulation and an application of this method. Section 3.5 discusses the performance of LLCT and finally, section 3.6 gives a summary and discusses future areas of genetic applications and methodological developments.

3.3 Method

3.3.1 Longitudinal Linear Combination Test (LLCT)

We propose a two-step method to analyze multiple longitudinal phenotypes when there is high dimensionality in either response or explanatory variables. In the first step, within-subject variation is analyzed. For each gene set, the changing trend of outcomes over time is estimated using an appropriate model for the structure and type of the data. In the second step, LCT is applied to analyze the between-subject variation. In this step, LCT is employed to examine if the maximum possible correlation between a linear combination of the time trends and a linear combination of the predictors given by the gene expressions is statistically significant. Our method is generalized to accommodate data generated by two complex study designs: time-course microarray studies and family-based studies. A time-course study measures gene expression repeatedly over time and is designed to find the correlation between time trajectory of gene-expressions and covariates. A family-based design collects the information from family members and examines the association between longitudinal phenotypes and gene expressions while taking care of the correlation between subjects within each family.

We borrowed the main idea of this method from the mixed effect modelling. Through mixed effect modelling, the variation in the longitudinal phenotype is modelled taking two steps: first step, the within-subject variation is modelled; in the second step, the between-subject variation is modeled using the coefficients estimated in the first step[119]. Roughly the same strategy is also practiced by Conesa et al. [47] in their microarray significant profiles (maSigPro) method.

The proposed method is designed to model continuous outcome variables. However, this method can be generalized to work with other type of data, such as binary or categorical

response variable, using an appropriate link function in the first step. This method is self-contained, designed to accommodate the correlations between genes in the gene sets, works well in the presence of missing data at random and is efficient to work with high dimensional data. It can also adjust for time-variant covariates. Next, we describe the two steps of the method, followed by two generalizations.

Analysis of Within-Subject Variation (Step 1): Consider a microarray study on I subjects where longitudinal phenotypes of size M is measured for n_i times for the i th subject, $i = 1, \dots, I$. Let Y_{mij} be the $j\ th$ measurement ($j = 1, \dots, n_i$) of the m th phenotype ($m = 1, \dots, M$) of the i th subject that happened at time t_{ij} and let $Y_{mi} = (Y_{mi1}, \dots, Y_{min_i})^T$ be the vector of n_i measurements of the m th phenotype for the ith subject ($\sum_{i=1}^{I} n_i = n_.$) and $Y_i = (Y_{1i}, \dots, Y_{Mi})$ be the matrix of phenotype measurements of the ith subject. We also consider that the study measured the expressions of a predefined set of P genes for the i th subject, $G_i = (G_{i1}, \dots, G_{iP})^T, i = 1, \dots, I$; and we define the vector of the expressions of gene p for I subjects as $G_p = \left(G_{1p}, \dots, G_{Ip}\right)^T, p = 1, \dots, P$. We are interested to test if there is a significant linear relationship between the gene set G and the longitudinal phenotype Y. The null hypothesis is that the changes in Y over time are not dependent to the expressions of the genes in the predefined gene set G.

In order to analyze within-subject correlation, we define the regression equation in matrix notation as below:

$$Y_{mi} = Z_i\beta_{mi} + W_i\gamma_{mi} + \varepsilon_{mi} \qquad (3.1)$$

In this equation, Z_i is $(n_i \times Q)$ matrix of Q potential time variables and it usually includes $t_i = \left(t_{i1}, \dots., t_{in_i}\right)$ and different polynomial functions of t_i (e.g. t_i^2, t_i^3) if required. W_i is

$(n_i \times Q')$ matrix of Q' potential time-variant covariates, and $\gamma_{mi}(Q' \times 1)$ represents their corresponding coefficients. Also, β_{mi} denotes a $(Q \times 1)$ vector of coefficients for each specific phenotype (m) with elements of β_{mqi}. We define β_ia $(Q \times M)$ matrix of regression coefficients generated by column-wise binding of β_{mi}s:$\beta_i = [\beta_{1i}|\beta_{2i}| \dots |\beta_{Mi}]$.

Analysis of Between-Subject Variation (Step 2): In our method, we used Linear Combination Test (LCT)[120] to detect significant gene sets associated with different trajectories of longitudinal phenotypes. If there is no gene set related variability in the subject-specific regression coefficient estimated in the first step, there will be no relationship between the gene set expressions and the changing trend of M longitudinal phenotypes. In other words, there is no linear combination of the columns of $\beta = [\beta_1^T| \dots |\beta_I^T]^T$ associated to any linear combination of gene set expression measurements. The null hypothesis is that there is no association between any of the linear combination of $G_1, \dots, G_P$ with any linear combination of columns of β.

Let G be a $((I \times Q) \times (P))$ matrix obtained by vertically merging the vectors of the gene expressions, G_ps, duplicating the rows for Q times. Then, let

$$Z(G, \mathrm{A}) = \begin{bmatrix} G_{11} & \dots & G_{1P} \\ \vdots & \ddots & \vdots \\ G_{11} & \dots & G_{1P} \\ G_{21} & \dots & G_{2P} \\ \vdots & \ddots & \vdots \\ G_{21} & \dots & G_{2P} \\ \vdots & \vdots & \vdots \\ G_{I1} & \dots & G_{IP} \\ G_{I1} & \dots & G_{IP} \end{bmatrix}_{(I.Q)\times(P)} \times \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_P \end{bmatrix}_{(P)\times 1} \tag{3.2}$$

be a linear combination of the columns of matrix G, and,

$$Z(\mathrm{B},\Gamma) = \beta_{(I.Q)\times M} \times \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_M \end{bmatrix}_{(M)\times 1} \tag{3.3}$$

a linear combination of columns of β. The null hypothesis can be written as an optimization problem, more precisely, identifying A and B to maximize the correlation of $Z(G,\mathrm{A})$ and (B,Γ) , and then test if this maximum correlation is significant or not.

Let $\Sigma_{\mathrm{G,G}} = cov(\mathrm{G},\mathrm{G})$ be the covariance matrix of G; and similarly, let $\Sigma_{\mathrm{B,B}} = cov(\mathrm{B},\mathrm{B})$ be the covariance matrix of B and $\Sigma_{\mathrm{G,B}} = cov(\mathrm{G},\mathrm{B})$ be the covariance matrix between G and B. This leads to the proposed test statistic:

$$T^2 = max_{A,B}|\rho\big(Z(\mathrm{G},\mathrm{A}),Z(\mathrm{B},\Gamma)\big)|^2 = max_{A,B}\frac{(\mathrm{A}^T\Sigma_{\mathrm{B,G}}\Gamma)^2}{\mathrm{A}^T\Sigma_{\mathrm{G,G}}\mathrm{A}.\ \Gamma^T\Sigma_{\mathrm{B,B}}\Gamma} \tag{3.4}$$

The problem of singularity of $\Sigma_{\mathrm{B,B}}$ and $\Sigma_{\mathrm{G,G}}$ emerges when the dimensions of B or G are large. This is very likely to happen as we usually measure the expressions of a large number of gene sets. A possible remedy for this problem is to utilize the shrinkage method [121]. Therefore, we need to replace the covariance matrices with their shrinkage versions, $\Sigma^*_{\mathrm{B,B}}$ and $\Sigma^*_{\mathrm{G,G}}$. T^{2*} which is the shrinkage version of T^2 is defined as below:

$$T^{2*} = max_{A,B}\frac{(\mathrm{A}^T\Sigma_{\mathrm{B,G}}\Gamma)^2}{\mathrm{A}^T\ \Sigma^*_{\mathrm{G,G}}\mathrm{A}.\ \Gamma^T\Sigma^*_{\mathrm{B,B}}\Gamma} \tag{3.5}$$

We use the permutation method to calculate the p-value corresponding to this statistic. When the permutation method is employed, it would be computationally inefficient to maximize the right-hand side of the equation above. The remedy could be using two groups of normalized orthogonal bases instead of using the original observation vectors G and B. We decomposed the

two shrinkage covariance matrices using eigenvalues ($\Sigma^*_{\mathrm{G,G}} = \Psi D_{\mathrm{G}} \Psi^T$ and $\Sigma^*_{\mathrm{B,B}} = \Omega D_{\mathrm{B}} \Omega^T$) in order to have two groups of orthogonal basis vectors $\widetilde{\mathrm{G}}$ and $\widetilde{\mathrm{B}}$. Thus, the test statistic becomes:

$$T^{2*} = \max_{\eta,\theta} \frac{(\eta^T \Sigma_{\widetilde{\mathrm{G}},\widetilde{\mathrm{B}}} \theta)^2}{\|\eta\|_2^2 . \|\theta\|_2^2} \tag{3.6}$$

where $\eta = D_{\mathrm{G}}^{1/2} \Psi^T \mathrm{A}$ and $\theta = D_{\mathrm{B}}^{1/2} \Omega^T \Gamma$. Optimizing this expression will be straightforward if we first optimize η given θ and then optimizing θ at the next step. The value of T^{2*} is equal to the largest eigenvalue of $\Sigma^T_{\widetilde{\mathrm{G}},\widetilde{\mathrm{B}}} \Sigma_{\widetilde{\mathrm{G}},\widetilde{\mathrm{B}}}$ (or $\Sigma^T_{\widetilde{\mathrm{B}},\widetilde{\mathrm{G}}} \Sigma_{\widetilde{\mathrm{B}},\widetilde{\mathrm{G}}}$).

The sample permutation method is employed to calculate p-values. The sample permutation changes neither the correlation structure within gene sets nor the correlation structure within phenotype. This feature brings a considerable computational advantage to the analysis because there is no need to repeat eigenvalue decomposition for each permuted version of the dataset.

3.3.2 Generalization 1: LLCT for Family-Based Data

Consider a microarray study on I subjects in which M longitudinal phenotype is measured for F families. Also consider that the number of subjects in family f is I_f and the number of repeated measurements for subject i in family f is n_{fi} so that $\sum_{f=1}^{F} \sum_{i=1}^{I_f} n_{fi} = n_{..}$ is the total number of observations in the study. Let Y_{mfij} be the $j\,th$ measurement ($j = 1, \dots, n_{fi}$) of the $m\,th$ phenotype ($m = 1, \dots, M$) recorded for the ith subject who belongs to family f and measured at time t_{fij}. Also consider that the study measured the expressions of a predefined set of P genes for the ith subject of family f, and we define the vector of the expressions of the pth gene as $G_p = \left(G_{11p}, \dots, G_{1I_1p}, \dots, G_{F1p}, \dots, G_{FI_Fp}\right)^T, p = 1, \dots, P$. We are interested to test if

there is a significant linear relationship between the gene set $G_{fi}$and the longitudinal trajectory of all longitudinal phenotypes Y_{mfij}. The null hypothesis is that the changes in Ys over time are not dependent to the expressions of the genes in the gene set of interest.

In the first step of our generalized method, we model the within-family variation for phenotypes separately using mixed effect model. So, in a matrix format, we define:

$$Y_m = X\beta_m + W\gamma_m + Zb_m + \varepsilon_m \quad (3.7)$$

In this equation, Y_m denotes a $(n_{\cdot\cdot} \times 1)$ vector of the mth phenotype measurements for all families and all subjects and $X\,(n_{\cdot\cdot} \times R)$ denotes a matrix of time variables and it usually includes a vector of t and different functions , such as t^2, t^3, and Z is $(n_{\cdot\cdot} \times (F.Q))$ matrix of the potential covariates for random effects with the format of $Z = \begin{pmatrix} \mathrm{Z}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathrm{Z}_F \end{pmatrix}$ where $Z_f \in \mathbb{R}^{I_f \times Q}$. Also, β_m is a $(R \times 1)$vector of family fixed regression coefficients for time variables corresponding to mth phenotype. $W\,(n_{\cdot\cdot} \times R')$ represents a matrix of potentially time-dependent and time-independent (but subject-variant) covariates for which the estimations are adjusted. The vector of coefficients is denoted by. $\gamma_m(R' \times 1)$. The $((F.Q) \times 1)$ vector of random effects is defined as b_m and varies by families. The $(n_{\cdot\cdot} \times 1)$ vector of residuals is ε_m and we have $\begin{pmatrix} b_m \\ \varepsilon_m \end{pmatrix} \sim N_{Q+n_{\cdot\cdot}}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Psi & 0 \\ 0 & \Omega \end{pmatrix}\right)$ where $\Omega = \begin{pmatrix} \Sigma_{n_{f1}} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_{n_{fI_f}} \end{pmatrix}$ and Ψ is the covariance matrix of random effects that must be estimated.

Gene expressions are also correlated within families. Therefore, we will take an additional step to accommodate the within-family association of gene expressions using random intercept model:

$$G_p = \Xi_p + \xi_p + \varepsilon_p^* \tag{3.8}$$

In this model, $G_p$is $(I \times 1)$ vector of gene expressions of pth gene for all families, $\varepsilon_p^*$is $(I \times 1)$ vector of residuals, $\Xi_p$is $(I \times 1)$ vector of fixed intercept, a constant for all families, $\xi_p$is $(I \times 1)$ vector of random intercepts and its elements vary for each family.

In the second step of our method, we will use LCT for multiple phenotypes to examine the between-family variations. If there is no gene set related variability in the family-specific regression coefficient, there will be no relationship between the gene set expressions and changing trend of M longitudinal phenotypes. In other words, there is no linear combination of family-specific phenotype trajectories $b = [b_1^T | \dots | b_M^T]^T$ associated to any linear combination of family-specific gene set expression measurements $\xi = [\xi_1 | \dots | \xi_P]$. The null hypothesis, here, is defined to be no association between any of the linear combination of $\xi_1, \dots, \xi_P$ with any linear combination of columns of .

Let G be a $((F.Q) \times P)$ matrix which is created by vertically merging the vectors of ξ_ps and duplicating each row for Q times. Then, let

$$Z(G, \mathrm{A}) = \begin{bmatrix} \xi_{11} & \cdots & \xi_{1P} \\ \vdots & \ddots & \vdots \\ \xi_{11} & \cdots & \xi_{1P} \\ \xi_{21} & \cdots & \xi_{2P} \\ \vdots & \ddots & \vdots \\ \xi_{21} & \cdots & \xi_{2P} \\ \vdots & \vdots & \vdots \\ \xi_{(F.Q)1} & \cdots & \xi_{(F.Q)P} \\ \xi_{(F.Q)1} & \cdots & \xi_{(F.Q)P} \end{bmatrix}_{(F.Q)\times P} \qquad GC_{(F)\times(P)} \times \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_P \end{bmatrix}_{(P)\times 1} \tag{3.9}$$

be a linear combination of the columns of matrix G, and,

$$Z(\mathrm{B},\Gamma) = b_{(F.Q)\times(M)} \times \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_M \end{bmatrix}_{(M)\times 1} \tag{3.10}$$

a linear combination of the columns of bs. As before, the null hypothesis can be written as an optimization problem, more precisely, identifying A and B to maximize the correlation of $Z(G,\mathrm{A})$ and (B,Γ) , and then test if this maximum correlation is significant or not.

3.3.3 Generalization 2: Time-Course Microarray Data Analysis

We considered this application as a special case of the general framework of analyzing multiple longitudinal data. In this special case, the longitudinal gene expressions measurements of a specific gene set are treated as multiple longitudinal phenotypes. Consider a microarray study on I subjects where gene expressions of a specific gene set are measured for n_i times for the ith subject, $i = 1, \dots, I$. Let G_{pij} be the $j\,th$ measurement $(j = 1, \dots, n_i)$ of the p th gene expression in the gene set $(p = 1, \dots, P)$ for the ith subject that happened at time t_{ij} and let $G_{pi} = (G_{pi1}, \dots, G_{pin_i})^T$ be the vector of n_i expression measurements of the p th gene for the ith subject $(\sum_{i=1}^{I} n_i = n_.)$ and $G_i = (G_{1i}, \dots, G_{Pi})$ the $(n_i \times P)$ matrix of phenotype measurements of the ith subject. We are interested to test if there is a significant linear relationship between the specific gene set G and a set of time-invariant covariates C. The null hypothesis is that the changes in predefined gene set G over time are not dependent to the covariates C. In other words, the genes in a specific gene set are not differentially expressed over time in response to the changes of covariates.

In this application we only modify the first step of our proposed LCT method to analyze within-subject variations. The second step, where LCT is employed to analyze between-subject variations, remains unchanged.

Consider the following model:

$$G_{ip} = Z_i\beta_{ip} + W_i\gamma_{ip} + \varepsilon_{ip} \quad (3.11)$$

where Z_i is $(n_i \times Q)$ matrix of the time variables and it usually includes $t_i = \left(t_{i1}, \ldots., t_{in_i}\right)$ and different functions of t_i (e.g. t_i^2, t_i^3). W_i is the matrix of potential time-dependent covariates for which we would like the estimations to be adjusted, with corresponding $(Q' \times 1)$ vector of coefficients of γ_{ip}. Also, β_{ip} denotes a $(Q \times 1)$ vector of coefficients of time variables corresponding to pth gene, with components denoted as β_{iqp}. We define β_ia $(Q \times P)$ matrix of regression coefficients generated by column-wise binding of β_{ip}s.

In the second step, we use LCT to examine the relationship between the covariates and the changing trend of the longitudinal gene set expressions. If there is no covariate related variability in the gene set-specific regression coefficient, there will be no relationship between the covariates and the changing trend of the specific gene set expressions. In other words, there is no linear combination of the columns of $\beta = [\beta_1^T | \ldots | \beta_I^T]^T$ associated to any linear combination of covariates measurements. Therefore, if $C_i = (C_{i1}, \ldots, C_{iU})$ is the vector of time-invariant covariates for the ith subject and $C_u = (C_{1u}, \ldots, C_{Iu}), u = 1, \ldots, U$, the null hypothesis can be formulated as" no association between any of the linear combination of $C_1, \ldots, C_U$ with any linear combination of columns of β.

Let C^* be a $((I.Q) \times (U))$ matrix obtained by duplicating the covariates measurements of subject i for Q times. Then, let

$$Z(C^*, \mathrm{A}) = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1U} \\ \vdots & \vdots & \ddots & \vdots \\ C_{11} & C_{12} & \cdots & C_{1U} \\ C_{21} & C_{22} & \ldots & C_{2U} \\ \vdots & \vdots & \ddots & \vdots \\ C_{21} & C_{22} & \cdots & C_{2U} \\ \vdots & \vdots & \ddots & \vdots \\ C_{I1} & C_{I2} & \cdots & C_{IU} \\ C_{I1} & C_{I2} & \cdots & C_{IU} \end{bmatrix}_{(I.Q)\times(U)} \times \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_U \end{bmatrix}_{U\times 1} \tag{3.12}$$

be a linear combination of the columns of matrix C^*, and,

$$Z(\mathrm{B}, \Gamma) = \beta_{(I.Q)\times P} \times \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_P \end{bmatrix}_{(P)\times 1} \tag{3.13}$$

be a linear combination of columns of β. As before, LCT can be used to find out the maximum correlation between Z(C*, A) and Z(B,Γ), and test its significance.

3.3.4 Design of Simulation Study

A simulation study was designed to evaluate the performance of LLCT method and compare its performance with PAVR proposed by Adewale et al [29]. Several simulations were carried out by varying number of subjects, gene set sizes, number of repeated measurements, within-gene set correlation, within-subject correlation and gene set effect sizes. The number of subjects and gene set size changed from 30,50 to 100.

For each gene set, gene expressions are simulated from $MVN(\mathrm{M}_G, \Sigma_G)$ where $\mathrm{M}_G$is the mean vector of gene expressions, taken from a truncated exponential distribution with $\lambda = 0.7$. Σ_G is

the variance-covariance matrix of genes within a gene set. The variances of the genes were set at $\sigma_G^2 = 0.5$ and the correlations between genes were set at $\rho_{G=}$.0.1, 0.5 or 0.7The effect of within-gene set correlation on the performance of the method was evaluated.

For each gene set, the longitudinal data was simulated based on the following model:

$$y_{ij} = \mathrm{B}_1 \times GS_i + \mathrm{B}_2 \times t_i + \mathrm{B}_3 \times GS_i \times t_i + b_{0i} + b_{1i} \times t_i + \varepsilon_{ij} \quad (3.14)$$

Where y_{ij} denotes the jth observation of the ith subject; GS_i is the vector of gene expression measurements for ith subject; $\mathrm{B}_1$is the vector of fixed effects of the genes on the longitudinal phenotype, with values of 0.05,0.1 and 0.2 for all the subjects; t_i is the measurement time vector of the ith subject varying from one subject to another. The length of t_i is set at 3,4 and 5 in different simulations, but the time points of measurement was uniformly distributed between 1 and 10. B_2 is the vector of fixed effect of time on phenotype, set at 0.3 for all the subjects. $\mathrm{B}_3$is the vector of fixed effects of interactions of gene expressions at time and was set at 0.25,0.05 and 0.1 for all subjects in different simulations. $b_{0i} \sim N(0,1)$ and $b_{1i} \sim N(0,2)$ are the random constant and the random effect of subject i, respectively and are assumed to be independent among subjects. ε_{ij} is the error term defining the variation of the jth observation of subject i. ε_{ij} is assumed to be correlated within subjects. In this simulation, the correlation structure of ε_{ij} is autoregressive and we assumed: $cor(\varepsilon_k, \varepsilon_l) = \rho_\varepsilon^{k-l}$ where $\rho_\varepsilon = 0.2,0.5\ or\ 0.7$.

For LLCT simulation, we simulated 1000 gene sets in each run and each p-value was calculated based on 1,000 permutations. In simulations of PAVR, the results are based on 50 permutation times.

3.4 Results

3.4.1 Simulation Study

We present here results of our simulation study on LLCT performance. Figures 3.1-3.5 show the power of LLCT analyzing diverse set of data, simulated by considering different within-gene-set and within-subject correlations, sample and gene set sizes and number of repeated measurements. For each plot, the type I error was constant at 0.05 and the simulated data were similar for all characteristics except the one mentioned at the top of the plot. The power was calculated at the presence of different B_3 values, determining the effect of each gene within specific gene set over time. The power of LLCT increased by higher within-gene-set correlation, sample size and gene set size (Figure 3.1-3.3). However, it remains unaffected by within-subject correlation and number of repeated measurements (Figure 3.4-3.5).

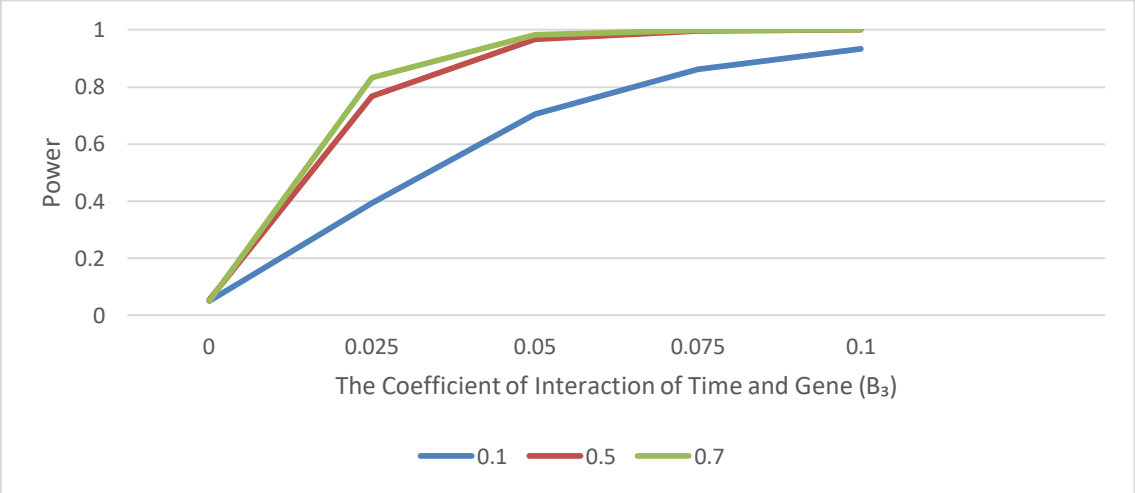


Figure 3.1 Calculation of the power of LLCT using simulated data generated with different within-gene set correlation. Type I error is set at 5%. For each plot, the simulation variables except the one mentioned on the title varies but remains comparable among the curves.

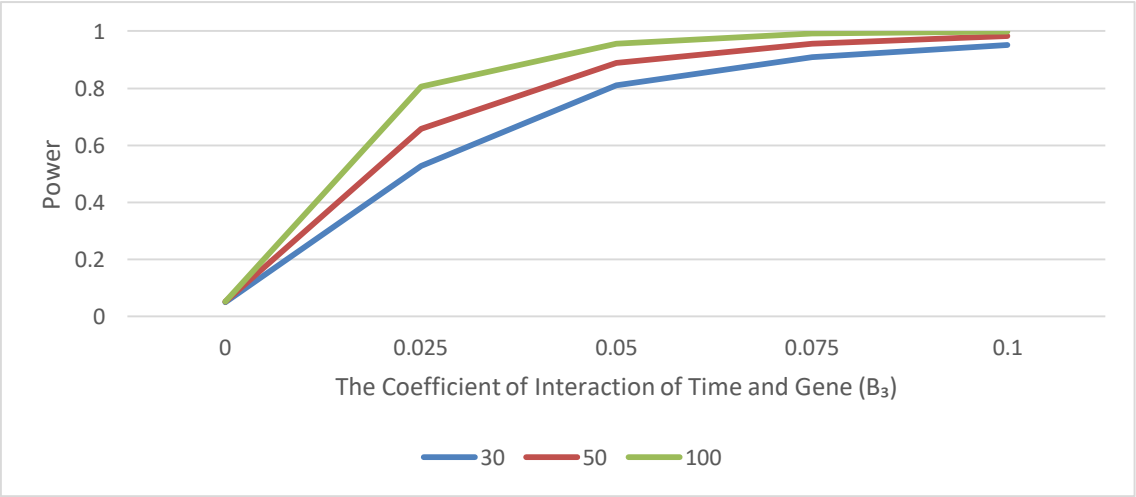


Figure 3.2 Calculation of the power of LLCT using simulated data generated with different sample size. Type I error is set at 5%. For each plot, the simulation variables except the one mentioned on the title varies but remains comparable among the curves.

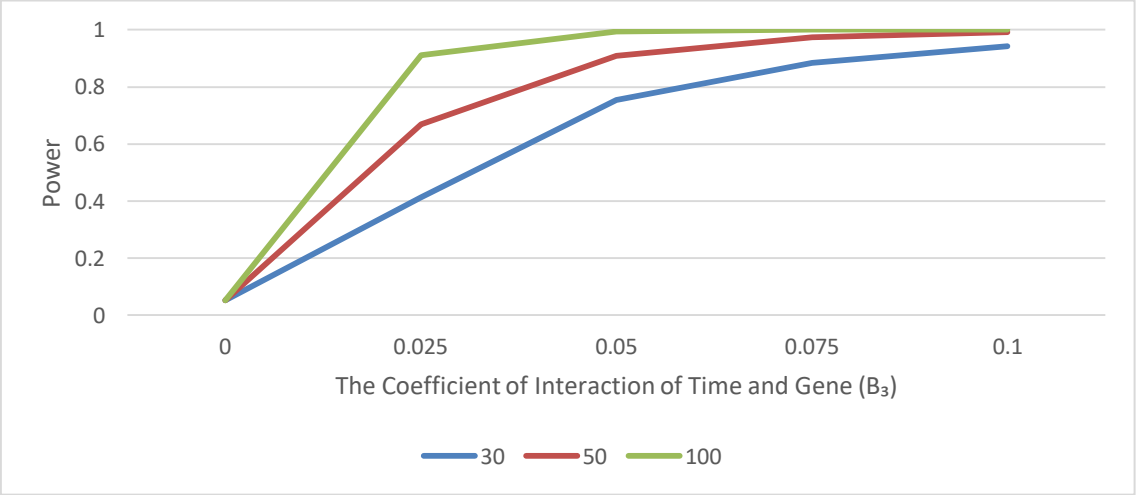


Figure 3.3 Calculation of the power of LLCT using simulated data generated with different gene set size. Type I error is set at 5%. For each plot, the simulation variables except the one mentioned on the title varies but remains comparable among the curves.

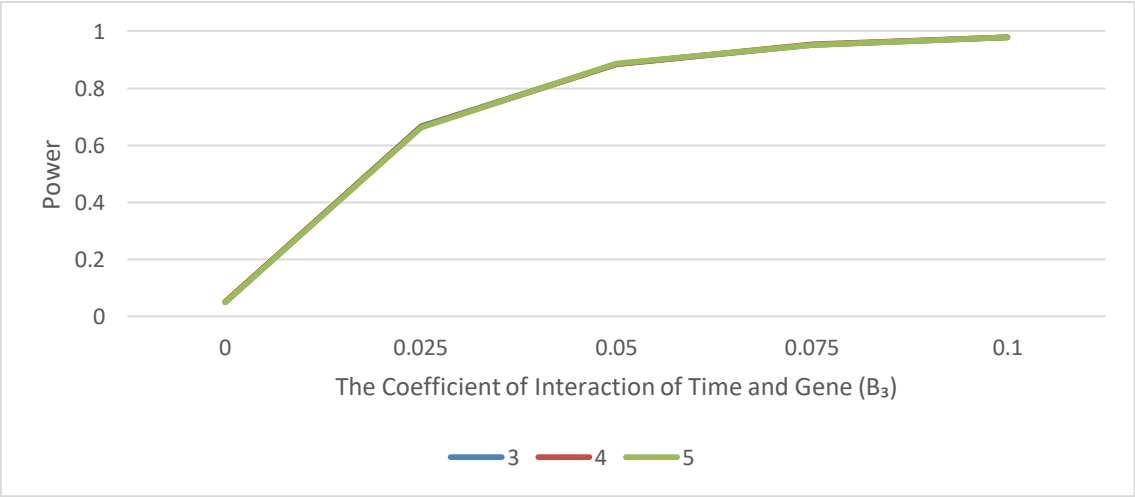


Figure 3.4 Calculation of the power of LLCT using simulated data generated with different number of repeated. Type I error is set at 5%. For each plot, the simulation variables except the one mentioned on the title varies but remains comparable among the curves.

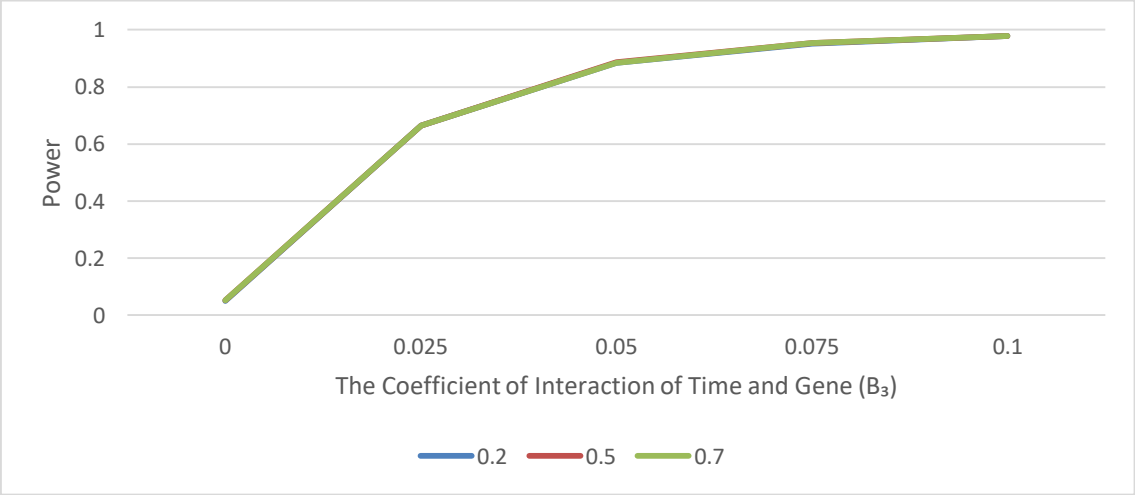


Figure 3.5 Calculation of the power of LLCT using simulated data generated with different within-subject correlation. Type I error is set at 5%. For each plot, the simulation variables except the one mentioned on the title varies but remains comparable among the curves.

The power of LLCT was compared with the power of PAVR in Figure 3.2 where we let within-gene set correlation, sample size, gene set size and number of repeated measurements change. PAVR does not distinguish between the gene effect and the gene effect over time. Therefore, two parameters of B_1 and B_3were set at different values (other than zero for both) to define alternative hypotheses for this method. However, the power of LLCT was consistent over different values of B_1 and altered by B_3 only. For small within-gene-set correlation values ($\rho < 0.5$), LLCT significantly outperformed PAVR. However, as the within-gene-set correlation increased, the difference between the power values of PAVR and LLCT became smaller (Figure 3.2 (A,B,C)). Comparing with LLCT, PAVR performed poorly when the sample was small (Figure 3.2 (D,E)). Furthermore, different gene set sizes did not make a considerable difference between the methods' powers (Figure 3.2 (F,G)). LLCT exhibited a better ability in dealing with large number of repeated measurements over time (Figure 3.2 (H,I)).

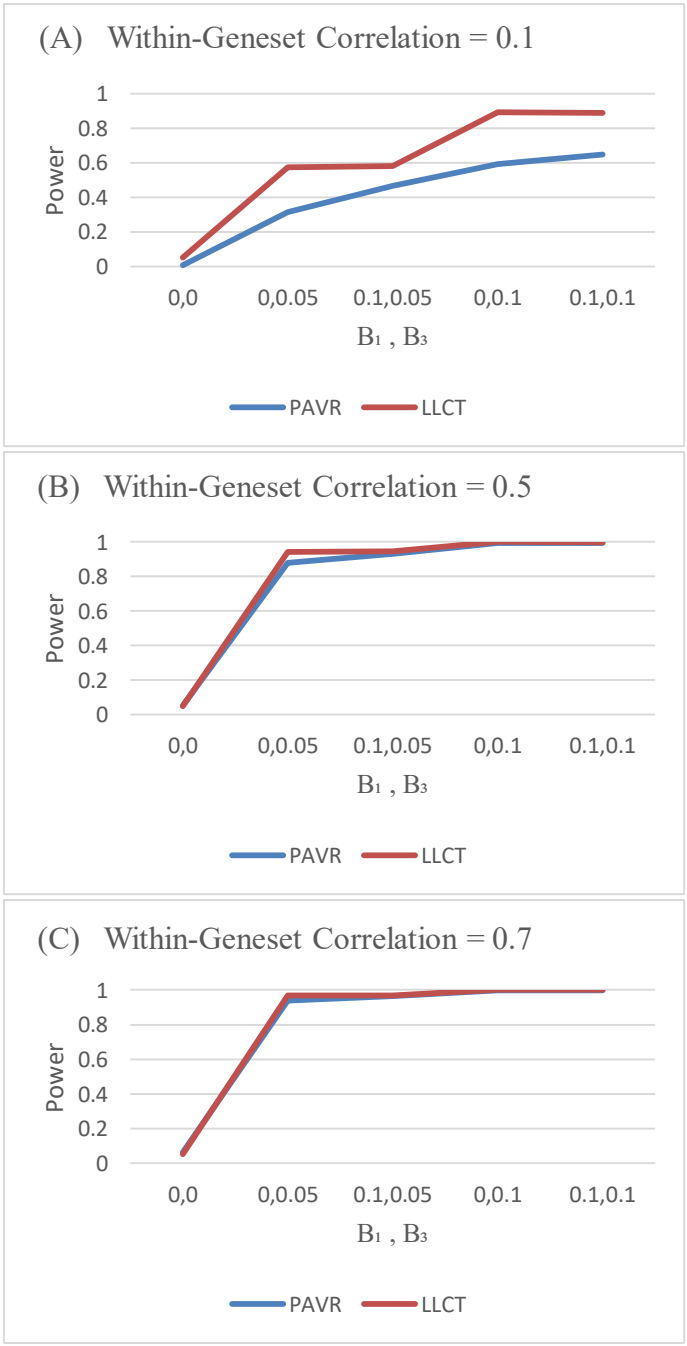


Figure 3.6 Comparison of powers of LLCT method and the method of pathway analysis via regression (PAVR) proposed by Adewale et a1, using simulated data generated with different within-gene-set correlation. $\mathbf{B_1}$ denotes the gene effect and $\mathbf{B_3}$ denotes the gene effect over time referring to equation 3.14.

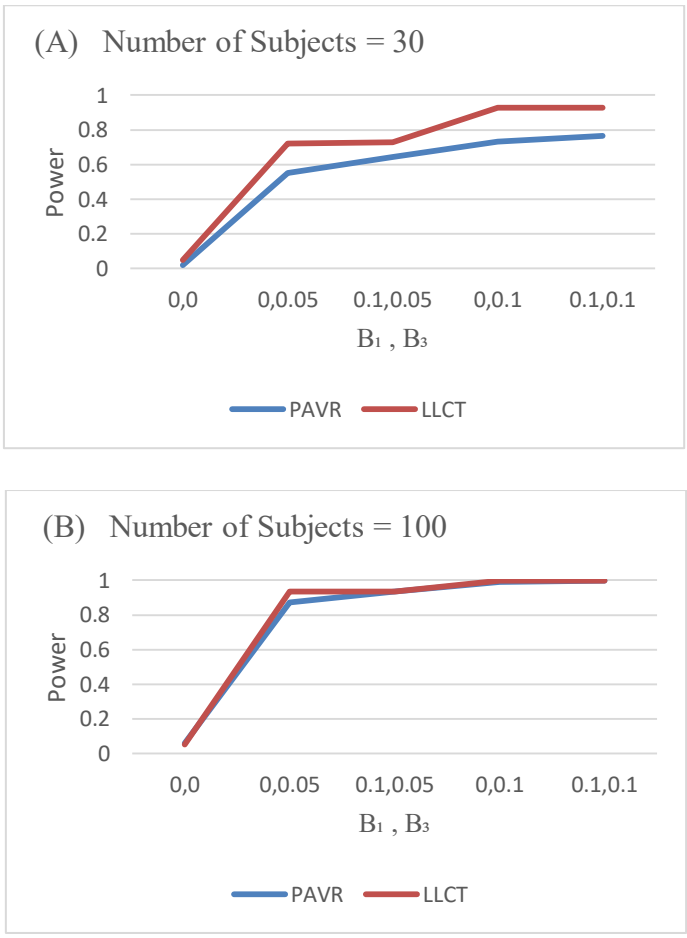
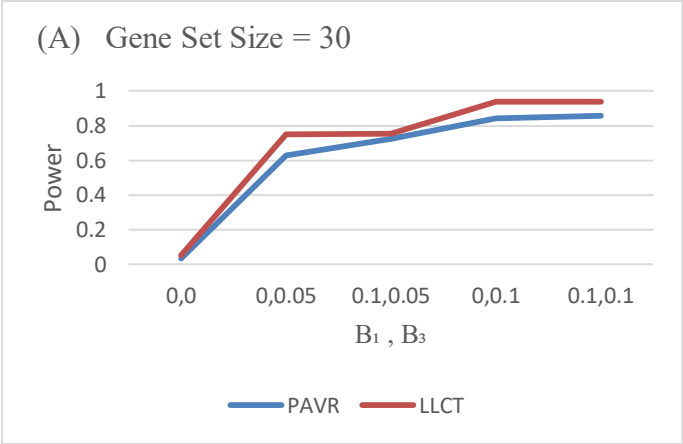


Figure 3.7 Comparison of powers of LLCT method and the method of pathway analysis via regression (PAVR) proposed by Adewale et a1, using simulated data generated with different sample size. $\mathbf{B_1}$ denotes the gene effect and $\mathbf{B_3}$ denotes the gene effect over time referring to equation 3.14.



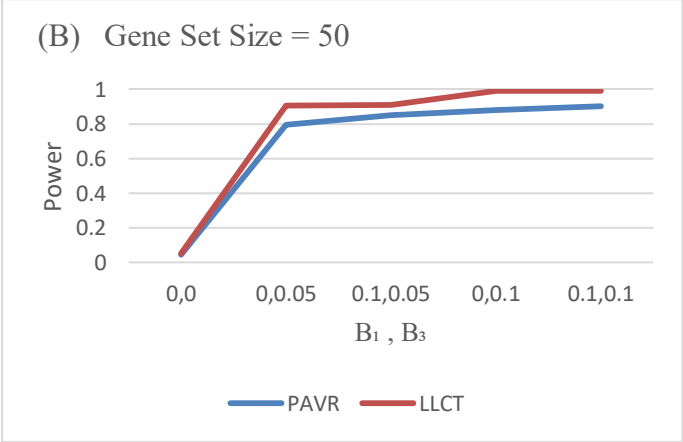
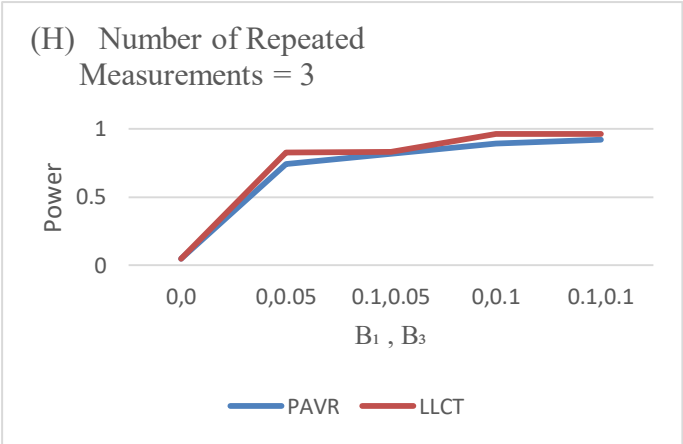


Figure 3.8 Comparison of powers of LLCT method and the method of pathway analysis via regression (PAVR) proposed by Adewale et a1, using simulated data generated with different gene set size. $\mathbf{B_1}$ denotes the gene effect and $\mathbf{B_3}$ denotes the gene effect over time referring to equation 3.14.



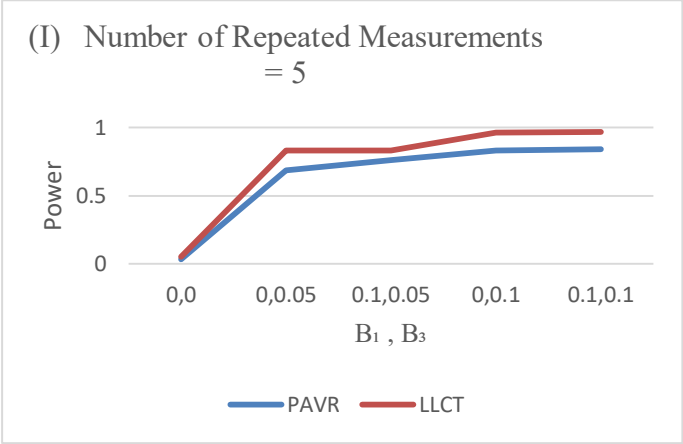


Figure 3.9 Comparison of powers of LLCT method and the method of pathway analysis via regression (PAVR) proposed by Adewale et a1, using simulated data generated with different number of repeated measurements. $\mathbf{B_1}$ denotes the gene effect and $\mathbf{B_3}$ denotes the gene effect over time referring to equation 3.14.

3.4.2 Application

Hypertension which affects more than a quarter of the world's adult population [122] annually adds a significant burden on healthcare systems. Long-term hypertension damages heart, kidney, brain, large blood vessels and retinal vessels[123] and explains about half of stroke and ischaemic heart diseases worldwide. Despite this high health risk, hypertension is

unknown for more than 30% of patients, untreated for 50% of them and uncontrolled for 75%[124].

Blood pressure is known as a highly-heritable complex trait[125] regulated by multiple environmental and genetic factors. The importance of understanding the genetics mechanism of blood pressure on identification of therapeutic and prevention targets has been emphasized in studies examining the variation of effectiveness of antihypertensive medications on different ancestral groups[126].

Hypertension is developed by small contributions of a large number of genes whose effects may be hard to detect. Facing this challenge, most studies on hypertension genetics failed to reach replication. Traditional approaches and small sample sizes may be the most probable reason explaining this deficiency. However, the novel statistical methods have come to help.

Genetic Analysis Workshops (GAWs) are designed to evaluate the performance of different statistical methods applied on high density genotype. Among them, GAW13[127], GAW16[128], GAW18[129] and GAW19[130] have focused on analysis of longitudinal datasets. GAW19 [130], the focus of our work, is based on data from San Antonio Family Heart Study (SAFHS), conducted to investigate the genetics of cardiovascular disease in Mexican Americans. GAW19 researchers were divided into different teams to work on heterogeneous statistical methods dealing with longitudinal datasets. For analysis of gene expressions, these teams independently worked on different areas of individual or pathway gene analysis, unrelated or family-based analysis and joint or separated analysis of phenotypes. However, utilizing heterogeneous statistical methods prevented them from replicating their findings.

The subjects of SAFHS were born in a large, multi-generational family and their stated pedigree relationships were verified. The transcriptional profile data of 647 people was recorded, including 16,383 gene expression measurements, for each individual. For each subject, systolic blood pressure (SBP), diastolic blood pressure (DBP), hypertension status (HTN), use of antihypertensive medications and smoking status were measured at four time points and the subjects' sex and age were recorded. By applying the proposed method to this family-based data, we detected differentially expressed gene sets significantly associated with blood pressure trajectories over time. We analyzed real dataset and considered DBP, SBP, pulse pressure (PP) (defined as PP = SBP − DBP), and hypertension (defined as blood pressure ≥140/90 mm Hg) as the outcome variables.

Table 3.1 Summary information (mean (standard deviation)) of covariates and outcomes at different time points: GAW19 application, studies of related and unrelated subjects

	Age	Antihypertensive Medication	Smoking Status	Systolic Blood Pressure (SBP)	Diastolic Blood Pressure (DBP)	Hypertension Status (HTN)
Related Subjects						
First visit	39.58 (16.88)	0.1(0.3)	0.23(0.42)	121.73(18.98)	71.48(9.99)	0.18(0.39)
Second visit	42.76(15.93)	0.19(0.39)	0.18(0.39)	124.96(19.34)	71.94(10.01)	0.28(0.45)
Third visit	46.34(15.10)	0.29(0.45)	0.2(0.4)	125.21(18.04)	70.73(10.02)	0.36(0.48)
Forth visit	50.88 (12.76)	0.43(0.5)	0.11(0.32)	128.24(17.63)	77.76(11.06)	0.52(0.5)
Unrelated Subjects						
First visit	53.84(14.77)	0.22(0.42)	0.25(0.43)	130.3(23.36)	72.96(9.48)	0.37(0.48)
Second visit	58.26(12.30)	0.36(0.48)	0.11(0.32)	135.01(20.17)	72.34(10.09)	0.59(0.49)
Third visit	59.52(10.85)	0.53(0.50)	0.17(0.38)	130.46(19.24)	69.14 (9.74)	0.59(0.49)
Forth visit	62.16(9.26)	0.63(0.49)	0.06(0.25)	135.5(23.44)	77.06(15.4)	0.71(0.46)

We first analyzed the unrelated subjects by selecting the subjects with no shared parents. In this part of analysis, the repeatedly measured expressions of 10,072 genes for 64 subjects, belonging to 5,898 gene sets were examined by LLCT for unrelated subjects. The gene sets are

defined by Gene Ontology database. The size of gene sets varied from 2 to 1,417 with median of 22.

In the second part of analysis, 647 related subjects in 17 family clusters were analyzed. The size of families varied from 21 to 62 with the median of 31. The total number of 10,072 genes contributing in 5,907 pathways was tested by LLCT for related subjects.

The test of association was conducted after adjustment for either smoking status or antihypertensive medications intake. As some subjects were measured for two times only, the method was unable to adjust for both time-dependent covariates at the same time, unless we restricted our subjects to those with more than 2 measurements.

Table 3.2 The number of significant gene sets found by LLCT at different levels of confidence, testing a variety of outcomes and datasets

Datasets	Type I Error	SBP	DBP	SBP& DBP*	SBP-DBP**	HTN
Adjusted for smoking status						
Related Subjects	1%	30	23	20	73	65
	5%	170	135	141	360	321
	10%	255	278	310	434	392
Unrelated Subjects	1%	12	3	5	27	5
	5%	136	39	60	389	82
	10%	408	78	245	735	162
Adjusted for antihypertensive medications						
Related Subjects	1%	98	13	63	127	12
	5%	402	127	271	541	99
	10%	413	242	390	614	159
Unrelated Subjects	1%	17	3	11	17	2
	5%	142	60	86	116	22
	10%	465	75	186	382	88

No Adjustment						
Related Subjects	1%	18	17	14	43	54
	5%	158	141	122	259	327
	10%	263	273	277	386	417
Unrelated Subjects	1%	9	2	3	17	2
	5%	234	37	70	273	71
	10%	537	68	231	682	168

*The multiple analysis of systolic and diastolic blood pressure. In this analysis, the outcome is a linear combination of SBP and DBP with the highest association with the linear combinations of gene expressions.

** Pulse pressure which is the difference between systolic and diastolic blood pressures.

LLCT was used to find the gene sets whose expressions are significantly associated with the outcome(s) and calculated 5,989 p-values for testing the gene sets in unrelated study and 5,907 p-values for analysis of family-based dataset. Table 3.2 shows the number of significant gene sets in testing each outcome and each dataset separately. The pathways that were significantly associated with both pulse pressure and linear combination of SBP and DBP, after adjusting for antihypertensive medication consumption, were selected and shown in Tables 3.3 and 3.4. Exposure to blood pressure medication, compared to smoking, showed more considerable effect in changing SBP and DBP trajectories and the best model is the one adjusting for this effect.

In Tables 3.3 and 3.4, the gene sets were classified based on their shared ancestral categories, derived from Gene Ontology Tree. We list here biological processes defined by gene sets identified significant by our studies of both unrelated, and related subjects: a few descendent pathways of immune system process, cellular response to stimulus, cell

communication, cellular metabolic process, multi-organism cellular process, multi-cellular organism process and metabolic process. Cell differentiation, cell activation, cell cycle, cellular component organization or biogenesis, biological regulation, system development, localization, metabolic process and response to stimulus are other parental classes of biological process with significant descending pathways in analysis of related dataset only. Aside from biological processes, few significant pathways in major classes of molecular function and cell components were found significant. The family-based analysis is expected to result in more accurate findings, as it works on the larger database.

Blood pressure is a complex phenotype that is controlled by multiple biological process, multiple molecular functions and multiple cell components. Comparing the results of analysis of multiple phenotypes, pulse pressure displayed higher level of robustness and was less affected by covariates. Also, HTN failed to reflect the changes of SBP and DBP and mostly failed to agree with the analysis results of other phenotypes. From statistical perspective, the result of HTN analysis is limited because the information is lost by dichotomizing the continuous variables. Also, many biological studies doubted the reliability of this one-size-fits-all stratification scheme[131]. The other noteworthy finding of this study was the difference between SBP and DBP trajectories in their association with gene expressions. There were larger number of pathways associated with SBP compared to DBP. This underlines the sensitivity of SBP, as a blood pressure measurement, to gene expression alterations.

By discussing the list of significant pathways in Table 3.4, insights can be gained into the genetic of hypertension. However, we admit that an in-depth biological interpretation of the findings is beyond the scope of this thesis. Below, we will discuss some processes underlying hypertension, whose presence was supported by more than one significant pathway in LLCT analysis.

Table 3.3 Results of LLCT of association between the expressions of different gene sets and various measures of blood pressure for UNRELATED subjects in GAW19 database

	Gene set size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Molecular function																	
Organic Hydroxy Compound Transmembrane Transporter Activity	32	p-value	0.19	0.78	0.146	0.029**	0.899	0.034**	0.768	0.039**	0.044**	0.519	0.09*	1	0.106	0.027**	0.848
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Snrna Binding	31	p-value	0.297	0.96	0.321	0.111	0.891	0.017**	0.632	0.017**	0.014**	0.293	0.103	0.99	0.143	0.063*	0.684
		q-value	0.307	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Voltage Gated Calcium Channel Activity	21	p-value	0.038**	0.62	0.044**	0.012**	0.532	0.026**	0.365	0.044**	0.033**	0.143	0.015**	0.48	0.026**	0.014**	0.425
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Cell Component																	
Copi Coated Vesicle	22	p-value	0.025**	0.093*	0.048**	0.052*	0.803	0.007***	0.044**	0.015**	0.009***	0.928	0.019**	0.072*	0.05*	0.071*	0.608
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.557	0.392	1.000	0.376	0.167	0.991
Synaptonemal Complex	17	p-value	0.082*	0.34	0.179	0.075*	0.259	0.013**	0.148	0.036**	0.02**	0.178	0.047**	0.16	0.121	0.123	0.296
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Organellar Small Ribosomal Subunit	25	p-value	0.044**	0.27	0.115	0.06*	0.297	0.033**	0.096*	0.044**	0.03**	0.404	0.023**	0.1	0.056*	0.068*	0.307
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Cytosolic Ribosome	97	p-value	0.205	0.83	0.313	0.117	0.577	0.03**	0.194	0.044**	0.026**	0.046**	0.071*	0.56	0.181	0.089*	0.451
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Biological Process																	
Immune System Process																	
Regulation Of Inflammatory Response To Antigenic Stimulus	15	p-value	0.016**	0.19	0.033**	0.007***	0.602	0.015**	0.118	0.018**	0.011**	0.626	0.004***	0.087*	0.019**	0.017**	0.710
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Negative Regulation Of Toll Like Receptor Signaling Pathway	16	p-value	0.392	0.49	0.653	0.55	0.412	0.018**	0.066*	0.045**	0.026**	0.251	0.169	0.27	0.352	0.379	0.259
		q-value	0.315	1.000	0.393	0.212	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.194	0.991
Negative Regulation Of Osteoclast Differentiation	16	p-value	0.059*	0.18	0.113	0.107	0.279	0.001***	0.03**	0.009***	0.007***	0.089*	0.031**	0.17	0.074*	0.071*	0.272
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Cellular Process																	

	Gene set size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Cellular Process: Cellular Response to Stimulus																	
Response To Ph	22	p-value	0.251	0.56	0.124	0.043**	0.623	0.034**	0.586	0.046**	0.039**	0.71	0.048**	0.96	0.048**	0.012**	0.328
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Regulation Of Chemotaxis	119	p-value	0.136	0.92	0.114	0.027**	0.465	0.04**	0.624	0.039**	0.028**	0.153	0.047**	1	0.068*	0.028**	0.341
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Cellular Process: Cell Communication																	
Cellular Process: Cell Communication: Cell-cell Signaling																	
Beta Catenin Destruction Complex Disassembly	16	p-value	0.094*	0.33	0.157	0.093*	0.424	0.027**	0.078*	0.032**	0.03**	0.33	0.038**	0.14	0.085*	0.109	0.375
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Cellular Process: Cell Communication: Signal Transduction																	
G Protein Coupled Receptor Signaling Pathway Coupled To Cyclic Nucleotide Second Messenger	71	p-value	0.068*	0.62	0.088*	0.03**	0.473	0.041**	0.379	0.045**	0.038**	0.271	0.034**	0.63	0.064*	0.029**	0.422
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Cellular Process: Cell Communication: Others																	
Cellular Response To Starvation	83	p-value	0.038**	0.28	0.079*	0.03**	0.766	0.019**	0.079*	0.028**	0.02**	0.582	0.02**	0.13	0.039**	0.032**	0.744
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Cellular Process: Cellular Metabolic Process																	
Regulation Of Protein Deacetylation	22	p-value	0.086*	0.67	0.095*	0.031**	0.381	0.018**	0.098*	0.041**	0.023**	0.202	0.045**	0.33	0.105	0.052*	0.463
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Regulation Of Receptor Internalization	23	p-value	0.267	0.95	0.277	0.128	0.948	0.015**	0.188	0.023**	0.009***	0.892	0.068*	0.53	0.14	0.05*	0.722
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Pteridine Containing Compound Metabolic Process	24	p-value	0.092*	0.77	0.127	0.037**	0.221	0.018**	0.294	0.038**	0.021**	0.691	0.055*	0.65	0.118	0.047**	0.245
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Peptidyl Lysine Trimethylation	21	p-value	0.086*	0.76	0.081*	0.019**	0.972	0.022**	0.306	0.031**	0.019**	0.931	0.021**	0.57	0.037**	0.009***	0.779
		q-value	0.306	1.000	0.367	0.141	0.972	0.578	1.000	0.531	0.579	0.557	0.392	1.000	0.376	0.167	0.991

	Gene set size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Cellular Process: Multi-organism Cellular Process																	
Multi Organism Organelle Organization	19	p-value	0.101	0.35	0.177	0.11	0.629	0.024**	0.155	0.039**	0.033**	0.833	0.04**	0.22	0.11	0.078*	0.761
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Cellular Process: Others																	
B Cell Proliferation	20	p-value	0.036**	0.22	0.075*	0.027**	0.395	0.017**	0.093*	0.044**	0.025**	0.481	0.019**	0.13	0.037**	0.046**	0.378
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Negative Regulation Of Muscle Cell Apoptotic Process	19	p-value	0.086*	0.73	0.095*	0.021**	0.807	0.027**	0.524	0.042**	0.038**	0.336	0.038**	0.63	0.069*	0.019**	0.877
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Positive Regulation Of T Helper Cell Differentiation	16	p-value	0.032**	0.3	0.055*	0.021**	0.942	0.027**	0.282	0.034**	0.036**	0.844	0.013**	0.19	0.033**	0.022**	0.982
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.993
Multicellular Organismal Process																	
Regulation Of Bone Resorption	18	p-value	0.538	0.2	0.07*	0.063*	0.950	0.023**	0.27	0.018**	0.013**	0.904	0.023**	0.98	0.04**	0.021**	0.387
		q-value	0.345	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Multicellular Organismal Process: System Process																	
Regulation Of Vasodilation	27	p-value	0.059*	0.25	0.125	0.088*	0.924	0.02**	0.222	0.025**	0.015**	0.838	0.023**	0.18	0.065*	0.056*	0.883
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991
Metabolic Process																	
Multicellular Organism Metabolic Process	42	p-value	0.132	0.52	0.228	0.135	0.851	0.009***	0.112	0.022**	0.012**	0.183	0.055*	0.3	0.145	0.107	0.716
		q-value	0.306	1.000	0.367	0.141	0.969	0.578	1.000	0.531	0.579	0.555	0.392	1.000	0.376	0.167	0.991

*Significance level of 0.1

**Significant level of 0.05

***Significance level of 0.001

† The multiple analysis of systolic and diastolic blood pressure measurements

‡The pulse pressure: difference between systolic and diastolic blood pressure values

Table 3.4 Results of LLCT of association between the expressions of different gene sets and various measures of blood pressure for RELATED subjects in GAW19 database

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP& DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Molecular fuction																	
Binding																	
Flavin Adenine Dinucleotide Binding	49	p-value	0.568	0.161	0.105	0.051*	0.099*	0.081*	0.218	0.031**	0.008***	0.323	0.507	0.171	0.152	0.063*	0.069*
		q-value	0.850	0.952	0.743	0.337	0.351	0.287	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Antigen Binding	65	p-value	0.063*	0.303	0.162	0.065*	0.17	0.006***	0.283	0.012**	0.008***	0.179	0.067*	0.293	0.177	0.087*	0.159
		q-value	0.850	0.952	0.743	0.341	0.355	0.234	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Basal Transcription Machinery Binding	24	p-value	0.78	0.014**	0.017**	0.081*	0.251	0.604	0.031**	0.026**	0.048**	0.236	0.738	0.023**	0.024**	0.125	0.187
		q-value	0.850	0.952	0.743	0.342	0.361	0.388	0.997	0.424	0.088	1.000	0.849	0.830	0.778	0.516	0.270
Single Stranded Dna Binding	74	p-value	0.417	0.007***	0.021**	0.218	0.055*	0.291	0.024**	0.025**	0.039**	0.236	0.436	0.015**	0.027**	0.243	0.051*
		q-value	0.850	0.952	0.743	0.370	0.351	0.332	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.517	0.269
Integrin Binding	62	p-value	0.066*	0.915	0.061*	0.009***	0.329	0.002***	0.833	0.001***	0.001***	0.568	0.087*	0.838	0.095*	0.017**	0.247
		q-value	0.850	0.952	0.743	0.334	0.369	0.182	0.997	0.267	0.050	1.000	0.849	0.831	0.778	0.516	0.279
Damaged Dna Binding	53	p-value	0.113	0.797	0.204	0.106	0.34	0.009***	0.776	0.031**	0.02**	0.605	0.141	0.798	0.307	0.134	0.263
		q-value	0.850	0.952	0.743	0.353	0.369	0.234	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.279
Snap Receptor Activity	35	p-value	0.207	0.161	0.012**	0***	0.567	0.085*	0.298	0.001***	0***	0.865	0.239	0.201	0.016**	0.003***	0.429
		q-value	0.850	0.952	0.743	0.000	0.391	0.290	0.997	0.267	0.000	1.000	0.849	0.830	0.778	0.516	0.292
Transcription Cofactor Binding	18	p-value	0.297	0.187	0.08*	0.019**	0.192	0.053*	0.32	0.016**	0.004***	0.872	0.31	0.234	0.094*	0.033**	0.161
		q-value	0.850	0.952	0.743	0.337	0.355	0.270	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.269
Growth Factor Activity	66	p-value	0.003***	0.37	0.012**	0.019**	0.417	0***	0.221	0***	0.011**	0.03**	0.002***	0.303	0.01**	0.015**	0.414
		q-value	0.850	0.952	0.743	0.337	0.375	0.000	0.997	0.000	0.085	1.000	0.849	0.830	0.778	0.516	0.291
Heat Shock Protein Binding	74	p-value	0.462	0.052*	0.084*	0.125	0.016**	0.181	0.118	0.049**	0.015**	0.505	0.425	0.068*	0.131	0.187	0.019**
		q-value	0.850	0.952	0.743	0.355	0.343	0.314	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Actin Binding	262	p-value	0.189	0.512	0.399	0.208	0.3	0.012**	0.519	0.049**	0.035**	0.669	0.205	0.558	0.442	0.199	0.198
		q-value	0.850	0.952	0.743	0.369	0.368	0.234	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.517	0.271

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP& DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Catalytic Activity																	
Hydrolase Activity Hydrolyzing N Glycosyl Compounds	19	p-value	0.49	0.507	0.294	0.076*	0.219	0.061*	0.676	0.049**	0.012**	0.633	0.482	0.593	0.313	0.086*	0.207
		q-value	0.850	0.952	0.743	0.342	0.355	0.280	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.273
Metalloendopeptidase Activity	62	p-value	0.224	0.265	0.086*	0.009***	0.35	0.214	0.301	0.042**	0.01**	0.902	0.294	0.254	0.082*	0.013**	0.297
		q-value	0.850	0.952	0.743	0.334	0.369	0.321	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.280
Cell Component																	
Organelle																	
Vesicle Membrane	337	p-value	0.414	0.271	0.229	0.04**	0.116	0.2	0.445	0.033**	0.002***	0.961	0.432	0.26	0.242	0.04**	0.115
		q-value	0.850	0.952	0.743	0.337	0.352	0.317	0.997	0.424	0.065	1.000	0.849	0.830	0.778	0.516	0.269
Organellar Large Ribosomal Subunit	31	p-value	0.141	0.137	0.041**	0.007***	0.35	0.047**	0.13	0.039**	0.005***	0.296	0.14	0.121	0.039**	0.005***	0.325
		q-value	0.850	0.952	0.743	0.334	0.369	0.268	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.282
Vacuolar Membrane	441	p-value	0.352	0.216	0.093*	0.023**	0.471	0.169	0.317	0.043**	0.017**	0.074*	0.355	0.24	0.131	0.044**	0.453
		q-value	0.850	0.952	0.743	0.337	0.382	0.314	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.294
cell: intrecellular																	
U2 Type Spliceosomal Complex	24	p-value	0.318	0.906	0.304	0.085*	0.14	0.013**	0.92	0.023**	0.012**	0.803	0.287	0.871	0.287	0.092*	0.122
		q-value	0.850	0.952	0.743	0.342	0.355	0.237	0.997	0.424	0.085	1.000	0.849	0.831	0.778	0.516	0.269
Inclusion Body	56	p-value	0.029**	0.701	0.052*	0.01**	0.619	0.004***	0.698	0.016**	0.009***	0.203	0.037**	0.685	0.068*	0.022**	0.604
		q-value	0.850	0.952	0.743	0.334	0.396	0.223	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.311
Coated Vesicle Membrane	108	p-value	0.06*	0.575	0.064*	0.004***	0.807	0.006***	0.621	0.008***	0.002***	0.18	0.073*	0.623	0.092*	0.018**	0.723
		q-value	0.850	0.952	0.743	0.316	0.426	0.234	0.997	0.424	0.065	1.000	0.849	0.830	0.778	0.516	0.327
Vacuolar Part	508	p-value	0.022**	0.222	0.045**	0.025**	0.516	0.01**	0.243	0.026**	0.024**	0.731	0.016**	0.181	0.043**	0.04**	0.45
		q-value	0.850	0.952	0.743	0.337	0.386	0.234	0.997	0.424	0.086	1.000	0.849	0.830	0.778	0.516	0.294

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Intrinsic Component Of Mitochondrial Inner Membrane	17	p-value	0.016**	0.656	0.075*	0.057*	0.148	0.007***	0.665	0.022**	0.021**	0.237	0.03**	0.714	0.086*	0.053*	0.151
		q-value	0.850	0.952	0.743	0.337	0.355	0.234	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Cyclin Dependent Protein Kinase Holoenzyme Complex	26	p-value	0.045**	0.602	0.078*	0.008***	0.216	0.016**	0.656	0.021**	0.009***	0.397	0.074*	0.542	0.088*	0.012**	0.223
		q-value	0.850	0.952	0.743	0.334	0.355	0.262	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.276
cell: intracellular: intracellular organelle																	
Intermediate Filament	52	p-value	0.001***	0.82	0.006***	0.003***	0.244	0.001***	0.781	0.003***	0.006***	0.266	0.001***	0.788	0.011**	0.002***	0.333
		q-value	0.850	0.952	0.743	0.316	0.359	0.144	0.997	0.424	0.081	1.000	0.849	0.830	0.778	0.516	0.283
Centrosome	378	p-value	0.071*	0.812	0.159	0.061*	0.516	0.005***	0.795	0.027**	0.018**	0.331	0.069*	0.855	0.179	0.074*	0.445
		q-value	0.850	0.952	0.743	0.337	0.386	0.234	0.997	0.424	0.085	1.000	0.849	0.831	0.778	0.516	0.293
Synaptonemal Complex	17	p-value	0.119	0.002***	0.004***	0.003***	0.013**	0.076*	0.011**	0.001***	0.001***	0.111	0.119	0.006***	0.009***	0.006***	0.016**
		q-value	0.850	0.952	0.743	0.316	0.343	0.285	0.997	0.267	0.050	1.000	0.849	0.830	0.778	0.516	0.269
Transcription Elongation Factor Complex	45	p-value	0.045**	0.485	0.097*	0.035**	0.628	0.008***	0.394	0.017**	0.015**	0.207	0.027**	0.406	0.089*	0.059*	0.579
		q-value	0.850	0.952	0.743	0.337	0.398	0.234	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.308
Cop9 Signalosome	30	p-value	0.041**	0.846	0.018**	0.002***	0.517	0.048**	0.8	0.032**	0.01**	0.979	0.046**	0.825	0.025**	0.004***	0.523
		q-value	0.850	0.952	0.743	0.257	0.386	0.268	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.301
endomembrane system																	
Platelet Alpha Granule	52	p-value	0.023**	0.7	0.059*	0.022**	0.699	0***	0.487	0.004***	0.008***	0.092*	0.031**	0.662	0.075*	0.028**	0.664
		q-value	0.850	0.952	0.743	0.337	0.408	0.000	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.319
Recycling Endosome Membrane	28	p-value	0.051*	0.846	0.146	0.037**	0.828	0.007***	0.744	0.03**	0.022**	0.096*	0.065*	0.82	0.18	0.069*	0.767
		q-value	0.850	0.952	0.743	0.337	0.430	0.234	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.334
Recycling Endosome	88	p-value	0.073*	0.812	0.099*	0.033**	0.5	0.004***	0.693	0.01**	0.004***	0.21	0.091*	0.752	0.148	0.065*	0.354
		q-value	0.850	0.952	0.743	0.337	0.385	0.223	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.283
membrane																	

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Clathrin Coat	42	p-value	0.147	0.583	0.033**	0.002***	0.203	0.064*	0.695	0.02**	0.001***	0.955	0.183	0.553	0.059*	0.006***	0.167
		q-value	0.850	0.952	0.743	0.257	0.355	0.280	0.997	0.424	0.050	1.000	0.849	0.830	0.778	0.516	0.269
Extrinsic Component Of Cytoplasmic Side Of Plasma Membrane	63	p-value	0.076*	0.093*	0.216	0.05*	0.013**	0.013**	0.256	0.014**	0.002***	0.253	0.087*	0.156	0.222	0.052*	0.023**
		q-value	0.850	0.952	0.743	0.337	0.343	0.237	0.997	0.424	0.065	1.000	0.849	0.830	0.778	0.516	0.269
Others																	
Excitatory Synapse	105	p-value	0.026**	0.906	0.044**	0.01**	0.95	0.012**	0.846	0.006***	0.004***	0.254	0.03**	0.911	0.064*	0.006***	0.92
		q-value	0.850	0.952	0.743	0.334	0.456	0.234	0.997	0.424	0.076	1.000	0.849	0.832	0.778	0.516	0.363
Lamellipodium	122	p-value	0.123	0.281	0.037**	0.002***	0.018**	0.012**	0.434	0.002***	0***	0.98	0.139	0.313	0.057*	0.007***	0.016**
		q-value	0.850	0.952	0.743	0.257	0.343	0.234	0.997	0.424	0.000	1.000	0.849	0.830	0.778	0.516	0.269
Intercellular Bridge	37	p-value	0.119	0.717	0.234	0.086*	0.176	0.012**	0.661	0.024**	0.02**	0.179	0.104	0.651	0.252	0.104	0.184
		q-value	0.850	0.952	0.743	0.342	0.355	0.234	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Presynapse	163	p-value	0.262	0.735	0.377	0.077*	0.17	0.006***	0.596	0.008***	0.004***	0.075*	0.249	0.688	0.387	0.098*	0.166
		q-value	0.850	0.952	0.743	0.342	0.355	0.234	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.269
Biological Process																	
developmental process: multicellular organismal process: system development																	
Cardiac Chamber Development	78	p-value	0.393	0.863	0.642	0.246	0.129	0.017**	0.716	0.046**	0.03**	0.396	0.395	0.828	0.653	0.278	0.111
		q-value	0.850	0.952	0.748	0.377	0.352	0.262	0.997	0.424	0.087	1.000	0.849	0.830	0.779	0.517	0.269
Endothelial Cell Development	36	p-value	0.237	0.838	0.46	0.256	0.342	0.016**	0.661	0.045**	0.044**	0.447	0.232	0.827	0.471	0.254	0.33
		q-value	0.850	0.952	0.743	0.378	0.369	0.262	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.517	0.282
Coronary Vasculature Development	24	p-value	0.34	0.64	0.225	0.035**	0.294	0.02**	0.699	0.027**	0.006***	0.409	0.331	0.653	0.235	0.066*	0.242
		q-value	0.850	0.952	0.743	0.337	0.368	0.264	0.997	0.424	0.081	1.000	0.849	0.830	0.778	0.516	0.279
Pituitary Gland Development	16	p-value	0.278	0***	0.003***	0.51	0.003***	0.092*	0***	0.004***	0.046**	0.006***	0.228	0.002***	0.006***	0.575	0.008***
		q-value	0.850	0.000	0.743	0.427	0.320	0.294	0.000	0.424	0.087	1.000	0.849	0.830	0.778	0.563	0.269

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP& DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Ventral Spinal Cord Development	19	p-value	0.688	0.083*	0.045**	0.045**	0.271	0.345	0.147	0.039**	0.018**	0.779	0.714	#N/A	0.054*	0.053*	0.283
		q-value	0.850	0.952	0.743	0.337	0.363	0.342	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.280
Digestive System Development	67	p-value	0.02**	0.397	0.066*	0.068*	0.593	0.007***	0.264	0.018**	0.049**	0.111	0.026**	0.313	0.057*	0.084*	0.662
		q-value	0.850	0.952	0.743	0.342	0.395	0.234	0.997	0.424	0.088	1.000	0.849	0.830	0.778	0.516	0.319
Embryonic Heart Tube Development	37	p-value	0.173	0.694	0.04**	0.007***	0.804	0.087*	0.657	0.038**	0.005***	0.989	0.19	0.651	0.051*	0.004***	0.698
		q-value	0.850	0.952	0.743	0.334	0.426	0.292	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.323
Exocrine System Development	26	p-value	0.408	0.165	0.156	0.054*	0.049**	0.187	0.207	0.036**	0.009***	0.9	0.467	0.149	0.16	0.06*	0.065*
		q-value	0.850	0.952	0.743	0.337	0.351	0.315	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Blood Vessel Morphogenesis	216	p-value	0.03**	0.912	0.028**	0.004***	0.642	0.008***	0.947	0.015**	0.005***	0.467	0.048**	0.919	0.049**	0.01**	0.643
		q-value	0.850	0.952	0.743	0.316	0.401	0.234	0.997	0.424	0.076	1.000	0.849	0.833	0.778	0.516	0.317
Organ Growth	37	p-value	0.001***	0.453	0.01**	0.002***	0.675	0.011**	0.537	0.03**	0.012**	0.509	0.005***	0.528	0.012**	0.001***	0.679
		q-value	0.850	0.952	0.743	0.257	0.404	0.234	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.320
Olfactory Lobe Development	17	p-value	0.055*	0.815	0.073*	0.005***	0.968	0.034**	0.823	0.048**	0.006***	0.845	0.083*	0.839	0.09*	0.013**	0.955
		q-value	0.850	0.952	0.743	0.334	0.460	0.266	0.997	0.424	0.081	1.000	0.849	0.831	0.778	0.516	0.371
Negative Regulation Of Developmental Process	463	p-value	0.188	0.426	0.084*	0.017**	0.944	0.035**	0.531	0.032**	0.006***	0.937	0.205	0.4	0.111	0.025**	0.92
		q-value	0.850	0.952	0.743	0.337	0.455	0.268	0.997	0.424	0.081	1.000	0.849	0.830	0.778	0.516	0.363
Developmental Process Involved In Reproduction	343	p-value	0.016**	0.347	0.056*	0.059*	0.525	0.011**	0.312	0.02**	0.037**	0.876	0.019**	0.333	0.066*	0.063*	0.532
		q-value	0.850	0.952	0.743	0.337	0.387	0.234	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.516	0.302
Regulation Of Keratinocyte Differentiation	18	p-value	0.188	0.287	0.066*	0.013**	0.032**	0.015**	0.42	0.006***	0***	0.632	0.169	0.275	0.097*	0.025**	0.049**
		q-value	0.850	0.952	0.743	0.334	0.343	0.259	0.997	0.424	0.000	1.000	0.849	0.830	0.778	0.516	0.269
Positive Regulation Of Muscle Tissue Development	28	p-value	0.009***	0.395	0.024**	0.011**	0.129	0.002***	0.423	0.006***	0.013**	0.026**	0.011**	0.395	0.038**	0.016**	0.172
		q-value	0.850	0.952	0.743	0.334	0.352	0.182	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Positive Regulation Of Dendritic Spine Development	23	p-value	0.94	0.136	0.055*	0.068*	0.025**	0.227	0.145	0.017**	0.001***	0.72	0.947	0.097*	0.07*	0.096*	0.021**
		q-value	0.850	0.952	0.743	0.334	0.352	0.182	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Embryonic Heart Tube	28	p-value	0.349	0.863	0.346	0.061*	0.25	0.019**	0.914	0.025**	0.007***	0.701	0.327	0.856	0.409	0.09*	0.215

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP& DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Morphogenesis		q-value	0.850	0.952	0.743	0.337	0.361	0.264	0.997	0.424	0.085	1.000	0.849	0.831	0.778	0.516	0.273
Positive Regulation Of Neuron Projection Development	146	p-value	0.005***	0.92	0.006***	0***	0.344	0.005***	0.976	0.02**	0.009***	0.757	0.012**	0.956	0.006***	0.003***	0.393
		q-value	0.850	0.952	0.743	0.000	0.369	0.234	0.997	0.424	0.085	1.000	0.849	0.835	0.778	0.516	0.289
Developmental Growth	199	p-value	0.058*	0.138	0.116	0.208	0.533	0.003***	0.096*	0.007***	0.041**	0.139	0.047**	0.13	0.087*	0.258	0.443
		q-value	0.850	0.952	0.743	0.369	0.387	0.210	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.517	0.293
localization																	
Regulation Of Telomerase Rna Localization To Cajal Body	15	p-value	0.083*	0.957	0.088*	0.015**	0.951	0.018**	0.921	0.037**	0.017**	0.566	0.091*	0.927	0.094*	0.022**	0.922
		q-value	0.850	0.953	0.743	0.337	0.457	0.264	0.997	0.424	0.085	1.000	0.849	0.833	0.778	0.516	0.363
Regulation Of Leukocyte Migration	105	p-value	0.126	0.76	0.06*	0.006***	0.315	0.052*	0.758	0.029**	0.006***	0.991	0.167	0.645	0.068*	0.008***	0.341
		q-value	0.850	0.952	0.743	0.334	0.368	0.270	0.997	0.424	0.081	1.000	0.849	0.830	0.778	0.516	0.283
Ameboidal Type Cell Migration	86	p-value	0.254	0.758	0.332	0.057*	0.397	0.009***	0.731	0.016**	0.003***	0.65	0.235	0.766	0.374	0.069*	0.321
		q-value	0.850	0.952	0.743	0.337	0.373	0.234	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.280
Negative Regulation Of Establishment Of Protein Localization	147	p-value	0.002***	0.212	0.019**	0.03**	0.667	0.002***	0.151	0.008***	0.021**	0.084*	0.003***	0.207	0.021**	0.027**	0.666
		q-value	0.850	0.952	0.743	0.337	0.404	0.182	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.319
Protein Localization To Chromosome	36	p-value	0.347	0.123	0.021**	0.008***	0.339	0.247	0.098*	0.014**	0.009***	0.647	0.424	0.099*	0.018**	0.008***	0.267
		q-value	0.850	0.952	0.743	0.334	0.369	0.326	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.280
Regulation Of Cellular Extravasation	17	p-value	0.178	0.869	0.184	0.024**	0.129	0.006***	0.928	0.016**	0.005***	0.709	0.183	0.864	0.23	0.047**	0.106
		q-value	0.850	0.952	0.743	0.337	0.352	0.234	0.997	0.424	0.076	1.000	0.849	0.831	0.778	0.516	0.269
Regulation Of Ryanodine Sensitive Calcium Release Channel Activity	15	p-value	0.101	0.773	0.216	0.106	0.216	0.003***	0.569	0.007***	0.022**	0.292	#N/A	0.746	0.236	0.134	0.183
		q-value	0.850	0.952	0.743	0.353	0.355	0.210	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Divalent Inorganic Cation Transport	152	p-value	0.177	0.78	0.2	0.025**	0.367	0.023**	0.831	0.029**	0.001***	0.924	0.181	0.76	0.166	0.027**	0.277
		q-value	0.850	0.952	0.743	0.337	0.371	0.264	0.997	0.424	0.050	1.000	0.849	0.830	0.778	0.516	0.280
metabolic process																	
Polysaccharide	16	p-value	0.057*	0.774	0.144	0.052*	0.251	0.001***	0.431	0.004***	0.009***	0.145	0.051*	0.633	0.125	0.068*	0.237

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP& DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Catabolic Process		q-value	0.850	0.952	0.743	0.337	0.361	0.144	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.278
Cellular Carbohydrate Catabolic Process	22	p-value	0.128	0.951	0.217	0.038**	0.059*	0***	0.899	0.002***	0.001***	0.318	0.14	0.944	0.245	0.061*	0.065*
		q-value	0.850	0.953	0.743	0.337	0.351	0.000	0.997	0.424	0.050	1.000	0.849	0.833	0.778	0.516	0.269
Gpi Anchor Metabolic Process	28	p-value	0.09*	0.898	0.123	0.032**	0.646	0.012**	0.758	0.027**	0.008***	0.356	0.121	0.845	0.163	0.051*	0.545
		q-value	0.850	0.952	0.743	0.337	0.402	0.234	0.997	0.424	0.085	1.000	0.849	0.831	0.778	0.516	0.304
Alcohol Metabolic Process	218	p-value	0.7	0.057*	0.09*	0.119	0.003***	0.468	0.098*	0.047**	0.02**	0.468	0.696	0.052*	0.103	0.178	0.003***
		q-value	0.850	0.952	0.743	0.355	0.320	0.367	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Fucosylation	15	p-value	0.176	0.606	0.062*	0.002***	0.092*	0.018**	0.787	0.008***	0.001***	0.885	0.243	0.651	0.098*	0.014**	0.068*
		q-value	0.850	0.952	0.743	0.257	0.351	0.264	0.997	0.424	0.050	1.000	0.849	0.830	0.778	0.516	0.269
Lipid Catabolic Process	137	p-value	0.285	0.163	0.01**	0.001***	0.937	0.234	0.195	0.017**	0.01**	0.861	0.355	0.196	0.028**	0.006***	0.851
		q-value	0.850	0.952	0.743	0.257	0.453	0.324	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.350
Ubiquitin Dependent Protein Catabolic Process Via The Multivesicular Body Sorting Pathway	15	p-value	0.624	0.515	0.333	0.119	0.071*	0.062*	0.654	0.039**	0.008***	0.88	0.644	0.526	0.391	0.175	0.065*
		q-value	0.850	0.952	0.743	0.355	0.351	0.280	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Regulation Of Autophasome Assembly	30	p-value	0.189	0.162	0.243	0.4	0.339	0.002***	0.075*	0.006***	0.044**	0.047**	0.146	0.12	0.182	0.468	0.278
		q-value	0.850	0.952	0.743	0.410	0.369	0.182	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.553	0.280
Protein O Linked Glycosylation	62	p-value	0.512	0.572	0.371	0.097*	0.125	0.026**	0.637	0.02**	0.005***	0.703	0.551	0.523	0.409	0.133	0.135
		q-value	0.850	0.952	0.743	0.350	0.352	0.264	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.269
Regulation Of Gluconeogenesis	31	p-value	0.121	0.838	0.171	0.071*	0.383	0.024**	0.821	0.049**	0.011**	0.42	0.144	0.82	0.218	0.084*	0.372
		q-value	0.850	0.952	0.761	0.487	0.372	0.390	0.997	0.514	0.152	1.000	0.849	0.830	0.793	0.589	0.297
Multicellular Organismal Macromolecule Metabolic Process	36	p-value	0.394	0.119	0.054*	0.019**	0.059*	0.283	0.223	0.022**	0.008***	0.619	0.418	0.145	0.069*	0.036**	0.065*
		q-value	0.850	0.952	0.743	0.337	0.351	0.332	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
immune system process																	
Negative Regulation Of Production Of Molecular Mediator Of Immune	19	p-value	0.258	0.392	0.097*	0.042**	0.06*	0.034**	0.496	0.019**	0.005***	0.986	0.285	0.367	0.12	0.039**	0.063*

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP& DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Response		q-value	0.850	0.952	0.743	0.337	0.351	0.266	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.269
Toll Like Receptor Signaling Pathway	71	p-value	0.151	0.739	0.046**	0.006***	0.082*	0.011**	0.669	0.005***	0.002***	0.712	0.183	0.687	0.074*	0.013**	0.089*
		q-value	0.850	0.952	0.743	0.334	0.351	0.234	0.997	0.424	0.065	1.000	0.849	0.830	0.778	0.516	0.269
Regulation Of Megakaryocyte Differentiation	20	p-value	0.07*	0.825	0.087*	0.022**	0.536	0.012**	0.762	0.023**	0.006***	0.352	0.087*	0.802	0.129	0.043**	0.425
		q-value	0.850	0.952	0.743	0.337	0.387	0.234	0.997	0.424	0.081	1.000	0.849	0.830	0.778	0.516	0.292
Positive T Cell Selection	17	p-value	0.264	0.572	0.092*	0.009***	0.2	0.019**	0.739	0.011**	0.002***	0.924	0.274	0.653	0.117	0.024**	0.176
		q-value	0.850	0.952	0.743	0.334	0.355	0.264	0.997	0.424	0.065	1.000	0.849	0.830	0.778	0.516	0.269
Negative Regulation Of Myeloid Leukocyte Differentiation	30	p-value	0.447	0.215	0.067*	0.018**	0.301	0.121	0.349	0.022**	0.009***	0.904	0.467	0.208	0.079*	0.042**	0.263
		q-value	0.850	0.952	0.743	0.337	0.368	0.302	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.279
Osteoclast Differentiation	20	p-value	0.739	0.126	0.109	0.044**	0.026**	0.083*	0.271	0.017**	0.004***	0.898	0.768	0.171	0.142	0.088*	0.023**
		q-value	0.850	0.952	0.743	0.337	0.343	0.289	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.269
Negative Regulation Of Osteoclast Differentiation	16	p-value	0.11	0.547	0.053*	0.007***	0.718	0.037**	0.69	0.021**	0.008***	0.782	0.111	0.575	0.063*	0.014**	0.544
		q-value	0.850	0.952	0.743	0.334	0.411	0.268	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.304
Regulation Of Production Of Molecular Mediator Of Immune Response	70	p-value	0.001***	0.396	0.009***	0.012**	0.868	0***	0.339	0.004***	0.023**	0.211	0.001***	0.374	0.008***	0.012**	0.942
		q-value	0.850	0.952	0.743	0.334	0.439	0.000	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.367
Regulation Of B Cell Differentiation	18	p-value	0.236	0.552	0.119	0.013**	0.356	0.057*	0.688	0.046**	0.013**	0.868	0.257	0.622	0.136	0.035**	0.285
		q-value	0.850	0.952	0.743	0.334	0.370	0.275	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.280
response to stimulus																	
Response To Acid Chemical	200	p-value	0.323	0.999	0.485	0.214	0.079*	0.014**	0.966	0.023**	0.014**	0.226	0.327	0.998	0.51	0.207	0.102
		q-value	0.850	0.960	0.743	0.369	0.351	0.248	0.997	0.424	0.085	1.000	0.849	0.845	0.778	0.517	0.269
Regulation Of Intracellular Estrogen Receptor Signaling Pathway	20	p-value	0.072*	0.991	0.059*	0.013**	0.322	0.012**	0.999	0.017**	0.004***	0.507	0.096*	0.974	0.083*	0.014**	0.314
		q-value	0.850	0.952	0.743	0.408	0.351	0.369	0.997	0.492	0.165	1.000	0.849	0.830	0.778	0.546	0.269
Response To Camp	62	p-value	0.021**	0.536	0.052*	0.044**	0.559	0.001***	0.39	0.01**	0.02**	0.334	0.018**	0.488	0.059*	0.053*	0.501
		q-value	0.850	0.952	0.743	0.337	0.391	0.144	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.299

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Detection Of Biotic Stimulus	18	p-value	0.691	0.2	0.16	0.097*	0.018**	0.128	0.327	0.02**	0.005***	0.712	0.644	0.221	0.234	0.112	0.016**
		q-value	0.850	0.952	0.743	0.350	0.343	0.303	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.269
Detection Of Stimulus	141	p-value	0.136	0.845	0.31	0.166	0.363	0.002***	0.771	0.015**	0.016**	0.648	0.166	0.866	0.334	0.206	0.389
		q-value	0.850	0.952	0.743	0.362	0.371	0.182	0.997	0.424	0.085	1.000	0.849	0.831	0.778	0.517	0.288
Response To Gonadotropin	17	p-value	0.091*	0.508	0.006***	0.001***	0.286	0.01**	0.615	0.001***	0***	0.965	0.136	0.524	0.015**	0.002***	0.222
		q-value	0.850	0.952	0.743	0.257	0.366	0.234	0.997	0.267	0.000	1.000	0.849	0.830	0.778	0.516	0.276
Regulation Of Camp Metabolic Process	69	p-value	0.082*	0.556	0.24	0.126	0.397	0.007***	0.33	0.022**	0.036**	0.099*	0.072*	0.503	0.202	0.116	0.414
		q-value	0.850	0.952	0.743	0.355	0.373	0.234	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.516	0.291
Cellular Defense Response	45	p-value	0.101	0.902	0.098*	0.012**	0.595	0.032**	0.948	0.043**	0.016**	0.631	0.121	0.893	0.125	0.026**	0.492
		q-value	0.850	0.952	0.743	0.334	0.395	0.264	0.997	0.424	0.085	1.000	0.849	0.831	0.778	0.516	0.298
Positive Chemotaxis	15	p-value	0.116	0.262	0.015**	0.001***	0.475	0.211	0.218	0.028**	0.013**	0.893	0.161	0.227	0.025**	0.006***	0.468
		q-value	0.850	0.952	0.743	0.257	0.383	0.321	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.297
cellular process																	
cellular process: cell communication: cell-cell signaling																	
Non Canonical Wnt Signaling Pathway	104	p-value	0.022**	0.603	0.052*	0.04**	0.278	0.003***	0.592	0.022**	0.018**	0.126	0.034**	0.638	0.09*	0.055*	0.284
		q-value	0.850	0.952	0.743	0.337	0.364	0.210	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.280
Excitatory Postsynaptic Potential	15	p-value	0.95	0.307	0.198	0.155	0.059*	0.236	0.434	0.041**	0.014**	0.936	0.968	0.257	0.234	0.198	0.044**
		q-value	0.856	0.952	0.743	0.359	0.351	0.324	0.997	0.424	0.085	1.000	0.858	0.830	0.778	0.517	0.269
Canonical Wnt Signaling Pathway	55	p-value	0.043**	0.763	0.039**	0.002***	0.381	0.027**	0.817	0.049**	0.011**	0.741	0.052*	0.754	0.052*	0.017**	0.395
		q-value	0.850	0.952	0.743	0.257	0.372	0.264	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.289
Signal Release	98	p-value	0.622	0.246	0.094*	0.04**	0.115	0.069*	0.489	0.024**	0.004***	0.914	0.625	0.324	0.145	0.06*	0.096*
		q-value	0.850	0.952	0.743	0.337	0.352	0.281	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.269
cellular process: cell communication: signal transduction																	

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP& DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Platelet Derived Growth Factor Receptor Signaling Pathway	25	p-value	0.135	0.955	0.194	0.043**	0.161	0.011**	0.927	0.021**	0.007***	0.372	0.123	0.956	0.211	0.044**	0.161
		q-value	0.850	0.952	0.759	0.417	0.355	0.392	0.997	0.480	0.127	1.000	0.849	0.830	0.794	0.546	0.269
Negative Regulation Of Erk1 And Erk2 Cascade	39	p-value	0.276	0.473	0.492	0.437	0.136	0.013**	0.305	0.038**	0.041**	0.284	0.242	0.448	0.435	0.425	0.142
		q-value	0.850	0.952	0.743	0.417	0.355	0.237	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.545	0.269
Negative Regulation Of Signal Transduction In Absence Of Ligand	22	p-value	0.46	0.448	0.245	0.091*	0.054*	0.034**	0.734	0.033**	0.008***	0.665	0.397	0.481	0.298	0.112	0.05*
		q-value	0.850	0.952	0.743	0.346	0.351	0.266	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Positive Regulation Of Erk1 And Erk2 Cascade	106	p-value	0.055*	0.852	0.062*	0.01**	0.159	0.001***	0.764	0***	0.002***	0.416	0.064*	0.818	0.088*	0.026**	0.136
		q-value	0.850	0.952	0.743	0.334	0.355	0.144	0.997	0.000	0.065	1.000	0.849	0.830	0.778	0.516	0.269
cellular process: cellular metabolic process																	
Phospholipid Dephosphorylation	22	p-value	0.031**	0.652	0.062*	0.029**	0.206	0.008***	0.564	0.024**	0.013**	0.197	0.032**	0.603	0.085*	0.033**	0.21
		q-value	0.850	0.952	0.743	0.337	0.355	0.234	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.273
Glutamate Metabolic Process	20	p-value	0.125	0.911	0.287	0.141	0.228	0.015**	0.76	0.048**	0.033**	0.276	0.118	0.907	0.284	0.114	0.209
		q-value	0.850	0.952	0.743	0.355	0.358	0.259	0.997	0.424	0.087	1.000	0.849	0.832	0.778	0.516	0.273
Ribonucleoside Triphosphate Biosynthetic Process	42	p-value	0.212	0.223	0.367	0.375	0.164	0.01**	0.173	0.035**	0.04**	0.172	0.209	0.206	0.283	0.417	0.153
		q-value	0.850	0.952	0.743	0.403	0.355	0.234	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.544	0.269
Positive Regulation Of Phosphorus Metabolic Process	659	p-value	0.15	0.334	0.028**	0.002***	0.335	0.067*	0.426	0.024**	0.003***	0.244	0.165	0.389	0.049**	0.006***	0.369
		q-value	0.850	0.952	0.743	0.257	0.369	0.281	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.285
Telomere Maintenance Via Recombination	25	p-value	0.024**	0.329	0.067*	0.021**	0.04**	0.002***	0.423	0.003***	0.001***	0.634	0.043**	0.359	0.128	0.033**	0.036**
		q-value	0.850	0.952	0.743	0.337	0.351	0.182	0.997	0.424	0.050	1.000	0.849	0.830	0.778	0.516	0.269
Positive Regulation Of Nucleotide Metabolic Process	73	p-value	0.063*	0.84	0.085*	0.012**	0.501	0.002***	0.718	0.009***	0.004***	0.582	0.071*	0.777	0.091*	0.015**	0.49
		q-value	0.850	0.952	0.743	0.334	0.385	0.182	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.298
Negative Regulation Of	49	p-value	0.637	0.182	0.134	0.097*	0.037**	0.17	0.243	0.048**	0.017**	0.799	0.674	0.152	0.156	0.128	0.042**

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP& DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Dephosphorylation		q-value	0.850	0.952	0.743	0.350	0.347	0.314	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Mrna Transcription	16	p-value	0.056*	0.215	0.146	0.212	0.342	0.001***	0.151	0.007***	0.023**	0.137	0.066*	0.184	0.131	0.242	0.324
		q-value	0.850	0.952	0.743	0.369	0.369	0.144	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.517	0.281
Positive Regulation Of Dephosphorylation	33	p-value	0.389	0.555	0.318	0.091*	0.136	0.046**	0.918	0.049**	0.014**	0.674	0.356	0.657	0.299	0.092*	0.146
		q-value	0.850	0.952	0.743	0.346	0.355	0.268	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
cellular process: cellular metabolic process: cellular macromolecule metabolic process: cellular protein metabolic process																	
Histone Methylation	64	p-value	0.157	0.94	0.131	0.017**	0.663	0.029**	0.968	0.037**	0.001***	0.94	0.159	0.957	0.163	0.028**	0.604
		q-value	0.850	0.952	0.743	0.337	0.403	0.264	0.997	0.424	0.050	1.000	0.849	0.835	0.778	0.516	0.311
Histone Monoubiquitination	20	p-value	0.722	0.243	0.109	0.047**	0.21	0.272	0.326	0.036**	0.011**	0.638	0.774	0.279	0.173	0.103	0.148
		q-value	0.850	0.952	0.743	0.337	0.355	0.328	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Protein Polyubiquitination	204	p-value	0.4	0.786	0.238	0.049**	0.099*	0.026**	0.942	0.009***	0.002***	0.941	0.411	0.789	0.292	0.066*	0.088*
		q-value	0.850	0.952	0.743	0.337	0.351	0.264	0.997	0.424	0.065	1.000	0.849	0.830	0.778	0.516	0.269
Peptidyl Glutamic Acid Modification	16	p-value	0.022**	0.9	0.044**	0.013**	0.351	0***	0.804	0.003***	0***	0.381	0.029**	0.908	0.048**	0.012**	0.304
		q-value	0.850	0.952	0.743	0.334	0.369	0.000	0.997	0.424	0.000	1.000	0.849	0.832	0.778	0.516	0.280
Protein Dephosphorylation	138	p-value	0.031**	0.735	0.088*	0.047**	0.907	0.003***	0.588	0.011**	0.016**	0.096*	0.032**	0.721	0.086*	0.04**	0.874
		q-value	0.850	0.952	0.743	0.337	0.448	0.210	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.354
cellular process: cellular response to stimulus																	
Intrinsic Apoptotic Signaling Pathway In Response To Dna Damage	56	p-value	0.441	0.214	0.109	0.036**	0.101	0.08*	0.29	0.029**	0.004***	0.601	0.514	0.204	0.149	0.072*	0.066*
		q-value	0.850	0.952	0.743	0.337	0.351	0.287	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.269
Nucleotide Excision Repair	103	p-value	0.696	0.045**	0.055*	0.119	0.073*	0.549	0.076*	0.037**	0.036**	0.417	0.653	0.033**	0.063*	0.155	0.065*
		q-value	0.850	0.952	0.743	0.355	0.351	0.379	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.516	0.269
Positive Regulation Of Dna Repair	29	p-value	0.316	0.171	0.053*	0.013**	0.177	0.127	0.303	0.021**	0.008***	0.986	0.313	0.181	0.062*	0.027**	0.158
		q-value	0.850	0.952	0.743	0.334	0.355	0.303	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP& DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Regulation Of Response To Reactive Oxygen Species	28	p-value	0.667	0.059*	0.016**	0.023**	0.23	0.357	0.126	0.036**	0.015**	0.49	0.698	0.078*	0.024**	0.028**	0.162
		q-value	0.850	0.952	0.743	0.337	0.358	0.344	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Cellular Response To Amino Acid Stimulus	37	p-value	0.72	0.065*	0.063*	0.068*	0.052*	0.226	0.123	0.029**	0.014**	0.601	0.711	0.071*	0.09*	0.11	0.048**
		q-value	0.850	0.952	0.743	0.342	0.351	0.324	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
cellular process: cell differentiation																	
Positive Regulation Of Fat Cell Differentiation	27	p-value	0.013**	0.702	0.045**	0.013**	0.087*	0.019**	0.737	0.042**	0.02**	0.415	0.027**	0.715	0.061*	0.017**	0.095*
		q-value	0.850	0.952	0.743	0.334	0.351	0.264	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Fat Cell Differentiation	74	p-value	0.062*	0.561	0.165	0.143	0.589	0.008***	0.475	0.033**	0.035**	0.187	0.066*	0.585	0.194	0.143	0.581
		q-value	0.850	0.952	0.743	0.356	0.394	0.234	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.516	0.308
cellular process: cell activation																	
B Cell Activation	93	p-value	0.019**	0.941	0.024**	0.002***	0.656	0.012**	0.897	0.031**	0.024**	0.847	0.025**	0.877	0.022**	0.008***	0.688
		q-value	0.850	0.952	0.743	0.257	0.402	0.234	0.997	0.424	0.086	1.000	0.849	0.831	0.778	0.516	0.321
Lipoprotein Biosynthetic Process	68	p-value	0.049**	0.805	0.068*	0.005***	0.755	0.014**	0.819	0.02**	0.009***	0.46	0.054*	0.782	0.073*	0.009***	0.739
		q-value	0.850	0.952	0.743	0.334	0.419	0.248	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.329
cellular process: cell cycle																	
Negative Regulation Of Mitotic Nuclear Division	26	p-value	0.439	0.634	0.216	0.052*	0.078*	0.038**	0.883	0.034**	0.002***	0.737	0.446	0.703	0.242	0.073*	0.067*
		q-value	0.850	0.952	0.743	0.337	0.351	0.268	0.997	0.424	0.065	1.000	0.849	0.830	0.778	0.516	0.269
Regulation Of Cell Cycle Phase Transition	261	p-value	0.185	0.32	0.087*	0.007***	0.235	0.043**	0.36	0.027**	0.008***	0.531	0.217	0.291	0.095*	0.009***	0.235
		q-value	0.850	0.952	0.743	0.334	0.358	0.268	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.278
cellular process: others																	
Chaperone Mediated Protein Folding	40	p-value	0.019**	0.437	0.065*	0.051*	0.636	0.004***	0.417	0.035**	0.045**	0.185	0.02**	0.421	0.061*	0.048**	0.592
		q-value	0.850	0.952	0.743	0.337	0.399	0.223	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.516	0.309

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Positive Regulation Of Protein Oligomerization	15	p-value	0.149	0.799	0.272	0.068*	0.439	0.009***	0.653	0.031**	0.013**	0.402	0.137	0.785	0.232	0.092*	0.443
		q-value	0.850	0.952	0.743	0.342	0.378	0.234	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.293
cellular component organization or biogenesis																	
Regulation Of Cell Size	117	p-value	0.849	0.071*	0.054*	0.071*	0.188	0.431	#N/A	0.037**	0.026**	0.669	0.889	0.071*	0.072*	0.092*	0.118
		q-value	0.850	0.952	0.743	0.342	0.355	0.360	0.997	0.424	0.086	1.000	0.851	0.830	0.778	0.516	0.269
Membrane Invagination	24	p-value	0.687	0.103	0.044**	0.032**	0.237	0.26	0.16	0.026**	0.016**	0.692	0.721	0.117	0.055*	0.051*	0.204
		q-value	0.850	0.952	0.743	0.337	0.358	0.328	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.272
Positive Regulation Of Protein Polymerization	64	p-value	0.427	0.693	0.373	0.088*	0.174	0.024**	0.746	0.035**	0.005***	0.754	0.441	0.702	0.391	0.131	0.159
		q-value	0.850	0.952	0.743	0.344	0.355	0.264	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.269
Lamellipodium Assembly	21	p-value	0.017**	0.815	0.021**	0.001***	0.801	0.017**	0.735	0.031**	0.001***	0.655	0.028**	0.754	0.026**	0.001***	0.704
		q-value	0.850	0.952	0.743	0.257	0.425	0.262	0.997	0.424	0.050	1.000	0.849	0.830	0.778	0.516	0.325
biological regulation																	
Regulation Of Monooxygenase Activity	41	p-value	0.186	0.488	0.102	0.019**	0.391	0.029**	0.59	0.021**	0.007***	0.676	0.176	0.509	0.125	0.032**	0.38
		q-value	0.850	0.952	0.743	0.337	0.373	0.264	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.287
Negative Regulation Of Nf Kappab Transcription Factor Activity	47	p-value	0.03**	0.888	0.041**	0.004***	0.985	0.008***	0.81	0.03**	0.004***	0.379	0.032**	0.841	0.069*	0.011**	0.971
		q-value	0.850	0.952	0.743	0.316	0.464	0.234	0.997	0.424	0.076	1.000	0.849	0.831	0.778	0.516	0.374
Endoplasmic Reticulum Calcium Ion Homeostasis	17	p-value	0.14	0.892	0.138	0.024**	0.449	0.012**	0.879	0.021**	0.005***	0.56	0.159	0.862	0.169	0.048**	0.459
		q-value	0.850	0.952	0.743	0.337	0.379	0.234	0.997	0.424	0.076	1.000	0.849	0.831	0.778	0.516	0.295
Regulation Of Heart Rate By Cardiac Conduction	15	p-value	0.135	0.867	0.113	0.026**	0.117	0.012**	0.972	0.023**	0.002***	0.874	0.145	0.879	0.154	0.044**	0.12
		q-value	0.850	0.952	0.743	0.337	0.352	0.234	0.997	0.424	0.065	1.000	0.849	0.831	0.778	0.516	0.269
Cardiac Muscle Cell Action Potential	17	p-value	0.084*	0.742	0.057*	0.01**	0.339	0.015**	0.712	0.021**	0.003***	0.508	0.091*	0.722	0.082*	0.015**	0.382
		q-value	0.850	0.952	0.743	0.334	0.369	0.259	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.287
multicellular organismal process																	

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP& DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Regulation Of Tumor Necrosis Factor Superfamily Cytokine Production	83	p-value	0.163	0.874	0.189	0.044**	0.65	0.027**	0.914	0.046**	0.014**	0.707	0.14	0.869	0.215	0.054*	0.512
		q-value	0.850	0.952	0.743	0.337	0.402	0.264	0.997	0.424	0.085	1.000	0.849	0.831	0.778	0.516	0.300
Positive Regulation Of Type I Interferon Production	64	p-value	0.87	0.166	0.129	0.075*	0.128	0.169	0.427	0.044**	0.007***	0.993	0.88	0.265	0.166	0.126	0.083*
		q-value	0.850	0.952	0.743	0.342	0.352	0.314	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.269
Regulation Of Tumor Necrosis Factor Biosynthetic Process	15	p-value	0.371	0.23	0.036**	0.003***	0.504	0.06*	0.23	0.013**	0*	0.361	0.427	0.233	0.06*	0.015**	0.445
		q-value	0.850	0.952	0.743	0.316	0.386	0.279	0.997	0.424	0.000	1.000	0.849	0.830	0.778	0.516	0.293
Fertilization	73	p-value	0.175	0.98	0.203	0.065*	0.197	0.007***	0.978	0.014**	0.003***	0.632	0.185	0.989	0.24	0.076*	0.211
		q-value	0.850	0.957	0.743	0.341	0.355	0.234	0.997	0.424	0.076	1.000	0.849	0.841	0.778	0.516	0.273
Negative Regulation Of Endothelial Cell Migration	25	p-value	0.196	0.83	0.299	0.071*	0.131	0.011**	0.86	0.019**	0.005***	0.749	0.186	0.853	0.308	0.096*	0.153
		q-value	0.850	0.952	0.743	0.342	0.354	0.234	0.997	0.424	0.076	1.000	0.849	0.831	0.778	0.516	0.269
multicellular organismal process: system process																	
Sensory Perception Of Mechanical Stimulus	75	p-value	0.079*	0.513	0.193	0.107	0.211	0.004***	0.329	0.015**	0.019**	0.215	0.086*	0.488	0.204	0.11	0.256
		q-value	0.850	0.952	0.743	0.353	0.355	0.223	0.997	0.424	0.085	1.000	0.849	0.830	0.778	0.516	0.279
Positive Regulation Of Smooth Muscle Contraction	18	p-value	0.042**	0.566	0.138	0.071*	0.76	0.004***	0.286	0.009***	0.037**	0.06*	0.054*	0.471	0.135	0.059*	0.698
		q-value	0.850	0.952	0.743	0.342	0.419	0.223	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.516	0.323
Regulation Of Vascular Permeability	19	p-value	0.638	0.102	0.046**	0.012**	0.503	0.32	0.14	0.038**	0.018**	0.594	0.668	0.127	0.057*	0.028**	0.422
		q-value	0.850	0.952	0.743	0.359	0.369	0.364	0.997	0.432	0.093	1.000	0.849	0.830	0.778	0.517	0.280
multi-organism process																	
Response To Protozoan	17	p-value	0.051*	0.606	0.144	0.144	0.291	0.004***	0.514	0.017**	0.036**	0.102	0.07*	0.578	0.188	0.188	0.24
		q-value	0.850	0.952	0.743	0.356	0.367	0.223	0.997	0.424	0.087	1.000	0.849	0.830	0.778	0.516	0.279
Negative Regulation Of Multi Organism Process	107	p-value	0.005***	0.482	0.035**	0.027**	0.681	0.001***	0.369	0.002***	0.004***	0.08*	0.014**	0.468	0.032**	0.032**	0.63

	GS size		adjusted for Smoking Status					Adjusted for Antihypertensive Medication					No Adjustment				
			SBP	DBP	SBP& DBP	SBP-DBP	HTN	SBP	DBP	SBP&DBP	SBP-DBP	HTN	SBP	DBP	DBP&SBP	SBP-DBP	HTN
Response To Fungus	31	q-value	0.850	0.952	0.743	0.337	0.405	0.144	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.314
		p-value	0.726	0.255	0.093*	0.048**	0.154	0.112	0.52	0.029**	0.005***	0.984	0.713	0.304	0.138	0.061*	0.119
		q-value	0.850	0.952	0.743	0.337	0.355	0.299	0.997	0.424	0.076	1.000	0.849	0.830	0.778	0.516	0.269
others																	
Positive Regulation Of Cell Matrix Adhesion	24	p-value	0.126	0.918	0.205	0.061*	0.242	0.023**	0.913	0.041**	0.011**	0.711	0.162	0.904	0.214	0.089*	0.233
		q-value	0.850	0.952	0.743	0.337	0.358	0.264	0.997	0.424	0.085	1.000	0.849	0.832	0.778	0.516	0.277
Regulation Of Cation Transmembrane Transport	118	p-value	0.371	0.503	0.099*	0.017**	0.107	0.027**	0.798	0.014**	0.001***	0.998	0.374	0.569	0.117	0.037**	0.106
		q-value	0.856	0.952	0.754	0.443	0.351	0.356	0.997	0.432	0.093	1.000	0.858	0.830	0.790	0.572	0.269

*Significance level of 0.1

**Significant level of 0.05

***Significance level of 0.001

† The multiple analysis of systolic and diastolic blood pressure measurements

‡The pulse pressure: difference between systolic and diastolic blood pressure values

Regulation of Smooth Muscle Contraction by signal transduction

Recent developments in blood pressure studies have highlighted the importance of the regulation of vascular smooth muscle contraction and vascular tone on regulation of blood pressure. The young blood vessels are contractible and plastic and as people age, they become synthetic and less contractible in response to proinflammatory stimuli, diet or other factors[132–134].

The significant pathways *negative and positive regulation of ERK1 and ERK2 cascade*, *negative and positive regulation of dephosphorylation*, *protein dephosphorylation*, *actin binding, response to camp* may reveal some biological processes behind the regulation of vascular smooth cell and its subsequent effect on blood pressure regulation. Previous studies have detected significant roles of these pathways and other related pathways in regulation of vascular smooth muscle contraction[135,136]. Brozovich et al. [137]provided a thorough description of these roles.

Regulation of Smooth Muscle Contraction by epigenetic mechanism

Epigenetic mechanism refers to heritable changes of gene expression which are not related to the genome sequence [138]. These mechanisms may contribute in changing plasticity of vascular smooth muscle by either altering the accessibility of transcription factors at DNA regulatory regions or changing the genetic translations [139]. Our study identified *histone methylation* as a significant pathway to alter accessibility of transcription factors by changing chromatin packaging of the cells. Also, significant pathways *messenger RNA transcription*, *basal transcription machinery binding*, *transcription cofactor binding* and *damaged DNA*

binding may reveal more epigenetic mechanisms causing differential transcription of smooth muscle cell.

Cell-cell Signalling: *WNT Signaling*

Non-canonical and canonical WNT pathways were found to be associated with trajectories of pulse pressure and multiple outcome of SBP and DBP. Massive literature has supported the association between *WNT* pathway and hypertension. The study of these pathways has been motivated by heterogeneity of hypertensive patient population in response to antihypertensive medications. Patients with type 2 diabetes mellitus responded poorly to the treatment compared to others.

Many Genome Wide Association Studies (GWAS) suggest the association between hypertension and *WNT3* that encodes a canonical *WNT* ligand and SOX proteins which interact with b-catenin and modulate the transcription of *WNT*-target genes [140–143]. In experiments, mice infused with angiotensin II have been diagnosed with activated b-catenin and proliferated vascular smooth muscle contraction. The other line of evidences supporting this relationship is the association of neurolocal regulation of blood pressure with interaction of insulin and *WNT signaling* [144].

DNA Damage and Genomic Instability

The association between age and development of cardiovascular diseases and hypertension can be explained by pathways related to DNA damage and repair. This result is in agreement with our earlier observation that biological processes of *intrinsic apoptotic signaling pathway in response to DNA damage*, *nucleotide excision repair*, *positive regulation of DNA repair* and

regulation of response to reactive oxygen species (ROS) are significantly associated with blood pressure trajectory over time. Below, there is a description of how these pathways collaborate to develop hypertension.

DNA is damaged by exposure to exogenous and endogenous agents, such as smoking and diabetes mellitus. Aging leads to prolonged exposure, accumulation of DNA damages and elevated production of ROS at the molecular level. In order to preserve genomic stability under ROS-induced stress, multiple pathways to repair or respond to the presence of DNA damage are employed by the cell and their functions may overlap, compromise or exceed the capability to repair DNA. A defective DNA repair system leads to genomic instability and can accelerate development of vascular problems, such as increased blood pressure, increased vascular stiffness and decreased vascular relaxation [145] Also, multiple lines of evidence have suggested the direct or indirect effect of increased ROS on hypertension incidence, affecting blood vessels (contraction, relaxation and growth), heart, kidney[146] and nervous system functions[147]. This path of investigation can promote antioxidant therapies and production of drugs enhancing genomic integrity.

Nervous system development: Pituitary development and ventral spinal cord development

Blood pressure changes can be related to nervous system development. In our study, we found pituitary development as a significant pathway affecting the pulse pressure and SBP&DBP trajectories. Endocrine hypertension, a special type of hypertension, is caused by the pituitary or adrenal gland producing too much or not enough of the hormones [148,149]. Secretion of Antidiuretic hormone (vasopressin) by pituitary gland plays an important role in water retention in kidneys and controlling blood pressure. Furthermore, the imbalanced

influence of the posterior and interior parts of pituitary gland is known to increase blood pressure [150].

The other significant nervous-system-related pathway in this study is spinal cord development. Higher prevalence of hypertension among patients with spinal cord injury as a result of the interruption in the autonomic nervous pathways supports our finding. Reduction in autonomic cardiovascular control of hypertension explains this result [151].

Heart and Blood Vessel Development

Our results are consistent with the significant influence of *cardiac chamber development*, *coronary vasculature development*, *embryonic heart tube development*, *embryonic heart tube morphogenesis* and *blood vessel morphogenesis* pathways on blood pressure trajectories.

The extra load on thin wall chamber or tube caused by increased blood pressure is normalized by an increase in wall thickness and/or by a reduction in chamber/lumen diameter. More specifically, left ventricle adopts its structure in response to imposed stress through remodelling or hypertrophy[152]. At the cellular level, cardiac gene expressions are altered in response to stress stimulus[153].

Overall, this study illustrated the application of LLCT on gene expression data measured on related and unrelated subjects. This was the first attempt to analyze gene sets when the blood pressure is repeatedly measured, and the dataset is clustered by families. Analysis at the gene set level improves interpretability of findings. Incorporating repeated measurements of outcome over time enables us to investigate the temporal progression of phenotype over time. These studies provide the opportunity to investigate genomics under an important assumption: the effects of the genes contributing to the underlying phenotype are persistent over time. Also, the potential genetic and environmental covariates are better controlled via longitudinal study

design. The family-based structure of data decreases heterogeneity leading to more precise investigations. The previous works in GAW19 never had these three features together. Although this study is unique in its kind, our findings have been shown to be mostly consistent with those of experimental or GWAS studies. However, we recognize that our study may not present the best set of pathways involved in blood pressure development because of the following limitations. The first limitation is very common among genomic studies. A single significant gene may lead to the significance of the whole pathway. Second, although we adjusted for anti-hypertensive intake and smoking status, there are many other uncontrolled covariates, such as diet, stress, physical activity [154]. Lack of availability of informative covariates such as behavioral recommendations that accompany medical prescriptions has also been mentioned as a general limitation of GAW19 studies in the summary provided by Chiu et al[155].

3.5 Discussion

The interest in temporal patterns of change in the patients' conditions is becoming increasingly popular, as it explains the complexity of biological systems. Longitudinal studies provide a possibility to study individual development of an outcome over time. They advance our understanding of disease progression or phenotype trajectory. Through longitudinal studies, the development of other variables can also be examined as determinants of the outcome trajectories. Therefore, incorporating longitudinal designs in genetic studies enable examination of genetic variants that affect phenotypes over time [155,156]. Moreover, longitudinal studies are more reliable as the subjects are closely followed up and the onset of the events is precisely

observed [155]. Obviously, there is higher certainty behind the existence of an effect that is detected to be continuously significant over time in the presence of many uncontrolled or unmeasured time-dependent covariates than an effect which is observed once. In other words, multiple measurements and significant trajectory over time provide more reliable evidence than what a single time point measurement and a cross-sectional effect can provide. Adding family structure to the study design weighs more this reliability by detecting a significant genetic effect in a family rather than an individual.

The main purpose of the current study was to develop a statistical method for high-dimensional data able to analyze repeatedly-measured outcomes and covariates. This method offers many interesting flexibilities to the analysis. It allows adjusting for potentially time-dependent covariates. While genetics and environment always interact to shape the phenotype, the result of genetics studies may be biased without taking the environmental factors into account. It also incorporates gene-gene correlations within a gene set into the test statistic. A very common drawback of many available GSA methods is the lack of ability to accommodate between-gene correlation. In addition, LLCT is a self-contained method proven to be powerful and computationally efficient compared to existing methods. This method can be applied to different classes of phenotypes, such as continuous, binary or categorical phenotype if an appropriate model is defined in the first stage. Furthermore, it is applicable to both unbalanced and incomplete data. In longitudinal studies, it is quite common that some subjects are lost to follow up. The evidence from the simulation study suggests higher power of LLCT in comparison to existing method, PAVR [29]. Aside from higher power of LLCT, there are two critical features that discriminate these two methods. First, LLCT is computationally far more efficient. Compared to LLCT, the run time is about 70 times longer for PAVR. For the same reason, we could not design a large simulation for evaluation of PAVR. Second, PAVR is

unable to test the interaction of time and covariate over time and it only tests the covariate effect. The interaction of time and covariate indicates if the covariate's effect varies over time and it is known as the most critical parameter of longitudinal analysis. Without considering this parameter, the longitudinal study resembles a cross-sectional study that takes advantage of multiple measurements for gaining higher accuracy of measurements. Our simulation study also showed that the power, and therefore the required sample size, is dependent on the gene set size and the within-gene-set correlation and it is independent on the number of repeated measurements and within-subjects correlation. Significance of a lower heterogeneity within a larger gene set can be achieved with a smaller sample.

Despite the strengths mentioned above, there are few limitations for this method that need to be considered. Our method dealing with longitudinal phenotype is unable to adjust for time-independent covariates. Including time-independent covariates in the second step of the method may result in misleading findings. As a self-contained method, LCT would identify a set as significant even if a small number of genes, or even if one single gene is associated with the phenotype. One way to address this limitation is to consider reducing the significant sets to their core members. In time-course microarray data analysis, this method can identify the gene sets which are differentially expressed over time in association with a set of covariates. However, our method is unable to distinguish the individual covariates responsible for this difference, unless we include one covariate at a time.

LLCT was applied to GAW19 data. As noted earlier, GAW19 has been analyzed before. However, significant differences across various methods used prevented a meaningful comparison of the results. There are four pedigree-based GAW19 studies exploring the association between phenotype and gene expressions via different methods: linear mixed models, nonparametric weighted U statistics, structural equation modeling, Bayesian unified

frameworks, and multiple regression. However, their results cannot be compared with ours. Here are the differences between our approach and theirs: 1. They incorporated the information of rare variants into their analysis 2. They did not include the priori information of gene pathways. 3. They did not take the longitudinal pattern of the phenotype into the account. There are seven GAW19 pathway-based analysis, three of which explored gene expression data[157]. There are three GAW19 studies with longitudinal analytical approaches, all of them examining genetic variants[155]. The longitudinal studies used generalized estimating equations (GEEs), latent class growth modeling (LCGM), linear mixed-effect (LME), and variance components (VC) in their analysis. Among all these studies, the study of Ziyatdinov et al. which is a gene ontology pathway and family-based enrichment analysis of gene-expression data came closest to the current study, but it is unpublished at the time of submission of this work. They used linear mixed models. GAW19 studies acknowledged higher power of longitudinal methods in detecting genetic effects, decreased trait heterogeneity and smaller standard error of effect estimates[155]. Also, they recognized identification of unique genetic-related trajectories of disease progression missed by the previous studies.

3.6 Conclusion

LLCT method can be used for analysis of complex genetic studies and may result in better reproducibility across studies. LLCT can be applied to a wide range of longitudinal genomics, transcriptomics, proteomics, metabolomics and microbiota data. A very important application of LLCT is to link omics over time, the approach that has been emphasized by recent studies for gaining better understanding of complex biological process. Linkage of omics over time requires a method that can handle large scale outcomes and predictors datasets, simultaneously,

which cannot be accommodated by most methods. Therefore, we think that our method had the potential to contribute efficiently in the future progression of genetic science.

Chapter 4

Discussion

4.1 The demand for Novel Statistical Methods

Complex study designs are crucial for answering complex biomedical and public health questions. The shift from classic toward advanced designs is particularly necessary for the studies with a background of heterogeneous findings, such as genetics. As a result, genetic researchers, today, focus on more complicated study designs such as pedigree-based or time-course studies. Examining the network of omics associations rather than evaluating each omics separately has received increased attention in recent genetic literature, as it improves our grasp of biological systems and may lead to development of personalized health care models. However, generating knowledge from these types of studies is not possible unless technology and statistical methodology are keeping up with the latest advances in study designs.

System biology investigations, and more specifically integration studies, require statistical methodology accommodating the challenging structure of the data. Omics are mostly measured on a continuous scale and they are collected in large numbers. Omics related to a specific biological pathway are correlated and therefore, the approach employed in GSA methods is more advantageous. as Although popular, GSEA does not address analytical challenges for linking omics data. Most importantly, GSEA cannot accommodate multiple, continuous phenotypes. GSEA is unable to detect a significant gene set consisting of multiple correlated genes, with weak to moderate correlations with the outcome[25]. Also, a gene set consisting of

genes whose expressions are not associated with the phenotype, is many times called as significant by GSEA, mostly due to the correlations of genes across the set. GSEA performance depends on the gene set size. The enrichment score of a larger gene set is always higher than the smaller one with a similar correlation ranking[87]. GSEA tends to cancel out positive and negative gene-phenotype associations, resulting in a small enrichment score and large, non-significant p-values. This is true for biological pathway with feedback loops. GSEA fails to identify such pathways as significant. With respect to the classification of competitive or self-contained, GSEA is a combination between the two approached, a hybrid method. Therefore, it is susceptible to important deficiencies of a competitive approach, such as subjectivity and relying on the untenable assumption of independence of genes across a set, as well as significance of a gene set depending on genes outside the set.

Many high-dimensional data analyses have been proposed. However, most of them are not designed for fitting the complex structure of integrated omics, especially within and between omics correlations. Many of the high dimensional existing methods suffer from computational efficiency and/or lack of inferential analysis. A collection of these methods was reviewed by Huang et al.[158], emphasizing the need for developing methods with the ability to consider the interactive relationship among different omics layers. LASSO, as the most popular method in dimension reduction, applicable to multivariate response analysis, primarily assumes no association among predictors and among dependent variables. In addition, LASSO suffers from two other noteworthy limitations. First, when genes are not performing independently, and they are linked via a genomic pathway, LASSO tends to randomly select one gene from each correlated pathway. Second, when the number of genes is larger than the sample size, the

largest number of genes selected by LASSO does not exceed the sample size. Given these limitations, the false negatives may be overrepresented in the findings[159].

LCT is suggested by this thesis as a suitable approach for analysis of omics linkage because of the following reasons. LCT is a self-contained GSA method. As such, it considers the gene-gene correlations within a gene set. Taking advantage of shrinkage method, LCT deals with the high dimensionality of the data and maintains a reasonable computational efficiency. LCT handles a combination of positive and negative associations by assigning optimal gene-specific weights. LCT is unique in case of omics linkage analysis, when the numbers of measurements for both omics are large.

It is now recognized that genetic studies benefit from longitudinal designs. The observation of temporal associations of omics-omics or omics-phenotype enhances our understanding about underlying biological processes. Through examination of the temporal associations, the findings from genetic studies are expected to achieve more consistency. LLCT is proposed here as an efficient analytical tool to identify the genes whose differential expressions lead to different temporal pattern of phenotype variations. LLCT is able to adjust for environmental variables affecting the temporal variations of the phenotype, leading to more accurate interpretations. LLCT takes care of the longitudinal correlation structure of the data, as shown in the simulation study. Changing the main elements of the longitudinal design, such as number of repeated measurements, and within-subject correlations, did not affect the performance of LLCT. Larger sample size, larger gene set and greater within-gene-set correlation increased LLCT power.

LLCT is a two-step method, modelling within-subject variation at the first step, followed by testing the between-subject variation at the second step. On the grounds of the unique flexibility of LCT in dealing with different complexities within a dataset and failure of other GSA methods to satisfy the requirements of this data structure, we selected LCT for the second step of our proposed method. The main idea of LLCT is derived from mixed effect models, which can be regarded as a two-step regression modeling of within and between subject variations. Alternatives to LCT in the second step were extension of SAM-GS to continuous phenotype [29] and Global Test[12], the self-contained GSA methods able to work with continuous phenotypes. However, LCT is a more powerful method compared to its alternatives[160].

In our simulation study, we compared the performance of LCT with that of PAVR method, proposed for uncensored dependent variables[29]. To the best of our knowledge, PAVR is the only method in the literature that can be applied to longitudinal phenotype. PAVR performed poorly when dealing with small sample size large number of repeated measurements, and low within-gene-set correlation. LCT shows a reasonable performance in all these scenarios. Aside from the lower power of PAVR, a critical drawback of this method is that it assumes that the effects of gene expressions remain constant over time. In other words, it does not allow the gene and time effect interaction. However, LLCT can detect differential temporal patterns of phenotype in association with differentially expressed genes, which is aligned to the objectives of longitudinal designs.

Another feature improving consistency of findings among genetic studies, is measuring multiple members of a family. Lower heterogeneity of pedigree-based genetic data improves the analysis power. However, since the subjects related to a family are correlated, the analysis is challenging. Our proposed method is able to deal with this additional correlation structure

imposed by this study design. We used multiple measurements from the same family to find family-specific trend of phenotype, which obliviously provides more accurate estimation of the temporal patterns, by diminishing the between-subject variations.

Simultaneous variation of omics and phenotypes is the subject of investigation in many genetic studies and was recognized to provide more knowledge about the underlying temporal biological process. The measurements of thousands of genes repeatedly over time, although possible by the advances of technology, are expensive. Thus, the sample size in these studies is usually very small. There is an intensive literature developing methods for time-course microarray data analysis. However, as reviewed in introduction section, existing methods are mostly defined to analyze binary or multiple discrete conditions and, therefore, are unable to examine the phenotypes measured on a continuous scale. Dichotomizing a continuous variable is discouraged by statisticians because of loss of information. The extension of LLCT for analysis of time-course omics data may play an important role in identifying genes differentially expressed over time in association with a continuous phenotype, such as blood pressure or cholesterol level.

4.2 Strengths

There are several noteworthy strengths of LCT method as applied to omics linkage. The omics data are measured on a continuous scale. Dichotomization of continuous variables to satisfy the requirements of a specific statistical method may be misleading, no matter how reliable the classification method is. Therefore, investigators of omics linkage should take advantages of methods able to accommodate continuous variables. LCT is able to analyze

continuous omics data. LCT is a powerful and computationally efficient method. In a simulation study conducted by Dinu et al., LCT outperformed SAM-GS (extended for continuous variables) and Global Test, in terms of power. The omics data are correlated within biological pathways. Ignoring this correlation structure may inflate type I error. As such, the GSA methods taking into the account the correlation within a priori defined genes are advantageous. Competitive methods are sensitive to genes outside the set of interest. LCT works on the genes within the set of interest. LCT is able to detect cumulative weak to moderate effects of two or multiple omics, a scenario where competitive methods fail.

LLCT provides a unique tool in analysis of longitudinal phenotype of omics data. Phenotype dynamic may be influenced by environmental factors. The first step of LLCT allows adjusting for time-dependent covariates who may affect the variation of phenotype over time. LLCT is applicable in the presence of non-linear temporal patterns. If the temporal pattern of gene expressions or phenotypes is non-linear and a slope does not explain all the variations, a polynomial model in the first step should be used. In the case polynomial models fail to indicate the variations, spline models should be employed in the first step of LLCT[30]. However, these modifications require a larger sample size. LLCT is a GSA method and therefore evaluates the association between a given phenotype trajectory and the set of genes sharing a biological function. LLCT assumes that the genes within a gene set are correlated, a crucial assumption in analysis at gene set level. LLCT is a self-contained method testing the null hypothesis of "No gene within a given gene set is differentially expressed in association with different trajectories of a phenotype". Thus, the results of LLCT on a certain gene set are not influenced by alterations to the collections of gene sets in the study. LLCT is computationally very efficient compared to existing methods. In our simulation analysis, we observed that it takes 70 times

longer for PAVR to run a single test, compared to LCT. LLCT can be applied to a variety of phenotypes, given an appropriate model is used in the first step. A considerable proportion of longitudinal observations are usually missing. This calls for statistical methods that can deal with missing values. LLCT is applicable to both unbalanced and incomplete dataset. When an observation is missing, the other observations related to the corresponding subject still contribute to the analysis. A reasonable power of LLCT was observed in our simulation study. LLCT does not perform poorly when the sample size and gene set sizes are small. LLCT performance remains robust by changing the number of repeated measurements and within-subject correlation, indicating its ability to take care of the longitudinal structure of the data. LLCT is able to incorporate family-based structure of data into the analysis. Family-based data imposes a complicated correlation structure into the data, which is addressed in the first step of our proposed method.

4.3 Limitations

Despite all the strengths of our proposed methods, we identified two important limitations discussed below.

Firstly, not all the genes within a significant gene set may contribute to its significance. Identification of a subset of genes that actually drive the significant association with the phenotypes improves our understanding of the biological system. Vatanpour et al. [105]proposed Linear Combination Test for Gene Set Reduction (LCT-GSR) method to find a core gene set for a continuous phenotype. A significant gene set identified by LCT undergoes a secondary analysis. SAM is used to reduce the set to its core subset, by eliminating the

redundant genes. LLCT can take advantage of this method to find the core gene sets significantly associated with the different temporal patterns of phenotype.

The second limitation is that time-independent covariates such as gender may confound the association between genes and the phenotype behaviour over time. In other words, the between-subject variations may be partially explained by subject-variant covariates. LLCT cannot adjust for such effects. The same limitation applies when analyzing family-based data. The family-dependent covariates may alter the gene-phenotype associations.

4.4 Conclusions and Public Health Implications

Lack of replication among genetic studies is widely observed. Incorporation of complex designs in genetic studies should improve consistency of the findings, by lowering potential epidemiological errors and analytical biases. Lack of appropriate statistical methods discourages investigators to benefit from these designs[161]. Consequently, complex designs accompanied by reliable statistical methods help generating better results and contribute to changes of clinical practices and enhancement of screening, preventive and therapeutic clinical outcomes. The statistical approaches proposed in this thesis facilitate the analysis of family-based and longitudinal omics data.

Integration of multiple omics data allowing molecular profiling of subjects helps development of personalized medical care. When human genome is regulated at multiple levels, examination of single omics provides limited information regarding the etiology of diseases.

A genetic study may benefit from family-based design by gaining higher quality assurance of data collection[162], boosting the motivation of participants when their family members are enrolling[163] and minimizing the effects of unmeasured subject-related confounders[164].

Longitudinal design monitors the process of disease development allowing for measurements of time-dependent covariates and increasing quality assurance by repeated measurements of subjects. This design allows for an increased understanding of the biological systems underlying progression of a disease.

4.5 Future Directions

The needs for practicing complex study designs in order to enhance the precision of findings have been highlighted by many researchers. Biotatisticians and bioinformaticians have developed sound methods for advanced study designs in the past two decades. In spite of availability of some review papers classifying and explaining these methods for time-course microarray data analysis, microbiota data analysis and omics integrative analysis, there are very few papers comparing the methods via simulation analyses and providing the researchers with a guide to indicate the situations where each method best applies. Therefore, future high-throughput data investigations may benefit from simulation-based reviews of the methodology. Otherwise, the attempts toward generating the high-quality knowledge may be wasted.

4.6 Software Package

We used R software version 3.4.3 to execute LCT, LLCT and data simulation. Free R codes for performing LCT for continuous phenotype is available at https://sites.ualberta.ca/~yyasui/software.html.

References

1. Johnstone IM, Titterington DM. Statistical challenges of high-dimensional data. Philos Transact A Math Phys Eng Sci. 2009;367: 4237–4253. doi:10.1098/rsta.2009.0159

2. Bickel PJ, Brown JB, Huang H, Li Q. An overview of recent developments in genomics and associated statistical methods. Philos Transact A Math Phys Eng Sci. 2009;367: 4313–4337. doi:10.1098/rsta.2009.0164

3. Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc Natl Acad Sci. 2003;100: 3351–3356. doi:10.1073/pnas.0530258100

4. McLachlan GJ, Bean RW, Ng A. Clustering of Microarray Data via Mixture Models. Statistical Advances in the Biomedical Sciences. Wiley-Blackwell; 2007. pp. 365–383. doi:10.1002/9780470181218.ch21

5. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970;12: 55–67. doi:10.1080/00401706.1970.10488634

6. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58: 267–288.

7. Dezeure R, Bühlmann P, Meier L, Meinshausen N. High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi. Stat Sci. 2015;30: 533–558. doi:10.1214/15-STS527

8. Wasserman L, Roeder K. High-dimensional variable selection. Ann Stat. 2009;37: 2178–2201. doi:10.1214/08-AOS646

9. Bühlmann P. Statistical significance in high-dimensional linear models. Bernoulli. 2013;19: 1212–1242. doi:10.3150/12-BEJSP11

10. Zhang Cun-Hui, Zhang Stephanie S. Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat Soc Ser B Stat Methodol. 2013;76: 217–242. doi:10.1111/rssb.12026

11. Fan Jianqing, Lv Jinchi. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol. 2008;70: 849–911. doi:10.1111/j.1467-9868.2008.00674.x

12. Goeman JJ, Geer SA van de, Kort F de, Houwelingen HC van. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. 2004;20: 93–99. doi:10.1093/bioinformatics/btg382

13. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. Stat Sci. 2003;18: 104–117. doi:10.1214/ss/1056397488

14. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98: 5116–5121. doi:10.1073/pnas.091062498

15. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34: 267–273. doi:10.1038/ng1180

16. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102: 15545–15550. doi:10.1073/pnas.0506580102

17. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol. 2015;19: A68–A77. doi:10.5114/wo.2014.47136

18. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30: 207–210.

19. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28: 27–30.

20. Nishimura D. BioCarta. Biotech Softw Internet Rep. 2001;2: 117–120. doi:10.1089/152791601750294344

21. Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat. 2007;1: 107–129. doi:10.1214/07-AOAS101

22. Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics. 2005;21: 1943–1949. doi:10.1093/bioinformatics/bti260

23. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci U S A. 2005;102: 13544–13549. doi:10.1073/pnas.0506577102

24. Delongchamp R, Lee T, Velasco C. A method for computing the overall statistical significance of a treatment effect among a group of genes. BMC Bioinformatics. 2006;7: S11. doi:10.1186/1471-2105-7-S2-S11

25. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, et al. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics. 2007;8: 242. doi:10.1186/1471-2105-8-242

26. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci. 2003;100: 9440–9445. doi:10.1073/pnas.1530509100

27. Bolli P, Hemmelgarn B, Myers MG, McKay D, Tremblay G, Tobe SW. High normal blood pressure and prehypertension: The debate continues. Can J Cardiol. 2007;23: 581–583.

28. Kitano H. Systems Biology: A Brief Overview. Science. 2002;295: 1662–1664.

29. Adewale A j., Dinu I, Potter J d., Liu Q, Yasui Y. Pathway Analysis of Microarray Data via Regression. J Comput Biol. 2008;15: 269–277. doi:10.1089/cmb.2008.0002

30. Gauderman WJ, Macgregor S, Briollais L, Scurrah K, Tobin M, Park T, et al. Longitudinal data analysis in pedigree studies. Genet Epidemiol. 25: S18–S28. doi:10.1002/gepi.10280

31. Kerner B, North KE, Fallin MD. Use of Longitudinal Data in Genetic Studies in the Genome-wide Association Studies Era: Summary of Group 14. Genet Epidemiol. 2009;33: S93–S98. doi:10.1002/gepi.20479

32. Beyene J, Hamid JS. Longitudinal Data Analysis in Genome-Wide Association Studies. Genet Epidemiol. 2014;38: S68–S73. doi:10.1002/gepi.21828

33. Zhou H, Zhou J, Sobel EM, Lange K. Fast genome-wide pedigree quantitative trait loci analysis using MENDEL. BMC Proc. 2014;8: S93. doi:10.1186/1753-6561-8-S1-S93

34. Wu Z, Hu Y, Melton PE. Longitudinal Data Analysis for Genetic Studies in the Whole-Genome Sequencing Era. Genet Epidemiol. 2014;38: S74–S80. doi:10.1002/gepi.21829

35. Chiu Y-F, Justice AE, Melton PE. Longitudinal analytical approaches to genetic data. BMC Genet. 2016;17 Suppl 2: 4. doi:10.1186/s12863-015-0312-y

36. Bar-Joseph Z. Analyzing time series gene expression data. Bioinformatics. 2004;20: 2493–2503. doi:10.1093/bioinformatics/bth283

37. Ruan L, Yuan M. Statistical Analysis of Time Course Microarray Data. Handbook of Statistical Bioinformatics. Springer, Berlin, Heidelberg; 2011. pp. 299–313. doi:10.1007/978-3-642-16345-6_14

38. Tai YC, Speed TP. Statistical Analysis of Microarray Time Course Data. Data Anal. : 28.

39. Ernst J, Nau GJ, Bar-Joseph Z. Clustering short time series gene expression data. Bioinformatics. 2005;21: i159–i168. doi:10.1093/bioinformatics/bti1022

40. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999;22: 281–285. doi:10.1038/10343

41. Schliep A, Steinhoff C, Schönhuth A. Robust inference of groups in gene expression time-courses using mixtures of HMMs. Bioinformatics. 2004;20: i283–i289. doi:10.1093/bioinformatics/bth937

42. Park T, Yi S-G, Lee S, Lee SY, Yoo D-H, Ahn J-I, et al. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. Bioinformatics. 2003;19: 694–703. doi:10.1093/bioinformatics/btg068

43. Hong F, Li H. Functional Hierarchical Models for Identifying Genes with Different Time-Course Expression Profiles. Biometrics. 62: 534–544. doi:10.1111/j.1541-0420.2005.00505.x

44. Bar-Joseph Z, Gerber G, Simon I, Gifford DK, Jaakkola TS. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. Proc Natl Acad Sci. 2003;100: 10146–10151. doi:10.1073/pnas.1732547100

45. Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. Proc Natl Acad Sci U S A. 2005;102: 12837–12842. doi:10.1073/pnas.0504609102

46. Ramoni MF, Sebastiani P, Kohane IS. Cluster analysis of gene expression dynamics. Proc Natl Acad Sci U S A. 2002;99: 9121–9126. doi:10.1073/pnas.132656399

47. Conesa A, Nueda MJ, Ferrer A, Talón M. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. Bioinformatics. 2006;22: 1096–1102. doi:10.1093/bioinformatics/btl056

48. Hejblum BP, Skinner J, Thiébaut R. Time-Course Gene Set Analysis for Longitudinal Gene Expression Data. PLOS Comput Biol. 2015;11: e1004310. doi:10.1371/journal.pcbi.1004310

49. Jonnalagadda S, Srinivasan R. Principal components analysis based methodology to identify differentially expressed genes in time-course microarray data. BMC Bioinformatics. 2008;9: 267. doi:10.1186/1471-2105-9-267

50. Nueda MJ, Conesa A, Westerhuis JA, Hoefsloot HCJ, Smilde AK, Talón M, et al. Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA. Bioinforma Oxf Engl. 2007;23: 1792–1800. doi:10.1093/bioinformatics/btm251

51. Hidden Markov Models for Microarray Time Course Data in Multiple Biological Conditions: Journal of the American Statistical Association: Vol 101, No 476 [Internet]. [cited 31 May 2018]. Available: https://amstat.tandfonline.com/doi/abs/10.1198/016214505000000394#.WxBBZkgvzIU

52. Haas GP, Delongchamps N, Brawley OW, Wang CY, Roza G de la. The Worldwide Epidemiology of Prostate Cancer: Perspectives from Autopsy Studies. Can J Urol. 2008;15: 3866–3871.

53. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136: 359. doi:10.1002/ijc.29210

54. Bray F, Ren J-S, Masuyer E, Ferlay J. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. Int J Cancer. 2013;132: 1133–1145. doi:10.1002/ijc.27711

55. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin. 2015;65: 87–108. doi:10.3322/caac.21262

56. Lima AR, Bastos M de L, Carvalho M, Guedes de Pinho P. Biomarker Discovery in Human Prostate Cancer: an Update in Metabolomics Studies. Transl Oncol. 2016;9: 357–370. doi:10.1016/j.tranon.2016.05.004

57. DeNicola GM, Karreth FA, Humpton TJ, Gopinathan A, Wei C, Frese K, et al. Oncogene-induced Nrf2 transcription promotes ROS detoxification and tumorigenesis. Nature. 2011;475: 106–109. doi:10.1038/nature10189

58. Ru P, Steele R, Nerurkar PV, Phillips N, Ray RB. Bitter melon extract impairs prostate cancer cell-cycle progression and delays prostatic intraepithelial neoplasia in TRAMP model. Cancer Prev Res Phila Pa. 2011;4: 2122–2130. doi:10.1158/1940-6207.CAPR-11-0376

59. Carlsson SV, Kattan MW. Prostate cancer: Personalized risk - stratified screening or abandoning it altogether? Nat Rev Clin Oncol. 2016;13: 140–142. doi:10.1038/nrclinonc.2016.11

60. Abrate A, Lughezzani G, Gadda GM, Lista G, Kinzikeeva E, Fossati N, et al. Clinical Use of [-2]proPSA (p2PSA) and Its Derivatives (%p2PSA and Prostate Health Index) for the Detection of Prostate Cancer: A Review of the Literature. Korean J Urol. 2014;55: 436–445. doi:10.4111/kju.2014.55.7.436

61. Dasgupta S, Srinidhi S, Vishwanatha JK. Oncogenic activation in prostate cancer progression and metastasis: Molecular insights and future challenges. J Carcinog. 2011;11: 4. doi:10.4103/1477-3163.93001

62. Zhang A, Yan G, Han Y, Wang X. Metabolomics Approaches and Applications in Prostate Cancer Research. Appl Biochem Biotechnol. 2014;174: 6–12. doi:10.1007/s12010-014-0955-6

63. Dowd TL, Kaplan BA, Gupta RK, Aisen P. Detection of malignant tumors: water-suppressed proton nuclear magnetic resonance spectroscopy of plasma. Magn Reson Med. 1987;5: 395–397.

64. Liesenfeld DB, Habermann N, Owen RW, Scalbert A, Ulrich CM. Review of mass spectrometry-based metabolomics in cancer research. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2013;22: 2182–2201. doi:10.1158/1055-9965.EPI-13-0584

65. Pavlova NN, Thompson CB. THE EMERGING HALLMARKS OF CANCER METABOLISM. Cell Metab. 2016;23: 27–47. doi:10.1016/j.cmet.2015.12.006

66. Fenner A. Prostate cancer: Sarcosine: the saga continues... Nat Rev Urol. 2011;8: 175–175. doi:10.1038/nrurol.2011.33

67. Spratlin JL, Serkova NJ, Eckhardt SG. Clinical applications of metabolomics in oncology: a review. Clin Cancer Res Off J Am Assoc Cancer Res. 2009;15: 431–440. doi:10.1158/1078-0432.CCR-08-1059

68. Vermeersch KA, Styczynski MP. Applications of metabolomics in cancer research. J Carcinog. 2013;12: 9. doi:10.4103/1477-3163.113622

69. Giskeødegård GF, Hansen AF, Bertilsson H, Gonzalez SV, Kristiansen KA, Bruheim P, et al. Metabolic markers in blood can separate prostate cancer from benign prostatic hyperplasia. Br J Cancer. 2015;113: 1712–1719. doi:10.1038/bjc.2015.411

70. Costello LC, Franklin RB. The clinical relevance of the metabolism of prostate cancer; zinc and tumor suppression: connecting the dots. Mol Cancer. 2006;5: 17. doi:1476-4598-5-17 [pii]

71. Roberts MJ, Schirra HJ, Lavin MF, Gardiner RA. Metabolomics: a novel approach to early and noninvasive prostate cancer detection. Korean J Urol. 2011;52: 79–89. doi:10.4111/kju.2011.52.2.79 [doi]

72. DeBerardinis RJ, Chandel NS. Fundamentals of cancer metabolism. Sci Adv. 2016;2. doi:10.1126/sciadv.1600200

73. Heiden MGV, DeBerardinis RJ. Understanding the intersections between metabolism and cancer biology. Cell. 2017;168: 657–669. doi:10.1016/j.cell.2016.12.039

74. Phan LM, Yeung S-CJ, Lee M-H. Cancer metabolic reprogramming: importance, main features, and potentials for precise targeted anti-cancer therapies. Cancer Biol Med. 2014;11: 1–19. doi:10.7497/j.issn.2095-3941.2014.01.001

75. Altman BJ, Stine ZE, Dang CV. From Krebs to clinic: glutamine metabolism to cancer therapy. Nat Rev Cancer. 2016;16: 619–634. doi:10.1038/nrc.2016.71

76. Rebello RJ, Pearson RB, Hannan RD, Furic L. Therapeutic Approaches Targeting MYC-Driven Prostate Cancer. Genes. 2017;8. doi:10.3390/genes8020071

77. Griffin JL, Shockcor JP. Metabolic profiles of cancer cells. Nat Rev Cancer. 2004;4: 551–561. doi:10.1038/nrc1390

78. Kaushik AK, Vareed SK, Basu S, Putluri V, Putluri N, Panzitt K, et al. Metabolomic profiling identifies biochemical pathways associated with castration-resistant prostate cancer. J Proteome Res. 2014;13: 1088–1100. doi:10.1021/pr401106h

79. Ciccarese C, Santoni M, Massari F, Modena A, Piva F, Conti A, et al. Metabolic Alterations in Renal and Prostate Cancer. Curr Drug Metab. 2016;17: 150–155.

80. Eidelman E, Twum-Ampofo J, Ansari J, Siddiqui MM. The Metabolic Phenotype of Prostate Cancer. Front Oncol. 2017;7. doi:10.3389/fonc.2017.00131

81. Priolo C, Pyne S, Rose J, Regan ER, Zadra G, Photopoulos C, et al. AKT1 and MYC Induce Distinctive Metabolic Fingerprints in Human Prostate Cancer. Cancer Res. 2014;74: 7198–7204. doi:10.1158/0008-5472.CAN-14-1490

82. Holmes E, Wilson ID, Nicholson JK. Metabolic phenotyping in health and disease. Cell. 2008;134: 714–717. doi:10.1016/j.cell.2008.08.026

83. Koh CM, Bieberich CJ, Dang CV, Nelson WG, Yegnasubramanian S, Marzo AMD. MYC and Prostate Cancer. Genes Cancer. 2010;1: 617–628. doi:10.1177/1947601910379132

84. Clegg NJ, Couto SS, Wongvipat J, Hieronymus H, Carver BS, Taylor BS, et al. MYC Cooperates with AKT in Prostate Tumorigenesis and Alters Sensitivity to mTOR Inhibitors. PLoS ONE. 2011;6. doi:10.1371/journal.pone.0017449

85. Liu Q, Dinu I, Adewale AJ, Potter JD, Yasui Y. Comparative evaluation of gene-set analysis methods. BMC Bioinformatics. 2007;8: 431. doi:10.1186/1471-2105-8-431

86. Wang X, Pyne S, Dinu I. Gene set enrichment analysis for multiple continuous phenotypes. BMC Bioinformatics. 2014;15: 260. doi:10.1186/1471-2105-15-260

87. Damian D, Gorfine M. Statistical concerns about the GSEA procedure. Nat Genet. 2004;36: 663–663. doi:10.1038/ng0704-663a

88. Wise DR, Thompson CB. Glutamine Addiction: A New Therapeutic Target in Cancer. Trends Biochem Sci. 2010;35: 427–433. doi:10.1016/j.tibs.2010.05.003

89. Dang CV. Rethinking the Warburg effect with Myc micromanaging glutamine metabolism. Cancer Res. 2010;70: 859–862. doi:10.1158/0008-5472.CAN-09-3556

90. Mannava S, Grachtchouk V, Wheeler LJ, Im M, Zhuang D, Slavina EG, et al. Direct role of nucleotide metabolism in C-MYC-dependent proliferation of melanoma cells. Cell Cycle Georget Tex. 2008;7: 2392–2400.

91. Hsieh AL, Walton ZE, Altman BJ, Stine ZE, Dang CV. MYC and metabolism on the path to cancer. Semin Cell Dev Biol. 2015;43: 11–21. doi:10.1016/j.semcdb.2015.08.003

92. Edmunds LR, Sharma L, Kang A, Lu J, Vockley J, Basu S, et al. c-Myc Programs Fatty Acid Metabolism and Dictates Acetyl-CoA Abundance and Fate. J Biol Chem. 2014;289: 25382–25392. doi:10.1074/jbc.M114.580662

93. Research AA for C. MYC-Overexpressing TNBCs Depend on Fatty Acid Oxidation. Cancer Discov. 2016; doi:10.1158/2159-8290.CD-RW2016-049

94. Menendez JA, Lupu R. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. Nat Rev Cancer. 2007;7: 763–777. doi:10.1038/nrc2222

95. Wise DR, DeBerardinis RJ, Mancuso A, Sayed N, Zhang X-Y, Pfeiffer HK, et al. Myc regulates a transcriptional program that stimulates mitochondrial glutaminolysis and leads to glutamine addiction. Proc Natl Acad Sci U S A. 2008;105: 18782–18787. doi:10.1073/pnas.0810199105

96. Fan Y, Dickman KG, Zong W-X. Akt and c-Myc Differentially Activate Cellular Metabolic Programs and Prime Cells to Bioenergetic Inhibition. J Biol Chem. 2010;285: 7324–7333. doi:10.1074/jbc.M109.035584

97. Balakumaran BS, Porrello A, Hsu DS, Glover W, Foye A, Leung JY, et al. MYC activity mitigates response to rapamycin in prostate cancer through eukaryotic initiation factor 4E-binding protein 1-mediated inhibition of autophagy. Cancer Res. 2009;69: 7803–7810. doi:10.1158/0008-5472.CAN-09-0910

98. Amente S, Zhang J, Lavadera ML, Lania L, Avvedimento EV, Majello B. Myc and PI3K/AKT signaling cooperatively repress FOXO3a-dependent PUMA and GADD45a gene expression. Nucleic Acids Res. 2011; gkr638. doi:10.1093/nar/gkr638

99. Fu Z, Tindall DJ. FOXOs, cancer and regulation of apoptosis. Oncogene. 2008;27: 2312–2319. doi:10.1038/onc.2008.24

100. Bouchard C, Marquardt J, Brás A, Medema RH, Eilers M. Myc-induced proliferation and transformation require Akt-mediated phosphorylation of FoxO proteins. EMBO J. 2004;23: 2830–2840. doi:10.1038/sj.emboj.7600279

101. Peck B, Ferber EC, Schulze A. Antagonism between FOXO and MYC Regulates Cellular Powerhouse. Front Oncol. 2013;3: 96. doi:10.3389/fonc.2013.00096

102. Yeh ES, Belka GK, Vernon AE, Chen C-C, Jung JJ, Chodosh LA. Hunk negatively regulates c-myc to promote Akt-mediated cell survival and mammary tumorigenesis induced by loss of Pten. Proc Natl Acad Sci. 2013;110: 6103–6108. doi:10.1073/pnas.1217415110

103. Gill RM, Gabor TV, Couzens AL, Scheid MP. The MYC-Associated Protein CDCA7 Is Phosphorylated by AKT To Regulate MYC-Dependent Apoptosis and Transformation. Mol Cell Biol. 2013;33: 498. doi:10.1128/MCB.00276-12

104. Zou Z, Chen J, Liu A, Zhou X, Song Q, Jia C, et al. mTORC2 promotes cell survival through c-Myc–dependent up-regulation of E2F1. J Cell Biol. 2015;211: 105–122. doi:10.1083/jcb.201411128

105. Vatanpour S. Public health risk assessment : validation of risk assessment matrix limitations and an analytical approach to gene set reduction for continuous phenotype in microarray studies. [2016]; 2016.

106. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98: 5116–5121. doi:10.1073/pnas.091062498

107. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102: 15545–15550. doi:10.1073/pnas.0506580102

108. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34: 267–273. doi:10.1038/ng1180

109. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinforma Oxf Engl. 2007;23: 980–987. doi:10.1093/bioinformatics/btm051

110. Nam D, Kim S-Y. Gene-set approach for expression pattern analysis. Brief Bioinform. 2008;9: 189–197. doi:10.1093/bib/bbn001

111. Fan R, Albert PS, Schisterman EF. A Discussion of Gene-Gene and Gene-Environment Interactions and Longitudinal Genetic Analysis of Complex Traits. Stat Med. 2012;31: 2565–2568. doi:10.1002/sim.5495

112. Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. Genet Res. 2001;77: 123–128.

113. Park T, Yi S-G, Lee S, Lee SY, Yoo D-H, Ahn J-I, et al. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. Bioinforma Oxf Engl. 2003;19: 694–703.

114. Turner JA, Bolen CR, Blankenship DM. Quantitative gene set analysis generalized for repeated measures, confounder adjustment, and continuous covariates. BMC Bioinformatics. 2015;16: 272. doi:10.1186/s12859-015-0707-9

115. Zhang K, Wang H, Bathke AC, Harrar SW, Piepho H-P, Deng Y. Gene set analysis for longitudinal gene expression data. BMC Bioinformatics. 2011;12: 273. doi:10.1186/1471-2105-12-273

116. Wijsman EM. Family-based approaches: design, imputation, analysis, and beyond. BMC Genet. 2016;17 Suppl 2: 9. doi:10.1186/s12863-015-0318-5

117. Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). Hum Mutat. 2011;32: 564–567. doi:10.1002/humu.21466

118. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol. 2012;8: e1002375. doi:10.1371/journal.pcbi.1002375

119. Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. John Wiley & Sons; 2012.

120. Dinu I, Wang X, Kelemen LE, Vatanpour S, Pyne S. Linear combination test for gene set analysis of a continuous phenotype. BMC Bioinformatics. 2013;14: 212. doi:10.1186/1471-2105-14-212

121. Schäfer J, Strimmer K. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. Stat Appl Genet Mol Biol. 2005;4. doi:10.2202/1544-6115.1175

122. Kearney PM, Whelton M, Reynolds K, Muntner P, Whelton PK, He J. Global burden of hypertension: analysis of worldwide data. Lancet Lond Engl. 2005;365: 217–223. doi:10.1016/S0140-6736(05)17741-1

123. Mancia G, Fagard R, Narkiewicz K, Redon J, Zanchetti A, Böhm M, et al. 2013 ESH/ESC Guidelines for the management of arterial hypertensionThe Task Force for the management of arterial hypertension of the European Society of Hypertension (ESH) and of the European Society of Cardiology (ESC). Eur Heart J. 2013;34: 2159–2219. doi:10.1093/eurheartj/eht151

124. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL, et al. Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Hypertens Dallas Tex 1979. 2003;42: 1206–1252. doi:10.1161/01.HYP.0000107251.49515.c2

125. Johnson T, Gaunt TR, Newhouse SJ, Padmanabhan S, Tomaszewski M, Kumari M, et al. Blood pressure loci identified with a gene-centric array. Am J Hum Genet. 2011;89: 688–700. doi:10.1016/j.ajhg.2011.10.013

126. Johnson JA. Ethnic Differences in Cardiovascular Drug Response: Potential Contribution of Pharmacogenetics. Circulation. 2008;118: 1383–1393. doi:10.1161/CIRCULATIONAHA.107.704023

127. Almasy L, Amos CI, Bailey-Wilson JE, Cantor RM, Jaquish CE, Martinez M, et al. Genetic Analysis Workshop 13: Analysis of Longitudinal Family Data for Complex Diseases and Related Risk Factors. BMC Genet. 2003;4: S1. doi:10.1186/1471-2156-4-S1-S1

128. Cupples LA, Beyene J, Bickeböller H, Daw EW, Fallin MD, Gauderman WJ, et al. Genetic Analysis Workshop 16: Strategies for genome-wide association study analyses. BMC Proc. 2009;3: S1. doi:10.1186/1753-6561-3-S7-S1

129. Bickeböller H, Bailey JN, Beyene J, Cantor RM, Cordell HJ, Culverhouse RC, et al. Genetic Analysis Workshop 18: Methods and strategies for analyzing human sequence and phenotype data in members of extended pedigrees. BMC Proc. 2014;8: S1. doi:10.1186/1753-6561-8-S1-S1

130. Engelman CD, Greenwood CMT, Bailey JN, Cantor RM, Kent JW, König IR, et al. Genetic Analysis Workshop 19: methods and strategies for analyzing human sequence and gene expression data in extended families and unrelated individuals. BMC Proc. 2016;10: 67–70. doi:10.1186/s12919-016-0007-z

131. Ziki MDA, Mani A. Atherosclerosis [Internet]. 2017. Available: http://www.atherosclerosis-journal.com/article/S0021-9150(17)30190-9/fulltext

132. Heusch G, Libby P, Gersh B, Yellon D, Böhm M, Lopaschuk G, et al. Cardiovascular remodelling in coronary artery disease and heart failure. Lancet Lond Engl. 2014;383: 1933–1943. doi:10.1016/S0140-6736(14)60107-0

133. Brown DI, Griendling KK. Regulation of signal transduction by reactive oxygen species in the cardiovascular system. Circ Res. 2015;116: 531–549. doi:10.1161/CIRCRESAHA.116.303584

134. Tabas I, García-Cardeña G, Owens GK. Recent insights into the cellular biology of atherosclerosis. J Cell Biol. 2015;209: 13–22. doi:10.1083/jcb.201412052

135. Hedman AC, Smith JM, Sacks DB. The biology of IQGAP proteins: beyond the cytoskeleton. EMBO Rep. 2015;16: 427–446. doi:10.15252/embr.201439834

136. Michael SK, Surks HK, Wang Y, Zhu Y, Blanton R, Jamnongjit M, et al. High Blood Pressure Arising from a Defect in Vascular Function. Proc Natl Acad Sci U S A. 2008;105: 6702–6707. doi:10.1073/pnas.0802128105

137. Brozovich FV, Nicholson CJ, Degen CV, Gao YZ, Aggarwal M, Morgan KG. Mechanisms of Vascular Smooth Muscle Contraction and the Basis for Pharmacologic Treatment of Smooth Muscle Disorders. Pharmacol Rev. 2016;68: 476–532. doi:10.1124/pr.115.010652

138. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003;33 Suppl: 245–254. doi:10.1038/ng1089

139. Kouzarides T. Chromatin modifications and their function. Cell. 2007;128: 693–705. doi:10.1016/j.cell.2007.02.005

140. Wain LV, Verwoert GC, O'Reilly PF, Shi G, Johnson T, Johnson AD, et al. Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. Nat Genet. 2011;43: 1005. doi:10.1038/ng.922

141. Adeyemo A, Gerry N, Chen G, Herbert A, Doumatey A, Huang H, et al. A Genome-Wide Association Study of Hypertension and Blood Pressure in African Americans. PLOS Genet. 2009;5: e1000564. doi:10.1371/journal.pgen.1000564

142. Iguchi H, Urashima Y, Inagaki Y, Ikeda Y, Okamura M, Tanaka T, et al. SOX6 Suppresses Cyclin D1 Promoter Activity by Interacting with β-Catenin and Histone Deacetylase 1, and Its Down-regulation Induces Pancreatic β-Cell Proliferation. J Biol Chem. 2007;282: 19052–19061. doi:10.1074/jbc.M700460200

143. Zorn AM, Barish GD, Williams BO, Lavender P, Klymkowsky MW, Varmus HE. Regulation of Wnt Signaling by Sox Proteins: XSox17α/β and XSox3 Physically Interact with β-catenin. Mol Cell. 1999;4: 487–498. doi:10.1016/S1097-2765(00)80200-2

144. Cheng P-W, Chen Y-Y, Cheng W-H, Lu P-J, Chen H-H, Chen B-R, et al. Wnt Signaling Regulates Blood Pressure by Downregulating a GSK-3β–Mediated Pathway to Enhance Insulin Signaling in the Central Nervous System. Diabetes. 2015;64: 3413–3424. doi:10.2337/db14-1439

145. Durik M, Kavousi M, van der Pluijm I, Isaacs A, Cheng C, Verdonk K, et al. Nucleotide excision DNA repair is associated with age-related vascular dysfunction. Circulation. 2012;126: 468–478. doi:10.1161/CIRCULATIONAHA.112.104380

146. Xu S, Touyz RM. Reactive oxygen species and vascular remodelling in hypertension: Still alive. Can J Cardiol. 2006;22: 947–951.

147. Lassègue B, Griendling KK. Reactive oxygen species in hypertension; An update. Am J Hypertens. 2004;17: 852–860. doi:10.1016/j.amjhyper.2004.02.004

148. Henstell H. The Pituitary Gland and the Maintenance of Blood Pressure. Yale J Biol Med. 1933;5: 531–544.

149. Das CJ, Baruah MP, Baruah UM. Radiological imaging in endocrine hypertension. Indian J Endocrinol Metab. 2011;15: S383–S388. doi:10.4103/2230-8210.86984

150. Hunter J, Haist RE. Hormonal Hypertension Resulting from Pituitary Imbalance. Can J Physiol Pharmacol. 1965;43: 269–278. doi:10.1139/y65-026

151. Mancia G, Grassi G. The Autonomic Nervous System and Hypertension. Circ Res. 2014;114: 1804–1814. doi:10.1161/CIRCRESAHA.114.302524

152. Mayet J, Hughes A. Cardiac and vascular pathophysiology in hypertension. Heart. 2003;89: 1104–1109.

153. Dirkx E, da Costa Martins PA, De Windt LJ. Regulation of fetal gene expression in heart failure. Biochim Biophys Acta BBA - Mol Basis Dis. 2013;1832: 2414–2424. doi:10.1016/j.bbadis.2013.07.023

154. Bress AP, Irvin MR, Muntner P. Genetics of Blood Pressure: New Insights Into a Complex Trait. Am J Kidney Dis Off J Natl Kidney Found. 2017;69: 723–725. doi:10.1053/j.ajkd.2017.02.365

155. Chiu Y-F, Justice AE, Melton PE. Longitudinal analytical approaches to genetic data. BMC Genet. 2016;17: S4. doi:10.1186/s12863-015-0312-y

156. Kerner B, North KE, Fallin MD. Use of longitudinal data in genetic studies in the genome-wide association studies era: summary of Group 14. Genet Epidemiol. 2009;33 Suppl 1: S93-98. doi:10.1002/gepi.20479

157. Kent JW. Pathway-based analyses. BMC Genet. 2016;17. doi:10.1186/s12863-015-0314-9

158. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet. 2017;8. doi:10.3389/fgene.2017.00084

159. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67: 301–320. doi:10.1111/j.1467-9868.2005.00503.x

160. Dinu I, Wang X, Kelemen LE, Vatanpour S, Pyne S. Linear combination test for gene set analysis of a continuous phenotype. BMC Bioinformatics. 2013;14: 212. doi:10.1186/1471-2105-14-212

161. Freely associating. Nat Genet. 1999;22: 1–2. doi:10.1038/8702

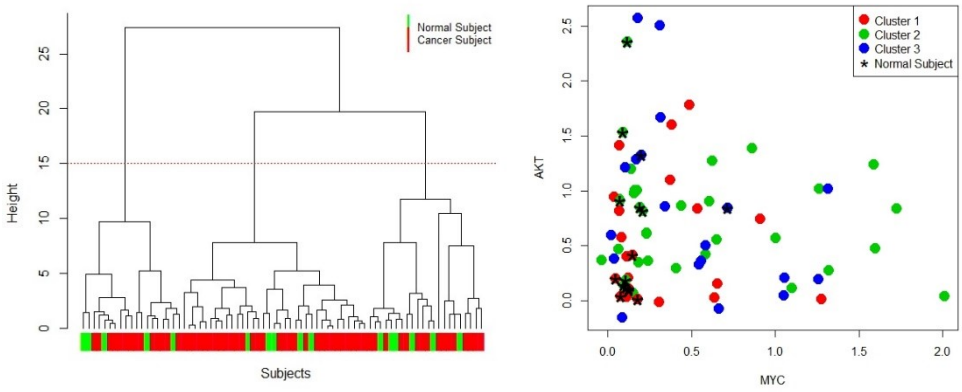
162. Molecular epidemiology [electronic resource] : applications in cancer and other human diseases / edited by Timothy R. Rebbeck, Christine B. Ambrosone, Peter G. Shields. [Internet]. New York: Informa Healthcare; 2008. Available: http://www.taylorfrancis.com/books/9781420052923

163. Witte JS, Gauderman WJ, Thomas DC. ORIGINAL CONTRIBUTIONS Asymptotic Bias and Efficiency in Case-Control Studies of Candidate Genes and Gene-Environment Interactions: Basic Family Designs.

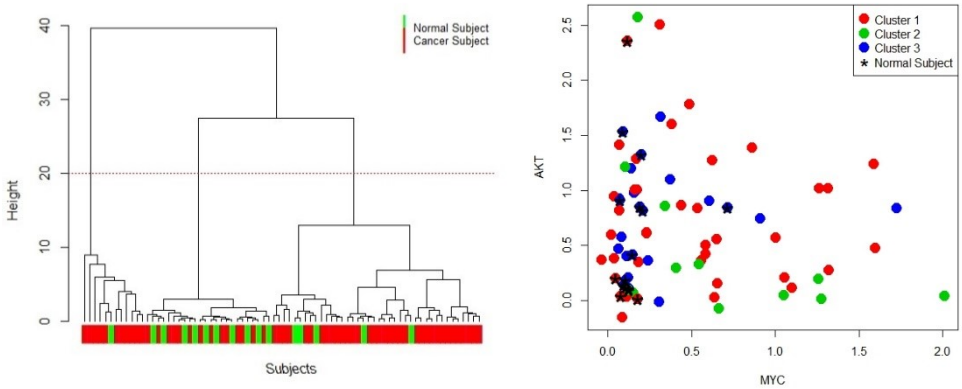
164. Thomas DC, Witte JS. Point: Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations? Cancer Epidemiol Prev Biomark. 2002;11: 505–512.

Appendices

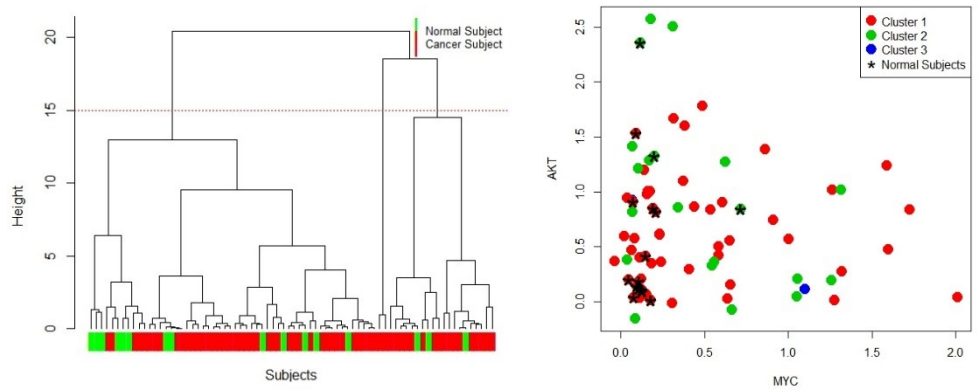
A. Nitrogen Metabolism



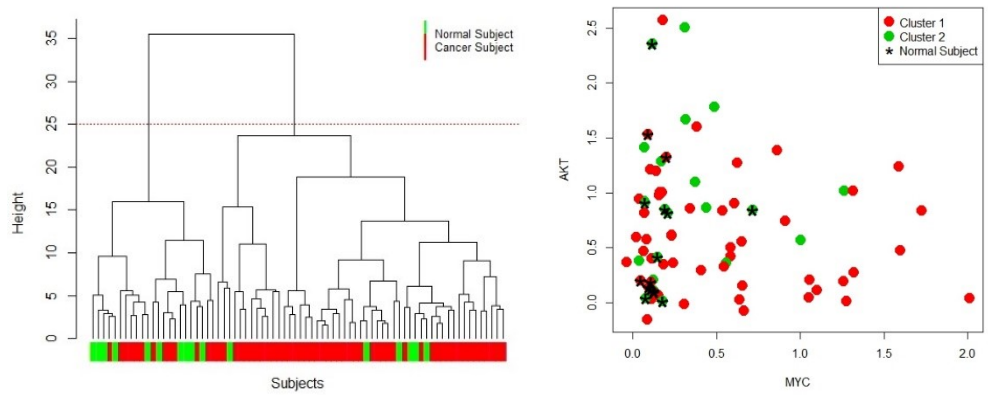
B. Fatty acid biosynthesis



C. D-Glutamine and D-glutamate metabolism



D. Purine Metabolism



Supplementary Figure 1. Cluster analysis of subjects based on the significant metabolite sets signature: Graphs on the left: dendrogram of subjects clustered based on their metabolite set signature, and scatterplots of AKT1 vs. MYC with cluster-specified observations.

Supplementary Table 1. P-values and q-values calculated by LCT for association between bivariate oncogenes (MYC, AKT1) or univariate oncogenes (MYC only or AKT1 only) and different metabolite sets in tumor samples

Metabolite Set	Size of Metabolite Set	p-value for (MYC, AKT1)	q-value for (MYC, AKT1)	P-value for (MYC, AKT1) Interaction	q-value for (MYC, AKT1) Interaction	p-value for MYC only	q-value for MYC only	p-value for AKT1 only	q-value for AKT1 only	Estimated linear combination of oncogenes
Alanine, aspartate and glutamate metabolism	11	0.363	0.829	0.221	0.686	0.405	0.799	0.755	0.887	0.64 MYC -1.88 AKT
Amino sugar and nucleotide sugar metabolism	7	0.190	0.716	0.248	0.686	0.240	0.799	0.626	0.887	-1.24 MYC -1.37 AKT
Arginine and proline metabolism	21	0.168	0.716	0.236	0.686	0.495	0.799	0.920	0.945	1.07 MYC -1.57 AKT
Ascorbate and aldarate metabolism	4	0.314	0.829	0.299	0.686	0.337	0.799	0.952	0.952	1.12 MYC -1.52 AKT
Benzoate degradation	2	0.747	0.855	0.806	0.831	0.514	0.799	0.504	0.887	-1.37 MYC -1.18 AKT
beta-Alanine metabolism	12	0.528	0.829	0.706	0.787	0.842	0.885	0.504	0.887	0.68 MYC -1.86 AKT
Biosynthesis of unsaturated fatty acids	13	0.148	0.716	0.069	0.500	0.389	0.799	0.209	0.887	0.4 MYC -1.97 AKT
Butanoate metabolism	9	0.876	0.898	0.680	0.787	0.595	0.799	0.747	0.887	-1.54 MYC -0.82 AKT
C5-Branched dibasic acid metabolism	5	0.634	0.840	0.477	0.686	0.317	0.799	0.517	0.887	-1.64 MYC +0.43 AKT
Chlorocyclohexane and chlorobenzene degradation	2	0.850	0.896	0.884	0.884	0.839	0.885	0.142	0.887	0.83 MYC -1.77 AKT
Citrate cycle (TCA cycle)	8	0.883	0.898	0.817	0.831	0.606	0.799	0.516	0.887	-1.64 MYC -0.44 AKT
Cyanoamino acid metabolism	5	0.554	0.829	0.692	0.787	0.565	0.799	0.326	0.887	0.11 MYC -2.03 AKT
Cysteine and methionine metabolism	11	0.802	0.878	0.611	0.770	0.514	0.799	0.457	0.887	1.32 MYC -1.26 AKT
D-Alanine metabolism	2	0.750	0.855	0.803	0.831	0.542	0.799	0.180	0.887	-1.48 MYC -0.97 AKT
D-Glutamine and D-glutamate metabolism	3	0.035*	0.472	0.01*	0.309**	0.070	0.799	0.619	0.887	1.17 MYC -1.46 AKT
Dioxin degradation	2	0.747	0.855	0.806	0.831	0.514	0.799	0.504	0.887	-1.37 MYC -1.18 AKT
Fatty acid biosynthesis	5	0.043*	0.472	0.014*	0.309**	0.062	0.799	0.462	0.887	1.06 MYC -1.58 AKT
Fatty acid metabolism	3	0.395	0.829	0.448	0.686	0.167	0.799	0.616	0.887	-1.64 MYC +0.45 AKT
Fructose and mannose metabolism	6	0.018*	0.348**	0.022*	0.319**	0.125	0.799	0.404	0.887	-0.92 MYC -1.7 AKT
Galactose metabolism	6	0.130	0.716	0.166	0.686	0.160	0.799	0.544	0.887	-0.07 MYC -2.03 AKT
Glutathione metabolism	12	0.398	0.829	0.403	0.686	0.169	0.799	0.730	0.887	-1.59 MYC+ 0.66 AKT
Glycerolipid metabolism	3	0.229	0.738	0.319	0.686	0.854	0.885	0.251	0.887	-0.27 MYC -2.01 AKT
Glycerophospholipid metabolism	9	0.393	0.829	0.343	0.686	0.326	0.799	0.054	0.887	-1.35 MYC -1.22 AKT
Glycine, serine and threonine metabolism	12	0.423	0.829	0.497	0.686	0.162	0.799	0.295	0.887	-1.68 MYC -0.05 AKT
Glyoxylate and dicarboxylate metabolism	8	0.444	0.829	0.471	0.686	0.455	0.799	0.448	0.887	-1.28 MYC -1.32 AKT

Histidine metabolism	9	0.568	0.829	0.445	0.686	0.339	0.799	0.762	0.887	-1.5 MYC+ 0.91 AKT
Inositol phosphate metabolism	2	0.119	0.716	0.147	0.686	0.246	0.799	0.405	0.887	1.04 MYC -1.59 AKT
Lysine biosynthesis	5	0.841	0.896	0.388	0.686	0.995	0.995	0.734	0.887	0.07 MYC -2.03 AKT
Lysine degradation	9	0.768	0.857	0.728	0.797	0.472	0.799	0.851	0.907	-1.55 MYC+ 0.8 AKT
Methane metabolism	7	0.376	0.829	0.408	0.686	0.163	0.799	0.550	0.887	-1.61 MYC -0.6 AKT
Nicotinate and nicotinamide metabolism	8	0.752	0.855	0.491	0.686	0.708	0.838	0.105	0.887	0.6 MYC -1.9 AKT
Nitrogen metabolism	5	0.057	0.472	0.037*	0.358	0.115	0.799	0.585	0.887	0.73 MYC -1.83 AKT
Novobiocin biosynthesis	2	0.423	0.829	0.588	0.758	0.641	0.808	0.466	0.887	0.46 MYC -1.95 AKT
Oxidative phosphorylation	7	0.644	0.840	0.482	0.686	0.318	0.799	0.834	0.907	-1.68 MYC -0.03 AKT
Pantothenate and CoA biosynthesis	10	0.717	0.855	0.377	0.686	0.678	0.819	0.735	0.887	0.18 MYC -2.02 AKT
Pentose and glucuronate interconversions	6	0.345	0.829	0.276	0.686	0.450	0.799	0.860	0.907	-1.32 MYC -1.26 AKT
Pentose phosphate pathway	7	0.476	0.829	0.519	0.700	0.196	0.799	0.680	0.887	-1.68 MYC -0.12 AKT
Peptidoglycan biosynthesis	3	0.652	0.840	0.635	0.784	0.545	0.799	0.501	0.887	-0.11 MYC -2.03 AKT
Phenylalanine, tyrosine and tryptophan biosynthesis	4	0.210	0.716	0.304	0.686	0.903	0.919	0.068	0.887	0.24 MYC -2.01 AKT
Phenylalanine metabolism	7	0.580	0.829	0.661	0.787	0.364	0.799	0.648	0.887	-1.3 MYC -1.29 AKT
Porphyrin and chlorophyll metabolism	4	0.672	0.847	0.357	0.686	0.371	0.799	0.458	0.887	-1.6 MYC+ 0.61 AKT
Propanoate metabolism	5	0.339	0.829	0.297	0.686	0.579	0.799	0.322	0.887	0.83 MYC -1.77 AKT
Purine metabolism	18	0.01*	0.29**	0.03*	0.348**	0.021*	0.799	0.929	0.945	1 MYC -1.63 AKT
Pyrimidine metabolism	12	0.004*	0.232**	0.016*	0.309**	0.837	0.885	0.780	0.887	0.58 MYC -1.91 AKT
Pyruvate metabolism	3	0.126	0.716	0.204	0.686	0.043*	0.799	0.543	0.887	-1.61 MYC+ 0.57 AKT
Riboflavin metabolism	3	0.576	0.829	0.079	0.509	0.373	0.799	0.767	0.887	-1.44 MYC -1.06 AKT
Sphingolipid metabolism	4	0.559	0.829	0.668	0.787	0.844	0.885	0.246	0.887	0.57 MYC -1.91 AKT
Starch and sucrose metabolism	4	0.190	0.716	0.275	0.686	0.067	0.799	0.390	0.887	-1.66 MYC -0.34 AKT
Sulfur metabolism	3	0.586	0.829	0.162	0.686	0.278	0.799	0.286	0.887	-1.68 MYC+ 0.08 AKT
Taurine and hypotaurine metabolism	7	0.405	0.829	0.311	0.686	0.245	0.799	0.534	0.887	1.45 MYC -1.03 AKT
Thiamine metabolism	4	0.578	0.829	0.425	0.686	0.321	0.799	0.634	0.887	-1.36 MYC -1.2 AKT
Toluene degradation	3	0.200	0.716	0.319	0.686	0.590	0.799	0.429	0.887	0.54 MYC -1.92 AKT
Tryptophan metabolism	6	0.646	0.840	0.484	0.686	0.778	0.885	0.649	0.887	0.87 MYC -1.74 AKT
Tyrosine metabolism	5	0.512	0.829	0.491	0.686	0.531	0.799	0.560	0.887	-1 MYC -1.63 AKT
Ubiquinone and other terpenoid-quinone biosynthesis	4	0.528	0.829	0.548	0.722	0.638	0.808	0.804	0.897	-0.42 MYC -1.97 AKT
Valine, leucine and isoleucine biosynthesis	5	0.154	0.716	0.180	0.686	0.555	0.799	0.573	0.887	-0.1 MYC -2.03 AKT
Valine, leucine and isoleucine degradation	3	0.051	0.472	0.067	0.500	0.761	0.883	0.543	0.887	0.42 MYC -1.97 AKT

Vitamin B6 metabolism	3	0.913	0.913	0.436	0.686	0.673	0.819	0.504	0.887	-1.47 MYC -0.99 AKT

*Associations significant at p-value<0.05.
**Associations significant at q-value<0.35.

Supplementary Table 2. P-values and q-values calculated by LCT for association between bivariate oncogenes (MYC, AKT1) or univariate oncogenes (MYC only or AKT1 only) and different metabolite sets in normal samples

Metabolite Set	Size of Metabolite Set	p-value for (MYC, AKT1)	q-value for (MYC, AKT1)	P-value for (MYC, AKT1) Interaction	q-value for (MYC, AKT1) Interaction	p-value for MYC only	q-value for MYC only	p-value for AKT1 only	q-value for AKT1 only	Estimated linear combination of oncogenes
Alanine, aspartate and glutamate metabolism	11	0.264	0.979	0.144	0.964	0.611	0.911	0.491	0.988	-0.24 MYC -2.01 AKT
Amino sugar and nucleotide sugar metabolism	7	0.743	0.979	0.785	0.964	0.781	0.911	0.464	0.988	0 MYC -2.03 AKT
Arginine and proline metabolism	21	0.802	0.979	0.595	0.964	0.545	0.911	0.809	0.988	-0.12 MYC -2.03 AKT
Ascorbate and aldarate metabolism	4	0.835	0.979	0.840	0.964	0.805	0.911	0.493	0.988	-0.28 MYC -2 AKT
Benzoate degradation	2	0.028*	0.667	0.057	0.964	0.686	0.911	0.021*	0.445	0.06 MYC -2.03 AKT
beta-Alanine metabolism	12	0.839	0.979	0.841	0.964	0.677	0.911	0.506	0.988	-0.03 MYC -2.03 AKT
Biosynthesis of unsaturated fatty acids	13	0.575	0.979	0.548	0.964	0.654	0.911	0.368	0.988	-0.37 MYC -1.98 AKT
Butanoate metabolism	9	0.575	0.979	0.488	0.964	0.707	0.911	0.667	0.988	-0.21 MYC -2.02 AKT
C5-Branched dibasic acid metabolism	5	0.636	0.979	0.584	0.964	0.526	0.911	0.680	0.988	-0.31 MYC -2 AKT
Chlorocyclohexane and chlorobenzene degradation	2	0.042*	0.667	0.070	0.964	0.584	0.911	0.023*	0.445	0.06 MYC -2.03 AKT
Citrate cycle (TCA cycle)	8	0.960	0.979	0.954	0.964	0.711	0.911	0.997	0.997	1.12 MYC -1.52 AKT
Cyanoamino acid metabolism	5	0.744	0.979	0.742	0.964	0.662	0.911	0.433	0.988	-0.14 MYC -2.02 AKT
Cysteine and methionine metabolism	11	0.738	0.979	0.585	0.964	0.893	0.955	0.700	0.988	-0.26 MYC -2.01 AKT
D-Alanine metabolism	2	0.814	0.979	0.684	0.964	0.633	0.911	0.667	0.988	-0.21 MYC -2.02 AKT
D-Glutamine and D-glutamate metabolism	3	0.286	0.979	0.255	0.964	0.467	0.911	0.283	0.988	-0.36 MYC -1.99 AKT
Dioxin degradation	2	0.028*	0.667	0.057	0.964	0.686	0.911	0.021*	0.445	0.06 MYC -2.03 AKT
Fatty acid biosynthesis	5	0.601	0.979	0.540	0.964	0.585	0.911	0.662	0.988	-0.6 MYC -1.9 AKT

Fatty acid metabolism	3	0.923	0.979	0.896	0.964	0.707	0.911	0.854	0.988	-0.41 MYC -1.97 AKT
Fructose and mannose metabolism	6	0.421	0.979	0.481	0.964	0.769	0.911	0.238	0.988	0.01 MYC -2.03 AKT
Galactose metabolism	6	0.655	0.979	0.688	0.964	0.577	0.911	0.408	0.988	-0.01 MYC -2.03 AKT
Glutathione metabolism	12	0.798	0.979	0.702	0.964	0.716	0.911	0.565	0.988	-0.06 MYC -2.03 AKT
Glycerolipid metabolism	3	0.570	0.979	0.541	0.964	0.407	0.911	0.318	0.988	0.01 MYC -2.03 AKT
Glycerophospholipid metabolism	9	0.373	0.979	0.267	0.964	0.227	0.911	0.191	0.988	-0.51 MYC -1.93 AKT
Glycine, serine and threonine metabolism	12	0.682	0.979	0.751	0.964	0.737	0.911	0.380	0.988	-0.12 MYC -2.03 AKT
Glyoxylate and dicarboxylate metabolism	8	0.977	0.979	0.964	0.964	0.739	0.911	0.985	0.997	0.04 MYC -2.03 AKT
Histidine metabolism	9	0.307	0.979	0.294	0.964	0.550	0.911	0.121	0.988	-0.15 MYC -2.02 AKT
Inositol phosphate metabolism	2	0.210	0.979	0.209	0.964	0.143	0.911	0.132	0.988	-0.24 MYC -2.01 AKT
Lysine biosynthesis	5	0.711	0.979	0.660	0.964	0.930	0.955	0.391	0.988	-0.03 MYC -2.03 AKT
Lysine degradation	9	0.159	0.979	0.190	0.964	0.939	0.955	0.041*	0.595	-0.04 MYC -2.03 AKT
Methane metabolism	7	0.855	0.979	0.773	0.964	0.772	0.911	0.618	0.988	-0.21 MYC -2.02 AKT
Nicotinate and nicotinamide metabolism	8	0.705	0.979	0.604	0.964	0.316	0.911	0.435	0.988	0.52 MYC -1.93 AKT
Nitrogen metabolism	5	0.320	0.979	0.246	0.964	0.510	0.911	0.214	0.988	-0.2 MYC -2.02 AKT
Novobiocin biosynthesis	2	0.719	0.979	0.671	0.964	0.342	0.911	0.748	0.988	-0.6 MYC -1.9 AKT
Oxidative phosphorylation	7	0.616	0.979	0.575	0.964	0.467	0.911	0.304	0.988	0.23 MYC -2.01 AKT
Pantothenate and CoA biosynthesis	10	0.963	0.979	0.903	0.964	0.907	0.955	0.937	0.988	-0.38 MYC -1.98 AKT
Pentose and glucuronate interconversions	6	0.972	0.979	0.952	0.964	0.792	0.911	0.843	0.988	-0.18 MYC -2.02 AKT
Pentose phosphate pathway	7	0.846	0.979	0.770	0.964	0.770	0.911	0.608	0.988	-0.3 MYC -2 AKT
Peptidoglycan biosynthesis	3	0.641	0.979	0.597	0.964	0.455	0.911	0.513	0.988	-0.38 MYC -1.98 AKT
Phenylalanine, tyrosine and tryptophan biosynthesis	4	0.625	0.979	0.568	0.964	0.313	0.911	0.814	0.988	-0.75 MYC -1.82 AKT
Phenylalanine metabolism	7	0.941	0.979	0.922	0.964	0.654	0.911	0.886	0.988	0.14 MYC -2.02 AKT
Porphyrin and chlorophyll metabolism	4	0.046*	0.667	0.018*	0.964	0.021*	0.911	0.498	0.988	-1.37 MYC -1.18 AKT
Propanoate metabolism	5	0.796	0.979	0.757	0.964	0.598	0.911	0.622	0.988	-0.15 MYC -2.02 AKT
Purine metabolism	18	0.629	0.979	0.649	0.964	0.862	0.943	0.361	0.988	-0.1 MYC -2.03 AKT
Pyrimidine metabolism	12	0.815	0.979	0.707	0.964	0.693	0.911	0.760	0.988	-0.26 MYC -2.01 AKT
Pyruvate metabolism	3	0.814	0.979	0.724	0.964	0.481	0.911	0.967	0.997	-0.53 MYC -1.93 AKT
Riboflavin metabolism	3	0.421	0.979	0.374	0.964	0.381	0.911	0.288	0.988	0.23 MYC -2.01 AKT
Sphingolipid metabolism	4	0.215	0.979	0.117	0.964	0.141	0.911	0.875	0.988	-0.89 MYC -1.72 AKT
Starch and sucrose metabolism	4	0.788	0.979	0.743	0.964	0.980	0.980	0.562	0.988	-0.1 MYC -2.03 AKT

Sulfur metabolism	3	0.890	0.979	0.813	0.964	0.729	0.911	0.771	0.988	-0.32 MYC -1.99 AKT
Taurine and hypotaurine metabolism	7	0.407	0.979	0.322	0.964	0.512	0.911	0.548	0.988	-0.42 MYC -1.97 AKT
Thiamine metabolism	4	0.979	0.979	0.917	0.964	0.817	0.911	0.929	0.988	-0.38 MYC -1.98 AKT
Toluene degradation	3	0.725	0.979	0.703	0.964	0.396	0.911	0.931	0.988	-1 MYC -1.64 AKT
Tryptophan metabolism	6	0.691	0.979	0.597	0.964	0.438	0.911	0.874	0.988	-0.42 MYC -1.97 AKT
Tyrosine metabolism	5	0.890	0.979	0.881	0.964	0.662	0.911	0.724	0.988	-0.34 MYC -1.99 AKT
Ubiquinone and other terpenoid-quinone biosynthesis	4	0.674	0.979	0.616	0.964	0.791	0.911	0.409	0.988	-0.34 MYC -1.99 AKT
Valine, leucine and isoleucine biosynthesis	5	0.779	0.979	0.659	0.964	0.395	0.911	0.909	0.988	-0.5 MYC -1.94 AKT
Valine, leucine and isoleucine degradation	3	0.822	0.979	0.806	0.964	0.663	0.911	0.731	0.988	-0.49 MYC -1.94 AKT
Vitamin B6 metabolism	3	0.602	0.979	0.519	0.964	0.729	0.911	0.395	0.988	-0.14 MYC -2.03 AKT

*Associations significant at p-value<0.05.
**Associations significant at q-value<0.35.