7-16-2014

Set Based Association Testing in High Dimensional Genomic Studies

Xueyuan Cao

Recommended Citation

Cao, Xueyuan, "Set Based Association Testing in High Dimensional Genomic Studies" (2014). *Electronic Theses and Dissertations*. 995.
https://digitalcommons.memphis.edu/etd/995

SET BASED ASSOCIATION TESTING IN HIGH DIMENSIONAL GENOMIC STUDIES

by

Xueyuan Cao

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Mathematical Sciences

The University of Memphis

August 2014

ACKNOWLEDGMENTS

I would like to sincerely thank my advisor and mentor, Dr. E. Olusegun George, for his wonderful guidance, supervision, support, and encouragement, during my graduate studies at the University of Memphis.

I am also very grateful for having an outstanding doctoral committee and wish to thank Dr. Dale B. Armstrong, Dr. Vinhthuy Phan, Dr. Stanley Pounds and Dr. Cheng Cheng for their support, suggestions and encouragements.

I am very grateful to having Dr. Stanley Pounds as my supervisor and mentor in the Department of Biostatistics, St. Jude Children’s Research Hospital. Under Dr. Pounds’ supervision, I am exposed to various clinical and genetic studies, which motivated me to pursue PhD study at the University of Memphis. Dr. Pounds always fully supported my research and PhD study. St. Jude Children’s Research Hospital is a great place to work with a great reason ’Finding Cures and Saving Children’. The Department of Biostatistics is especially a warm family with great colleagues from whom I have enjoyed and learned so much. I also want to thank Dr. Rubnitz, Dr. Lamba and Dr. Downing for allowing me to use their data.

I am especially grateful to my parents, Yingxian Cao and Jingai Liu, for raising and teaching me, and sisters for their hearty support. This journey cannot be successful and enjoyable without my loved ones: my wife, Mingjuan Wang, and my sons, Kevin and Kerry. My study and life cannot be in the proper order and balance without them.

ABSTRACT

Cao, Xueyuan, Ph.D. The University of Memphis, August, 2014. Set Based Association Testing in High Dimensional Genomic Studies. Major Professor: E. Olusegun George.

The last decade has ushered in an era of high dimensional, high volume data. In particular with the biotechnological revolution of the era, high-dimensional genomic studies of various designs have provided investigators with the tools to study thousands or even millions of genomic features simultaneously. These studies have shed new light on the underlying mechanisms of complex diseases. The accumulated knowledge of these complex relationship between genes has led scientists to formalize pathways and graphical networks that visually and succinctly give descriptions of the geometry of these relationships. With such knowledge, it has become possible to develop procedures for statistical inference, not just at the individual genes level, but at the more meaningful gene-set level. The focus of this thesis is the development of new statistical procedures for such gene-set analysis.

After presenting an overview at the introduction, we give a comprehensive review of the literature relevant developments in the thesis in Chapter 2. In Chapter 3, we develop a Bayesian procedure that incorporates information contained in a gene graphical network, viewed as a directed graph, into the construction of prior distributions and we use the derived posterior distributions to construct statistical tests at the gene-set level. Our procedure extends the work of Pan (2006) and Wei and Pan (2008) which did not use the direction as information in the graphical network, but rather used undirected graphs and assumed a mixture model for the distribution to generate the posterior distribution of the mixing parameters via the use of a Markov random field. We demonstrate the gain in statistical power of our procedure over Pan and Wei's in an application to detect differentially expressed genes, and gene-sets by analyzing a data set that compares favorable risk and poor risk defined by cytogenetics in adults with acute myeloid leukemia (AML).

To enhance comprehension of the vast and complex information in high-dimensional data from genomic studies, it is sometimes useful and desirable to have a procedure that relates such data to specific endpoints. In this regards, association tests are highly desirable. In Chapter 4, we propose a procedure which we label ‘Projection onto Orthogonal Space Testing (POST)’ as a flexible method for testing association of gene sets and pathways with specific phenotypic endpoints while adjusting for other factors and variables as needed. In a simulation study, we demonstrate that POST has better operating characteristics than other methods recently developed to address the same objective. Thus we feel that POST does not only help to better understand treatment responses, but also prioritizes pathways for further study. We expect that POST will be especially valuable in clinical studies where cohorts with moderate to large sample sizes have rich high-dimensional data.

Another new procedure for association testing which we label ’Locus Based Integrated Testing(LOCIT)’ and an extension of the procedure -LOCITO- are introduced in Chapter 5. LOCIT is designed to test association of multiple forms of genomic data within a locus with an endpoint of interest in genomic studies. Given different forms of genomic data such as SNP genotypes, gene expression, and methylation levels, LOCIT performs one test per locus, taking several features at the locus into consideration. To illustrate the efficacy of LOCIT, we apply the procedure to a set consisting of SNP genotypes and gene profiling in an AML cohort to identify loci /genes that are associated with clinical outcomes.

In chapter 6, we summarize our development of gene-set level association tests and outline future directions of our research in this area.

TABLE OF CONTENTS

Contents **Pages**

LIST OF TABLES

Tables **Pages**

LIST OF FIGURES

Figures **Pages**

Chapter 1

Introduction

The pioneering “working draft” DNA sequencing of the human genome was completed in 2000 through an international effort of the Human Genome Project (HGP, http://www.genome.gov/). Since then, partial or whole genomic sequences of many species of animals, such as nematode worm (Sequencing Consortium and others, 1998)[1], fruit fly (Adams *et al.*, 2000)[2], mouse (Chinwalla *et al.*, 2002)[3], and plants such as Arabidopsis (Arabidopsis Genome Initiative and others, 2000)[4], rice (Goff *et al.*, 2002)[5] and maize have been published. Together with the efforts on sequencing of Expression Sequence Tag (EST) these explosion of sequencing the genome of living organisms has unlocked the door to the study of genetics and biology at the genome wide level, and launched scientific research to a level previously unattainable.

One of the goals in the post genomic era is to decipher genetic information encoded in the genomic sequences and use this information to formulate and test hypotheses. At the same time, new technologies and methods have evolved to acquire and analyze data that generate new biological and biomedical hypotheses. One such new technology is microarray gene profiling. In a gene profiling experiment, the expression levels of thousands of genes are measured simultaneously using micro-chip such as GeneChip by Affymetrix (http://www.affymetrix.com/, available in dozens of models and commercial species) and BeadArray by Illumina (http://www.illumina.com/, human and mouse). Recently, a second-generation sequencing based technology, RNAseq, has emerged for measuring gene expression. The RNAseq is an advancement that is considered to be more accurate than the older methods, not only for measuring gene expression levels, but also for detecting alternative splicing events (http://www.illumina.com/). Besides gene expression profiling, some of these micro-chip and sequencing technologies have been used to study micro RNA (miRNA), for the purpose of identifying activities such as single nucleotide polymorphisms (SNPs), and epigenetic phenomenon such as DNA methylation, in genetic samples.

In a typical GeneChip microarray profiling experiment, mRNA or total RNA strands, isolated from experimental units (cells, tissue etc), are reversely transcribed to single-strands of cDNA (complementary DNA) which are then synthesized to double-stranded cDNA. Biotin-labeled cRNA strands are then transcribed from the double-stranded cDNA, fragmented, and hybridized to a GeneChip microarray. After undergoing washing and staining, the hybridized mircoarray is then scanned by a laser and the scanned signals are processed by MAS5.0 (Statistical Algorithms Description Document (2002) Affymetrix Inc.) or other robust multi-array average (RMA) methods (Irizarry *et al.*, 2003)[6] to obtain expression values. Statistical analysis of these expression values across experimental units are performed in accordance to the design of the experiment. Accompanying the development of these new technologies are ongoing research to develop new software to process, generate data and perform statistical inference on the high dimensional data set generated by these processes.

One of the first applications of these new technologies was to study differentially expressed genes (DEGs) in probe-set (gene) level under different treatments (for example between normal and disease samples). At its most rudimentary, the methods to test the null hypothesis of no group mean difference include the two-sample t-test. More sophisticated test procedures have emerged in recent years to deal with experiments involving multiple treatments, involving thousands of genes. The statistical test at the gene level is called individual gene analysis (IGA). When considering thousands of tests performed simultaneously, adjustments must be made to control error rate. After the necessary adjustment for multiple testing, a list of genes/probes are declared to be significantly differentially expressed at certain level of false discovery rate (FDR) and provided to investigators. Khatri and Drăghici (2005)[7] provide an extensive review of IGA approaches.

Given a long list of differentially expressed genes at a specified FDR, investigators rely heavily on bioinformatic databases or tools to annotate the gene list in order to prioritize the genes and formulate working hypothesis. For the purpose of prioritizing genes and formulating hypotheses, gene set based analysis is used for formal testing and interpreting. Geoman and Buhlmann 2007[8], Nam and Kim 2008[9] reviewed some of the methods and recommended guidelines to be used for analyzing gene expression data at gene-set level. This topic will be extensively studied in the next chapter.

High throughput technologies have been evolving to study different biological mechanisms of model systems and diseases. Fundamental to biological systems and the inception of diseases is the micro RNA (miRNA), a small non-coding RNA molecule which functions in transcriptional and post-transcriptional levels to regulate gene expressions. The human genome encodes over 1000 miRNA which have been estimated to target about 60% of genes (Bentwich, *et al.*, 2005)[10]. They are abundant in many cell types and are involved in many biological processes and diseases. The levels of miRNA can be measured by micro-chips or by direct sequencing. Epigenetics is a phenomenon that attributes gene expression or occurrence of cellular phenotype to activities of other mechanisms other than changes in the underlying DNA sequence. Such activities includes histone modification, DNA methylation and RNA editing. Histone modifications have been studied using ChIP-Chip method (chromatin imunoprecipitation with microarray technology, Lieb *et al.*, 2001)[11] and recently Chip-seq (a next-generation sequencing based technology, Johnson *et al.*, 2007)[12]. DNA methylation levels are measured by micro-chip such as Illumina Infinium Methylation array, or by pyrosequencing.

In clinical trials, patient samples are extremely valuable for elucidating the mechanisms of diseases and for evaluation of treatment outcomes. In clinical trials, multiple types of genetic data are collected from the patient samples. In addition to multiple presenting features, these include various treatment outcome related variables and genetic data such as gene expression, miRNA, SNP, DNA methylation, histone modification. In addition to relating each type of genetic data with sample phenotypes such as presenting features, short- or long-term treatment responses or outcomes, integrating all these rich genetic data in a unified test is challenging. However, understanding the information encoded in theses data is of great interests to investigators.

In this dissertation, methods to incorporate prior knowledge of pathways into genetic studies are proposed. A flexible set-based procedure is proposed to evaluate the association of gene sets with diverse phenotypes. An integrated analysis approach based on predefined sets is also proposed to take advantage of the rich genomic data from multiple sources in a clinical trial setting. These methods may be adapted and extended to address other more or less complicated applications. We discuss such potential applications in the summary and future research section.

Chapter 2

GENE SET ANALYSIS

Since the introduction of high throughput expression profiling and genotyping, the primary interest has been to identify differential presentation of the genomic features and to elucidate the underlying biology. Many methods have been proposed to facilitate the interpretation this profiling in the context of clustering genes into gene sets and identifying gene pathways. In this chapter, we give a comprehensive review of the bioinformatics and statistical literatures in this context.

2.1 Biological Pathways

It has long been well known that genes and proteins do not function in isolation. Genes are organized on chromosomes, expressed, and function in a complex dependent manner under cellular context (Figures 4.2 and 4.3). The accumulated functional dependence can be described by graphical networks and biological pathways. Most of these pathways are metabolic, regulatory or signal transduction pathways.

For diverse organisms and at various genetic levels, many public and commercial databases have been developed to structure, store and characterize the dependent relationships between genes and proteins in these pathways and sets. Among the most widely accessed public databses is the Gene Ontology (GO). GO describes relationship between genes in term of unified ontology using directed acyclic graph (DAG) with a hierarchical structure (http://www.geneontology.org/; Ashburner *et al.*, 2000[13]). Another repository of genetic databases is NetPath. NetPath holds 20 manually curated human signaling pathways, including 10 immune signaling pathways and 10 cancer signaling pathways (Kandasamy *et al.*, 2010)[14]. Another is BioCarta which represents molecular or cellular pathways by interactive graphic models (http://www.biocarta.com/). Others include Reactome, an online database developed by Croft *et al.* 2011[15], authored by expert biologists in collaboration with Reactome editorial staffs, and the National Center for Biotechnology Information (NCBI) and Pathway Interaction Database (PID) which contains 137 human pathways curated by NCI-Nature and 322 pathways imported from BioCarta and Reactome (http://pid.nci.nih.gov/). These are a few examples and there are many more publicly available ones.

Besides the publicly available pathway databases, there are a few commercialized databases. For example, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database represents current knowledge of molecular interactions and reaction networks related to metabolism (metabolic pathways), signal transduction, cellular processes and human diseases using graphical representation (http://www.genome.jp/kegg/pathway.html; Kanehisa and Goto 2000[16]; Kanehisa *et al.*, 2006[17]). Ingenuity is another commercial web based pathway analysis tool (http://www.ingenuity.com/). Both KEGG and Ingenuity are widely used in the literature.

According to Pathguide, there are currently 159 pathway-related databases with more than 150K pathway entities. To alleviate the burden of using pathways across different databases, Yu *et al.* (2012) proposed to integrate various pathway databases for building a unified database. hiPathDB[18] was developed incorporating KEGG, Reactome, PID by NCI-Nature, and BioCarta (http://hipathdb.kobic.re.kr/). Since these databases are dynamic, it is necessary to adopt a flexible definition of gene sets in accordance to developments in statistical methodology.

2.2 Set level analysis in gene profiling

One of these developments in statistical procedures is the construction of a single test for the differential expression of a gene set, rather than multiplicity of tests of each gene in the gene set. Goeman and Buhlmann 2007[8], Nam and Kim 2008[9] divided methods for gene set analysis according to various definitions of null hypothesis and mechanisms for permutation in the calculation of *p*-values. Goeman and Buhlmann noted that these procedures can be classified into three groups. One group summarizes the over-representation of a gene set in a differential gene list by a 2x2 contingency table. However, this representation has a drawback, namely that it requires a strict *p*-value cut-off for declaring differential expression. Another uses a statistic based on the whole vector of *p*-values. A third group uses original expression data instead of *p*-value. Pounds (2013) noted that these methods do not always lead to meaningful inference because they have no information about the study design. Allison *et al.* 2006[19] also questioned the validity of some these methods.

The definition of global null hypothesis for a gene set and the calculation of *p*-values by permutation provide useful guide for developing statistically sound procedures for gene set analysis. In terms of the null hypothesis, a gene set test can be described as competitive or self-contained. A competitive test compares a gene set to a standard defined by the complement of the gene set. A drawback of competitive test is that it penalizes the gene set in zero-sum-game manner if the complement of that gene set has highly differential expression (Allison *et al.*, 2006). A self-contained test compares the gene set to a fixed standard which does not depend on the measurement of genes outside of the gene set. Thus a self-contained test evaluates the whole set to address the global null of no difference in an experiment, while a competitive test does not. *p*-values for these tests are usually computed by permutation of subject labeling or genes.

There are two major mechanisms for calculating permutation-based *p*-value: subject sampling (random assignment of group labels) and gene sampling. In subject sampling it is assumed that the measurements of different subjects are independent and identically distributed, while the measurements within a subject could be correlated. In contrast, gene sampling assumes that genes are random samples that are independently and identically distributed, a reversal of the roles of samples and genes relative to classical statistical setup. Subject sampling produces valid *p*-values and interpretation of the *p*-values is straightforward (Pounds 2013).

In conducting gene set analysis, Goeman and Buhlmann (2007) recommended testing a self-contained null hypothesis and basing the calculation of *p*-values on subject sampling. Based on these criteria, Nam and Kim (2008) provided a detailed list of methods for gene set analysis with guideline for self-contained versus competitive tests and the use of gene versus sample randomization for computing P-values. A few commonly used methods are selected for further study.

2.3 Some gene set analysis methods

In this section, we review several methods which have had high impact on the analysis of gene sets. In this thesis, we extend these methods and develop new procedures to address some of the existing shortcomings in these methods.

2.3.1 Gene set enrichment analysis (GSEA)

Gene set enrichment analysis (GSEA) was first proposed by Mootha *et al.* (2003)[20] and Subramanian *et al.* (2005)[21] for interpreting gene expression data. GSEA considers data from randomized experiments or observational studies with two groups. Based on the existence of correlation between gene expression and phenotype (evidence of association of gene expression with groups), a ranked gene list is generated. GSEA attempts to determine whether the members of a set S defined a priori, are randomly distributed across the ranked gene list or lie primarily on the top or bottom of the gene list. From the ranked gene list, an enrichment score (ES) is calculated by walking down the gene list, increasing a running-sum statistics when encountering a gene in the gene set S and decreasing it when encountering a gene not in S. The ES is the maximum deviation from zero encountered in this random walk (similar to Kolmogorov-Smirnov statistic). Subsequently, the null distribution of normalized ES (NES) is approximated by permutations of the class labels, and *p*-value is calculated by using observed NES under the null distribution. The implementation of this method provides an option of permuting genes. This method is applied to 4 data sets and its advantages are demonstrated.

In this initial GSEA, the gene set test statistic (ES) is competitive, not self-contained. It penalizes a gene set when genes out of the gene set are strong correlated with the phenotype. Tian *et al.* (2005) [22] and Kim and Volsky (2005)[23] proposed an extension to GSEA, in which a two-sample statistic, such as a t-statistic, is used instead of enrichment score. The test statistic for a gene set is the aggregate of per gene test statistics of its members and significance is determined by permutation. The test statistics is self-contained.

Efron and Tibshirani (2007)[24] extended the GSEA by using an alternative summary statistics for gene-sets and restandardization based on row randomization. They showed that the *maxmean* statistic is more powerful than the original GSEA. An R package *GSA* for implementing the procedure described by Efron and Tibshirani is available on CRAN and it is more user friendly than the original GSEA.

Jiang and Gentleman (2007)[25] pursued an extension to original GSEA. In GSEA, the definition of gene set statistics has three components: (1) per-gene statistics: measurement of association between genes and a phenotype, (2) relationship of genes with gene sets and (3) per-gene set summarization function. Jiang and Gentleman extended the GSEA approach in all these three components. Besides using the two-sample t-statistics, the per-gene statistic was extended to a linear model setting, adjusting for covariates as needed. It was also extended to use posterior probability as per-gene statistics. In terms of the per-gene set summarization function, the mean was extended to the median. To tackle the problem of overlap between gene sets, three gene sets were constructed from two gene sets with significant overlap. This extension helped to isolate and identify the gene set associated with the phenotype. This extension to three components led to a number of GSEA approaches which are self-contained and the use of permutation of subject labels led to legitimate p-values. Jiang and Gentleman attempted to apply principal component analysis (PCA) to gene sets to identify substructure of the sets. In a sense, the extension of GSEA could be considered as gene set testing instead of enrichment analysis.

2.3.2 Significance analysis of functional categories (SAFE)

Barry *et al.* (2005)[26] proposed SAFE (significance analysis of functional categories) procedure to test association of predefined gene sets with a phenotype. SAFE has two statistics: local statistic (per-gene statistic measuring association of a gene with the phenotype) and global statistic (per-gene set statistic measuring the difference between genes within a gene set and those outside of the gene set). The local statistics is derived from various models based on the experimental design. This flexibility of local statistics is valuable and applicable to many experimental designs, similar to one of the extensions in GSEA by Jiang and Gentleman (2007). The global statistic assesses how the distribution of local statistics within a gene set differs from the local statistics outside of the gene set. The significance of global statistics is determined by permuting subjects' labeling with the experimental design taken into account.

The global test statistic for SAFE treats genes within or outside of a gene set as independent samples, which is generally not valid as genes are usually correlated within a gene set. The null hypothesis is hard to explicitly define. The global test result of a gene set is influenced or penalized by the genes outside of the gene set. It does not test the global null of no association of gene expression in the experiment with the phenotype. It does not stably test a gene set consisting of one or a few genes, either.

2.3.3 Hotellings T^2 test

GSEA and its extensions, and SAFE are not multivariate analyses of predefined gene sets, although these methods take the relationship of membership in the set into account. Lu *et al.* (2005)[27] attempted to directly test association of gene sets with treatments. A multiple forward search (MFS) algorithm was proposed to select genes in a gene set using the maximum Hotellings T^2 statistic between the two groups. A re-sampling technique was used to obtain robust mean estimate of the Hotellings T^2 statistics with lower and upper 5% quantiles removed. The authors used the *p*-value from the Hotellings T^2 test of a gene set as *p*-value for all the selected genes in the gene set. However, this is not valid and represents a misinterpretation of Hotellings T^2 test. The *p*-value of Hotellings T^2 test is the probability of obtaining a statistic as extreme as the observed under global null that all the mean expression levels of member genes in the gene set are equal between two groups. This is dramatically different from *p*-value obtained from testing that each of the mean expression levels of member genes in the test is different between two groups.

The MFS introduces selection bias and the re-sampling stabilization of Hotellings T^2 statistics makes the test statistics intractable and hard to interpret statistically. However, the method has been demonstrated to have the ability to predict group labeling. This feature could be expected from the MFS algorithm. However, the prediction accuracy should be demonstrated in an independent experiment. The test itself is for global null of gene sets and should not be treated for individual genes selected by MFS.

Srivastava *et al.* (2007)[28] proposed the use of Hotellings T^2 statistic to measure difference in mean vectors between two groups in compositional data. The significance is determined by permuting group labeling. This permeation-based Hotellings T^2 can be applied to high dimensional genomic data with two treatments/groups. Although the Hotellings T^2 lacks power for high dimensional data, the statistic is an appropriate measure of difference between two groups. As *p*-value is determined by permutation of group labeling, the procedure might work well in this scenario.

2.3.4 MRPP test

Nettleton *et al.* (2008)[29] proposed a nonparametric multivariate analysis approach to identify differentially expressed gene categories (sets) between two or more treatment groups. MRPP (multi response permutation procedure) was proposed to test the null of equal multivariate distribution of a gene set across treatment groups. The coherence of a gene set in a treatment group is measured using all the Euclidean distances between pairs of data vectors from the treatment group. The MRPP statistics is the average of the coherence measurement across treatment groups, weighted by sample sizes of corresponding treatment groups. The MRPP statistics can be scaled to have common variance for each gene. The *p*-value is assessed by permuting sample group labels, taking experimental design into account if needed.

MRPP test is self-contained and uses subject sampling. It produces valid *p*-values and is easily interpreted. It has power to detect gene sets that are different between/among treatment groups in multivariate space, but not marginally. It is non-parametric and has hence the advantage of fewer assumptions. However, it is difficult to extend to more complex experimental design or to adjust for other covariates. It is not applicable in many cohort studies, in which presenting features or prognostic factors need to be adjusted for.

2.3.5 Sparse canonical correlation analysis

Canonical correlation is widely used in psychology to test the association between two sets of variables such as assessing agreement of items in instruments. The traditional canonical correlation is hard to apply to high dimensional genetic data. Karkhomenko *et al.* (2009)[30] proposed using sparse canonical correlation to test association between two set of variables in genetic studies. To account for experimental design or other factors, the residuals after a linear model with other factors as predictors are used as starting data. Karkhomenko *et al.* proposed to use soft threshold of left and right eigenvectors to reduce or select subset of variables in each set to maximize first-order approximation of correlation matrix. Adaptive sparse canonical correlation was employed to select even small set of variable with penalty similar to LASSO.

The method was demonstrated to select two manageable sets of genetic variables for hypothesis development. The example provided was for a whole study with both expression data and SNP genotype data. The methods are mainly for feature selection instead of testing gene sets. The method can be extended to gene set or locus based data for selecting subset of variables with high first-order correlation in a locus or gene set. These selected coherent variables can then be used to test for association with phenotypes such as outcomes and presenting features. So, the (adaptive) sparse canonical correlation method is a potentially useful feature selection tool for an integrated analysis with two types of genetic data in which one type regulates/influences the other. SNP and gene profiling, methylation and gene profiling, or microRNA and gene profiling are potentially suitable data types.

The idea of maximization of correlation between two sets of variables can be applied to maximization of correlation between one phenotype variable and a set of genetic variables. The genetic features selected can be used to test for association with the phenotype or other phenotypes if biologically warranted.

Witten *et al.* (2009)[31] proposed a penalized matrix decomposition (PMD) to approximate a matrix. The approximation approach was then applied to sparse canonical correlation setting resulting in penalized CCA (canonical correlation analysis) using $L1$-constrain or fused LASSO constrain on the so-called canonical variates. The proposed method was applied to a breast caner data to identify gene expressions that are associated with genomic gain/loss.

2.3.6 SKAT

Sequence Kernel Association Test was first proposed to test association of SNPs in a genomic locus or gene in case/control or with continuous variables (Wu *et al.* 2011[32], Lee *et al.* 2012[33]). In a linear or logistic regression framework, the phenotype is modeled with known covariates and SNPs in a set (gene or chromosomal region) as predictors. If the SNPs coefficients follow an arbitrary distribution with mean 0 and a variance of $w_j\tau$, SKAT uses variance-component score statistic to test $\tau = 0$, which is equivalent to requiring all coefficients of SNPs equal to be 0.

The variance-component score statistic of SKAT contains two parts: the deviation of phenotype from that predicted under null and a kernel function to measure genetic similarity among subjects. SKAT provides several options for assigning weights, based on minor allele frequency, under different assumptions of SNPs effects. The functional form of kernel function can be also extended to a more flexible function, allowing more complex models.

The significance of variance-component score statistic of SKAT can be analytically approximated by generalized chi-square distribution with available methods. The SKAT method was proposed for testing association SNPs with phenotypes. It can be generalized for use with other types of data.

2.3.7 PROMISE

In genomic studies, each experimental unit or subject may have multiple phenotypes measured. The phenotypes measured on the same subject are usually correlated. The PROMISE procedure proposed by Pounds *et al.* (2009)[34] tests a predefined projection of individual association statistics with each of the phenotypes.

The projection used is based on biological knowledge of the relationships among the phenotype variables. The association with individual phenotype can be measured by various models. The method handles a variety of endpoint of interests including categorical, continuous, and time to event variable. Compared to other available methods such as overlap approach, canonical correlation, principal component analysis, result from PROMISE is more biologically motivated and has meaningful and easier interpretation.

A test of PROMISE was also performed on gene sets. The method was demonstrated to have great power to detect association of predefined association pattern with multiple related phenotypes and has been successfully applied to a few studies (Lamba *et al.* 2011[35]).

2.4 Remarks

Various biological pathway databases have been developed to integrate and present accumulated gene-gene interaction, regulation and biological processes. These pathways can be used to formulate gene sets or biological processes of interest in an *a priori* manner. Although, these databases cannot be viewed as complete and are on-going in nature, successful applications are emerging.

It is generally agreed that a method for gene set analysis should have two features in term of the hypotheses to be tested and the calculation of *p*-value: It should be self-contained instead of competitive, and *p*-values calculation should be based on subject permutation instead of gene permutation, if permutations are needed. Some of the available gene set analysis methods have these features, while others do not.

The above review of methods for gene set analysis shows that many of procedures are not applicable to complex design and not malleable enough to allow for variables adjustments. However some of these methods can be improved or modified to address these shortcomings. This thesis addresses some of the needs for flexible general procedure to handle various phenotypes in complex design specially in biomedical field.

Chapter 3

APPLYING GENE NETWORK PRIOR KNOWLEDGE IN GENOMIC TESTING

3.1 Introduction

In an organism, genes are organized on chromosomes, expressed, and function in a complex interdependent manner. The accumulated functional dependence can be summarized succinctly as gene networks. Several public and commercial databases have been developed to structure and store the biological knowledge. Gene Ontology (GO) describes the relationship of genes in term of unified ontoloy terms using directed acyclic graph (DAG) with hierarchy structure. When the graph is cut at different levels, various gene sets are formed. Kyoto Encyclopedia of Genes and Genomes (KEGG)[16] depicts current knowledge of molecular interactions and reaction networks related to metabolism, cellular processes and human diseases using graphical networks. NCBI Pathway Integration Database (http://pid.nci.nih.gov) contains 137 human pathways curated by NCI-Nature and 322 pathways imported from BioCarta and Reactome.

A gene network is a set of genes represented by a graph of which the nodes denote genes and the edges represent relationships between genes. While undirected edges are used to represent conditional dependence, directed edges often represent causal relationships between genes. Directed acyclic graphs (DAGs) are graphs in which all the edges are directed and the graph has no cycles. Figure 3.1 shows yeast MAP kinase pathway derived from KEGG with DAG representation.

A number of methods have been proposed to incorporate the information of gene networks into joint analysis of gene expression data. Based on the null hypotheses, Goeman and Buhlmann (2007), Nam and Kim (2008) classified these methods into three groups: self-contained, competitive and mixed. Most of these methods involve two distinct steps: The differential expressions are tested at probe (gene) level separately, and then the gene level testing results are extended to gene set level by assessing the

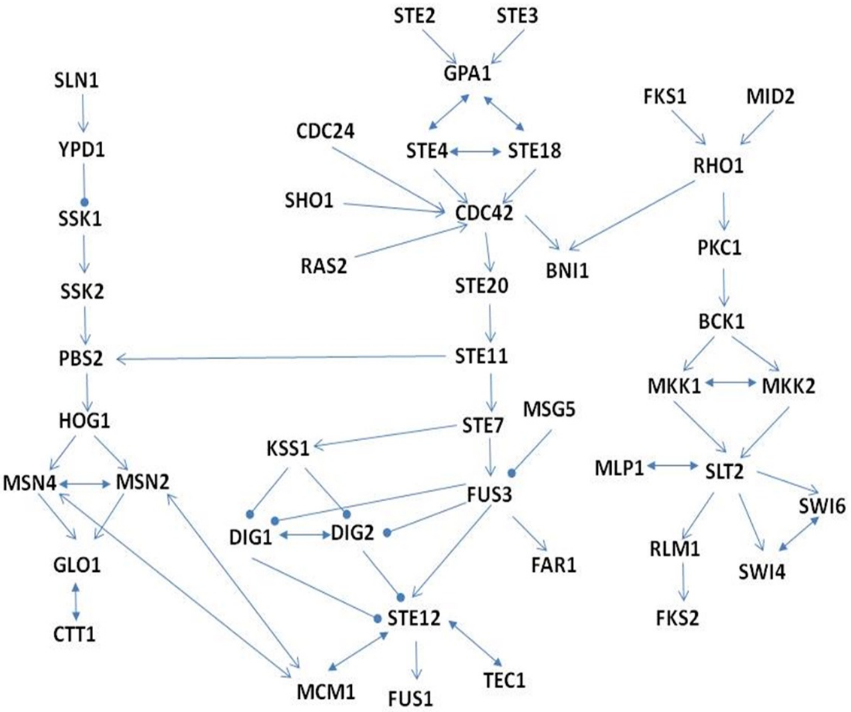


Fig. 3.1: Yeast MAP kinase pathway derived from KEGG database.

The pointed arrows denote positive regulation or activation, while the dotted denote negative regulation or repression. The physical interactions are shown by bi-directional edges. Gene names are translated from the KEGG gene names to the ones used by Affymetrix.

over-representation of differentially expressed genes in each gene set. Another family of methods directly perform multivariate tests of differential expression for a group of genes belonging to a gene set.

In order to incorporate the information contained in gene networks into prior distributions for Bayesian inference, Pan (2006) [36], Wei and Pan (2008) [37] proposed using a Markov random field of first-order dependence to model the dependence of member genes in a pathway. They used a logistic model to represent the probability that a gene is expressed or inhibited through latent Gaussian Markov random field variables of the gene. Through the use of these techniques to generate prior distribution, they were able to demonstrate gain in gene ranking with the more interesting genes appearing at the top of the gene list in a Yeast experiment. However, they used normal transformation of p-values from tests of gene expression (Z-scores) as data, and their procedure lacked dependences among genes in a pathway. Moreover, they did not take directions into consideration.

In this thesis, we propose the use of a more informative prior that incorporates gene network derived from pathway databases into gene level testing. By using the Markov random field on directed graphs, the procedure improves on the method proposed by Wei and Pan (2008). In contrast to Wei and Pan, we also use raw gene expression values, rather the transformed p-values that are obtained from tests of differential expression. Finally, we directly test differential expression along the directed graphs.

3.2 Model Specification

For a gene network/pathway described by a graph $G = (V, E)$, where V corresponds to the set of nodes with p elements and $E \subset V \times V$ is the set of edges, the nodes of the graph represent the genes and the edges capture the relationship among them. If $(i, j) \in E$ implies $(j, i) \notin E$, then the edge is directed. A directed acyclic graph is a graph with no bi-directed edges. In a directed acyclic graph, starting from a node v, there is no way to follow a sequence of directed edges and loop back to the node v again. An adjacency matrix can be used to represent which nodes have edges with other nodes. For an undirected graph, the adjacency matrix is symmetric. A directed acyclic graph (DAG) can be represented by a binary adjacency matrix $\boldsymbol{A}$, where each entry a_{ij} is either 0 or 1. A zero entry, $a_{ij} = 0$ indicates the absence of an edge between node i and node j; while, if $a_{ij} = 1$, there is a directed edge from node i to node j. The sum of i^{th} row and column of adjacency matrix $\boldsymbol{A}$ denotes the number of children and parents of the node i, respectively. From now on, we use adjacency matrix $\boldsymbol{A}$ to denote a gene network with a directed graph and DAG to denote a directed acyclic graph.

3.2.1 Model Z Scores

Assume that the data in a genomic study have been summarized to Z_i for each genomic feature $i, i = 1, ..., p$. Z_i could be a statistic to measure difference of a genomic feature between experiment conditions or groups of subjects with certain phenotypic feature, or a probability integral transformation of the statistical significance level to reject null hypothesis (of p-value), namely Z-score. The distribution of Z_i is assumed to be a mixture of a null component, a negative component and a positive component, that is

$$f(Z_i) = \pi_{0i} f_0(Z_i) + \pi_{1i} f_1(Z_i) + \pi_{2i} f_2(Z_i) \tag{3.1}$$

Under normal framework, $f_j(Z_i)$ are assumed to be a normal density function for $j = 0, 1, 2$, then

$$f(Z_i) = \pi_{0i}\Phi(Z_i; 0, \sigma_0^2) + \pi_{1i}\Phi(Z_i; \mu_1, \sigma_1^2) + \pi_{2i}\Phi(Z_i; \mu_2, \sigma_2^2) \tag{3.2}$$

$\Phi(Z_i; \mu_j, \sigma_j^2)$ is normal density with mean μ_j and variance σ_j^2, where $j = 0, 1, 2$.

Assume that a directed acyclic gene network is given and represented by adjacency matrix $\boldsymbol{A}$. The row sum of $\boldsymbol{A}$ is the number of children for a gene and column sum of $\boldsymbol{A}$ is the number of parents for the gene. The sum of row sum and column sum is the total number of directed edges that a gene has in the gene network $\boldsymbol{A}$. We assume that the children of a gene carry less information relative to its parents in an adjacency matrix $\boldsymbol{A}$ with $0 < w_i < 1$. Let m_i be number of information weighted edges for i^{th} gene, which is number of its parents plus w_i times of number of its children. Mathematically,

$$m_i = w_i\mathbf{A}[i,]\underline{\mathbf{1}} + \underline{\mathbf{1}}'\mathbf{A}[,i] = (w_i\mathbf{A}[i,] + \mathbf{A}'[,i])\underline{\mathbf{1}} \tag{3.3}$$

and $\mathbf{A}[i,] = [\mathbf{A}(i,1), \ldots, \mathbf{A}(p,i)]$

$$\mathbf{A}[,i] = \begin{bmatrix} \mathbf{A}(1,i) \\ \vdots \\ \mathbf{A}(p,i) \end{bmatrix}; \underline{\mathbf{1}} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}; w\underline{\mathbf{1}} = \begin{bmatrix} w \\ \vdots \\ w \end{bmatrix}$$

Given a gene network with adjacency matrix $\boldsymbol{A}$, we assume that the prior probability $\Pi_i' = (\pi_{0i}, \pi_{1i}, \pi_{2i})$ of Z_i coming from the j^{th} component is related to three latent Markov random field variables X_{ji}, for $j = 0, 1, 2$, through following logistic transformations for i^{th} gene:

$$\begin{aligned} \pi_{0i} &= \frac{1}{1 + e^{X_{1i} - X_{0i}} + e^{X_{2i} - X_{0i}}} \\ \pi_{1i} &= \frac{1}{1 + e^{X_{0i} - X_{1i}} + e^{X_{2i} - X_{1i}}} \\ \pi_{2i} &= \frac{1}{1 + e^{X_{0i} - X_{2i}} + e^{X_{1i} - X_{2i}}} \end{aligned} \tag{3.4}$$

We further assume that $\mathbf{X}_j = (X_{j1}, \ldots, X_{jp})'$ is distributed according to an intrinsic Gaussian conditional auto regression model (ICAR). That is the distribution of each latent variable $X_{j,i}$, conditional on $X_{j,(-i)} = \{X_{j,k}, k \neq i\}, i = 1, \ldots, p; j = 0, 1, 2$, depends only on its first-order neighbors on the adjacency matrix $\boldsymbol{A}$ with m_i defined as in equation 3.3.

$$X_{j,i} | X_{j,(-i)} \sim N\left(\frac{1}{m_i}(w_i \mathbf{A}[i,]\mathbf{X}_j + \mathbf{X}_j'\mathbf{A}[,i]), \frac{\sigma_{C_j}^2}{m_i}\right) \tag{3.5}$$

Where $\sigma_{C_j}^2$ is a hyperparameter that controls the strength of dependence of latent variable $X_{j,i}$ on its neighbors.

Let $\mathbf{L}_0, \mathbf{L}_1, \mathbf{L}_2$ be vectors of 0's or 1's, which indicate whether Z_i comes from null $f_0(Z_i)$, negative component $f_1(Z_i)$ or positive component $f_2(Z_i)$, respectively. Given

$\mathbf{\Pi}_j, j = 0, 1, 2$, the i^{th} component of $(\mathbf{L}_0, \mathbf{L}_1, \mathbf{L}_2)$ has a multinomial distribution with $\mathbf{p}_i = (\pi_{ji}, j = 0, 1, 2)$. Given $\mathbf{L}_j$, each Z_i has a normal distribution.

We use non-informative prior distributions for the parameters. The prior distributions for the means of negative and positive components are assumed to be truncated normal. For i^{th} gene, $\mu_{0i} = 0$; $\mu_{1i} \sim N(0, \sigma^2)I(a, 0)$, a normal distribution between a and 0, where $I(,)$ is an indicator function; and $\mu_{2i} \sim N(0, \sigma^2)I(0, b)$. a and b are chosen as $a = min(Z_1, \ldots, Z_p) - 3std, b = max(Z_1, \ldots, Z_p) + 3std$, where std is the standard deviation of $(Z_1, \ldots, Z_p)$. The hyperparameter σ^2 is set to 10^6 to be vague. The variances of null, negative and positive components are assumed to have inverse gamma distributions: $\sigma^2_{ji} \sim IG(\alpha, \beta)$ for $j = 0, 1, 2$, where the hyperparameters, α, β, are chosen to be 0.1; and $\sigma^2_{C_j} \sim IG(\alpha_C, \beta_C)$ for $j = 0, 1, 2$, where $\alpha_C = \beta_C = 0.01$. We further assume $w = w_1, \ldots, w_p$ for all genes. The overall structure of the model is as in figure 3.2. The joint posterior

$$\Pi(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\sigma}^2_0, \boldsymbol{\sigma}^2_1, \boldsymbol{\sigma}^2_2, \mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \mathbf{L}_0, \mathbf{L}_1, \mathbf{L}_2, \sigma^2_{c0}, \sigma^2_{c1}, \sigma^2_{c2} \mid \mathbf{Z}, \mathbf{A} \propto$$
$$f(\mathbf{Z}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\sigma}^2_0, \boldsymbol{\sigma}^2_1, \boldsymbol{\sigma}^2_2, \mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \mathbf{L}_0, \mathbf{L}_1, \mathbf{L}_2, \sigma^2_{c0}, \sigma^2_{c1}, \sigma^2_{c2} \mid \mathbf{A}, w, \sigma^2, \alpha, \beta, \alpha_c, \beta_c, a, b)$$

where

$$
\begin{aligned}
&f(\mathbf{Z}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\sigma}_0^2, \boldsymbol{\sigma}_1^2, \boldsymbol{\sigma}_2^2, \mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \mathbf{L}_0, \mathbf{L}_1, \mathbf{L}_2, \sigma_{c0}^2, \sigma_{c1}^2, \sigma_{c2}^2 \mid \mathbf{A}, w, \sigma^2, \alpha, \beta, \alpha_c, \beta_c, a, b) = \\
&\quad \prod_{i=1}^{p} \left(\frac{1}{\sqrt{2\pi} \times \sigma_{1i}} e^{-\frac{1}{2\sigma_{1i}^2}(Z_i - \mu_{1i}))^2} \right)^{L_{1i}} \left(\frac{1}{\sqrt{2\pi} \times \sigma_{0i}} e^{-\frac{1}{2\sigma_{0i}^2} Z_i^2} \right)^{L_{0i}} \left(\frac{1}{\sqrt{2\pi} \times \sigma_{2i}} e^{-\frac{1}{2\sigma_{2i}^2}(Z_i - \mu_{2i})^2} \right)^{1-L_{1i}-L_{0i}} \\
&\quad \prod_{i=1}^{p} \binom{1}{L_{0i}, L_{1i}} \left(\frac{1}{1+e^{X_{2i}-X_{1i}} e^{X_{0i}-X_{1i}}} \right)^{L_{1i}} \left(\frac{1}{1+e^{X_{1i}-X_{0i}} e^{X_{2i}-X_{0i}}} \right)^{L_{0i}} \left(\frac{1}{1+e^{X_{1i}-X_{2i}} e^{X_{0i}-X_{2i}}} \right)^{1-L_{1i}-L_{0i}} \\
&\quad \prod_{i=1}^{p} \frac{\sqrt{m_i}}{\sqrt{2\pi}\sigma_{C1}} e^{-\frac{m_i}{2\sigma_{C1}^2} \left(X_{1i} - \frac{\mathbf{X}_1' \mathbf{A}[,i] + w * \mathbf{A}[i,] \mathbf{X}_1}{m_i} \right)^2} \frac{\sqrt{m_i}}{\sqrt{2\pi}\sigma_{C0}} e^{-\frac{m_i}{2\sigma_{C0}^2} \left(X_{0i} - \frac{\mathbf{X}_0' \mathbf{A}[,i] + w * \mathbf{A}[i,] \mathbf{X}_0}{m_i} \right)^2} \\
&\quad \prod_{i=1}^{p} \frac{\sqrt{m_i}}{\sqrt{2\pi}\sigma_{C2}} e^{-\frac{m_i}{2\sigma_{C2}^2} \left(X_{2i} - \frac{\mathbf{X}_2' \mathbf{A}[,i] + w * \mathbf{A}[i,] \mathbf{X}_2}{m_i} \right)^2} \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu_{1i}^2}{2\sigma^2}} I(a,0) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu_{2i}^2}{2\sigma^2}} I(0,b) \\
&\quad \prod_{i=1}^{p} \frac{1}{\Gamma(\alpha)\beta^\alpha} \left(\sigma_{1i}^2\right)^{-\alpha-1} e^{\frac{1}{\beta\sigma_{1i}^2}} \frac{1}{\Gamma(\alpha)\beta^\alpha} \left(\sigma_{0i}^2\right)^{-\alpha-1} e^{\frac{1}{\beta\sigma_{0i}^2}} \frac{1}{\Gamma(\alpha)\beta^\alpha} \left(\sigma_{2i}^2\right)^{-\alpha-1} e^{\frac{1}{\beta\sigma_{2i}^2}} \\
&\quad \prod_{i=1}^{p} \left(\frac{1}{\Gamma(\alpha_C)\beta_C^{\alpha_C}} \left(\sigma_{C1}^2\right)^{-\alpha_C-1} e^{\frac{1}{\beta_C \sigma_{C1}^2}} \right) \left(\frac{1}{\Gamma(\alpha_C)\beta_C^{\alpha_C}} \left(\sigma_{C0}^2\right)^{-\alpha_C-1} e^{\frac{1}{\beta_C \sigma_{C0}^2}} \right) \\
&\quad \prod_{i=1}^{p} \left(\frac{1}{\Gamma(\alpha_C)\beta_C^{\alpha_C}} \left(\sigma_{C2}^2\right)^{-\alpha_C-1} e^{\frac{1}{\beta_C \sigma_{C2}^2}} \right)
\end{aligned}
$$

In order to use Gibbs sampler to draw posterior samples for parameters of interest, the fully conditional distributions for each set of parameters are derived from the joint distribution of data and parameters given above. The fully conditional posterior for μ, σ^2 are given as in the following equations:

For $j = 1, 2$; $c = a, d = 0$ if $j = 1$, and $c = 0, d = b$ if $j = 2$

$$
\begin{gathered}
\mu_{ji} | Z_i, \sigma_{ji}^2, \sigma^2, L_{ji} = 0 \sim N(0, \sigma_{ji}^2) I(c,d) \\
\mu_{ji} | Z_i, \sigma_{ji}^2, \sigma^2, L_{ji} = 1 \sim N(\tfrac{\sigma^2 Z_i}{\sigma^2 + \sigma_{ji}^2}, \tfrac{\sigma^2 \sigma_{ji}^2}{\sigma^2 + \sigma_{ji}^2}) I(c,d)
\end{gathered}
$$

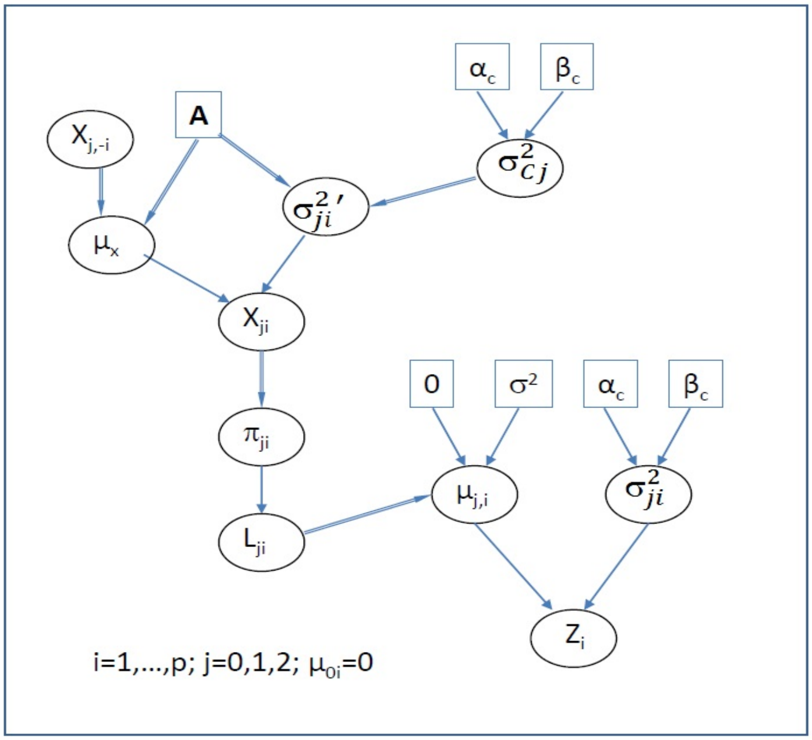


Fig. 3.2: Graphic representation of the DG Markov random field model with Z score.

The nodes in boxes are constants. The nodes in ellipses are stochastic nodes with distributions or deterministic nodes with logical function of other nodes. A solid arrow indicates a stochastic dependence while a hollow arrow indicates a logical function.

For $j = 0, 1, 2$

$$\sigma^2_{ji}|Z_i, \mu_{ji}, L_{ji} = 0 \sim IG(\alpha, \beta)$$

$$\sigma^2_{ji}|Z_i, \mu_{ji}, L_{ji} = 1 \sim IG(\alpha + \tfrac{p}{2}, \tfrac{2\beta}{2+\beta(Z_i-\mu_{ji})^2})$$

$$\sigma^2_{Cj}|\mathbf{X}_j, \mathbf{A}, \alpha_{Cj}, \beta_{Cj} \sim IG\left(\alpha_{Cj} + \tfrac{p}{2}, \frac{2\beta_{Cj}}{2+\sum_{i=1}^n m_i(X_{ji}-\frac{\mathbf{X}_j'\mathbf{A}[,i]+w\times\mathbf{A}[i,]\mathbf{X}_j}{m_i})^2}\right)$$

$$X_{ji}|\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \mathbf{A}, \sigma^2_{Cj}, L_{li} = 1 \propto \pi_{ji} \times \sqrt{\frac{m_i}{\sigma^2_{Cj}}} \times e^{-\frac{m_i}{2\sigma^2_{Cj}}(X_{ji}-\frac{\mathbf{X}_j'\mathbf{A}[,i]+w\times\mathbf{A}[i,]\mathbf{X}_j}{m_i})^2}$$

where, π_{ji} are as defined in equations (3.4).

All the fully conditional distributions, except for X_{ji}, are standard conjugate distributions from which posterior samples could be easily drawn. The following is a sampling scheme to draw a posterior sample for X_{ji}. (1) Obtain $X_{ji} = x_{ji}$ from the normal distribution $N(\frac{\mathbf{X}_j'\mathbf{A}[,i]+w\times\mathbf{A}[i,]\mathbf{X}_j}{m_i}, \frac{\sigma^2_{Cj}}{m_i})$; (2) Generate an independent $uniform(0, 1)$ random variable, ν; (3) Accept x_{ji} if $\nu \leq \frac{1}{1+ce^{-x_{ji}}}$, where $c = e^{x_{1i}} + e^{x_{2i}}$ if $j = 0$, $c = e^{x_{0i}} + e^{x_{2i}}$ if $j = 1$ $c = e^{x_{0i}} + e^{x_{1i}}$ if $j = 2$.

3.2.2 Two-Sample Model

In modeling the Z-scores, the raw expression data are not directly modeled and the prior knowledge was shown to dominate the posterior result in a simulation study (see next section). When the data of original expression values are available, the analysis could be formulated to incorporate both prior knowledge of gene networks and the expression data. A two-sample test with assumption of normality can be derived as follows: In an experiment with two treatment groups or a factor with two levels, let n_k be the number of subjects in group $k, k = 1, 2$ and let g_{1ih} and g_{2il} be the values of i^{th} genomic feature $(i = 1, ..., p)$ measured for h^{th} and l^{th} subjects respectively, where $h = 1, ..., n_1$ and $l = 1, ..., n_2$. Let **G** represent all the values of expression data. Suppose that in 1^{st}

treatment group, each expression value is normally distributed with mean μ_{0i} and variance σ_{0i}^2 (equation 3.6), and in 2^{nd} treatment group, each expression value follows a mixture distribution of three components: same expression level as group 1, lower expression and higher expression level compared to group 1, as in equation 3.7. Similarly, given a gene network represented by an adjacency matrix $\boldsymbol{A}$, the relationship between $\Pi_j, j = 0, 1, 2,$ the prior probability of group 2 coming from the j^{th} component, with latent Gaussian Markov random field variables is as described in equation 3.4.

$$f(g_{1ih}) = f(g_{1ih}, \mu_{0i}, \sigma_{0i}^2) \tag{3.6}$$

$$f(g_{2il}) = \pi_{0i} f_0(g_{2il}, \mu_{0i}, \sigma_{0i}^2) + \pi_{1i} f_1(g_{2il}, \mu_{0i} + \mu_{1i}, \sigma_{1i}^2) + \pi_{2i} f_2(g_{2il}, \mu_{0i} + \mu_{2i}, \sigma_{2i}^2) \tag{3.7}$$

As before, assume vague prior distributions for μ's: Specifically, for i^{th} gene, $\mu_{0i} = 0$; $\mu_{1i} \sim N(0, \sigma^2) I(a, 0)$, where $I(,)$ is an indicator function and $a = -max(\mathbf{G})$, a normal distribution between a and 0; and $\mu_{2i} \sim N(0, \sigma^2) I(0, b)$, where $b = max(\mathbf{G})$. σ^2, σ_{Cj}^2 for $j = 0, 1, 2$, α, β, α_C, and β_C, are same as in the model of Z-scores. The overall structure of the two-sample model is as in figure 3.3.

The joint likelihood function of data and parameters, given gene network described

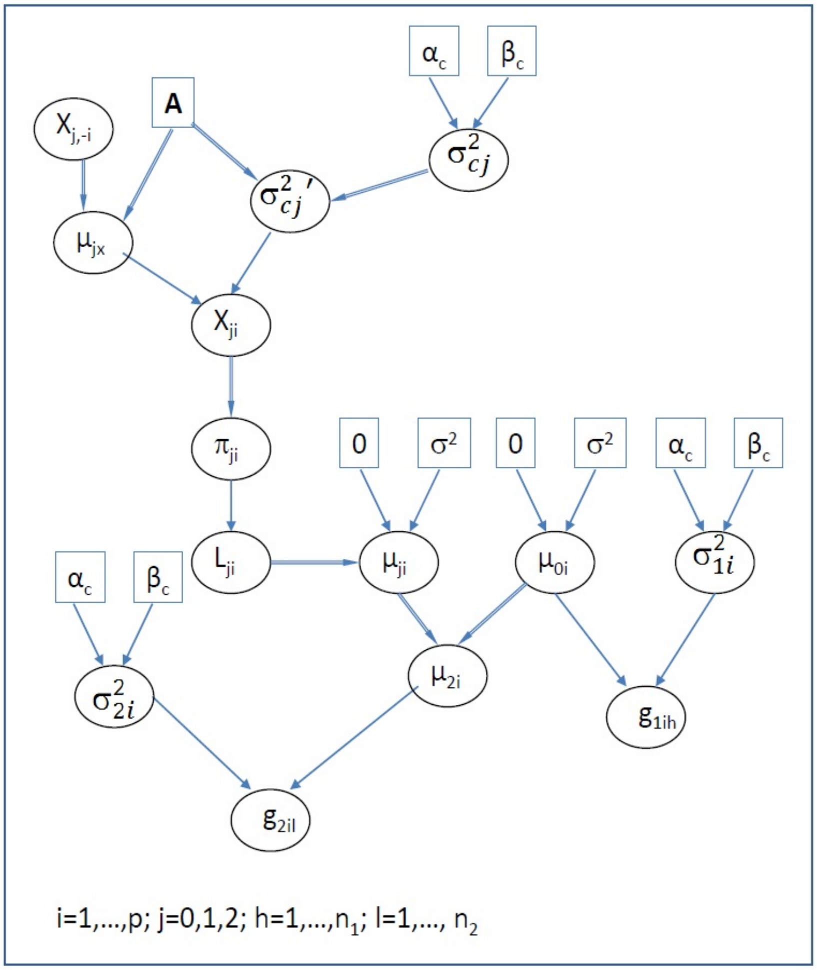


Fig. 3.3: Graphic representation of the DG Markov random field model for two-sample problem.

The nodes and links are interpreted as in figure 3.2.

in adjacency matrix $\boldsymbol{A}$, parameters and hyper-parameters, is given by

$$
\begin{aligned}
&f(\mathbf{G}, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\sigma}_0^2, \boldsymbol{\sigma}_1^2, \boldsymbol{\sigma}_2^2, \mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, L_0, \mathbf{L}_1, \mathbf{L}_2, \sigma_{c0}^2, \sigma_{c1}^2, \sigma_{c2}^2 \mid \mathbf{A}, w, \sigma^2, \alpha, \beta, \alpha_c, \beta_c, a, b) \\
&= \prod_{i=1}^{p} \prod_{h=1}^{n_1} \frac{1}{\sqrt{2\pi}\times\sigma_{0i}} e^{-\frac{1}{2\sigma_{0i}^2}(g_{1ih}-\mu_{0i})^2} \times \prod_{i=1}^{p} \prod_{l=1}^{n_2} \left(\frac{1}{\sqrt{2\pi}\times\sigma_{0i}} e^{-\frac{1}{2\sigma_{0i}^2}(g_{2il}-\mu_{0i})^2} \right)^{L_{0i}} \\
&\quad \prod_{i=1}^{p} \prod_{l=1}^{n_2} \left(\frac{1}{\sqrt{2\pi}\times\sigma_{1i}} e^{-\frac{1}{2\sigma_{1i}^2}(g_{2il}-(\mu_{1i}+\mu_{0i}))^2} \right)^{L_{1i}} \left(\frac{1}{\sqrt{2\pi}\times\sigma_{2i}} e^{-\frac{1}{2\sigma_{2i}^2}(g_{2il}-(\mu_{2i}+\mu_{0i}))^2} \right)^{1-L_{1i}-L_{0i}} \\
&\quad \prod_{i=1}^{p} \binom{1}{L_{1i}, L_{0i}} \left(\frac{1}{1+e^{X_{2i}-X_{1i}} e^{X_{0i}-X_{1i}}} \right)^{L_{1i}} \left(\frac{1}{1+e^{X_{1i}-X_{0i}} e^{X_{2i}-X_{0i}}} \right)^{L_{0i}} \left(\frac{1}{1+e^{X_{1i}-X_{2i}} e^{X_{0i}-X_{2i}}} \right)^{1-L_{1i}-L_{0i}} \\
&\quad \prod_{i=1}^{p} \frac{\sqrt{m_i}}{\sqrt{2\pi}\sigma_{C1}} e^{-\frac{m_i}{2\sigma_{C1}^2}\left(X_{1i} - \frac{\mathbf{X}_1' \mathbf{A}[,i] + w * \mathbf{A}[i,] \mathbf{X}_1}{m_i} \right)^2} \frac{\sqrt{m_i}}{\sqrt{2\pi}\sigma_{C0}} e^{-\frac{m_i}{2\sigma_{C0}^2}\left(X_{0i} - \frac{\mathbf{X}_0' \mathbf{A}[,i] + w * \mathbf{A}[i,] \mathbf{X}_0}{m_i} \right)^2} \\
&\quad \prod_{i=1}^{p} \frac{\sqrt{m_i}}{\sqrt{2\pi}\sigma_{C2}} e^{-\frac{m_i}{2\sigma_{C2}^2}\left(X_{2i} - \frac{\mathbf{X}_2' \mathbf{A}[,i] + w * \mathbf{A}[i,] \mathbf{X}_2}{m_i} \right)^2} \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu_{1i}^2}{2\sigma^2}} I(a,0) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu_{2i}^2}{2\sigma^2}} I(0,b) \\
&\quad \prod_{i=1}^{p} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \left(\sigma_{1i}^2\right)^{-\alpha-1} e^{\frac{1}{\beta\sigma_{1i}^2}} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \left(\sigma_{0i}^2\right)^{-\alpha-1} e^{\frac{1}{\beta\sigma_{0i}^2}} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \left(\sigma_{2i}^2\right)^{-\alpha-1} e^{\frac{1}{\beta\sigma_{2i}^2}} \\
&\quad \prod_{i=1}^{p} \left(\frac{1}{\Gamma(\alpha_C)\beta_C^{\alpha_C}} \left(\sigma_{C1}^2\right)^{-\alpha_C-1} e^{\frac{1}{\beta_C \sigma_{C1}^2}} \right) \left(\frac{1}{\Gamma(\alpha_C)\beta_C^{\alpha_C}} \left(\sigma_{C0}^2\right)^{-\alpha_C-1} e^{\frac{1}{\beta_C \sigma_{C0}^2}} \right) \\
&\quad \prod_{i=1}^{p} \left(\frac{1}{\Gamma(\alpha_C)\beta_C^{\alpha_C}} \left(\sigma_{C2}^2\right)^{-\alpha_C-1} e^{\frac{1}{\beta_C \sigma_{C2}^2}} \right)
\end{aligned}
$$

The fully conditional distributions of unknown parameters (μ_{ji} and σ_{ji}) are given in following set of distributions:

$$
\begin{aligned}
\mu_{1i} | \sigma^2, L_{1i} = 0 \quad &\sim \quad N\left(0, \sigma^2\right) I(a,0) \\
\mu_{1i} | \mathbf{g}_{2i}, \mu_{0i}, \sigma_{1i}^2, \sigma^2, L_{1i} = 1 \quad &\sim \quad N\left(\frac{\sigma^2 \sum_{l=1}^{n_2}(g_{2il}-\mu_{0i})}{n_2\sigma^2+\sigma_{1i}^2}, \frac{\sigma^2\sigma_{1i}^2}{n_2\sigma^2+\sigma_{1i}^2} \right) I(a,0) \\
\mu_{0i} | \mathbf{g}_{1i}, \mathbf{g}_{2i}, \mu_{1i}, \sigma^2, \sigma_{0i}^2, \sigma_{1i}^2, L_{1i} = 1 \quad &\sim \quad N\left(\frac{\sigma^2\sigma_{1i}^2 \sum_{h=1}^{n_1} g_{1ih} + \sigma^2\sigma_{0i}^2 \sum_{l=1}^{n_2}(g_{2il}-\mu_{1i})}{n_1\sigma^2\sigma_{1i}^2 + n_2\sigma^2\sigma_{0i}^2 + \sigma_{1i}^2\sigma_{0i}^2}, \frac{\sigma^2\sigma_{1i}^2\sigma_{0i}^2}{n_1\sigma^2\sigma_{1i}^2 + n_2\sigma^2\sigma_{0i}^2 + \sigma_{1i}^2\sigma_{0i}^2} \right) \\
\mu_{0i} | \mathbf{g}_{1i}, \mathbf{g}_{2i}, \sigma^2, L_{0i} = 1 \quad &\sim \quad N\left(\frac{\sigma^2\left(\sum_{h=1}^{n_1} g_{1ih} + \sum_{l=1}^{n_2} g_{2il}\right)}{(n_1+n_2)\sigma^2+\sigma_{0i}^2}, \frac{\sigma^2\sigma_{0i}^2}{(n_1+n_2)\sigma^2+\sigma_{0i}^2} \right) \\
\mu_{0i} | \mathbf{g}_{1i}, \mathbf{g}_{2i}, \mu_{2i}, \sigma^2, \sigma_{0i}^2, \sigma_{2i}^2, L_{2i} = 1 \quad &\sim \quad N\left(\frac{\sigma^2\sigma_{2i}^2 \sum_{h=1}^{n_1} g_{1ih} + \sigma^2\sigma_{0i}^2 \sum_{l=1}^{n_2}(g_{2il}-\mu_{2i})}{n_1\sigma^2\sigma_{2i}^2 + n_2\sigma^2\sigma_{0i}^2 + \sigma_{2i}^2\sigma_{0i}^2}, \frac{\sigma^2\sigma_{2i}^2\sigma_{0i}^2}{n_1\sigma^2\sigma_{2i}^2 + n_2\sigma^2\sigma_{0i}^2 + \sigma_{2i}^2\sigma_{0i}^2} \right) \\
\mu_{2i} | \sigma^2, L_{2i} = 0 \quad &\sim \quad N\left(0, \sigma^2\right) I(0,b) \\
\mu_{2i} | \mathbf{g}_{2i}, \mu_{0i}, \sigma_{2i}^2, \sigma^2, L_{2i} = 1 \quad &\sim \quad N\left(\frac{\sigma^2 \sum_{l=1}^{n_2}(g_{2il}-\mu_{0i})}{n_2\sigma^2+\sigma_{2i}^2}, \frac{\sigma^2\sigma_{2i}^2}{n_2\sigma^2+\sigma_{2i}^2} \right) I(0,b)
\end{aligned}
$$

$$\begin{array}{lcl}
\sigma_{1i}|L_{1i}=0 & \sim & IG\left(\alpha,\beta\right) \\
\sigma_{1i}|\mathbf{g}_{i2},\mu_{0i},\mu_{1i},L_{1i}=1 & \sim & IG\left(\alpha+\frac{n_1}{2},\frac{2\beta}{2+\beta\sum_{h=1}^{n_1}(g_{1ih}-\mu_{0i})^2}\right) \\
\sigma_{0i}|\mathbf{g}_{i1},\mathbf{g}_{i2},\mu_{0i},L_{0i}=0 & \sim & IG\left(\alpha+\frac{n_1}{2},\frac{2\beta}{2+\beta\sum_{l=1}^{n_2}(g_{2il}-(\mu_{1i}+\mu_{0i}))^2}\right) \\
\sigma_{0i}|\mathbf{g}_{i1},\mathbf{g}_{i2},\mu_{0i},L_{0i}=1 & \sim & IG\left(\alpha+\frac{n_1+n_2}{2},\frac{2\beta}{2+\beta\left(\sum_{h=1}^{n_1}(g_{1ih}-\mu_{0i})^2+\sum_{l=1}^{n_2}(g_{2il}-\mu_{0i})^2\right)}\right) \\
\sigma_{2i}|L_{2i}=0 & \sim & IG\left(\alpha,\beta\right) \\
\sigma_{2i}|\mathbf{g}_{i2},\mu_{0i},\mu_{2i},L_{2i}=1 & \sim & IG\left(\alpha+\frac{n_2}{2},\frac{2\beta}{2+\beta\sum_{l=1}^{n_2}(g_{2il}-(\mu_{0i}+\mu_{2i}))^2}\right)
\end{array}$$

The fully conditional distributions of L_{ji}, X_{ji}, and σ^2_{Cj} for $j=0,1,2$ are same as those in the model of Z-scores. Similarly, posterior samples could be easily drawn from these standard fully conditional distributions and X_{ji} are drawn using same acception-rejection algorithm as above.

The weight, w, was arbitrarily set to 0.5 in above two models with the assumption that children of a gene carry less information on its expression relative to its parents. The effect of w with various choices was further studied in the simulation studies.

3.2.3 Directed Graphs for prior distribution in the two-sample problem

In gene expression data analysis, it could be assumed that the expression of a gene directly depends on the expression levels of their neighbors. Suppose that g_{1ih} and g_{2il} are the values of i^{th} genomic feature ($i=1,...,p$) measured for h^{th} and l^{th} subjects respectively, where $h=1,...,n_1$ and $l=1,...,n_2$. Let **G** represent all the values of the genomic features. Suppose that in 1^{st} treatment group, each genomic feature is normally distributed with mean μ_{0i} and variance σ^2_{1i}, and in 2^{nd} treatment group, each genomic feature follows normal distribution with $\mu_{0i}+\mu_{di}$ and variance σ^2_{2i}, where μ_{di} is the mean difference between group 1 and group 2. The distribution of each element of **G** is given in following equation:

$$f(g_{1ih})=f_1(g_{1ih},\mu_{0i},\sigma^2_{0i});f(g_{2il})=f_2(g_{2il},\mu_{0i}+\mu_{di},\sigma^2_{2i}) \tag{3.8}$$

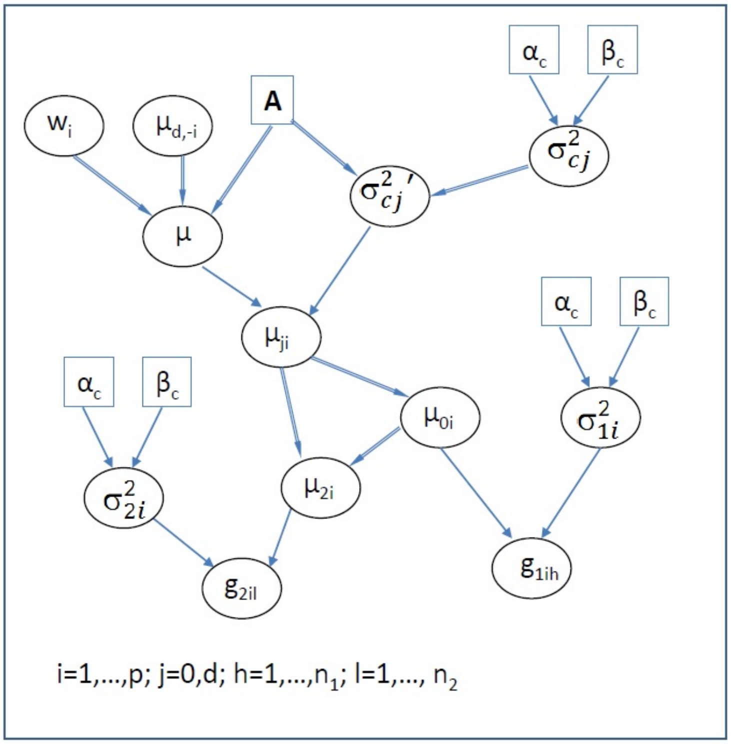


Fig. 3.4: Graphic representation of the directed graphs for prior distribution in the two-sample problem.

The nodes and links are interpreted as in figure 3.2.

Prior Distributions: Given a gene network with adjacency matrix $\boldsymbol{A}$, the prior distribution of $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_d$ are given as following: The distribution of each $\mu_{(j,i)}$, conditional on $\mu_{(j,-i)} = \{\mu_{(j,k)}, k \neq i\}$, depends only on it first-order neighbors on an adjacency matrix $\boldsymbol{A}$ with m_i defined in equation 3.3.

$$\mu_{(j,i)}|\mu_{(j,-i)} \sim N\left(\frac{1}{m_i}(w_i * \mathbf{A}[i,]\boldsymbol{\mu}_j + \boldsymbol{\mu}_j'\mathbf{A}[,i]), \frac{\sigma^2_{C_j}}{m_i}\right); j = 0, d \tag{3.9}$$

Non-informative prior distributions are used for other parameters as above. We further assume that $w_i \sim uniform(0,1)$. The overall structure of the model is as in figure 3.4.

From the distribution of data and prior distributions of parameters, we can derive the joint distribution of data and parameters. Given gene network with adjacency matrix $\boldsymbol{A}$ and hyperparameters of the priors, the fully joint distribution is given by

$$
\begin{aligned}
&f(\mathbf{G},\mathbf{w},\boldsymbol{\mu}_0,\boldsymbol{\mu}_d,\boldsymbol{\sigma}_1^2,\boldsymbol{\sigma}_2^2,\boldsymbol{\sigma}_{C0}^2,\boldsymbol{\sigma}_{Cd}^2 \mid \mathbf{A},n_1,n_2,\alpha,\beta,\alpha_c,\beta_c) \\
&= \prod_{i=1}^{p}\prod_{h=1}^{n_1} \tfrac{1}{\sqrt{2\pi}\times\sigma_{1i}} e^{-\frac{1}{2\sigma_{1i}^2}(g_{1ih}-\mu_{0i})^2} \times \prod_{i=1}^{p}\prod_{l=1}^{n_2} \tfrac{1}{\sqrt{2\pi}\times\sigma_{2i}} e^{-\frac{1}{2\sigma_{2i}^2}(g_{2il}-(\mu_{0i}+\mu_{di}))^2} \\
&\quad \prod_{i=1}^{p} \tfrac{\sqrt{\vec{1}*\mathbf{A}[,i]+w_i*\mathbf{A}[i,]*\vec{1}}}{\sqrt{2\pi}\sigma_{C0}} e^{-\frac{\vec{1}*\mathbf{A}[,i]+w_i*\mathbf{A}[i,]*\vec{1}}{2\sigma_{C0}^2}\left(\mu_{0i}-\frac{\boldsymbol{\mu}_0'\mathbf{A}[,i]+w_i*\mathbf{A}[i,]\boldsymbol{\mu}_0}{\vec{1}*\mathbf{A}[,i]+w_i*\mathbf{A}[i,]*\vec{1}}\right)^2} \\
&\quad \prod_{i=1}^{p} \tfrac{\sqrt{\vec{1}*\mathbf{A}[,i]+w_i*\mathbf{A}[i,]*\vec{1}}}{\sqrt{2\pi}\sigma_{Cd}} e^{-\frac{\vec{1}*\mathbf{A}[,i]+w_i*\mathbf{A}[i,]*\vec{1}}{2\sigma_{Cd}^2}\left(\mu_{di}-\frac{\boldsymbol{\mu}_d'\mathbf{A}[,i]+w_i*\mathbf{A}[i,]\boldsymbol{\mu}_d}{\vec{1}*\mathbf{A}[,i]+w_i*\mathbf{A}[i,]*\vec{1}}\right)^2} \\
&\quad \prod_{i=1}^{p} \tfrac{1}{\Gamma(\alpha)\beta^{\alpha}} \left(\sigma_{1i}^2\right)^{-\alpha-1} e^{\frac{1}{\beta\sigma_{1i}^2}} \tfrac{1}{\Gamma(\alpha)\beta^{\alpha}} \left(\sigma_{2i}^2\right)^{-\alpha-1} e^{\frac{1}{\beta\sigma_{2i}^2}} \\
&\quad \prod_{i=1}^{p} \tfrac{1}{\Gamma(\alpha_C)\beta_C^{\alpha_C}} \left(\sigma_{C0}^2\right)^{-\alpha_C-1} e^{\frac{1}{\beta_C\sigma_{C0}^2}} \tfrac{1}{\Gamma(\alpha_C)\beta_C^{\alpha_C}} \left(\sigma_{Cd}^2\right)^{-\alpha_C-1} e^{\frac{1}{\beta_C\sigma_{Cd}^2}}
\end{aligned}
$$

In order to use Gibbs sampler to draw posterior samples for interesting parameters, the fully conditional distributions for parameters are given in following set of distributions:

$$
\begin{aligned}
&\mu_{0i}|g_{i1},g_{i2},\boldsymbol{\mu}_0,\boldsymbol{\mu}_d,\sigma_{1i}^2,\sigma_{2i}^2,\sigma_{C0}^2,\mathbf{A},w_i \sim \\
&\qquad N(\tfrac{\sigma_{C0}^2\sigma_{2i}^2\sum_{j=1}^{n_1} g_{1ij}+\sigma_{C0}^2\sigma_{1i}^2\sum_{l=1}^{n_2}(g_{2il}-\mu_{di})+\sigma_{1i}^2\sigma_{2i}^2\left(\boldsymbol{\mu}_0\mathbf{A}[,i]+w_i\mathbf{A}[i,]\boldsymbol{\mu}_0\right)}{n_1\sigma_{C0}^2\sigma_{2i}^2+n_2\sigma_{C0}^2\sigma_{1i}^2+\sigma_{1i}^2\sigma_{2i}^2\left(\boldsymbol{\mu}_0\mathbf{A}[,i]+w_i\mathbf{A}[i,]\boldsymbol{\mu}_0\right)}, \\
&\qquad\qquad \tfrac{\sigma_{C0}^2\sigma_{1i}^2\sigma_{2i}^2}{n_1\sigma_{C0}^2\sigma_{2i}^2+n_2\sigma_{C0}^2\sigma_{1i}^2+\sigma_{1i}^2\sigma_{2i}^2\left(\boldsymbol{\mu}_0\mathbf{A}[,i]+w_i\mathbf{A}[i,]\boldsymbol{\mu}_0\right)}) \\
&\mu_{di}|g_{i2},\mu_0,\mu_d,\sigma_{1i}^2,\sigma_{2i}^2,\sigma_{Cd}^2,A,w_i \sim \\
&\qquad N\left(\tfrac{\sigma_{C0}^2\sum_{l=1}^{n_2}(g_{2il}-\mu_{0i})+\sigma_{2i}^2(\mu_d\mathbf{A}[,i]+w_i\mathbf{A}[i,]\mu_d)}{n_2\sigma_{C0}^2+\sigma_{2i}^2(\mu_d\mathbf{A}[,i]+w_i\mathbf{A}[i,]\mu_d)}, \tfrac{\sigma_{C0}^2\sigma_{2i}^2}{n_2\sigma_{Cd}^2+\sigma_{2i}^2(\mu_d\mathbf{A}[,i]+w_i\mathbf{A}[i,]\mu_d)}\right)
\end{aligned}
$$

$$
\begin{aligned}
&w_i|\mathbf{w}, \mathbf{A}, \boldsymbol{\mu}_0, \boldsymbol{\mu}_d, \sigma^2_{C0}, \sigma^2_{Cd} \sim \\
&\quad N(\tfrac{\sigma^2_{Cd}\mu_{0i}\mathbf{A}[i,]\left(\vec{m}[i]\mu_{0i}-\mathbf{A}[,i]\boldsymbol{\mu}_0\right)+\sigma^2_{C0}\mu_{di}\mathbf{A}[i,]\left(\vec{m}[i]\mu_{di}-\mathbf{A}[,i]\boldsymbol{\mu}_d\right)}{\sigma^2_{Cd}\left(\boldsymbol{\mu}_0\mathbf{A}[i,]\right)^2+\sigma^2_{C0}\left(\boldsymbol{\mu}_d\mathbf{A}[i,]\right)^2}, \\
&\qquad\qquad \tfrac{\vec{m}[i]\sigma^2_{C0}\sigma^2_{Cd}}{\sigma^2_{Cd}\left(\boldsymbol{\mu}_0\mathbf{A}[i,]\right)^2+\sigma^2_{C0}\left(\boldsymbol{\mu}_d\mathbf{A}[i,]\right)^2})I(0,1) \\
&\sigma_{C0}|\boldsymbol{\mu}_j, \mathbf{A}, \mathbf{w}, \alpha_C, \beta_C \sim \\
&\quad IG\left(\alpha_C + \tfrac{p}{2}, \tfrac{2\beta_C}{2+\beta_C\sum_{i=1}^{m}\left(\frac{\left(\vec{m}[i]*\mu_{ji}-\left(\mu_{ji}\mathbf{A}[,i]+\vec{m}[i]*\mathbf{A}[i,]\mu_{ji}\right)\right)^2}{\vec{m}[i]}\right)}\right) \\
&\quad j = 0, d
\end{aligned}
$$

where $\vec{m} = colsum(A) + w * rowsum(A)$

$$
\begin{aligned}
\sigma^2_{0i}|g_{1i}, \mu_{0i}, n_1, \alpha, \beta &\sim IG\left(\alpha + \tfrac{n_1}{2}, \tfrac{2\beta}{2+\beta\sum_{j=1}^{n_1}(g_{ij}-\mu_{0i})^2}\right) \\
\sigma^2_{2i}|g_{2i}, \mu_{0i}, \mu_{di}, n_2, \alpha, \beta &\sim IG\left(\alpha + \tfrac{n_2}{2}, \tfrac{2\beta}{2+\beta\sum_{l=1}^{n_1}(g_{il}-(\mu_{0i}+\mu_{di}))^2}\right)
\end{aligned}
$$

3.3 Simulation Studies

To investigate the behavior of the three models proposed in above section, simulation studies were performed. Suppose that there is a well defined pathway as illustrated in figure 3.5. In the directed graph, there are 26 nodes and 27 edges with 2 disjoint subgraphs. Each node has minimum of 1 edge and at most 4 edges with average around 2 edges. There is no singleton in this graph and the graph is very sparse.

3.3.1 Simulation on Z-scores

First, consider the model of Z-scores proposed in above section. The Z-scores were randomly generated through following mechanism based on the DG in figure 3.5: (1) Start with constant $\boldsymbol{\mu}$, each component of $\boldsymbol{\mu}_1$ was draw from normal distribution with with mean being weighted average of neighbors and weighted standard error in range $(-max(\mathbf{Z}), 0)$, and each component of $\boldsymbol{\mu}_2$ was draw from normal distribution with mean being weighted

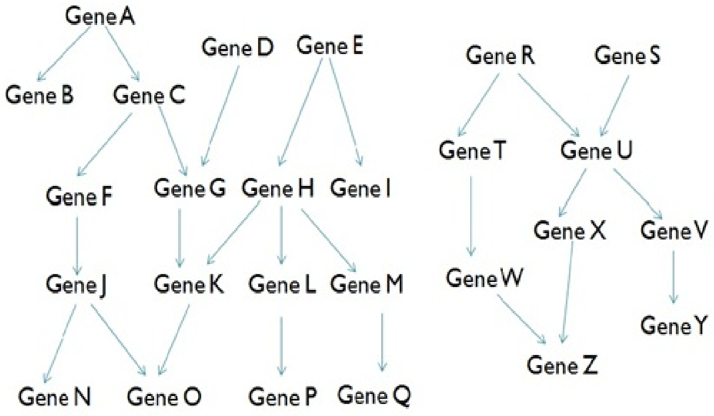


Fig. 3.5: A directed graph represents directional relationship of 26 putative genes in two disjoint subgraphs

average of neighbors and weighted standard error in range $(0, max(\mathbf{Z}))$; the process was repeated 20 times to introduce correlation among the genes according to the DG; (2) Start with constant $\mathbf{X}$, each component of $\mathbf{X}_0, \mathbf{X}_1$ and $\mathbf{X}_2$ was drawn from normal distribution with mean being weighted average of $\mathbf{X}$ of neighbors and weighed standard error. The process was repeated 50 times; (3) Π were calculated according to equation 3.4; (4) A multinomial sample $\mathbf{ll}_i$ was drawn with parameter $\mathbf{p}_i$ equal to i^{th} component of Π in step 3; (5) For i^{th} gene, μ_i was set to be μ_{1i} if $ll_{1i} = 1$, 0 if $ll_{0i} = 1$, or μ_{2i} if $ll_{2i} = 1$; (6) Z_i was drawn from a normal distribution with mean μ_i and a fixed variance.

Using hyperparameters as described above, and randomly generating initial values, chain was run for 4000 burn-in. 4000 posterior samples were drawn from the posterior distributions with thin equal to 3 (keep one out of three samples). Figure 3.6 shows the trace plots of posterior samples for selected nodes. The first 3 nodes are from the first disjoint subgraph in figure 3.5, last 3 nodes are from the second disjoint subgraph. The posterior samples of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ (first two rows in figure 3.6) varied in the truncated parameter space. The trace plots of posterior Π_0, Π_1 and Π_2 were similar among genes in

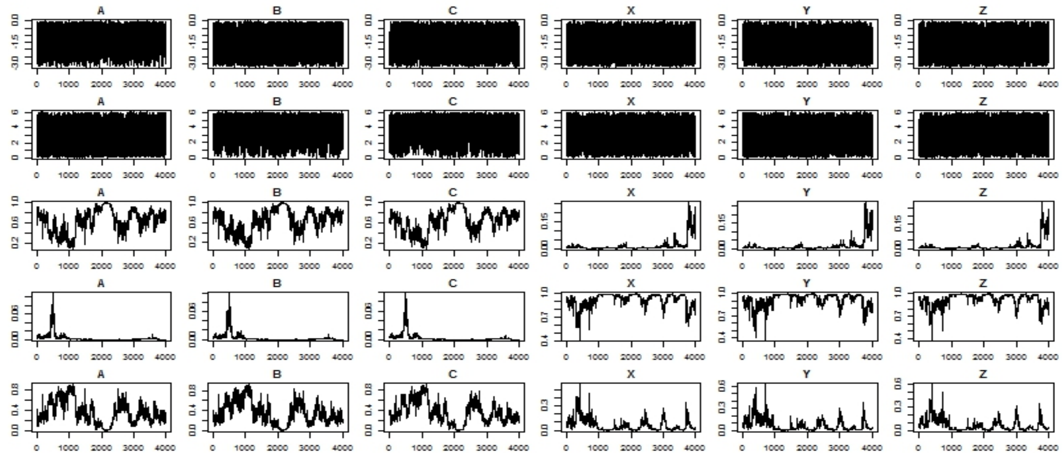


Fig. 3.6: Trace plot of posterior samples of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Pi_1, \Pi_0$ and Π_2 in Model of Z-scores. Posterior samples of first 3 nodes (the left 3 columns) and last 3 nodes (the 3 columns on right) of the DG as shown in figure 3.5. Row 1 to 5 are trace plots for $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Pi_1, \Pi_0$, and Π_2, respectively

same disjoint subgraph, indicating that the fraction of Z_i in a graph or subgraph came from negative (Π_1) or positive component (Π_3) together, but not in individual gene level.

We further summarized the posterior samples of $\boldsymbol{\mu}_j$ for $j = 1, 2$. For i^{th} gene, the posterior samples were chosen to be from negative component if π_{1i} was biggest, positive component if π_{2i} was the biggest, or 0 component otherwise. The standard error of posterior sample mean was computed according to Albert and Chib (1993)[38]. The posterior samples were batched to equal size of 100 and batch means were computed. The standard error of posterior mean was calculated as the standard deviation of the batch means divided by the square root of number of batches. Table 3.1 showed the posterior summarization of the 26 genes. For each gene, the simulated z-scores, the batched mean of posterior, standard error of batched mean and range of posterior were shown.

For gene R to Z, which compose an isolated subgraph in figure 3.5, all came from 0 component with $\pi_{0s} = 0.78$, even gene Q to V had big Z-scores among the 26 genes. For gene A to Q, which compose the first disjoint subgraph, all the 17 genes came from positive component with $\pi_{2s} = 0.79$. For gene L, the original Z-score is -3.15, a negative

Table 3.1: Summary of posterior samples in the simulation study on Z-scores

Gene	Zscore	Mean	StdErr	Min	Max
A	-2.145	2.4257	1.6464	2e-04	5.6451
B	5.197	4.1722	1.3814	0	5.6462
C	-1.782	2.3145	1.6499	0.0012	5.6435
D	4.661	3.9317	1.3037	0.0028	5.6458
E	3.865	3.4896	1.2112	0.0039	5.6416
F	-1.571	2.3084	1.6241	0.0013	5.6452
G	5.314	4.2233	1.3786	0.0047	5.6465
H	5.647	4.132	1.4637	5e-04	5.6472
I	-1.811	2.2674	1.6375	9e-04	5.6446
J	-2.698	2.4027	1.6372	4e-04	5.6468
K	4.271	3.6804	1.2916	6e-04	5.6466
L	-3.15	2.4382	1.6156	5e-04	5.645
M	4.41	3.7747	1.2919	0.0034	5.646
N	0.346	1.8728	1.5761	1e-04	5.6439
O	0.34	1.8952	1.6089	2e-04	5.6445
P	5.573	4.1398	1.477	5e-04	5.647
Q	-0.025	1.9653	1.663	0.0023	5.6433
R	5.491	0	0	0	0
S	-0.056	0	0	0	0
T	-0.222	0	0	0	0
U	-0.073	0	0	0	0
V	5.151	0	0	0	0
W	-0.666	0	0	0	0
X	0.386	0	0	0	0
Y	0.861	0	0	0	0
Z	0.039	0	0	0	0

Zscore is the original simulated Z-scores. For each gene, summary statistics of posterior samples are provided. For gene R to Z, the posterior Π_0 is near 0.78, indicating the Z_i for these nine genes came from null component and summary statistics were all set to 0. StdErr is the standard error of posterior mean component, the posterior mean is 2.44, a positive component. In general, the posterior means of genes in a disjoint subgraph are regressed to the mean in the disjoint subgraph. The prior knowledge of the DG describing the relationship among genes dominates the posterior results as the original expression values are not modeled in the model.

3.3.2 Simulation on Two-sample model

To study the two sample model proposed as in above section, we simulated expression values according to the following mechanism based on the DG in figure 3.5 with w set to be 0.5: (1) Start with constant $\boldsymbol{\mu}$, each component of $\boldsymbol{\mu}_0$ or $\boldsymbol{\mu}_2$ was drawn from positively truncated normal distribution with mean being weighted average of neighbors and weighed standard error, and similarly, $\boldsymbol{\mu}_1$ was drawn from negatively truncated normal distribution; the process was repeated 20 times to introduce correlation among the genes according to figure 3.5; (2) For the first 17 genes corresponding to the first disjoint subgraph, group 2 mean, μ_{g2i} is set to $\mu_{0i} + \mu_{2i}$ for $i = 1, \ldots, 17$ (differentially expressed); and for the rest 9 genes corresponding to the second disjoint subgraph, group 2 mean μ_{g2i} is set to μ_{0i} for $i = 18, \ldots, 26$; and (3) For each gene, 10 random samples was drawn from a normal distribution with mean μ_{0i} for group 1 and μ_{g2i} for group 2 with fixed variances.

Under hyperparameters as in above section and randomly generating initial values, the chain was run 4000 for burn-in. 4000 posterior samples were drawn from posterior distributions with thin equal to 3 (keep one out of three samples). Figure 3.7 showed the trace plots of posterior samples for 6 selected nodes: the first 3 nodes, from first disjoint subgraph and last 3 nodes, from second disjoint subgraph. For the 6 nodes, the chain converged well for $\boldsymbol{\mu}_0$ (1^{st} row) and $\boldsymbol{\sigma}_0^2$ (7^{th} row), indicating that the overall mean of the two groups could be stably estimated. For gene A, B and C, posterior distribution of $\boldsymbol{\mu}_2$ and $\boldsymbol{\sigma}_2^2$ converged, while those of $\boldsymbol{\mu}_1$ and $\boldsymbol{\sigma}_1^2$ were diffused, suggesting that gene A to C of group 2 came from a positive component relative to group 1. From the trace plots, the convergence of posterior for each component in $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\sigma}_0^2$ $\boldsymbol{\sigma}_1^2$ and $\boldsymbol{\sigma}_2^2$ was consistent with

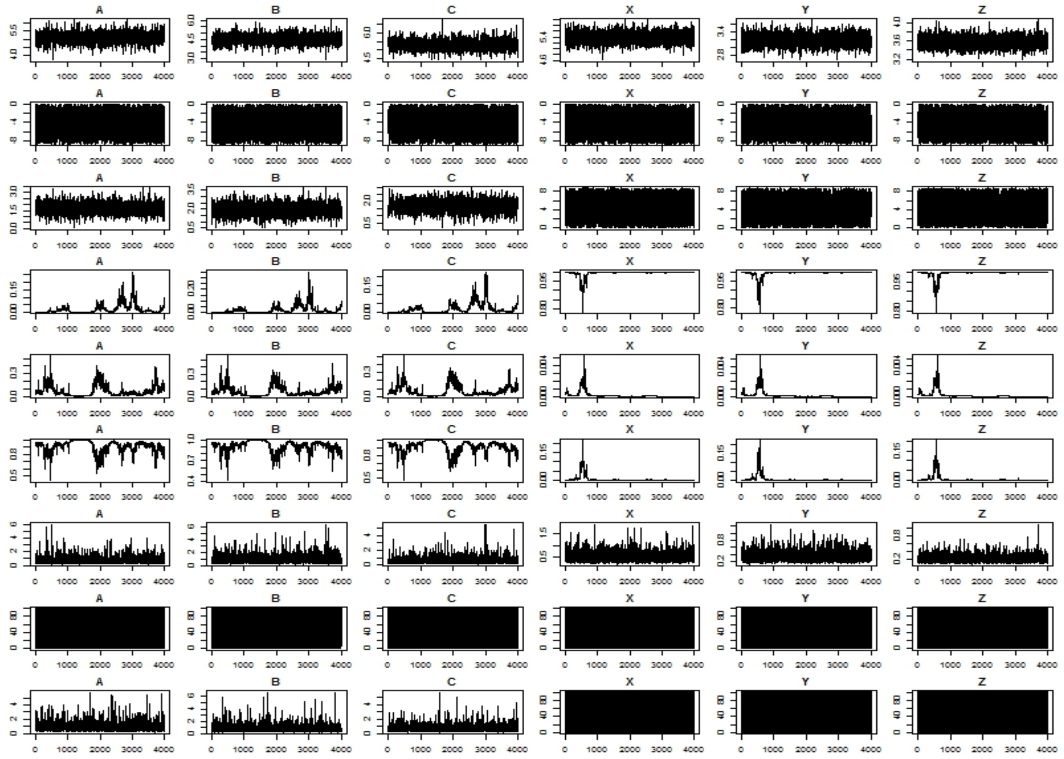


Fig. 3.7: Trace plots of posterior samples of $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Pi_0, \Pi_1, \Pi_2, \boldsymbol{\sigma}_0^2, \boldsymbol{\sigma}_1^2$ and $\boldsymbol{\sigma}_2^2$ in the simulation of two-sample model.

Posterior samples of first 3 nodes and last 3 nodes of the DG in figure 3.5 are shown. Row 1 to 3 are trace plots for posterior samples of $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$; row 4 to 6 are trace plots for Π_0, Π_1 and Π_2; row 7 to 9 are for $\boldsymbol{\sigma}_0^2, \boldsymbol{\sigma}_1^2$ and $\boldsymbol{\sigma}_2^2$, respectively

corresponding component of the posterior distributions, such as the mean of posterior Π_2 is the biggest among the three Πs for gene A to C. In contrast, gene X, Y, and Z were all from null component (no difference between group 1 and 2) as posterior $\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2, \boldsymbol{\mu}_2$ and $\boldsymbol{\sigma}_2^2$ were all diffused for these three genes. In fact, all the first 17 genes were from positive component with posterior distribution of $\boldsymbol{\mu}_2$ and $\boldsymbol{\sigma}_2^2$ converged and rest 9 genes were from 0 component with $\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2, \boldsymbol{\mu}_2$ and $\boldsymbol{\sigma}_2^2$ in these 9 genes diffused (trace plots not shown). This is consistent with the simulation.

The posterior samples of $\boldsymbol{\mu}_j$ for $j = 0, 1, 2$ were further summarized through following mechanism in concordance to trace plots. For i^{th} gene, if $\pi_{ji}, j = 0, 1, 2$ is

Table 3.2: Posterior estimates for the 26 genes in the simulation of the two-sample model.

Gene	Mu1	Mu2	MuEstGrp1	MuStdGrp1	MuEstGrp2	MuStdGrp2	MuEstDiff	MuStdDiff
A	7.0	8.7	7.030	0.019	8.932	0.019	1.902	0.027
B	7.9	8.5	8.018	0.012	8.658	0.014	0.64	0.021
C	7.1	7.7	7.026	0.014	7.742	0.012	0.715	0.021
D	7.4	7.9	7.394	0.009	7.867	0.013	0.473	0.015
E	6.9	8.5	6.839	0.013	8.620	0.022	1.782	0.026
F	7.0	9.2	6.980	0.016	9.246	0.016	2.266	0.024
G	6.8	8.7	6.922	0.020	8.636	0.014	1.713	0.026
H	6.3	8.5	6.325	0.021	8.449	0.013	2.124	0.029
I	6.5	8.2	6.486	0.010	8.281	0.017	1.796	0.02
J	7.2	7.6	7.301	0.026	7.777	0.522	0.476	0.505
K	7.5	8.3	7.509	0.011	8.144	0.022	0.635	0.026
L	7.2	8.6	7.063	0.020	8.623	0.013	1.56	0.024
M	7.1	8.7	7.183	0.014	8.792	0.013	1.609	0.017
N	6.8	9.0	6.773	0.015	9.221	0.017	2.448	0.023
O	6.7	8.8	6.738	0.024	8.797	0.011	2.058	0.029
P	6.5	8.3	6.278	0.011	8.318	0.013	2.04	0.018
Q	6.3	8.8	6.119	0.014	8.840	0.014	2.721	0.024
R	6.3	6.3	6.212	0.007	6.212	0.007		
S	6.2	6.2	6.279	0.011	6.279	0.011		
T	6.3	6.3	6.351	0.009	6.351	0.009		
U	6.0	6.0	6.046	0.008	6.046	0.008		
V	6.1	6.1	6.075	0.011	6.075	0.011		
W	6.2	6.2	6.227	0.012	6.227	0.012		
X	6.0	6.0	6.043	0.008	6.043	0.008		
Y	5.8	5.8	5.761	0.010	5.761	0.010		
Z	6.4	6.4	6.333	0.007	6.333	0.007		

Mu1 and Mu2 are the original simulated group means. For each gene, the sample averages and standard deviations of posterior samples of group mean are provided. MuEstDiff and MuStdDiff are the posterior estimates of mean difference between group 2 and group 1. A negative value indicates the group 2 has lower expression compared to group 1 and a positive value indicates a positive component. If posterior of both negative and positive are diffused, the mean difference is not summarized: from null component.

Table 3.3: The summary of difference between simulated (true) means with sample averages or posterior sample averages in the simulation of the two-sample model.

Groups	Sample_Avg	PAvg_P5	PAvg_P1	PAvg_P25	PAvg_P75	PAvg_1
Group1 Diff Mean	0.0744	0.0667	0.0673	0.0671	0.0688	0.0697
Group1 Diff Std	0.0951	0.0865	0.0869	0.0865	0.0889	0.0912
Group2 Diff Mean	0.0752	0.0778	0.0844	0.0711	0.0846	0.1011
Group2 Diff Std	0.0933	0.0935	0.104	0.087	0.1073	0.1452

Note: The posterior sample average tends to have smaller difference and variation in group 1, but not in group 2. PAvg: posterior average, P5, P1, P25, P75 and 1 were corresponding to $w = 0.5, 0.1, 0.25, 0.75, 1$

bigger than 0.99, the posterior samples of the gene come from the corresponding component, otherwise, the chain for each component is checked for convergence (small variation compared to range of posterior samples). If both negative and positive component fail convergence for a gene, the gene of group 2 comes from a null component, which means no difference between the two groups. Table 3.2 showed the summarization of the posterior samples. The first two columns are the simulated group means. All the 17 genes in the first disjoint subgraph have over-expression in group 2. From table 3.2, all the 17 genes from positive component were correctly summarized. We also observed that the posterior mean difference of genes between the two groups within disjoint subgraph tends to be bigger than the mean of true mean differences. In contrast with modeling Z-score, the direction of difference (negative or positive component) could be correctly obtained from the posterior samples. This suggested the advantage of working with original expression values instead of Z-scores.

As the true mean of each gene for both groups are known, sample average and posterior summarized mean for each gene were compared to the true mean. As shown in table 3.3 (first 3 columns), the difference of true mean with posterior averages had smaller average difference and variation than those with sample average for group 1 and slightly bigger in group2. In this scenario with the dependence of genes following the gene network, the Markov random field model incorporated dependence structure into the posterior and improved the accuracy of posterior estimates, especially for group 1. In contrast, the simple sample average ignored the dependence structure and had a loss in efficiency of the estimates of group means.

The Bayesian two-sample model was compared with t-test using the simulated raw data. For each gene, we performed two-sample t-test and ordered the genes according to the p-values. We also computed absolute value of standardized posterior mean (the absolute value of posterior mean divided by standard deviation of the batched means). Significant absolute value of standardized posterior mean indicates that the gene is differentially expressed. The genes were reversely ranked by the absolute value of standardized posterior mean, from large to small value. The result is shown in table 3.4 (first 5 columns). In the two sample t-test, 16 genes had p-values less than 0.05 and 14 genes were found to be differentially expressed after Bonferroni correction of the p-value cutoff of 0.05. The Bayesian two-sample model identified all the 17 differentially expressed genes. Gene J was detected to be differentially expressed in the Bayesian two-sample model, although its p-value in t-test was 0.22. This supports the advantage of taking the dependency structure of gene network into the genetic testing.

Several genes were ranked differently in the two methods. Gene O was ranked # 6 in t-test and # 8 in the Bayesian model, with weight w predefined to be 0.5. We further studied the influence of weight on the posterior samples. Using simulated data with weight equal to 0.5, we investigated the impact of arbitrary weights 0.1, 0.25, 0.75 and 1, where weight $= 1$ is equivalent to an undirected graph. Table 3.4 summarizes the results. There were 17 genes declared to be differentially expressed in first disjoint subgraph with weights 0.1, 0.25 and 0.5, out of which one gene was not significant in t-test at 0.05 level. Weight 0.75 identified 16 genes to be differentially expressed with ranks slightly changed. For weight $= 1$, equivalent to an undirected graph, 16 genes were declared to be differential, and gene B was not. The ranks of differential expressions differed from those for weights 0.25 and 0.5. We also compared the summarized posterior mean for each gene with its true mean under different weights. As shown in table 3.3 the undirected graph (weight = 1) had bigger average difference and variation in difference between posterior means and true means, compared to weights 0.1, 0.25, 0.5, and 0.75. This indicates the possible robustness of weight selection and advantage of known direction in gene relationships.

The prior knowledge of pathway is usually not complete. We further studied the robustness of the incomplete prior knowledge on posterior samples. We randomly added 3 directed edges into the first disjoint subgraph in figure 3.5. Using the simulated data as before, table 3.4 shows that the 16 genes were declared to be significant with slight change in ranking and with gene J left out. This indicates a degree of robustness to the directed graph by incorrect inclusion of extra edges. With only 27 directed edges in the original graph, 3 edges represent 11% of the total edges. Simulating a graph in which 8 (30%) directed edges were randomly inserted, the number of differentially expressed genes and ranking did not change much (see column 8 and 11 in table 3.5).

3.3.3 Simulation on DG model for two-sample problem

Gene expression data was simulated according to following mechanism based on the DG in figure 3.5 with $\mathbf{w} = 0.5$: (1) Start with constant $\boldsymbol{\mu}_j$ for $j = 0, 1, 2$, each component of $\boldsymbol{\mu}_0$ was drawn from positively truncated normal distribution with mean being weighted average of neighbors and weighted standard error; similarly, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ was drawn from a negatively and positively truncated normal distribution, respectively; the process was repeated 20 times to introduce correlation among the genes according to figure3.5; (2) For the first 17 genes corresponding to the first disjoint subgraph, μ_{di} was set to μ_{2i} for $i = 1, \ldots, 17$, and for the rest 9 genes corresponding to the second disjoint subgraph, μ_{di} was set to μ_{0i} for $i = 18, \ldots, 26$; and (3) For each gene, 10 random samples was drawn from a normal distribution with mean μ_{0i} for group 1 and $\mu_{g2i} = \mu_{0i} + \mu_{di}$ for group 2 with fixed variance.

With the same hyperparameters as in section 3.2.1 and randomly generating initial values, the chain was run 4000 for burn-in. 4000 posterior samples were drawn from posterior distributions with thin equal to 3 (keep one out of three samples). Figure 3.8 shows the trace plots of posterior samples for 8 selected nodes: the first 4 nodes from first disjoint subgraph in figure 3.5 and last 4 nodes from the second disjoint subgraph. For the 8 nodes, the chain converged for both $\boldsymbol{\mu}_0$ (1^{st} row) and $\boldsymbol{\sigma}_0^2$ (4^{th} row), indicating stability in the sampling of the overall mean of the two groups. The same observation holds for both $\boldsymbol{\mu}_d$ (2^{nd} row) and $\boldsymbol{\sigma}_2^2$ (5^{th} row). The chain for weight, $\mathbf{w}$, also converged. Out of the 8 nodes, 3 nodes converged around 1 and one converged around 0, for nodes B, Y, Z, the

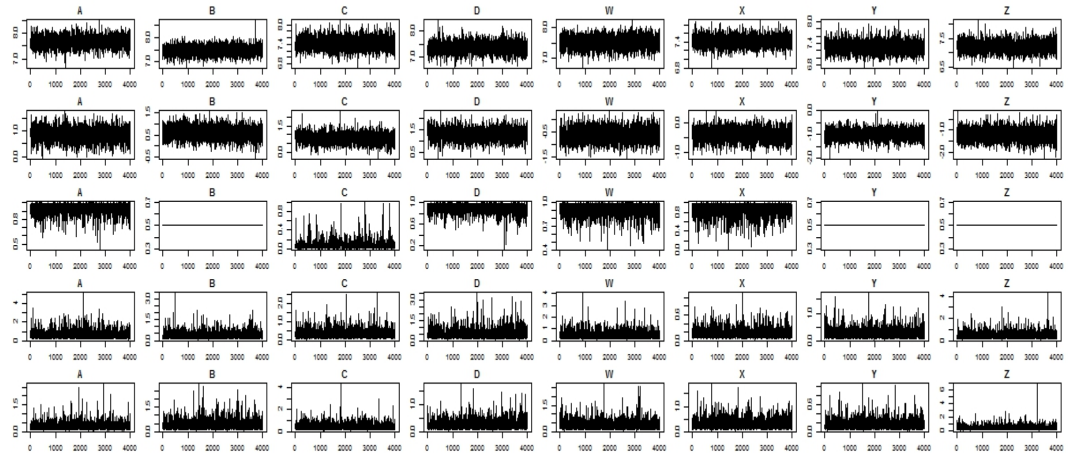


Fig. 3.8: Trace plot of posterior samples of $\boldsymbol{\mu}_0, \boldsymbol{\mu}_d, \mathbf{w}, \boldsymbol{\sigma}_1^2$ and $\boldsymbol{\sigma}_2^2$ in simulation of the DG model.

Posterior samples of first 4 nodes (left 4 columns) and last 4 nodes (4 columns on the right) of the DG as shown in figure 3.5. Row 1 and 2 are trace plots for posterior samples of $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_d$; row 3 are trace plots for $\mathbf{w}$; row 4 and 5 are for $\boldsymbol{\sigma}_1^2$ and $\boldsymbol{\sigma}_2^2$, respectively

weights were set to constant as these nodes do not have children and weight will have no effect. The trace plots of $\mathbf{w}$ for the remaining 18 nodes are shown in figure 3.9.

The standard error of posterior sample mean was computed according to Albert and Chib (1993). Table 3.6 shows the summarization of the posterior samples. The first 3 columns are the simulated group means and mean difference respectively. All the 17 genes in first disjoint subgraph have over-expression in group 2. The other 9 genes have

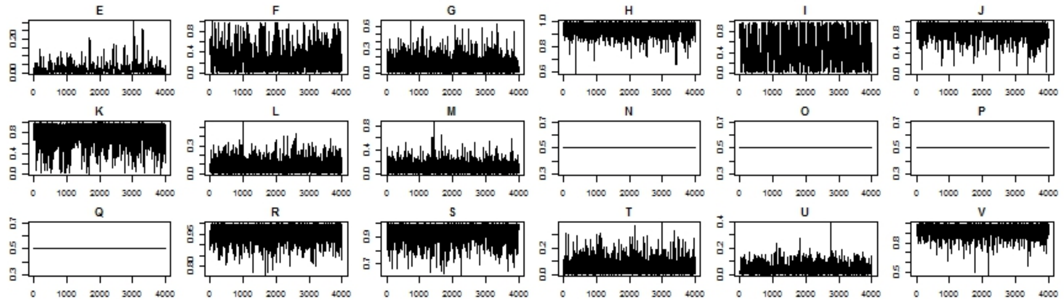


Fig. 3.9: Trace plots of posterior samples of $\mathbf{w}$ in simulation study of the DG model.

under-expression in group 2. As the true mean of each gene for both groups are known, sample averages and posterior summarized means were compared to their true mean. As shown in table 3.7 (the first 4 columns), the difference of true mean with posterior averages had smaller average difference and variation than those with sample averages for group 1 and slightly bigger in group2. In this scenario with the dependence of genes following the gene network, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_d$ incorporated dependence structure into the posterior and improved the accuracy of posterior estimates, especially for group 1. In contrast, the simple sample average ignored the dependence structure with loss of efficiency to estimate group means.

We also compared the DG model with two-sample t-test. For each gene, we performed two-sample t-test and ordered the genes according to *p*-values. The absolute value of standardized posterior mean was computed as the absolute value of posterior mean divided by standard error of batched means. The genes were ordered from the largest to the smallest according to the absolute value of standardized posterior mean. The result is shown in table 3.8 (first 5 columns). In the two-sample t-test, 22 genes had *p*-values less than 0.05. We also ranked the gene according to the absolute value of simulated mean difference: $\boldsymbol{\mu}_d$, and compared this rank to the rank of t-test and rank of DG model. The simulated mean difference had smaller rank difference from the DG model than that from t-test.

We now assume that weight has a uniform(0,1) prior distribution. To study the effect of random weight on the posterior samples, we compared the posterior samples with those from fixed weights of 0.1, 0.25, 0.5, 0.75 and 1 using the same simulated data. As shown in table 3.8, the rank of genes did not differ much using different fixed weights or prior distribution of $uniform(0, 1)$, all of which had less discrepancy from the ranks with simulated mean difference. We also compared the summarized posterior mean for each gene with its true mean under different weights. As shown in table 3.7 (column 5 to 8), the DG model with random weight had smaller average and variation in difference between posterior means and true means, compared to models with fixed weights 0.1, 0.25, 0.5, and 0.75 and 1. This demonstrated advantage of prior information of weight and advantage of known direction in gene relationships.

We further studied the robustness of using noninformative prior of gene relationships on posterior samples. We randomly added 3 or 8 directed edges into the first disjoint subgraph in figure 3.5. Using the simulated data as before, table 3.9 shows the summarization of the results. The rank did not change much for DG with 3 edges inserted. The rank did differ for DG with 8 random edges inserted. As shown in table 3.10, the mean and standard error of difference between posterior samples and simulated data did not differ in group 1 for all the three models, but did differ for group 2 or group difference in DG model with 8 random edges added. This indicates that the model is robust to moderate miss-specification and may not be robust for a higher degree of miss-specification.

3.4 Application

The Cancer Genome Atlas project has collected multiple forms of high throughput data for various cancer types. The mRNA expression data for 172 adult Acute Myeloid Leukemia (AML) patients is publicly available. The patients were classified into favorable, intermediate or poor risk groups according to cytogenetics of leukemia cells. Out of the 172 patients, 37 patients were in favorable risk group and 42 were in poor risk group. One interesting question is which set of genes are differentially expressed between these two groups of patients. This dataset is used to demonstrate the utility of the proposed models in above section.

The current pathway databases, such as KEGG, use different names for a gene. To derive a directed graph representing a pathway, human MAPK pathway in KEGG is used as an example, where gene names were matched to Affymetrix gene names through GeneCard (http://www.genecards.org/). The directed graph of MAPK pathway is shown in figure 3.10. In this graph, there are 102 genes with 1 to 14 directed edges per gene (median 2 and mean 2.7 per gene). The two-sample t-test and each of the proposed methods were applied to these 102 genes.

To apply our model with Z-scores to the AML data, one sided two-sample t-tests were performed on each gene to compare mean expression levels between poor and favorable risk adult AML, and the p-values were transformed to obtain Z-scores (range: $-4.27, 4.98$, mean: 0.03). The posterior of $\pi_j, j = 0, 1, 2$ indicated that all the 102 came from null component (no differential expression was declared), which is consistent with the distribution of Z-scores. The Bayesian two-sample model was used on the same data set. Similar to the model with Z-score, all the 102 genes were found to come from the null component.

In the DG model, the posterior samples of μ_d were summarized as batched mean and standard error according to Albert and Chib (1993). The genes were ordered from large to small according to the absolute values of ratio between batched mean and its standard error. The top 30 genes are shown in table 3.11. The key genes in MAPK pathway are in the top of the gene list, such as CUTL1 and MAP4K3.

In the application of MAPK pathway derived from the KEGG to the AML dataset, using Z-statistics as data (model I) and the two-sample model with raw expression value (model II) failed to detect differential expression in the 102 genes between favorable and poor risk AML. The distribution of z-statistics from two-sample t-tests does not support that the genes come from either positive or negative component. One possible reason is that the MAPK pathway derived from KEGG is a prior knowledge from various sources and tissues. Some of the relationships might not be applicable to AML. We decided to tailor the MAPK pathway based on correlations of genes in adult AML of intermediate risk. Two-way correlations of the 102 genes were calculated using expression values from the 93 intermediate risk AMLs. An arbitrary p-value cutoff of 0.1 was used to indicate potential correlation between genes. Out of the 102 genes in the MAPK pathway in figure 3.10, 37 genes were singletons after tailoring based on the correlation in AML of intermediate risk (genes in black and edges in grey in figure 3.10). In the rest 65 genes, there were 57 edges left (range: 1 to 6 edges, mean: 1.8 edges; genes in red and edges in blue in figure 3.10).

The tailored MAPK pathway with 65 genes was applied to the AML data to model differential expression between poor and favorable risk AML using the three models. The summary of posterior samples is given in table 3.12 for the model with z-statistics as data for top 20 genes. Out of the 20 genes, most were from negative component (AML of poor risk had lower expression compared to favorable risk AML). PPP2CB was on the top of the list based on the ratio of batched mean and its standard deviation. MAP3K1 was the number 3 and from the positive component. In the original MAPK pathway, all the 102 genes were from null component, this tailored MAPK pathway with AML of intermediate risk detected differential expression in AML, demonstrating the advantage of tailoring pathway using appropriate data.

The summary of posterior samples for two-sample model is shown in table 3.13 for genes with p-values of t-test less than 0.1 or high posterior ratio. CUTL1 is on the top of the differentially expressed genes in both t-test and based on the ratio of batched mean and standard deviation of posterior difference. PRKCA is ranked 9^{th} in t-test and 2^{nd} based on posterior. PAK1 was ranked 2^{nd} in t-test and 31^{st} based on posterior, in the middle of the list. This indicated that both original expression data and prior knowledge captured in the tailored pathway played important role in the Bayesian two-sample model.

When the DG model was applied to the AML data using the tailored MAPK pathways, both CUTL1 and PRKCA were on the top 2 based on posterior, similar to the Bayesian two-sample model. A few others are also on the top of the list based on t-test, posterior estimates in the DG model and in the Bayesian two-sample model. It is very likely that CUTL1 was under expressed in favorable AML compared to poor AML.

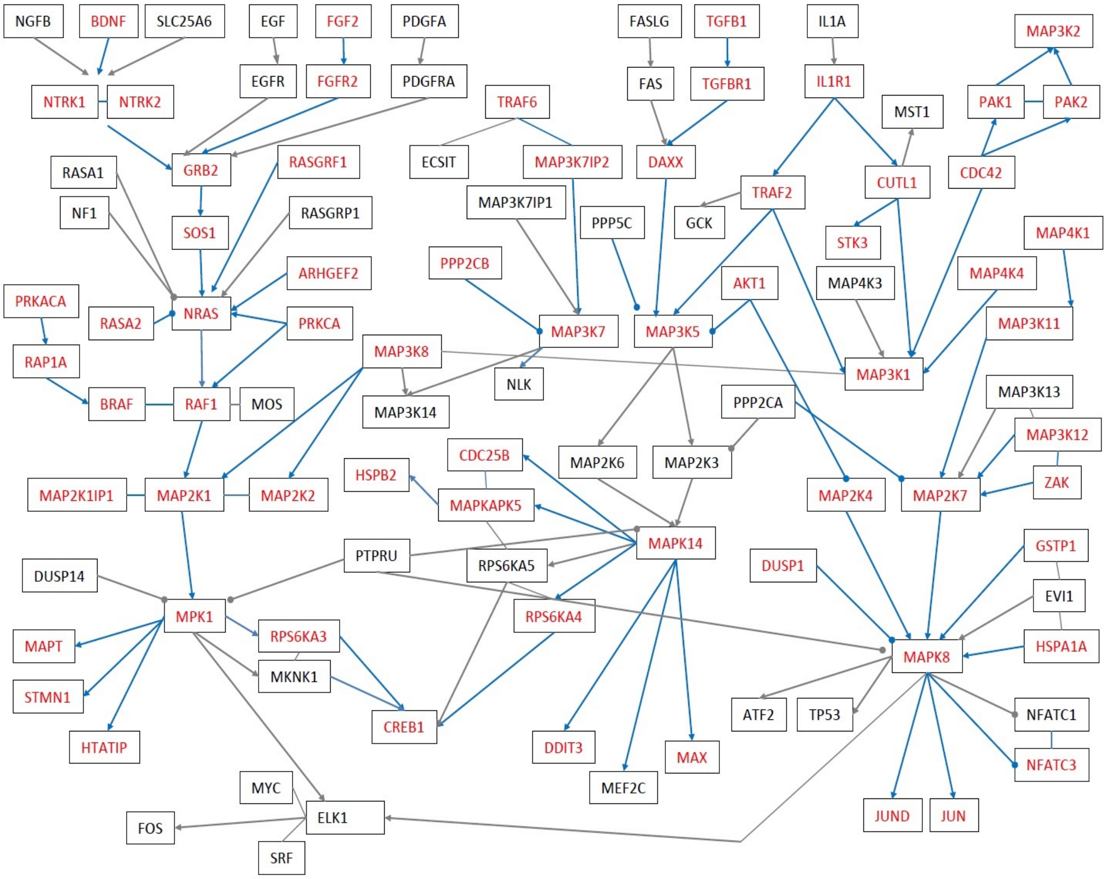


Fig. 3.10: Human MAPK pathway derived from KEGG database

Note: Lines with pointed arrows indicate positive regulation, lines with rounded end indicate negative regulation or repression. Gene names are translated from the KEGG to the ones used by Affymetrix. The genes in red and edges in blue are retained after tailoring with intermediate risk adult AML

3.5 Summary and Discussion

Genes and proteins regulate, interact and cross talk among each other, forming complex networks in a cellular context. This information is captured partially in gene network databases. In this study, we proposed incorporating prior knowledge of gene relationships in gene networks to statistical inference by Bayesian approaches.

In the model with Z-scores, it is flexible to include most statistical testing with *p*-values or testing statistics (standard normal transformation). Based on the simulation study, the prior knowledge in the directed graph will dominate the distribution of Z-scores and reorder the rank of genes in the posterior samples. Small and coherent directed graphs are desirable as the genes in such graphs are coherently regulated and expressed.

In an experiment with two treatment groups or conditions, it is not hard to formulate the Markov random field model with experimental data. In this case, the prior knowledge will not dominate the posterior distribution and presumably the Bayesian model combines both the prior knowledge and data. The model is robust for both weight miss-specification and moderate directed graph miss-specification. The true weight is not known. Furthermore, the prior knowledge of pathways represented by directed graphs is usually incomplete and potentially has miss specification for certain experiment conditions. The robust feature of the Bayesian model is desirable.

For gene expression data, expression values could be aligned to similar scales after proper data transformation or manipulation. It is reasonable to assume that the expression level of a gene depends on its regulators or partners in a pathway. In the Bayesian model with direct dependence on a directed graph, we demonstrated the feasibility through simulation studies. The model is robust to the selection of weights, but not to large miss-specification of the directed graph itself. It is not contradictory to intuition as expression levels of a gene in the model are assumed to be normally distributed with mean being weighted average of first-order neighbors. In this case, a valid directed graph of prior knowledge under the experiment condition is required.

Although gene and protein relationships are captured and represented in network databases, they are not easily transformed to adjacency matrix format directly usable to statisticians. We derived directed graphs of yeast and huma MAPK kinase pathway from KEGG database. We noticed the following difficulties: (1) A dozen genes in the original MAPK kinase pathway are not represented by Affymetrix microarray chip; (2) The gene names used in Affymetrix array are frequently different from the ones used in KEGG; (3) Genes are represented by multiple probes. In conquering these difficulties, efforts are demanding to put gene network relationships into statistician accessible format such as adjacency matrix.

Another issue of gene network data is that the relationships are extracted and integrated from diverse cellular contexts and experiment conditions. Part of the relationships in network database might not be valid for a specific experiment condition such as cell types, microarray platform *et al.* To account for this fact, it would be desirable to derive application specific prior gene network. If supposing gene profiling data are available in cell lines, patients, disease that are compatible to current experiment units or conditions, the correlation structure from this data set could be roughly estimated. Based on the correlation structure, the overall gene network represented by adjacency matrix could be tailored. This tailored gene network contains prior knowledge of both integrated information and specific information relevant to current experiment units or conditions.

In the application to the AML data using MAPK pathway derived from KEGG, model with Z-scores and two-sample model failed to detect any differential expression (all genes were from null component). The distribution of Z-scores also indicated that genes in the MAPK pathway were not all relevant in AML. After being tailored using adult AML with intermediate risk, several important genes were detected to be differentially expressed across the three models. This indicates that pathways in public database indeed need to be tailored to specific tissues and experiment conditions in order to be useful in these Markov random field models. More efforts are needed in this area in near future.

Table 3.4: Comparison of the two-sample model with t-test and weight effects on the two-sample model.

gene	TTest_pval	TT_Ord	PostDif	PostOrd	PDif_P1	POrd_P1	PDif_P25	POrd_P25	PDif_P75	POrd_P75	PDif_1	POrd_1
A	1.54E-07	11	70.49	9	70.56	10	68.58	10	67.84	11	67.54	11
B	0.0014	14	31.03	15	31.98	13	33.37	14	3.26	13		
C	2.14E-04	13	34.29	13	6.28	16	35.02	13	2.7	15	1.33	15
D	0.0067	15	32.44	14	26.92	15	27.6	15	1.35	16	28.49	13
E	2.48E-07	12	69.13	10	74.21	8	67.15	12	79.03	7	83.73	5
F	1.00E-10	3	94.13	4	115.98	4	90.94	7	92.62	3	78.34	7
G	2.62E-08	9	64.94	11	70.57	9	78.59	9	69.93	10	75.42	10
H	2.00E-10	5	74.23	7	78.88	7	98.4	5	73.64	9	78.07	8
I	2.60E-09	8	89.05	6	105.35	5	101.51	4	88.42	5	80.92	6
J	0.22	19	0.94	17	0.77	17	6.42	16			0.62	16
K	0.0068	16	24.73	16	27.04	14	4.01	17	2.82	14	27.3	14
L	3.16E-08	10	64.32	12	53.48	12	67.8	11	55.99	12	58.92	12
M	2.30E-09	7	92.55	5	70.07	11	97.05	6	78.35	8	76.56	9
N	1.00E-10	4	105.29	3	117.88	3	106.4	3	100.32	2	89.7	4
O	1.00E-09	6	71.1	8	82.45	6	84.89	8	89.43	4	94.65	3
P	1E-16	1	113.59	2	125.86	1	121.79	2	82.78	6	117.76	2
Q	1E-16	2	115.4	1	120.25	2	132.69	1	109.48	1	124.35	1
R	0.21	18										
S	0.84	25										
T	0.18	17										
U	0.53	21										
V	0.63	24										
W	0.55	22										
X	0.99	26										
Y	0.62	23										
Z	0.48	20										

Two-sample t-test was performed and order of gene based on p-value was in column 2. Post_Diff or PDiff is the absolute value of mean of posterior difference divided by standard error of batched means. Post_Ord or POrd is rank of genes based on Post_Diff or PDiff. P1, P25, P75 and 1 were corresponding to $w = 0.1, 0.25, 0.75, 1$, respectively.

Table 3.5: Impact of misspecification of directed graph by randomly inserting directed edges in the two-sample model.

Gene	MuDiff	MuEstDiff	MuStdDiff	Post_Ord	MuEstDiff_3	MuStdDiff_3	Post_Ord_3	MuEstDiff_8	MuStdDiff_8	Post_Ord_8
A	1.7	1.902	0.027	9	1.904	0.03	12	1.903	0.031	11
B	0.6	0.64	0.021	15	0.849	0.481	14	0.814	0.495	14
C	0.6	0.715	0.021	13	0.788	0.366	13	0.754	0.13	13
D	0.5	0.473	0.015	14	0.846	0.873	16	0.792	0.769	16
E	1.6	1.782	0.026	10	1.785	0.022	8	1.793	0.021	7
F	2.2	2.266	0.024	4	2.265	0.022	2	2.27	0.021	2
G	1.9	1.713	0.026	11	1.714	0.026	11	1.711	0.027	10
H	2.2	2.124	0.029	7	2.11	0.024	5	2.12	0.024	6
I	1.7	1.796	0.02	6	1.796	0.022	7	1.794	0.019	5
J	0.4	0.476	0.505	17						
K	0.8	0.635	0.026	16	0.856	0.536	15	0.834	0.687	15
L	1.4	1.56	0.024	12	1.564	0.022	10	1.559	0.025	12
M	1.6	1.609	0.017	5	1.608	0.022	9	1.611	0.023	9
N	2.2	2.448	0.023	3	2.448	0.025	4	2.443	0.023	3
O	2.1	2.058	0.029	8	2.052	0.024	6	2.056	0.026	8
P	1.8	2.04	0.018	2	2.039	0.015	1	2.038	0.018	1
Q	2.5	2.721	0.024	1	2.714	0.027	3	2.727	0.026	4
R	0.0									
S	0.0									
T	0.0									
U	0.0									
V	0.0									
W	0.0									
X	0.0									
Y	0.0									
Z	0.0									

Note: MuDiff is the simulated Mu difference in the two groups. MuEstDiff is average of posterior mean difference, MuStdDiff is the standard error of estimate of posterior mean differece and Post_Ord is the rank of gene considered both MuEstDiff and MuStdDiff. _3 and _8 are for graphs with 3 and 8 directed edges randomly inserted, respectively.

Table 3.6: Posterior estimates for the 26 genes in simulation of the the DG model.

Gene	Mu1	Mu2	MuDiff	MuEstGrp1	MuStdGrp1	MuEstGrp2	MuStdGrp2	MuEstDiff	MuStdDiff
A	7.55	8.27	0.72	7.683	0.028	8.512	0.013	0.829	0.036
B	7.45	8.21	0.76	7.471	0.021	8.056	0.016	0.586	0.030
C	7.45	8.22	0.77	7.425	0.020	8.226	0.022	0.801	0.029
D	7.60	8.53	0.93	7.377	0.027	8.666	0.017	1.289	0.036
E	7.05	7.08	0.03	6.988	0.016	7.312	0.021	0.325	0.029
F	7.26	8.03	0.77	7.337	0.018	7.827	0.019	0.491	0.027
G	7.19	8.24	1.05	7.034	0.016	8.123	0.015	1.089	0.024
H	7.02	8.64	1.62	7.175	0.011	8.614	0.018	1.439	0.019
I	6.91	8.64	1.73	6.870	0.023	8.491	0.019	1.621	0.030
J	7.56	8.36	0.80	7.404	0.018	8.156	0.019	0.752	0.024
K	7.22	7.87	0.65	7.296	0.021	7.944	0.017	0.648	0.029
L	6.72	7.03	0.31	6.814	0.015	7.161	0.022	0.346	0.030
M	6.78	6.97	0.19	6.680	0.021	6.997	0.019	0.317	0.029
N	7.45	8.88	1.43	7.313	0.024	8.724	0.024	1.411	0.037
O	7.12	8.50	1.38	7.218	0.017	8.397	0.019	1.179	0.025
P	6.90	8.53	1.63	7.035	0.009	8.651	0.021	1.616	0.022
Q	6.94	9.06	2.12	6.876	0.014	8.861	0.013	1.985	0.018
R	7.67	7.15	-0.52	7.531	0.014	6.969	0.015	-0.562	0.023
S	7.75	7.15	-0.60	7.746	0.035	7.115	0.013	-0.632	0.040
T	6.70	5.36	-1.34	6.922	0.020	5.504	0.013	-1.419	0.027
U	7.04	5.89	-1.15	7.163	0.014	5.833	0.013	-1.330	0.020
V	7.90	7.21	-0.69	7.900	0.038	7.273	0.015	-0.627	0.046
W	7.40	7.00	-0.40	7.523	0.024	6.904	0.017	-0.619	0.026
X	7.63	6.99	-0.64	7.424	0.013	7.035	0.017	-0.388	0.022
Y	7.23	6.27	-0.96	7.294	0.017	6.299	0.011	-0.995	0.021
Z	7.22	5.78	-1.44	7.220	0.019	6.017	0.018	-1.202	0.027

Mu1 and Mu2 are the original simulated group means. For each gene, the sample averages and standard deviations of posterior samples of group mean are provided. MuEstDiff and MuStdDiff are the posterior estimates of mean difference between group 2 and 1. A negative value indicates the group 2 has negative component and a positive value indicates a positive component. If posterior of both negative and positive diffused, the mean difference is not summarized: from 0 component.

Table 3.7: Summary of difference of simulated (true) means with sample averages or with posterior sample averages in simulation of the DG model.

Groups	Sample_Avg	PAvg	PAvg_P1	PAvg_P25	PAvg_P5	PAvg_P75	PAvg_1
Group1 Diff Mean	0.1214	0.1027	0.1078	0.1081	0.1087	0.1092	0.1099
Group1 Diff Std	0.144	0.1233	0.124	0.125	0.127	0.1287	0.1301
Group2 Diff Mean	0.1172	0.1215	0.1237	0.1232	0.1226	0.1221	0.1217
Group2 Diff Std	0.1373	0.143	0.1465	0.1462	0.146	0.1459	0.1458
MuD Diff Mean	0.1277	0.1278	0.1306	0.1303	0.1297	0.1298	0.1306
MuD Diff Std	0.1639	0.1661	0.1665	0.1672	0.1684	0.1695	0.1705

Sample_Avg: difference between sample average and simulated mean. PAvg: difference between posterior sample mean with random weight and simulated mean. PAvg_P1,PAvg_P25, PAvg_P5, PAvg_P75 and PAvg_1 are difference between simulated mean and posterior estimates with fixed weight 0.1, 0.25, 0.5, 0.75 and 1, respectively.

Table 3.8: Comparison of DG model with t-test and weight effects on the DG model.

gene	TTest_pval	TT_Ord	PDif	POrd	PDif_P1	POrd_P1	PDif_P25	POrd_P25	PostDif	PostOrd	PDif_P75	POrd_P75	PDif_1	POrd_1
A	0.011	17	23.26	17	16.15	23	17.29	22	18.35	21	19.16	20	19.84	19
B	0.021	20	19.68	19	22.02	17	22.35	17	22.63	17	22.69	17	22.67	17
C	0.0026	13	28.05	14	25.26	16	26.28	15	27.38	13	28.07	13	28.57	13
D	6.05E-05	10	35.94	12	30	12	31.38	12	32.69	12	33.43	12	33.98	12
E	0.14	24	11.06	25	10.5	26	11.39	25	12.47	25	13.2	24	13.74	24
F	0.037	22	17.9	20	18.5	21	18.76	20	19.29	19	19.65	19	19.80	20
G	6.35E-06	7	45.69	9	43.19	10	43.63	10	44.02	8	44.36	8	44.60	8
H	2.84E-07	3	74.76	2	54.68	3	55.7	3	56.59	3	57.51	3	58.14	3
I	1.86E-07	2	53.49	5	49.74	4	50.73	4	52.04	4	53.02	4	53.80	4
J	0.0081	15	31.14	13	25.48	15	26.23	16	26.96	14	27.44	14	27.82	14
K	0.022	21	22.61	18	26.86	13	26.43	14	25.73	16	25.43	16	25.30	16
L	0.24	25	11.43	24	13.36	24	13.22	24	13.01	24	13.12	25	13.26	25
M	0.32	26	10.94	26	10.9	25	10.58	26	10.29	26	10.39	26	10.57	26
N	1.33E-04	11	38.61	11	44.18	9	44.29	9	43.96	9	43.5	10	43.15	10
O	5.02E-05	9	47.87	7	46.8	6	46.31	5	45.96	5	45.45	6	44.89	7
P	1.71E-06	4	74.34	3	85.75	2	86.33	2	87.1	2	87.8	2	88.31	2
Q	4.00E-10	1	108.31	1	98.7	1	98.88	1	98.73	1	98.19	1	97.60	1
R	0.011	18	24.48	15	26.38	14	26.52	13	26.44	15	26.23	15	26.07	15
S	0.0081	16	15.61	22	16.6	22	16.45	23	16.01	22	15.79	22	15.71	22
T	2.23E-06	5	52.56	6	46.38	7	44.82	8	43.13	10	44.2	9	45.17	6
U	2.99E-06	6	68.04	4	47.23	5	46.22	6	45.19	6	46.41	5	47.33	5
V	0.0039	14	13.63	23	18.53	20	17.33	21	15.95	23	15.55	23	15.45	23
W	0.012	19	23.61	16	19.93	18	20.01	18	20.11	18	20.23	18	20.39	18
X	0.06	23	17.57	21	19.47	19	19.33	19	19.1	20	18.88	21	18.75	21
Y	2.87E-05	8	47.38	8	35.6	11	34.72	11	34.03	11	34.2	11	34.27	11
Z	1.67E-04	12	44.87	10	46.04	8	45.33	7	44.45	7	44.53	7	44.43	9

Two-sample t-test was performed and order of gene based on p-value was in column 3. Post_Dif or PDif is the absolute value of mena of posterior group difference divided by standard error of batched means. PostOrd or POrd is rank of genes based on Post_Dif or PDif. P1, P25, P75 and 1 are for weight 0.1, 0.25, 0.75 and 1, respectively.

Table 3.9: Impact of misspecification of directed graph by randomly inserting directed edges in then DG model.

Gene	MuDiff	MuEstDiff	MuStdDiff	Post_Ord	MuEstDiff_3	MuStdDiff_3	Post_Ord_3	MuEstDiff_8	MuStdDiff_8	Post_Ord_8
A	0.72	0.829	0.036	17	0.925	0.037	15	1.058	0.027	12
B	0.76	0.586	0.03	19	0.603	0.028	18	0.627	0.027	19
C	0.77	0.801	0.029	14	0.832	0.024	12	0.929	0.024	13
D	0.93	1.289	0.036	12	1.289	0.031	9	1.281	0.035	14
E	0.03	0.325	0.029	25	0.322	0.026	25	0.328	0.025	25
F	0.77	0.491	0.027	20	0.497	0.03	20	0.515	0.03	22
G	1.05	1.089	0.024	9	1.098	0.023	6	1.135	0.019	5
H	1.62	1.439	0.019	2	1.418	0.024	4	1.429	0.022	4
I	1.73	1.621	0.03	5	1.633	0.027	3	1.575	0.021	2
J	0.80	0.752	0.024	13	0.759	0.03	14	0.763	0.026	16
K	0.65	0.648	0.029	18	0.651	0.027	16	0.665	0.019	15
L	0.31	0.346	0.03	24	0.362	0.027	24	0.353	0.035	26
M	0.19	0.317	0.029	26	0.343	0.031	26	0.321	0.023	24
N	1.43	1.411	0.037	11	1.406	0.043	13	1.424	0.032	9
O	1.38	1.179	0.025	7	1.171	0.029	10	1.18	0.022	6
P	1.63	1.616	0.022	3	1.548	0.022	2	1.572	0.024	3
Q	2.12	1.985	0.018	1	1.971	0.019	1	1.905	0.022	1
R	-0.52	-0.562	0.023	15	-0.57	0.026	17	-0.571	0.022	17
S	-0.60	-0.632	0.04	22	-0.646	0.041	22	-0.634	0.04	23
T	-1.34	-1.419	0.027	6	-1.407	0.033	8	-1.411	0.032	10
U	-1.15	-1.33	0.02	4	-1.317	0.023	5	-1.33	0.025	7
V	-0.69	-0.627	0.046	23	-0.64	0.04	21	-0.644	0.026	18
W	-0.40	-0.619	0.026	16	-0.626	0.03	19	-0.631	0.03	21
X	-0.64	-0.388	0.022	21	-0.395	0.026	23	-0.394	0.019	20
Y	-0.96	-0.995	0.021	8	-0.984	0.024	11	-0.995	0.024	11
Z	-1.44	-1.202	0.027	10	-1.19	0.026	7	-1.19	0.025	8

MuDiff is the simulated Mu difference in the two groups. MuEstDiff is the average of posterior group difference, MuStdDiff is the standard error of estimate of posterior group difference and Post_Ord is the rank of gene considered both MuEstDiff and MuStdDiff. _3 and _8 are for graphs with 3 and 8 directed edges randomly inserted, respectively.

Table 3.10: Summary of difference of simulated means with sample averages or with posterior sample averages in the DG model.

Groups	Sample_Avg	PAvg	PAvg_3	PAvg_8	PAvg_1
Group1 Diff Mean	0.1214	0.1027	0.1009	0.1024	0.1099
Group1 Diff Std	0.144	0.1233	0.1221	0.1248	0.1301
Group2 Diff Mean	0.1172	0.1215	0.1238	0.1234	0.1217
Group2 Diff Std	0.1373	0.143	0.1458	0.145	0.1458
MuD Diff Mean	0.1277	0.1278	0.1371	0.1465	0.1306
MuD Diff Std	0.1639	0.1661	0.1716	0.1824	0.1705

Note: Sample_Avg: difference between sample average and simulated mean. PAvg: difference between posterior sample mean with random weight and simulated mean. PAvg_3 and PAvg_8 are difference between simulated mean and posterior estimates with 3 and 8 random edges inserted, respectively.

Table 3.11: Summary of posterior samples of MAPK genes in adult AML in the DG model

Gene	AvgGrp1	StdGrp1	AvgGrp2	StdGrp2	MuEstGrp1	MuStdGrp1	MuEstGrp2	MuStdGrp2	MuEstDiff	MuStdDiff
CUTL1	13.32	0.42	12.72	0.64	13.282	0.008	12.784	0.009	-0.498	0.013
MAP4K3	10.49	0.45	10.89	0.67	10.495	0.007	10.875	0.008	0.380	0.010
JUND	15.22	0.62	15.69	0.57	15.248	0.011	15.660	0.008	0.412	0.013
IL1A	10.12	0.53	9.61	0.75	10.111	0.008	9.627	0.013	-0.484	0.016
TGFB1	12.2	0.98	12.91	0.92	12.255	0.021	12.867	0.013	0.612	0.022
MAX	14.98	0.46	14.64	0.56	14.950	0.005	14.672	0.009	-0.278	0.010
PPP2CB	13.04	0.5	13.38	0.58	13.045	0.009	13.369	0.010	0.324	0.012
PAK2	13.15	0.36	12.83	0.47	13.114	0.007	12.875	0.007	-0.239	0.010
MAP2K4	10.99	0.3	10.84	0.41	10.989	0.004	10.841	0.007	-0.147	0.007
STMN1	12.74	0.56	12.36	0.57	12.702	0.010	12.394	0.009	-0.308	0.015
MAP4K4	12.21	0.61	11.89	0.66	12.206	0.011	11.890	0.009	-0.316	0.016
EGF	6.45	1.73	7.67	1.97	6.599	0.028	7.492	0.034	0.893	0.046
BRAF	9.61	0.48	9.31	0.66	9.585	0.008	9.356	0.010	-0.229	0.012
MST1	10.3	1.95	9.49	1.03	10.182	0.033	9.524	0.015	-0.658	0.035
GSTP1	13.57	0.64	13.85	0.66	13.570	0.011	13.845	0.010	0.275	0.015
MAP3K1	10.1	1.34	9.22	0.87	9.700	0.017	9.375	0.015	-0.324	0.018
PRKCA	11.4	0.95	10.65	1.31	11.355	0.016	10.737	0.029	-0.618	0.035
MEF2C	12.19	1.87	13.65	0.96	12.682	0.043	13.527	0.017	0.845	0.048
MOS	7.85	1.07	7.32	1.33	7.788	0.018	7.425	0.022	-0.363	0.021
PDGFA	8.49	1.06	7.91	1.29	8.463	0.023	7.945	0.018	-0.518	0.031
RPS6KA5	11.75	0.53	12.05	0.57	11.787	0.009	12.004	0.009	0.217	0.013
MAP3K13	6.55	1.08	7.32	1.33	6.707	0.019	7.129	0.018	0.422	0.026
RASGRF1	8.26	0.95	7.81	1.14	8.251	0.016	7.815	0.022	-0.436	0.027
FAS	11.74	0.57	11.44	0.66	11.695	0.009	11.488	0.008	-0.207	0.013
TP53	12	0.62	11.66	0.94	11.971	0.009	11.720	0.015	-0.251	0.016
PPP2CA	13.69	0.43	13.48	0.59	13.686	0.009	13.494	0.009	-0.192	0.013
NLK	10.55	1.12	10.96	0.89	10.602	0.021	10.924	0.013	0.322	0.022
MAP3K11	11.61	0.62	11.83	0.52	11.621	0.010	11.824	0.009	0.203	0.014
MAP3K7IP2	12.4	0.53	12.65	0.64	12.414	0.008	12.627	0.012	0.213	0.015
FGF2	6.39	1.47	5.88	1.43	6.328	0.029	5.920	0.020	-0.408	0.029

Avg and Std are the sample average and standard deviation, respectively. MuEst and MuStd are batched mean and standard deviation of posterior samples, respectively.

Table 3.12: Summary of posterior samples in the model with Z-scores using intermediate AML tailored MAPK pathway

Gene	Zscore	Mean	StdDev	Min	Max	MeanDstd
PPP2CB	-2.796	-2.436	0.071	-3.7243	-0.0028	-34.3
MAPK1	-1.852	-1.851	0.062	-3.7248	-5e-04	-29.9
MAP3K1	3.442	3.102	0.108	0.0016	4.9835	28.7
MAP3K14	-2.464	-2.249	0.08	-3.7225	-4e-04	-28.1
JUND	-3.488	-2.684	0.096	-3.7249	-0.004	-28.0
MAP3K7IP2	-1.925	-1.918	0.075	-3.7248	-0.004	-25.6
PAK2	3.45	3.124	0.13	0.0025	4.9829	24.0
MAP4K4	2.224	2.297	0.103	0.0161	4.9752	22.3
RASGRF1	1.928	2.143	0.106	8e-04	4.9826	20.2
MAPT	-1.57	-1.689	0.084	-3.7239	-1e-04	-20.1
RPS6KA4	1.746	1.989	0.1	0.0029	4.9797	19.9
GSTP1	-1.89	-1.887	0.095	-3.7231	-0.0066	-19.9
GRB2	1.554	1.859	0.103	0.0012	4.9821	18.0
MAP2K1IP1	-1.327	-1.538	0.09	-3.722	-1e-04	-17.1
NRAS	1.714	1.996	0.129	9e-04	4.9771	15.5
MAP2K1	0.042	-1.049	0.068	-3.7235	-2e-04	-15.4
RPS6KA3	-1.23	-1.455	0.095	-3.7239	-6e-04	-15.3
MAX	2.978	-1.559	0.105	-3.7236	-6e-04	-14.8
PRKCA	2.915	-1.545	0.105	-3.7249	0	-14.7
HSPA1A	0.613	-1.222	0.083	-3.7237	-2e-04	-14.7

Zscore: normal transformed z from two-sample t test; Mean and StdDev are the batched mean and standard deviation of posterior sample; MeanDstd: ratio of bached mean and standard deviation.

Table 3.13: Summary of posterior samples in the two-sample model using intermediate AML tailored MAPK pathway

gene	TTest_pval	TT_Ord	Post_Diff	Post_Ord	AvgGrp1	StdGrp1	AvgGrp2	StdGrp2	MuEstGrp1	MuStdGrp1	MuEstGrp2	MuStdGrp2	MuEstDiff	MuStdDiff
CUTL1	0.00000	1	47.673	1	13.32	0.42	12.72	0.64	13.320	0.008	12.721	0.012	-0.599	0.013
PAK1	0.00040	2	0.158	31	9.34	1.24	10.29	0.99	9.850	0.014	9.850	0.014	0.000	0.000
JUND	0.00082	3	0.161	24	15.22	0.62	15.69	0.57	15.467	0.009	15.467	0.008	0.000	0.000
PAK2	0.00092	4	0.161	22	13.15	0.36	12.83	0.47	12.977	0.006	12.977	0.006	0.000	0.003
MAP3K1	0.00105	5	0.158	40	10.1	1.34	9.22	0.87	9.635	0.013	9.635	0.013	0.000	0.000
TGFB1	0.00165	6	0.282	8	12.2	0.98	12.91	0.92	12.553	0.094	12.602	0.081	0.049	0.174
MAX	0.00388	7	0.280	9	14.98	0.46	14.64	0.56	14.797	0.005	14.790	0.023	-0.007	0.023
STMN1	0.00464	8	0.000	54	12.74	0.56	12.36	0.57	12.539	0.007	12.539	0.007	0.000	0.000
PRKCA	0.00470	9	29.639	2	11.4	0.95	10.65	1.31	11.398	0.013	10.659	0.021	-0.739	0.025
PPP2CB	0.00652	10	0.010	53	13.04	0.5	13.38	0.58	13.219	0.006	13.219	0.014	0.000	0.012
NTRK2	0.01526	11	0.158	32	8.46	1	7.84	1.23	8.133	0.013	8.132	0.014	-0.001	0.004
MAP3K14	0.01678	12	0.161	23	9.09	1.22	9.65	0.72	9.389	0.010	9.389	0.010	0.000	0.000
MAP3K5	0.01820	13	0.217	15	12.3	0.67	12.71	0.86	12.520	0.011	12.520	0.011	0.000	0.000
BRAF	0.02517	14	0.000	65	9.61	0.48	9.31	0.66	9.453	0.007	9.453	0.007	0.000	0.000
MAP4K4	0.02910	15	0.159	26	12.21	0.61	11.89	0.66	12.039	0.007	12.039	0.007	0.000	0.000
RASGRF1	0.05758	16	0.158	29	8.26	0.95	7.81	1.14	8.020	0.011	8.020	0.011	0.000	0.000
MAP3K7IP2	0.05789	17	0.011	52	12.4	0.53	12.65	0.64	12.535	0.007	12.535	0.015	0.000	0.013
MAP3K8	0.06089	18	0.307	7	11.71	0.77	11.3	1.14	11.492	0.013	11.491	0.014	-0.001	0.005
GSTP1	0.06263	19	0.158	43	13.57	0.64	13.85	0.66	13.720	0.006	13.720	0.006	0.000	0.000
MAP2K4	0.06523	20	0.197	17	10.99	0.3	10.84	0.41	10.911	0.005	10.909	0.010	-0.002	0.009
MAPK1	0.06806	21	0.000	60	11.85	0.49	12.11	0.74	11.991	0.008	11.991	0.008	0.000	0.000
RPS6KA4	0.08485	22	0.000	55	11.34	0.67	11.06	0.74	11.193	0.008	11.193	0.008	0.000	0.000
PRKACA	0.08715	23	0.000	58	10.28	0.97	10.62	0.74	10.456	0.008	10.456	0.008	0.000	0.000
MAP3K11	0.08852	24	0.158	39	11.61	0.62	11.83	0.52	11.730	0.006	11.730	0.006	0.000	0.000
NRAS	0.09066	25	0.158	33	11.28	0.69	10.96	0.92	11.111	0.010	11.111	0.010	0.000	0.000

TTest: two-sample t test; Post_Diff: absolute value of ratio of batched mean and standard deviation of posterior difference; Avg and Std are the sample average and standard deviation, respectively. MuEst and MuStd are batched mean and standard deviation of posterior samples, respectively.

Table 3.14: Summary of posterior samples in the DG model using intermediate AML tailored MAPK pathway

gene	TTest_pval	TT_Ord	Post_Diff	Post_Ord	AvgGrp1	StdGrp1	AvgGrp2	StdGrp2	MuEstGrp1	MuStdGrp1	MuEstGrp2	MuStdGrp2	MuEstDiff	MuStdDiff
CUTL1	0.00000	1	41.789	1	13.32	0.42	12.72	0.64	13.305	0.007	12.746	0.011	-0.559	0.013
PAK1	0.00040	2	10.322	28	9.34	1.24	10.29	0.99	9.780	0.029	10.082	0.013	0.302	0.029
JUND	0.00082	3	25.365	4	15.22	0.62	15.69	0.57	15.250	0.013	15.658	0.008	0.408	0.016
PAK2	0.00092	4	22.389	8	13.15	0.36	12.83	0.47	13.113	0.007	12.872	0.008	-0.241	0.011
MAP3K1	0.00105	5	18.396	10	10.1	1.34	9.22	0.87	9.809	0.021	9.327	0.015	-0.482	0.026
TGFB1	0.00165	6	25.838	3	12.2	0.98	12.91	0.92	12.252	0.021	12.860	0.016	0.608	0.024
MAX	0.00388	7	22.728	6	14.98	0.46	14.64	0.56	14.950	0.007	14.669	0.010	-0.281	0.012
STMN1	0.00464	8	22.705	7	12.74	0.56	12.36	0.57	12.701	0.010	12.396	0.010	-0.305	0.013
PRKCA	0.00470	9	29.861	2	11.4	0.95	10.65	1.31	11.385	0.018	10.681	0.019	-0.704	0.024
PPP2CB	0.00652	10	24.082	5	13.04	0.5	13.38	0.58	13.045	0.008	13.371	0.009	0.326	0.014
NTRK2	0.01526	11	15.547	16	8.46	1	7.84	1.23	8.360	0.017	7.970	0.018	-0.390	0.025
MAP3K14	0.01678	12	8.388	34	9.09	1.22	9.65	0.72	9.340	0.026	9.581	0.012	0.241	0.029
MAP3K5	0.01820	13	17.071	13	12.3	0.67	12.71	0.86	12.343	0.011	12.645	0.013	0.302	0.018
BRAF	0.02517	14	17.806	12	9.61	0.48	9.31	0.66	9.594	0.009	9.343	0.010	-0.251	0.014
MAP4K4	0.02910	15	16.495	15	12.21	0.61	11.89	0.66	12.215	0.013	11.885	0.012	-0.330	0.020
RASGRF1	0.05758	16	16.627	14	8.26	0.95	7.81	1.14	8.256	0.016	7.816	0.019	-0.440	0.026
MAP3K7IP2	0.05789	17	18.078	11	12.4	0.53	12.65	0.64	12.412	0.008	12.625	0.010	0.213	0.012
MAP3K8	0.06089	18	7.417	36	11.71	0.77	11.3	1.14	11.663	0.016	11.400	0.025	-0.262	0.035
GSTP1	0.06263	19	13.736	19	13.57	0.64	13.85	0.66	13.587	0.012	13.830	0.010	0.243	0.018
MAP2K4	0.06523	20	19.015	9	10.99	0.3	10.84	0.41	10.990	0.005	10.837	0.006	-0.153	0.008
MAPK1	0.06806	21	15.353	17	11.85	0.49	12.11	0.74	11.874	0.008	12.062	0.010	0.188	0.012
RPS6KA4	0.08485	22	12.695	23	11.34	0.67	11.06	0.74	11.335	0.013	11.067	0.012	-0.268	0.021
PRKACA	0.08715	23	13.364	21	10.28	0.97	10.62	0.74	10.286	0.019	10.616	0.013	0.330	0.025
MAP3K11	0.08852	24	12.568	24	11.61	0.62	11.83	0.52	11.622	0.014	11.825	0.008	0.203	0.016
NRAS	0.09066	25	10.556	27	11.28	0.69	10.96	0.92	11.215	0.010	11.061	0.013	-0.154	0.015
AKT1	0.10248	26	11.772	26	11.74	0.45	11.55	0.53	11.732	0.010	11.560	0.009	-0.173	0.015
MAPT	0.12045	27	14.517	18	6.68	1.13	7.12	1.39	6.742	0.016	7.075	0.017	0.333	0.023
FGF2	0.12180	28	13.411	20	6.39	1.47	5.88	1.43	6.352	0.029	5.904	0.020	-0.448	0.033
HSPB2	0.14277	30	11.779	25	5.43	1.15	5.86	1.41	5.496	0.020	5.789	0.018	0.292	0.025
RPS6KA3	0.22245	35	13.346	22	12.01	0.63	12.2	0.72	12.009	0.011	12.198	0.011	0.189	0.014

TTest: two-sample t test; Post_Diff: absolute value of ratio of batched mean and standard deviation of posterior difference; Avg and Std are the sample average and standard deviation, respectively. MuEst and MuStd are batched mean and standard deviation of posterior samples, respectively.

Chapter 4

POST: A FRAMEWORK FOR SET BASED ASSOCIATION ANALYSIS IN HIGH DIMENSIONAL DATA

4.1 Introduction

Gene profiling with microarray technology has enabled investigators to simultaneously measure gene expression levels of thousands of genes in biological specimens. Subsequently, statistical analyses are performed to test association of individual measurements with an endpoint of interest. As thousands of tests are performed simultaneously, the problems posed by multiple testing should be addressed before declaring which list of genes/features that are associated with the endpoint of interest. Pounds and Cheng (2006)[39] reviewed methods to address the multiple testing problems for estimating and controlling the false discovery rates (FDR). Most of these FDR controlling methods assume that the *p*-values are independent or weakly dependent, an assumption which is often violated. Benjamini and Yekutieli (2001)[40] and Storey (2003)[41] have shown that small clumpy dependencies are usually negligible and the procedures of Benjamini and Hochberg (1995) and Storey (2001) methods have good properties under certain dependency structures. A gene may be represented by multiple probes and genes in a gene network/pathway tend to be co-regulated. In widely used Affymetrix expression array, genes are represented by 1 to more than 10 probes sets (features). Statistical analysis of these data often leave investigators a long list of genes/features that show significant association with an endpoint. A selection of the most promising candidates for follow-up depends heavily on the biology.

To facilitate the selection process, one strategy is to reduce the selection pool, namely, the list of features that are associated with an endpoint. Specifically, instead of studying association with an endpoint of interest at individual gene/feature level, one could focus on the association between gene set or pathways first. The benefits of doing so are several folds: by performing association tests at gene sets or pathway level first, the number of tests is significantly reduced; and statistical dependence between p-values could be reduced. These simplify the process of FDR controlling. Moreover, a shorter list of significant association with better biological knowledge is produced. After selecting significantly associated gene sets or pathways for follow-up, investigators could perform second round of association testing feature by feature in the selected genes or pathways and determine which functional forms to follow.

Recently, a few dozens gene-set procedures have been developed for testing differential expression in gene profiling data analyses. However, these methods are designed for differential expression and most are not suitable for association testing with complex modeling. In association testing, the data structure could be very complex with known variable adjustment, stratification, and multiple dependent endpoints that include continuous, binary, ordinal, or censored variables. Goeman and Buhlmann (2007)[8] and Nam and Kim (2008)[9] have provided an extensive review of these methods and made recommendations on self-containedness and randomization strategy for obtaining p-values. Gene set enrichment analysis (GSEA) can be applied to association testing using the feature level p-values, however this procedure lacks self-containedness. Constructing flexible self-contained association testing is challenging.

To address this challenge, we investigate the development of a procedure based on empirical orthogonal functions (EOF) analysis or principal component analysis (PCA). PCA has been widely used to identify spatial and temporal patterns in meteorology, genetic patterns, and population structure in gemome wide associations (GWAS). PCA is also widely used for dimension reduction in high dimensional datasets. Tomfohr *et al.* (2005)[42] used a t-test after reducing the gene set to its first principal component, and pointed out that PCA could be used to reduce the dimensionality of variables entered in the Hotellings T^2 statistic in two-sample multivariate testing with decent sample size.

In this chapter, we propose a new procedure, labeled Projection onto the Orthogonal Space Testing (POST) as a flexible method for identifying gene sets or pathways that show association with an endpoint of interest. POST performs a hypothesis test for each gene pathway, thus reducing the number of tests performed. In the process, this reduces potential dependence between p-values, leading to more accurate FDR control and less misleading follow-up study. POST is a multivariate testing procedure, and it is potentially more powerful than competing procedures for detecting association of genes or pathways in which genomic features are jointly related to an endpoint. Moreover, POST is flexible enough to test association with various endpoints under several model structures. In section 4.2, we describe the POST procedure and in section 4.3 evaluate the procedure by simulation studies. In section 4.4, we apply POST to several real datasets. Finally, Section 4.5 provides the discussion and concluding remarks.

4.2 The POST method

For $j = 1, 2, \ldots, k$, let S_j be a collection of pathways or gene sets based on data from n subjects. Suppose that S_j has m_j genomic features. Let Y_{ig} represent the value of genomic feature g for subject i, and let C_{il} represents the value of covariate l fro subject i, $i = 1, \ldots, n$. Denote by $\mathbf{C}_l$, the vector $(C_{1l}, \ldots, C_{nl})'$ and by $\mathbf{Y}_g$, the vector $(Y_{1g}, \ldots, Y_{ng})'$, $g = 1, \ldots, m_j$. The objective of the method to be proposed is to explain observed endpoints on the n subjects by the genomic features, after adjusting for the covariates. Towards this objective, let X_i be the value of an endpoint measured (observed) on subject i. Let $\mathbf{X} = (X_1, \ldots, X_n)'$. Finally, let $\boldsymbol{Y}_j = (\mathbf{Y}_1, \ldots, \mathbf{Y}_{m_j})$ be the $m_j \times n$ matrix of genomic features.

The variance-covariance matrix Σ_j of $\boldsymbol{Y}_j$

$$\Sigma_j = E\left[(\boldsymbol{Y}_j - E[\boldsymbol{Y}_j])(\boldsymbol{Y}_j - E[\boldsymbol{Y}_j])'\right] \tag{4.1}$$

will be estimated by its sample covariance matrix

$$\widehat{\Sigma_j} = (\boldsymbol{Y}_j - \bar{\boldsymbol{Y}_j})(\boldsymbol{Y}_j - \bar{\boldsymbol{Y}_j})' \tag{4.2}$$

where $\bar{\boldsymbol{Y}}_j = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Y}_j$, the column mean of $\boldsymbol{Y}_j$

We assume that Σ_j is positive definite with ordered eigenvalues $\lambda_{1_j} \geq \ldots \geq \lambda_{m_j}$, and corresponding eigenvectors $\mathbf{e}_{1_j}, \ldots, \mathbf{e}_{m_j}$.

Based on the fact that the largest eigenvalues explain most of the variation among the genomic values $\boldsymbol{Y}_j$, let $t_j \leq m_j$ represent an integer such that $\lambda_{1_j} \geq \ldots \geq \lambda_{t_j}$ explain $100\delta\%$ of the total variation in genomic variation for some predefined $0 < \delta < 1$. Let

$$\mathbf{P}_j = \left(P_{1_j}, \ldots, P_{t_j}\right) = \sum_{k=1}^{m_j} \mathbf{Y}'_k \mathbf{e}_{ik} = \boldsymbol{Y}'_j \mathbf{E}_j \tag{4.3}$$

where $\mathbf{E}_j$ is the matrix $(\mathbf{e}_{1_j}, \ldots, \mathbf{e}_{t_j})$.

For many gene sets or pathways with a large number of genomic features, t_j is usually small in comparison to m_j. Thus, the dimension of $\mathbf{P}_j$ is usually small, representing a significant reduction in data without much loss of information. For a given set of endpoints from the individual subjects, we now use this reduced genomic feature data to explain the variation in those endpoints by regressing the endpoint variables on the genomic features as independent variables, while adjusting for covariates $\mathbf{C}_l$.

This procedure is flexible enough to use linear, generalized linear regression or Cox proportional hazard model for time to event endpoints. Let Z_{r_j} be the Z-statistics from the model associating the endpoint variable $\mathbf{X}$ with $\mathbf{P}_{r_j}$ and let

$$\mathbf{Z}_j = \left(Z_{1_j}, Z_{2_j}, \ldots, Z_{t_j}\right)' \tag{4.4}$$

We wish to determine if set S_j has significant association with observed endpoints. Under the assumption that the projected vector with larger eigenvalues carry more information about the association with endpoints, a reasonable choice of statistic is the

$$\mathbf{T}_j = \sum_{r=1}^{t_j} \lambda_{r_j} Z_{r_j}^2 = \mathbf{Z}'_j \Lambda_j \mathbf{Z}_j \tag{4.5}$$

where Λ_j is the $t_j \times t_j$ diagonal matrix with $\lambda_{1_j}, \ldots, \lambda_{t_j}$ as diagonal elements.

The POST statistic will thus be defined by equation (4.5). We expect that under reasonable conditions, $\mathbf{T}_j$ will have an asymptotic null distribution which is a linear combination of Chi-squared distributions. An argument for this conjecture may be constructed as follow. Under null of no association between set S_j with endpoint $\mathbf{X}$, $\mathbf{Z}_j$ is a multivariate normal vector with mean $\mathbf{0}$ and variance-covariance matrix Σ_{Z_j}, $\mathbf{Z}_j \sim N\left(\mathbf{0}, \Sigma_{Z_j}\right)$. According to Duchesne and Lafaye de Micheaux (2010)[43], let matrix $\mathbf{C}$ be the Cholesky decomposition of Σ_{Z_j} satisfying $\mathbf{C}'\mathbf{C} = \Sigma_{Z_j}$ and $\mathbf{U}$ be such that $\mathbf{U}\mathbf{U}' = \mathbf{I}_{t_j}$ that diagonalizes $\mathbf{C}\Lambda_j\mathbf{C}'$ as $\mathbf{U}\mathbf{C}\Lambda_j\mathbf{C}'\mathbf{U}' = \mathbf{D} = diag(\lambda_1, \ldots, \lambda_{t_j})$. Assuming $\lambda_1 \geq \ldots \geq \lambda_r > 0$ and $\lambda_{r+1} = \ldots = \lambda_{t_j} = 0$ and letting $\mathbf{Y} = \mathbf{U}\mathbf{C}'^{-1}\mathbf{Z}_j$,

$$E\left(\mathbf{Y}\right) = E\left(\mathbf{U}\mathbf{C}'^{-1}\mathbf{Z}_j\right) = \mathbf{U}\mathbf{C}'^{-1}E\left(\mathbf{Z}_j\right) = \mathbf{0} \tag{4.6}$$

$$\begin{aligned} Var\left(\mathbf{Y}\right) &= \mathbf{U}\mathbf{C}'^{-1}\mathbf{Z}_j\left(\mathbf{U}\mathbf{C}'^{-1}\mathbf{Z}_j\right)' = \mathbf{U}\mathbf{C}'^{-1}\mathbf{Z}_j\mathbf{Z}_j'(\mathbf{C}'^{-1})'\mathbf{U}' \\ &= \mathbf{U}\mathbf{C}'^{-1}\Sigma_{Z_j}(\mathbf{C}'^{-1})'\mathbf{U}' = \mathbf{U}\mathbf{C}'^{-1}\mathbf{C}'\mathbf{C}\mathbf{C}^{-1}\mathbf{U}' \\ &= \mathbf{U}\mathbf{U}' = \mathbf{I}_{t_j} \end{aligned} \tag{4.7}$$

$\mathbf{Y}$ is distributed as $N\left(\mathbf{0}, \mathbf{I}_{t_j}\right)$. Each component of $\mathbf{Y}'\mathbf{D}\mathbf{Y}$ is a weighted standard χ^2 distribution, independent of the rest components as $\mathbf{D}$ is diagonal matrix. So, the distribution of $\mathbf{Y}\mathbf{D}\mathbf{Y}'$ is a weighted sum of chi-square random variables. We can show that $\mathbf{Y}'\mathbf{D}\mathbf{Y}$ and $\mathbf{Z}_j'\Lambda_j\mathbf{Z}_j$ are equivalent:

$$\begin{aligned} \mathbf{Y}'\mathbf{D}\mathbf{Y} &= \left(\mathbf{U}\mathbf{C}'^{-1}\mathbf{Z}_j\right)'\mathbf{D}\left(\mathbf{U}\mathbf{C}'^{-1}\mathbf{Z}_j\right) \\ &= \mathbf{Z}_j'(\mathbf{C}'^{-1})'\mathbf{U}'\mathbf{D}\mathbf{U}\mathbf{C}'^{-1}\mathbf{Z}_j \\ &= \mathbf{Z}_j'\mathbf{C}^{-1}\mathbf{U}'\mathbf{U}\mathbf{C}\Lambda_j\mathbf{C}'\mathbf{U}'\mathbf{U}\mathbf{C}'^{-1}\mathbf{Z}_j \\ &= \mathbf{Z}_j'\Lambda_j\mathbf{Z}_j \end{aligned} \tag{4.8}$$

So, the quadratic form in equation (4.5) can be expressed as a weighted sum of chi-square random variables:

$$\mathbf{T}_j = \mathbf{Z}_j' \Lambda_j \mathbf{Z}_j = \mathbf{Y}'\mathbf{D}\mathbf{Y} = \sum_{s=1}^{r} \lambda_s \chi_s^2 \tag{4.9}$$

Farebrother (1984)[44] and others have derived algorithm to calculate $Pr(\mathbf{T}_j > \mathbf{t}_j)$ for quadratic form in equation (4.9). In practice, we do not know the correlation structure of elements in $\mathbf{Z}_j$ and variance-covariance matrix Σ_{Z_j}. One way to get an estimate of Σ_{Z_j} is by bootstrap re-sampling. We sample $\mathbf{P}_j$ with replacement B times to get B bootstrap samples $\mathbf{P}_j^{*1}, \mathbf{P}_j^{*2}, \ldots, \mathbf{P}_j^{*B}$. For each bootstrap sample $\mathbf{P}_j^*$, parametric models are fit for each component of $\mathbf{P}_j^*$ to obtain $\mathbf{Z}_j^*$. We get $\mathbf{Z}_j^* = \left(\mathbf{Z}_j^{*1}, \mathbf{Z}_j^{*2}, \ldots, \mathbf{Z}_j^{*B}\right)$ and the estimate of Σ_{Z_j} is the variance-covariance matrix of $\mathbf{Z}_j^*$.

$$\widehat{\Sigma_{Z_j}} = cov(\mathbf{Z}_j^{*\prime}) \tag{4.10}$$

and $\mathbf{T}_j$ is approximated by

$$\mathbf{T}_j = \mathbf{Z}_j' \mathbf{A} \mathbf{Z}_j \approx \sum_{s=1}^{r} \widehat{\lambda}_s \chi_s^2 \tag{4.11}$$

Where, $\widehat{\lambda_s}$ are derived as above with $\widehat{\Sigma_{Z_j}}$ replacing Σ_{Z_j}.

In a biological system, a gene could be involved in multiple biological processes. It is very likely that gene sets defined for biological system or derived from biological databases are not disjoint. So, POST requires a multiple testing procedure that remains effective when the tests are dependent.

Yekutieli and Benjamni (1999)[45] proposed a resampling-based FDR control procedure for dependent test statistics. The FDR control method has been used in several gene set analysis methods. We also provide this method as an option to control FDR in POST analysis. Besides the B bootstrap samples to estimate Σ_{Z_j}, D permutations of the subject labeling in the original genomic data are performed for resampling-based FDR control. Let $\mathbf{T}_j$ denote the POST statistic in equation (4.5) computed from the j^{th} set. Let $\mathbf{T}_{jd}$ be the value of $\mathbf{T}_j$ computed using the d^{th} permuted samples. We follow the convention that permutation 1 represents the original data; thus, $\mathbf{T}_j = \mathbf{T}_{j1}$ for all $j = 1, \ldots, k$. For any $j = 1, \ldots, k$ and $d = 1, \ldots, D$, we obtain p_{jd} by same generalized chi-square approximation as in equation (4.11). For a given rejection region, $[0, p]$, the estimate of the FDR is given:

$$\widehat{FDR}(p) = min_{p',:p' \geq = p} \left(\frac{1}{D-1} \sum_{d=2}^{D} \frac{R_d(p')}{R_d(p') + S(p')} \right) \tag{4.12}$$

Where

$$R_d(p') = \sum_{j=1}^{k} I(p_{jd} \leq p') \tag{4.13}$$

and

$$S_d(p') = R_1(p') - \frac{1}{D-1} \sum_{d=2}^{D} R_d(p') \tag{4.14}$$

In some applications, the sets (pathway/genes) might be disjoint or have very weak overlap. In these cases, the resampling-based FDR controlling could be dispensable. False discovery rate can be estimated using methods such as FDR estimates by Benjamini and Hochberg (1995)[46] or robust FDR estimates by Pounds and Cheng (2006)[39].

The POST method consists following steps:

1. Calculate covariance matrix $\widehat{\Sigma_j}$ using feature level signals in set S_j and perform eigenvalue decomposition.

2. Select first few eigenvalues to explain δ fraction of total variance and project feature level data to selected orthogonal space to get projected data $\mathbf{P}_j$.

3. Perform parametric modeling using projected data to get association vector $\mathbf{Z}_j$.

4. Define the POST statistic to be quadratic form as in equation (4.5).

5. Bootstrap to estimate covariance $\widehat{\Sigma_{zj}}$ under null as in equation (4.10).

6. Determine *p*-value based on generalized chi-square.

7. Perform 1 to 6 steps for each set.

8. Perform multiple testing adjustments, either resampling-based or other FDR controlling methods.

4.3 Simulation studies

POST procedure can be applied to association analyses with various endpoints of interest including continuous, binary, categorical, and time to event endpoints. To compare the statistical power of POST procedure to that of other approaches, simulations were performed in a simple setting with two treatment groups, where the other approaches could be easily applied. Nine disjoint gene sets with sample size 20 in each of the two treatment groups were generated as in Table 4.1

In the nine hypothetical gene sets, three sets (A to C) were small gene sets with 10 members. Members in Set B had moderate increase of mean in group 2 compared to group 1. Two members in Set C had large increase of mean in group 2, and rest 8 members had no difference in mean expression level. Three sets (D to F) had moderate size (30 members) and set G and H had large size (100). In set H, 2 members had large mean difference, 5 had moderate mean difference and rest 93 members had no difference in mean expression level. Set I had a large size of 500. Variance-covariance matrix for each gene set or subset was drawn from a Wishart distribution with Toeplitz matrix and number of members plus 10 as the parameters. The variance-covariance with Toeplitz structure instead of Identity matrix was used to introduce correlation structure among genes within a gene set or subset. The variance-covariance matrixes were further scaled with diagonal

Table 4.1: Set sizes and group means of nine gene sets in a simulation study.

Gene Set	Size	Sub Size	Group 1	Group 2
SetA	10	10	0	0
SetB	10	10	0	1
SetC	10	2	0	3
		8	0	0
SetD	30	30	0	0
SetE	30	30	0	1
SetF	30	4	0	3
		8	0	0
SetG	100	100	0	0
SetH	100	2	0	3
	100	5	0	1
	100	93	0	0
SetI	500	500	0	0

elements around 1. For each gene set and subset, random samples were drawn from multivariate normal distribution with mean in table 4.1 and variance-covariance generated as above. One thousand simulated data sets were generated. The POST procedure with $\delta = 0.8, 0.95$ and 1, SAFE (Barry *et al.*, 2005), MRPP test (Nettleton *et al.* 2008) and GSA (Efron and Tibshirani, 2007) were applied to each of the 1000 data sets. The power and type 1 error of the four methods were summarized in table 4.2.

In the POST procedure, the choice of δ is arbitrary. In the simulation study, we choose δ to be 0.8 (at least 80% genetic variable is retained), 0.95 (retaining most genetic variation), and 1 (retaining all the genetic variation). In the four gene sets without differential expression, POST method maintained the nominal alpha level of 5% (3.3% to 5.1%); SAFE procedure also maintained the nominal alpha level of 5% (0.2% to 3%). However, MRPP procedure was slightly loose on nominal alpha level maintenance (3.9% to 6%), and GSA was more loose (5.2% to 7.7%). In the 5 gene sets with differential expression, set B and E have moderate differential expression across probe sets, POST, SAFE and MRPP test had good power to detect the differential expression with MRPP

Table 4.2: Summary of simulation results on 9 gene sets with sample size 20 in both groups.

GeneSet	Differential	POST			SAFE	MRPP	GSA
		$\delta = 0.8$	$\delta = 0.95$	$\delta = 1$			
SetA	0	5.1	5	5	2.2	6	5.8
SetD	0	5.2	5.1	5.1	3	5.6	7.7
SetG	0	3.3	3.3	3.3	1.9	3.9	5.7
SetI	0	4.4	4.4	4.4	0.2	4.8	5.2
SetB	1	92.7	93.3	93.4	81.8	93.6	19.4
SetC	1	87.5	88.3	88.1	6.4	100	7.1
SetE	1	94.4	94.7	94.6	87.7	95.3	24.6
SetF	1	81.7	84.1	83.7	5.2	100	5.4
SetH	1	15.3	25.4	23.2	3.5	83.3	4.7

Notes: Differential: 0: no differential expression, 1: differential expression between the two groups; For gene sets without differential expression, the false positive percentages (type 1 error) are shown; For gene sets with differential expression, the true positive percentages (power) are shown.

being 94.5%, POST (93.9%) and SAFE (84.8%). The 3 choices of δ for POST procedure gave similar power in these two gene sets. For gene set C and F, in which only a small portion genes had relatively high differential expression, SAFE method lost power to detect differential expression (5.8%), while POST still had decent power (85.6%). In set H (100 genes), both POST and SAFE significant lost power to detect differential expression, although POST performed much better than SAFE (15.3% to 25.4% vs 3.5%). In this gene set, the choice of $\delta = 0.95$ performed better than $\delta = 0.8$ and slightly better than $\delta = 1$. It seems that $\delta = 1$ retained noise association and $\delta = 0.8$ lost too much genetic information. In the three gene sets (C, F, and H), MRPP had greater power to detect differential expression. GSA had little power to detect differential expression in all the five gene set (4.7% to 24.6%).

From the simulation, MRPP had better power to detect gene sets with any differential expression, especially with large difference in part of the gene set with reasonable maintenance of nominal alpha level. This result is consistent with the fact that MRPP test is designed to detect any differential expression in multivariate spaces. Unfortunately, the MRPP test is hard to extend to complicated models such as adjusting for known factors or other types of endpoints of interest including censored time to event variable. GSA lacked power to detect differential expression with loose nominal alpha level control in this simulation study. GSA is an enrichment test and might not work well on data sets with small number of genomic features. Both POST and SAFE methods could be applied to complicated statistical modeling with various phenotypes. POST method performed better in all the nine gene sets than SAFE (nine types of gene sets) and the method showed robustness in choice of δ with large δ. Taking statistical power, nominal alpha level control and flexibility into account, POST method outperforms the other three methods.

4.4 Applications

The example application used data of a combined cohort from the St. Jude AML83, AML87, AML91 and AML97 clinical trials. Affymetrix U133A microarray was used to measure gene expression in the leukemic cells of diagnostic bone marrow samples of 105 subjects in this combined cohort (Ross *et al.*, 2004[47]): 7 subjects were from AML83, 27 subjects from AML87, 29 subjects from AML91 and 42 subjects from AML97. The clinical trials, sample selection, and method for gene expression profiling were described in Ross *et al.*, 2004. Normalized expression signals were determined using the Affymetrix Microarray Suite (MAS) 5.0 algorithm and log transformed to be better represented by normality. There were several presenting features available for each subject, such as cytogenetic karyotype, FAB subtype, race, white blood count (WBC) and age at diagnosis. In the original paper, the primary interest of the authors was to use gene expression profiling to discriminate the known major prognostic subtypes. Although the experiment was not designed for testing association of gene expression with treatment outcome, we were interested in the biological processes that are associated with treatment response or outcome. We were also interested in demonstrating the flexibility and utility of POST procedure in association studies and potentially use this method for our ongoing or future gene profiling studies in current AML trials. Several prognostic factors: core binding factors (CBF: inv(16) and t(8;21)), age at diagnosis, other 11q23 translocation, M7 without t(1;22) and FLT3-ITD, have been explored and are associated with clinical outcome in a more recent trial (Rubnitz *et al.*, 2010[48]). Some of these prognostic features need to be adjusted for in association with clinical outcome in gene profiling analyses.

In Affymetrix U133A annotations, 1057 biological processes are represented by at least 5 probe sets, and up to 2641 with mean 52 and median of 13 probe sets. We were interested in the biological processes that are associated with various clinical outcomes, such as event-free survival (EFS) and risk of relapse, or associated with presenting features such as core binding factor (CBF) vs other. EFS was defined as the time elapsed from enrollment to induction failure, withdrawal, relapse, secondary malignancy, or death, with those living and event-free censored at last follow-up. From the available methods, there are no methods able to deal with all these phenotypes using original expression data in a gene set level.

Association with survival outcome

In the first application, we applied POST procedure to survival analysis setting. We were interested in biological processes associated with EFS. In previous studies, CBF has been shown to be a favorable prognostic factor. Here, we were interested in the association adjusting for CBF and stratified by study protocols. The treatment protocols had tremendous effect on treatment outcome, especially the combined cohort spanned two decades. The treatment regimens and available drug were different among the trials, and supportive care were also improved in recent trials. We could perform Cox proportional hazard regression with feature level signals and CBF as predictors stratified by study protocols, then perform GSEA analysis on the obtained *p*-values, rank the biological processes and perform FDR control. Instead, we performed POST test, FDR control and subsequent feature level testing if needed. For each of biological process, variance-covariance of the feature-level signals was calculated and eigenvalue decomposition were performed. δ was set to 0.95, ie 95$\%$ variation of feature-level signals could be explained by selected eigenvalues. The feature-level signals were then projected to the selected orthogonal space. EFS was then modeled with Cox proportional hazards model with each projected vector and CBF as covariates, stratified by study protocols to obtain z statistics for each selected projected vector. The POST test statistics was calculated according to equation (4.5). The original projected vectors were re-sampled $B = 200$ times to obtain 200 random samples, which were then fit into the same Cox proportional hazards model to estimate the covariance of $\mathbf{Z}$. The p-value was determined by generalized chi-square using the algorithm by Farebrother (1984) implemented in R (R package: ”CompQuadForm”). The computing resource demanding was not heavy. The whole analysis could be completed in 8 hours with one central process unit (cpu).

Figure 4.1 showed the p-values of association of biological processes with EFS adjusting for CBF and stratified by treatment protocols. From QQ plot in the left panel, a few processes were on the border line above the null area of no association. Another indication from QQ plot was that the biological processes were not disjoints. Some of these biological processes could be associated with EFS. We noticed that in one biological process, the first projections explained at least 95$\%$ of variation. The numbers of PCAs were from 1 to 91 with mean 17 and median 9, significantly reduced dimensionality. There were no special patterns in p-values vs. number of PCAs used in the sets (Figure 4.1, left panel). This indicated that the POST procedure does not bias to sets with low or high numbers of PCAs.

The biological processes defined in the AFFY annotation were not disjoint, some probes were in multiple biological processes. So, the p-values were not independent. Most methods for FDR control assume independence of p-values, such as Benjamin Hochberg

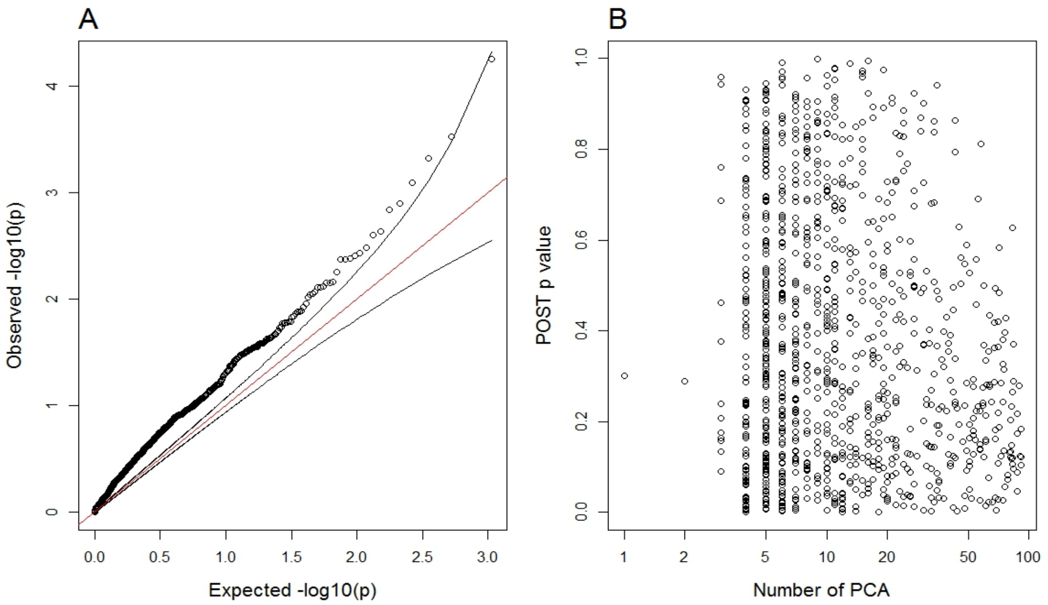
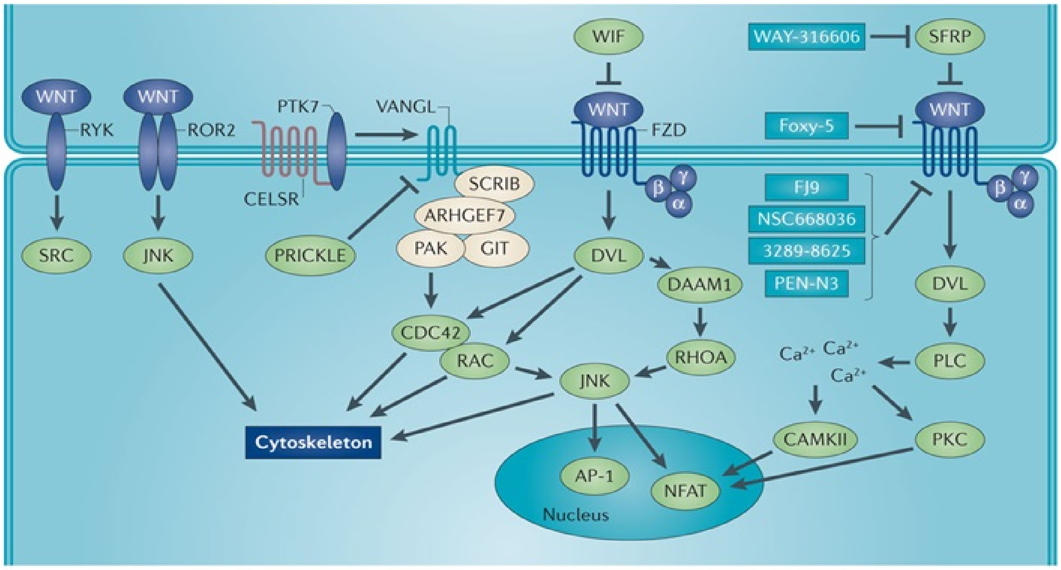


Fig. 4.1: Association of Biological Process with EFS in AML

Panel A: QQ plot of *p*-values of association of Biological Process with EFS. Panel B: POST *p*-values vs. number of PCAs in the tests.



Nature Reviews | Cancer

Fig. 4.2: CTNNB1-independent WNT signaling pathways.

Jamie N. Anastas & Randall T. Moon, Nature Reviews Cancer 13, 11-26 (January 2013)

Table 4.3: Top biological processes associated with EFS in AML

Biological Process	nProbe	nPCA	POSTStat	p-value	q-value
Regulation of Wnt receptor signaling pathway	18	12	82.1	5.6e-05	0.12
Mitochondrial electron transport; NADH to ubiquinone	29	18	31.9	3.0e-04	0.16
Coenzyme A biosynthesis	5	4	13.5	4.8e-04	0.19
Positive regulation of glucose import	5	4	17.6	8.1e-04	0.23
Establishment and/or maintenance of chromatin architecture	56	33	127.7	1.3e-03	0.25
Lipid catabolism	81	44	221.3	1.5e-03	0.29
Arginine catabolism	9	6	37.3	2.3e-03	0.31
Activation of JNK activity	20	14	50.7	2.5e-03	0.32
Nitric oxide mediated signal transduction	9	6	55.0	3.3e-03	0.33
Regulation of heart contraction	48	30	170.5	3.7e-03	0.34
Secretory pathway	10	7	25.1	4.0e-03	0.35
Response to toxin	13	10	59.2	4.2e-03	0.36
Morphogenesis	192	69	526.3	4.2e-03	0.37
Protein import into mitochondrial matrix	6	4	7.6	4.2e-03	0.37
Regulation of dephosphorylation	7	5	17.7	5.6e-03	0.37
Transmission of nerve impulse	5	4	27.3	6.9e-03	0.38
Positive regulation of gluconeogenesis	6	5	18.5	7.0e-03	0.38
Negative regulation of protein biosynthesis	18	12	42.9	7.0e-03	0.38
Negative regulation of cytokine and chemokine mediated signaling pathway	5	4	7.2	7.7e-03	0.39
Acetylcholine receptor signaling; muscarinic pathway	8	7	30.3	7.7e-03	0.39
Regulation of neuron differentiation	7	5	26.8	7.8e-03	0.39
Very-long-chain fatty acid metabolism	6	5	36.0	8.6e-03	0.40

Notes: Twenty-two biological processes are selected with FDR 0.4; important signal transduction pathways are highlighted in red.

1995. Here, the robust FDR estimation method proposed by Pounds and Cheng (2006) was used to control FDR.

With q-value (FDR) cut off 0.4, 22 biological processes were selected to be associated with EFS adjusting for CBF and stratified by treatment protocols (Table 4.3). Out of the 22 biological processes, five were signaling pathways. In particular, regulation of Wnt receptor signaling pathway was ranked as the number one among the 1057 pathways studied (p = 0.000056). Wnts are secreted glycoproteins that regulate multiple signaling pathways through both β-catenin (CTNNB1)-dependent and CTNNB1-indpendent mechanism (Figure 4.2). Wnt signaling pathway plays an important role in normal and leukemic hematopoietic stem cells and is an important target in several leukemogenic pathways (Mikesch *et al.*, 2007)[49]. Wnts and WNT pathway components are frequently over or under-expressed in different human cancers. WNT signaling pathway is well studied and has been shown involved in many development processes, and various types of cancer such as stomach, soft tissue pancreas, liver, ovary, kidney, and so on (reviewed by Anastas *et al.*, 2012[50]). It recently become one of the target pathways to develop therapeutic drugs. JNK singling pathway is another important pathway that is involved in apoptosis and cancer development (reviewed by Wagner and Nebreda 2009[51], Figure 4.3) and showed significant association with EFS in AML (p = 0.0025). JNK and p38 mitogen-activated protein kinases have import roles in signaling mechanism that regulates cellular responses to stresses, cell proliferation, survival in a cell-type specific manner. Their expressions and activities are altered in human tumors and cancer cells. Several phase I and II clinical trials are testing drugs directly targeting JNK in multiple cancers. The identification of regulation of JNK activity process in association with survival outcome further implicates the importance of JNK pathway. On the other hand, the identification also indicates the utility of the method.

Nitric oxide mediated signal transduction was ranked number 9 among the 1057 biological processes associated with EFS in AML. Nitric oxide has mixed effect in

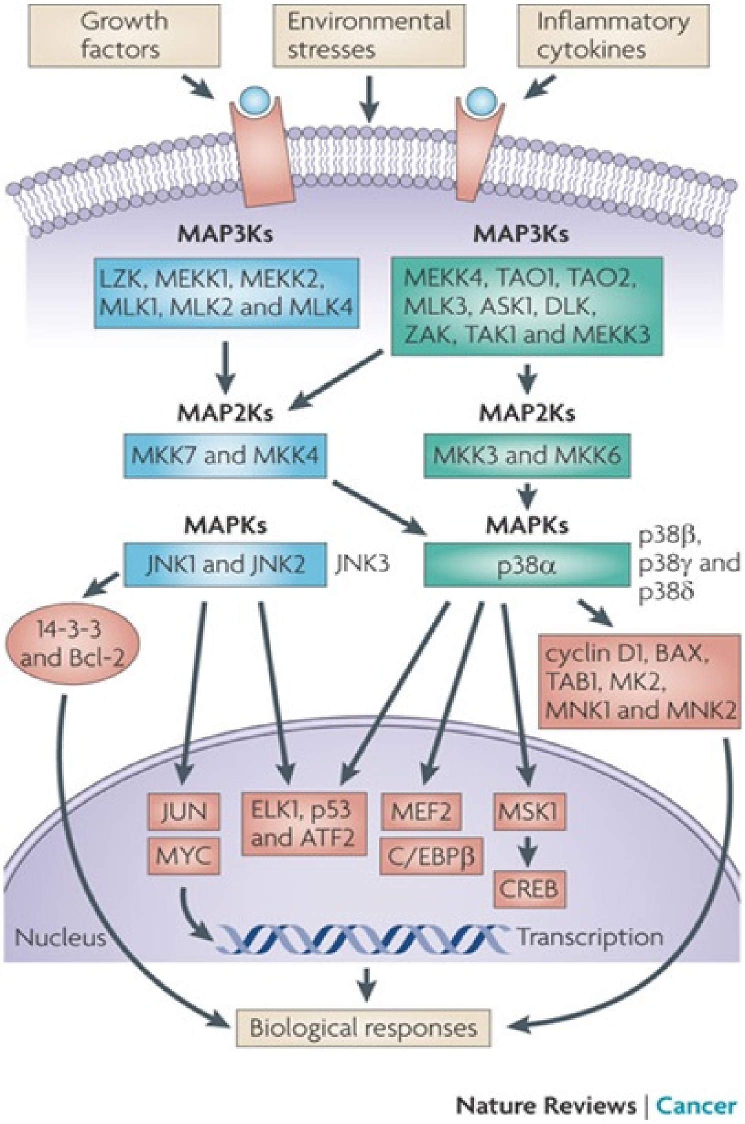


Fig. 4.3: Human JNK signaling pathways.

Erwin F. Wagner & Angel R. Nebreda Nature Reviews Cancer 9, 537-549 (August 2009) carcinogenesis. It could both cause DNA damage and protect cells from cytotoxicity; could inhibit and stimulate cell proliferation; and could be both pro- and anti-apoptotic (Hussain *et al.*, 2003[52]). The biological process of establishment and/or maintenance of chromatin architecture was also significantly associated with EFS (ranked #5, p = 0.0013). Chromosomal trans-location, inversions, chromosomal deletion or amplification were frequently observed in many cancers. Nambiar *et al.* (2008[53]) provided an extensive review on chromosomal translocations in AML, ALL and other more than 20 cancers. These important cancer related pathways were picked up in the POST analysis, further indicating the utility of POST procedure.

Association with categorical features

In the second application, we investigated the association of biological processes with core binding factors (CBF). The subjects were classified as with CBF (1) or without CBF (0). Using logistic regression model with study protocol as one of the covariates, we performed POST analysis similar to EFS to explore the association between biological processes with CBF.

With FDR 0.1, 113 biological processes were selected to be significantly associated with CBF. Out of the 113 biological processes, 26 were related to signaling transduction pathways or regulating a biological process (Table 4.4). These included many important signaling transduction pathways such as: cell surface receptor linked signal transduction, transmembrane receptor protein tyrosine kinase signaling pathway, integrin-mediated signaling pathway, intracellular signaling cascade, cell-cell signaling, and regulation of transcription through various mechanisms. These results suggest that AMLs with CBF and those without CBF are dramatically distinct diseases in term of underlying disease biology.

Association with time to events with competing events

Despite significant progress in treating patients with pediatric acute myeloid leukemia (AML), 20%-33% of patients still experienced relapse. Here, we were interested

Table 4.4: Summary of signaling and regulation biological processes associated with CBF in AML.

Biological Process	nProbe	nPCA	POSTstat	p-value	q-value
Regulation of cell growth	152	60	251.9	0.0000	0.000
Androgen receptor signaling pathway	70	35	116.3	0.0000	0.000
Regulation of apoptosis	177	57	261.5	0.0000	0.000
Positive regulation of I-kappaB kinase/NF-kappaB cascade	149	53	249.3	0.0000	0.000
Regulation of transcription	646	84	2469.1	0.0000	0.000
Negative regulation of progression through cell cycle	222	67	574.7	0.0000	0.000
Negative regulation of transcription; DNA-dependent	71	35	67.2	0.0000	0.000
Positive regulation of transcription from RNA polymerase II promoter	80	36	678.5	0.0000	0.000
Regulation of transcription; DNA-dependent	2641	91	7417.5	0.0000	0.000
Regulation of transcription from RNA polymerase II promoter	357	76	711.8	0.0000	0.000
Regulation of translation	104	43	102.3	0.0000	0.000
Signal transduction	2347	91	6949.9	0.0000	0.000
Cell surface receptor linked signal transduction	310	74	734.3	0.0000	0.000
Transmembrane receptor protein tyrosine kinase signaling pathway	146	61	542.7	0.0000	0.000
G-protein coupled receptor protein signaling pathway	679	87	1953.6	0.0000	0.000
Integrin-mediated signaling pathway	102	49	496.8	0.0000	0.000
Intracellular signaling cascade	614	83	1540.2	0.0000	0.000
Small GTPase mediated signal transduction	318	73	621.7	0.0000	0.000
Cell-cell signaling	445	82	1797.5	0.0000	0.000
Regulation of progression through cell cycle	432	78	1242.9	0.0000	0.000
Positive regulation of cell proliferation	203	67	645.6	0.0000	0.000
Negative regulation of cell proliferation	284	74	727.1	0.0000	0.000
Negative regulation of lymphocyte proliferation	6	1	9.8	0.0018	0.017
Negative regulation of JNK activity	6	5	0.1	0.0023	0.023
Glutamate signaling pathway	21	17	6.7	0.0071	0.067
DNA damage response; signal transduction resulting in induction of apoptosis	8	7	0.6	0.0086	0.080

Out of the 113 biological processes that associated with CBF, 26 are signaling andor regulation biological processes.

in biological processes that are associated with risk of relapse. Risk of relapse was defined as time elapsed from enrollment to relapse with induction failure, withdrawal, secondary malignancy and death treated as competing events, patients without any event were censored at last follow up. For each biological process, the time to relapse was modeled with each projected vector in the biological process, CBF, treatment protocol as independent predictors in Fine and Grays (1999) competing risk regression model, which is similar to EFS.

Table 4.5 showed the top 20 biological processes that were associated or risk of relapse, with a false discovery rate of 0.45. Out of the 20 biological processes, three were important signaling pathways: Nitric oxide mediated signal transduction (p = 0.0016), Regulation of Wnt receptor signaling pathway (p = 0.0037), and Activation of JNK activity (p = 0.0051). These three signal transduction pathways were also associated with event-free survival in AML.

4.5 Discussion

POST is a general procedure designed for set-based association studies. It is very flexible and can be adapted to many types of endpoints of interest. In the example applications in section 4.4, we demonstrated how to perform POST test for binary endpoint (logistic regression) and time to event endpoint (Cox proportional hazard model). Applying POST to continuous normal endpoint is trivial (linear regression). POST also could be applied to time to event endpoint with competing events. In this case, Fine and Gray (1999) competing risk regression can be applied to model event of interest with competing events after orthogonal projection. In fact, any parametric models with z-type of statistics (standard normal under null) could be applied to the POST test. After orthogonal projection, parametric modeling is applied to each projected vector. Certainly, non-parametric model could also be used as long as the statistics for each projected vector is standard normal under null. So, POST can handle most types of endpoints of interest in practice, which makes POST procedure very attractive.

Table 4.5: Summary of biological processes associated with risk of relapse in AML.

Biological Process	nProbe	nPCA	POSTstat	p-value	q-value
Positive regulation of glucose import	5	4	18.9	0.00053	0.35
Nitric oxide mediated signal transduction	9	6	64.8	0.00160	0.35
Mitochondrial electron transport; NADH to ubiquinone	29	18	28.4	0.00160	0.35
Phagocytosis; engulfment	8	6	54.8	0.00240	0.35
Regulation of dephosphorylation	7	5	20.7	0.00250	0.35
Arginine catabolism	9	6	37.4	0.00260	0.36
Protein import into mitochondrial matrix	6	4	8.3	0.00270	0.37
Positive regulation of gluconeogenesis	6	5	22.3	0.00280	0.37
Acetyl-CoA biosynthesis	5	5	11.9	0.00320	0.39
SRP-dependent cotranslational protein targeting to membrane	6	4	26.0	0.00340	0.41
Regulation of Wnt receptor signaling pathway	18	12	50.7	0.00370	0.41
Activation of JNK activity	20	14	48.6	0.00510	0.42
Fructose 2;6-bisphosphate metabolism	6	5	22.4	0.00510	0.44
Very-long-chain fatty acid metabolism	6	5	40.3	0.00530	0.45
Negative regulation of protein biosynthesis	18	12	47.0	0.00600	0.45
Transmission of nerve impulse	5	4	28.5	0.00710	0.45
Insulin secretion	5	4	9.2	0.00770	0.45
Secretory pathway	10	7	23.0	0.00780	0.45
Lipid catabolism	81	44	208.0	0.00810	0.45
Porphyrin biosynthesis	13	8	27.8	0.00850	0.45

Empirical orthogonal function projection or PCA projects the original data in a set to an orthogonal subspace spanned by eigenvectors. In each projected vector, the variation is maximized besides maintaining orthogonality. This potentially increases the power of detecting significant association of a set with an endpoint of interest, especially in the circumstance where feature level data in a set are weakly associated with the endpoint marginally but jointly show strong association. Gene sets such as pathways could be arbitrary with wide range in sizes. The dimensions are significantly reduced without much loss of information and potentially remove noise after orthogonal projection. In the example applications, some of the biological processes are with sizes of hundreds to 2000, which could be reduced to the sizes of dozens to a hundred with 95% variation among features retained. The choice of δ is arbitrary and should be predefined before the analysis. As shown in the simulation study, POST method is robust to the choice of δ for most gene sets as it is greater than 90%. The set definitions are usually derived from available databases or prior knowledge, and should be determined before the analysis. We do not suggest modifying set definition during analysis. Theoretically, any set can be shown significant association after certain modification. Modifying set definitions during analysis makes the test's validity questionable.

POST also has other desirable attributes. POST is self-contained. POST test for one set will not be influenced by genes in other gene sets. Geoman and Buhlmann 2007 pointed out that self-contained null hypothesis testing in gene set analysis gives advantage of valid *p*-values and easy interpretability. The POST test is model based and does not require permutation to determine *p*-values, although it does require resampling technique to estimate one parameter. The POST test is still computationally efficient. In the example application of association with EFS, the whole POST procedure took about 8 hours on a single CPU with 105 subjects and 200 bootstraps.

POST test statistic is defined as a quadratic form with the corresponding eigenvalues as diagonal elements. This choice of weight assumes that more variation among features in a set carries more information of association with an endpoint of interest. In most cases, this is a reasonable assumption. However, this assumption need to be further investigated in some circumstances. Alternative weighting strategy could be used and should be determined before analysis. We do not recommend searching for optimal weighting to extend POST in practice. We do not recommend different weighting schema for different sets in one application either.

To derive the generalized Chi-square distribution of POST statistics, we assume that vector $\mathbf{z}$ is a multivariate normal vector with mean $\mathbf{0}$ and an unknown variance-covariance matrix under the null hypothesis of no association between the set and an endpoint of interest. This assumption is valid in general. We need to use bootstrap resampling to estimate the unknown variance-covariance matrix. Usually, 200 or more bootstrap samples are desirable. As resampling technique used, we suggest applying POST procedure to data set with decent sample size. POST is design for test association in large clinical or biomedical studies. We do not recommend applying POST test to data set with less than 30 subjects.

POST is motivated by analysis of gene profiling data generated by microarray chips. However, POST is not limited to gene profiling data analysis. It also can be applied to other high dimension data as long as the data can be assumed to be normally distributed or after normal transformation in predefined sets. After careful data preparation, POST procedure can be applied to RNA-seq data for gene profiling, DNA methylation data either from next generation sequencing or from methylation array for studying epigenetic effects. Currently, we are trying to apply POST procedure to DNA methylation data.

In summary, POST is a general, very flexible procedure for association analysis in high dimension data. It can be easily adapted to various types of endpoints and data generation mechanisms.

Chapter 5

LOCIT: A FRAMEWORK FOR LOCUS-BASED INTEGRATED ASSOCIATION TEST IN HIGH DIMENSIONAL DATA

5.1 Introduction

High throughput technologies have enabled researchers to study thousands of genomic features of living organisms. Starting from gene expression profiling studies using microarray technology, high throughput technologies have been developed to understand the biological mechanisms of diseases and other biological phenomena.

Micro RNA (miRNA) is a small non-coding RNA molecule, which functions in transcriptional and post-transcriptional regulation of gene expressions. The human genome encodes over 1000 miRNA which may target about 60% of genes (Bentwich, *et al.*, 2005)[10]. miRNA are abundant in many cell types and are involved in many biological processes. Expression levels of miRNA are measured by microchips and by direct sequencing techniques. Epigenetic is a phenomenon in which gene expression and other cellular phenotypes are influenced by mechanisms other than changes in the underlying DNA sequence, such as histone modification, DNA methylation and RNA editing. Histone modifications are studied using ChIP-Chip method (chromatin imunoprecipitation with microarray technology, Lieb *et al.* 2001[11]) and, more recently, Chip-seq (based on next-generation sequencing technology, Johnson *et al.*, 2007[12]). DNA methylation levels are measured by microchip such as Illumina Infinium Methylation array, or pyrosequencing. It has been shown that DNA methylation variations are associated with multiple complex diseases. SNP arrays and direct sequencing have been widely used to study germline or disease polymorphisms associated with disease predisposition or treatment outcomes.

For many diseases and other biological outcomes, a variety of data are usually collected. For example, gene expression, DNA methylation, micro RNA, and SNP data could be available for each subject in a study cohort. To analyze these data, traditional association testing could be performed to relate each data type with an endpoint of interest. For each form of association procedure, FDR controls are usually performed to account for multiplicity of tests, before declaring a list of genes to be significantly associated with an endpoint of interest. Following such testings, the result for each data type can be used by investigators to obtain an overlapping list of genes. It is not uncommon to find that there is little overlaps between the genes on various lists. Given all forms of data, one question that investigators tend to ask is which genes are significantly associated with the endpoint of interest and should be followed in a future study. If there are some overlaps between result lists, the FDR control of the multiple testing for the significance of such overlaps is usually quite challenging. To address this challenge, an integrated association test approach is needed.

One approach that has been widely adopted in various fields, including meteorology to identify spatial and temporal patterns, genetic patterns analysis, and identification population structure in GWAS, is use of empirical orthogonal functions (EOF) analysis or principal component analysis (PCA). The important property of EOF and PCA is dimension reduction for high dimensional data. For a gene or locus set, the dimension are usually large. EOF and PCA are dimension reduction tools which also capture most information of the data.

In this chapter, we propose another approach: the Locus-based Integrated Test (LOCIT), as a flexible statistical procedure to test association of a locus with the endpoint of interest given multiple sources of high dimensional data. In the LOCIT procedure, we perform one hypothesis testing with multiple sources of data for a locus. This reduces number of tests and therefore results in better FDR control. As will be shown, the LOCIT procedure is flexible and can handle data from various endpoints to be adapted to different model structures. In section 5.2, we describe the LOCIT procedure. Section 5.3 presents the results from real applications and Section 5.4 presents simulation studies. Finally, Section 5.5 provides the discussion and concluding remarks.

5.2 Methods

5.2.1 The LOCIT method

Suppose that $g = 1_{cj}, \ldots, w_{cj}$ genomic features from source $c = 1, \ldots, m$ in the j^{th} locus S_j for $j = 1, \ldots, k$ are measured for $i = 1, \ldots, n$ subjects. Also, suppose that data of endpoint variables and $l = 1, \ldots, v$ covariates are available for these subjects. For $i = 1, \ldots, n$, $g = 1_{cj}, \ldots, w_{cj}$ and $c = 1, \ldots, m$, let y_{ig} represent the value of genomic feature g for subject i. Let $\mathbf{y}_g$ represent the vector $(y_{1g}, y_{2g}, \ldots y_{ng})$ of values for genomic variable g for all subjects and let $\mathbf{Y}_j$ represent the set of all $\mathbf{y}_g$ for locus S_j. For $l = 1, \ldots, v$, let q_{il} represent the value of covariate l for subject i, and $\mathbf{Q}_l$ be vector of $(q_{1l}, q_{2l}, \ldots q_{nl})$. Additionally, let x_i represent the value of endpoint for subject i and $\mathbf{X}$ represent the vector $(x_1, \ldots, x_n)$.

For $g = 1_{cj}, \ldots, w_{cj}$, we perform traditional parametric testing with endpoint variables $\mathbf{X}$ as dependent variables and $\mathbf{y}_g$ as independent variables, adjusting for covariates $\mathbf{Q}_l$. The model structure could be linear, generalized linear, or Cox proportional hazard model when the endpoint variables are survival times. Let z_g be the z-statistics retrieved from the model measuring association between gene expression measurements $\mathbf{y}_g$ and endpoint variable $\mathbf{X}$, and let

$$\mathbf{z}_j = \left(z_{1_{1j}}, z_{2_{1j}}, \ldots, z_{w_{1j}}, \ldots, z_{w_{mj}}\right)' \tag{5.1}$$

be the vector of z-statistics measuring association of $\mathbf{y}_g$ for $g = 1_{cj}, \ldots, w_{cj}$ and $c = 1, \ldots, m$ with the endpoint $\mathbf{X}$. The main interest is whether the locus S_j, for $j = 1, \ldots, k$, has significant association with the endpoint of interest (with adjustment made for covariates). Under the assumption that all the genomic features in the locus carrying same information of association with the endpoint, we use as a test statistic

$$\mathbf{T}_j = \mathbf{z}_j'\mathbf{z}_j \tag{5.2}$$

to measure the association between S_j with $\mathbf{X}$. We label this statistic as LOCIT statistic.

Next we need to determine the p-value of observing as extreme as $\mathbf{t}_j$ under null hypothesis where S_j is not associated with endpoint: $Pr(\mathbf{T}_j > \mathbf{t}_j)$. Under null of no association between set S_j with endpoint $\mathbf{X}$, $\mathbf{z}_j$ is a multivariate normal vector $N\left(\mathbf{0}, \Sigma_{z_j}\right)$ with mean $\mathbf{0}$ and variance-covariance matrix Σ_{z_j}. According to Duchesne and Lafaye de Micheaux (2010), let matrix $\mathbf{C}$ be the Cholesky decomposition of Σ_{z_j} satisfying $\mathbf{C}'\mathbf{C} = \Sigma_{z_j}$ and $\mathbf{U}$ be such that $\mathbf{U}\mathbf{U}' = \mathbf{I}_{w_{mj}}$ and that diagonalizes $\mathbf{CIC}'$, $\mathbf{UCIC}'\mathbf{U}' = \mathbf{D} = diag(\lambda_1, \ldots, \lambda_{w_{mj}})$. Assuming $\lambda_1 \geq \ldots \geq \lambda_r > 0$ and $\lambda_{r+1} = \ldots = \lambda_{w_{mj}} = 0$ and letting $\mathbf{Y} = \mathbf{UC}'^{-1}\mathbf{Z}_j$, $\mathbf{Y}$ is distributed as $N\left(\mathbf{0}, \mathbf{I}_{w_{mj}}\right)$. The quadratic form in equation (5.2) can be expressed as a weighted sum of chi-square random variables:

$$\mathbf{T}_j = \mathbf{z}_j'\mathbf{z}_j = \mathbf{Y}'\mathbf{D}\mathbf{Y} = \sum_{s=1}^{r} \lambda_s \chi_s^2 \tag{5.3}$$

Farebrother (1984) and others have derived algorithm to calculate $Pr(\mathbf{T}_j > \mathbf{t}_j)$ for quadratic form in equation (5.3). In practice, we do not know the correlation structure of elements in $\mathbf{z}_j$ and variance-covariance matrix Σ_{z_j}. One way to get an estimate of Σ_{z_j} is by bootstrap re-sampling. We sample $\mathbf{Y}_j$ with replacement B times to get B bootstrap samples $\mathbf{Y}_j^{*1}, \mathbf{Y}_j^{*2}, \ldots, \mathbf{Y}_j^{*B}$. For each bootstrap sample $\mathbf{Y}_j^{*}$, parametric models are fit for each component of $\mathbf{Y}_j^{*}$ to obtain $\mathbf{z}_j^{*}$. We get $\mathbf{Z}_j^{*} = \left(\mathbf{z}_j^{*1}, \mathbf{z}_j^{*2}, \ldots, \mathbf{z}_j^{*B}\right)$ and the estimate of Σ_{z_j} is the variance-covariance matrix of $\mathbf{Z}_j^{*}$

$$\widehat{\Sigma_{z_j}} = cov(\mathbf{Z}_j^{*\prime}) \tag{5.4}$$

and $\mathbf{T}_j$ is approximated by

$$\mathbf{T}_j = \mathbf{z}_j'\mathbf{z}_j \approx \sum_{s=1}^{r} \widehat{\lambda}_s \chi_s^2 \tag{5.5}$$

Where, $\widehat{\lambda}_s$ are derived as above with $\widehat{\Sigma_{z_j}}$ replacing Σ_{z_j}.

In Biological system, genes do not function by their own, and depend on each other to function as biological processes. So, loci could have profound dependence and LOCIT requires a multiple testing procedure that remains effective when the tests are dependent.

Yekutieli and Benjamni (1999) proposed a resampling-based FDR control procedure for dependent test statistics. The FDR control method has been used in several gene set analysis methods. We also provide this method as an option to control FDR in LOCIT analysis. Besides the B bootstrap samples to estimate Σ_{Z_j}, D permutations of the subject labeling are performed for resampling-based FDR control. Let $\mathbf{T}_j$ denote the LOCIT statistic in equation (5.2) computed from the j^{th} locus. Let $\mathbf{T}_{jd}$ be the value of $\mathbf{T}_j$ computed using the d^{th} permuted samples. We follow the convention that permutation 1 represents the original data; thus, $\mathbf{T}_j = \mathbf{T}_{j1}$ for all $j = 1, \ldots, k$. For any $j = 1, \ldots, k$ and $d = 1, \ldots, D$, we obtain p_{jd} by same generalized chi-square approximation as in equation (5.5). For a given rejection region, $[0, p]$, the estimate of the FDR is given:

$$\widehat{FDR}(p) = min_{p',:p' \geq p} \left(\frac{1}{D-1} \sum_{d=2}^{D} \frac{R_d(p')}{R_d(p') + S(p')} \right) \tag{5.6}$$

Where

$$R_d(p') = \sum_{j=1}^{k} I(p_{jd} \leq p') \tag{5.7}$$

and

$$S_d(p') = R_1(p') - \frac{1}{D-1}\sum_{d=2}^{D} R_d(p') \tag{5.8}$$

The FDR also could be controlled by Benjamini and Yekutieli (2001) method with dependency. In applications that loci are weakly correlated, the resampling-based FDR controlling could be dispensable. False discovery rate could be estimated using methods such as FDR estimates by Benjamini and Hochberg (1995) or robust FDR estimates by Pounds and Cheng (2006).

The LOCIT method is implemented by following steps:

1. Perform parametric modeling to get association vector $\mathbf{z}_j$.
2. Define the LOCIT statistic to be quadratic form as in equation (5.2).
3. Bootstrap to estimate covariance $\widehat{\Sigma_{zj}}$ under null as in equation (5.4).
4. Determine *p*-value based on generalized chi-square.
5. Perform same 1 to 4 steps for each locus.
6. Perform multiple testing adjustments, either resampling-based or other FDR controlling methods.

5.2.2 The LOCIT extension

The LOCIT test assumes that most of the genomic features in a locus show similar association with endpoint of interest. However, this is usually not the case. The LOCIT test with equal weight puts high penalty on a locus where most genomic features are not associated with the endpoint of interest. One potential strategy is to use different weights to define the LOCIT statistics. Another strategy is to perform feature selection prior to LOCIT test. It is also possible to combine these two strategies.

We implement a strategy of using different weights in defining the LOCIT statistics. As there are more than one form of genomic features for a given locus, we assume that each form of genomic feature makes the same contribution to the overall association with an endpoint of interest. For this purpose, the weights are simply set as 1. If there are more than one genomic features in a given form, an orthogonal projection is performed on a sub-orthogonal space that explains predefined fraction of total variation of the features and the weights are assigned according to the eigenvalues. This can be regarded as an extension of POST to multiple forms of genomic data in a locus. We label this procedure LOCITO (LOCIT with orthogonal projection).

Given the setting described in section 5.2.1, suppose there are $c = 1, \dots, m$ forms of genomic data. For $c = 1, \dots, m$, we perform following: For $g = 1_{cj}, \dots, w_{cj}$and $w_{cj} > 1$, compute a sample estimate of covariance $\widehat{\Sigma_{cj}}$ using equation (5.9). Now apply an eigenvalue decomposition to $\widehat{\Sigma_{cj}}$ to obtain eigenvalues: $\lambda_{1_{cj}}, \dots, \lambda_{w_{cj}}$ in descending order and corresponding eigenvectors: $\mathbf{e}_{1_{cj}}, \dots, \mathbf{e}_{w_{cj}}$. Let $t_{cj} \leq w_{cj}$ represent the least number of eigenvalues explaining a predefined fraction, $0 < \delta \leq 1$, of total variation in the genomic variables for cj, and project $\mathbf{Y}_{cj}$ to the orthogonal subspace spanned by eigenvectors $\mathbf{e}_{1_{cj}}, \dots, \mathbf{e}_{t_{cj}}$ as given in equation (5.10). The selected eigenvalues $\Lambda_{cj} = (\lambda_{1_{cj}}, \dots, \lambda_{t_{cj}})'$ are rescaled according to equation 5.11. If $w_{cj} = 1$, $\mathbf{P}_{cj}$ is the original measurement of the genomic feature and $\lambda'_{cj} = 1$.

$$\widehat{\Sigma_{cj}} = (\mathbf{Y}_{cj} - \bar{\mathbf{Y}_{cj}})(\mathbf{Y}_{cj} - \bar{\mathbf{Y}_{cj}})' \tag{5.9}$$

$$\mathbf{P}_{cj} = \left(P_{1_{cj}}, \dots, P_{t_{cj}}\right) = \mathbf{Y}'_{cj}\left(\mathbf{e}_{1_{cj}}, \dots, \mathbf{e}_{t_{cj}}\right) \tag{5.10}$$

$$\Lambda'_{cj} = \Lambda_{cj} / \sum \Lambda_{cj} \tag{5.11}$$

Let $\mathbf{P}_j = (\mathbf{P}_{1j}, \dots, \mathbf{P}_{mj})$, $\Lambda'_j = \left(\Lambda'_{1j}, \dots, \Lambda'_{mj}\right)$ and $\mathbf{\Lambda}'_j$ be a diagonal matrix with Λ'_j as diagonal elements. We perform traditional parametric testing with endpoint variable

$\mathbf{X}$ as dependent variable and each projected vector of $\mathbf{P}_j$ as independent variable, adjusting for covariates $\mathbf{Q}_l$. Let

$$\mathbf{z}_j = \left(z_{1j}, \ldots, z_{t_{1j}}, \ldots, z_{t_{mj}}\right) \tag{5.12}$$

be the vector of z-statistics measuring association of each projected vector in $\mathbf{P}_j$ with the endpoint variable $\mathbf{X}$. The statistics measuring association between j^{th} locus with the endpoint variable $\mathbf{X}$ is defined as in equation 5.13

$$\mathbf{T}_j = \mathbf{z}_j' \mathbf{\Lambda}_j \mathbf{z}_j \tag{5.13}$$

The $Pr(\mathbf{T}_j > \mathbf{t}_j)$ is approximated by generalized chi-square approximation as in equation (5.5). Resampling-based FDR control procedure can be carried out according to above section.

The LOCITO procedure is implemented by the following steps:

1. For j^{th} locus, perform orthogonal projection

   (a) Calculate covariance matrix of c^{th} form of genomic features and perform eigenvalue decomposition.

   (b) Select first few eigenvalues to explain δ fraction of total variance and project feature level data to the selected orthogonal space to obtain projected data.

   (c) Scale the selected eigenvalues according to equation 5.11.

   (d) Repeat (a) to (c) for each form of genomic features and get projected data $\mathbf{P}_j$ and scaled Λ_j'

2. Perform parametric modeling for each projected vector to get association vector $\mathbf{z}_j$.

3. Define the LOCITO statistic to be quadratic form as in equation (5.13).

4. Bootstrap to estimate covariance $\widehat{\Sigma_{zj}}$ under null as in equation (5.4).

5. Determine *p*-value based on generalized chi-square.

6. Perform same 1 to 5 steps for each locus.

7. Perform multiple testing adjustments, either resampling-based or other FDR controlling methods.

5.3 Applications

To illustrate the performance of LOCIT and LOCITO, we use a dataset of AML02 clinical trial (Rubnitz *et al.*, 2010[48]). Dense SNPs were genotyped in 187 patients by targeted genotyping in 37 genes[54, 55] and corresponding expressions of these genes were obtained from gene profiling data measured for these subjects by U133A microarray. The 37 genes include genes in araC pathway (see figure 5.1) and other key genes in drug metabolism. The prodrug araC is up-taken into cytoplasm by transporter hENT1, phosphorylated to active drug araCTP by three kinases sequentially; araCTP is then transported into nucleus and incorporated to DNA/RNA during synthesis which blocks DNA/RNA synthesis leading to apoptosis (program cell death); the active drug araCTP is also metabolized to inactive forms by dephosphorylation and deaminase, and is competed with dCTP. The understanding of these genes in patient responses to araC treatment will help tailoring treatments for patients in future. From the previous analyses, several SNPs were significantly associated with event-free survival (EFS). Here we investigate to see if any genes were significantly associated with EFS and ranked high given SNPs and gene expression data.

In addition to testing association of each SNP and expression probe with EFS, an integrated LOCIT test was performed for each of the 37 genes. EFS was modeled with Cox proportional hazards model with each SNP or gene expression as covariates, stratified by treatment arm to obtain z statistics. The LOCIT statistics was calculated according to equation (5.2). The subject labels were re-sampled B=200 times to obtain 200 random samples, which were then fit into the same Cox proportional hazards model to estimate the covariance of $\mathbf{z}$. The p-values were determined by generalized chi-square using the algorithm by Farebrother (1984) implemented in R. The results were shown in Table 5.1

Table 5.1: Top target genes that were associated with EFS in AML

Gene	N Probe	LOCIT Statistics	p-value
RRM2	7	19.89	0.017
SLC29A1	4	10.50	0.041

In the individual SNP association study, SNPs of RRM2 have been shown to associated with EFS (Cao *et al.*, 2013[55]). The LOCIT test ranked the RRM2 as top one gene associated with EFS among the 37 genes. RRM2 is the small subunit of holoenzyme of Ribonucleotide reductase, which is key enzyme involved in the biosynthesis of deoxynucleotides. The second gene in the ranking was SLC29A1, a drug transporter to transfer nucleotides into cells. The genetics of RRM2 and SLC29A1(SNP and expression) were associated with EFS outcome in AML patients treated with prodrug of araC. This result is consistent with the treatment model, in which prodrugs are transported into cells and metabolized by enzymes into active drug, eventually integrated into DNA/RNA synthesis resulting in program cell death.

Noting that the LOCIT incurs penalty on genes with genomic features most of which are not associated with endpoint of interest. LOCIT with orthogonal projection to reduce dimension and put heavier weight on import features (LOCITO) was applied to this dataset. δ was set to 0.99 and $B = 200$. The result of LOCITO test is shown in Table 5.2. We find that besides RRM2 and SLC29A1, both CDA and SOCS3 show association with EFS in AML at 0.05 alpha level. Orthogonal projection has reduced the dimension of

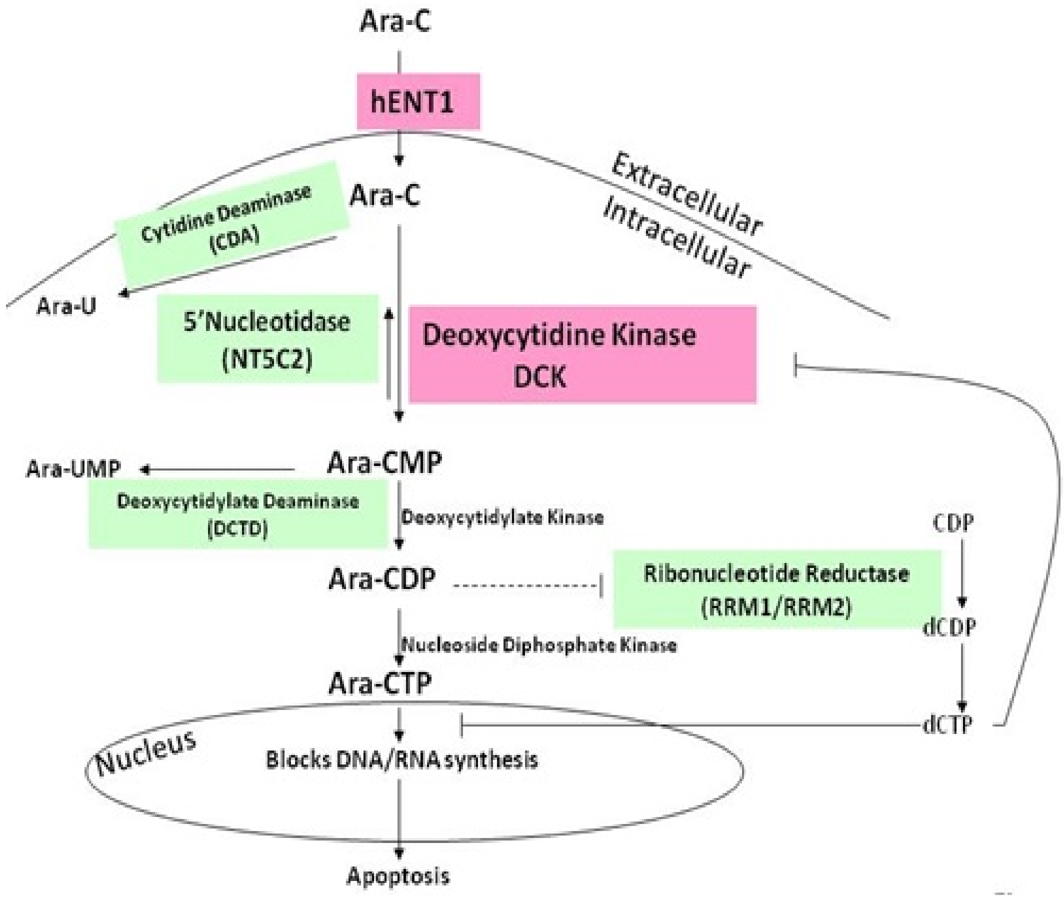


Fig. 5.1: Human araC pathway

Prodrug araC is up-taken into cytoplasm by transporter hENT1, phosphorylated to active drug araCTP by kinases, incorporated to DNA/RNA during synthesis which blocks DNA/RNA synthesis and leads to apoptosis; araCTP is also metabolized to inactive forms by dephosphorylation and deaminase, and compete with dCTP. (Courtesy by Dr. Lamba)

CDA from 10 to 8 and left the rest 3 genes untouched. However, the projection of genomic features in SOCS3 has increased the power to detect association of SOCS3 with EFS.

Table 5.2: Top genes associated with EFS in AML with orthogonal projection

Gene	Nprobe	Ndim	Stat	Pvalue
RRM2	6	6	7.01	0.016
SLC29A1	4	4	5.32	0.041
CDA	10	8	5.52	0.042
SOCS3	4	4	5.69	0.043
RRM2B	1	1	2.8	0.094

Similarly, LOCITO procedure was performed to test association of the 37 genes with day 22 MRD (present or absent of minimal residual disease) and over-all survival (OS) in this AML data set. The results with $\alpha \leq 0.1$ are shown in Table 5.3 and 5.4, respectively. Day 22 MRD is a measurement of early response of araC treatment in AML. SOCS3 was on the top of the list and RRM2 the second. DCK (deoxycytidine kinase) was also significantly associated with day 22 MRD (p =0.035). DCK phosphorylates araC to araCMP, an intermediate to active drug araCTP. Although DCK was not associated with EFS, it seems to be important for early response in treatment with araC. CDA (Cytidine Deaminase) was identified to be significantly associated with OS (p = 0.019). From these analyses, enzymes both activating the prodrug and deactivating the drug were associated with early response or long term outcome.

5.4 Simulation Study

LOCIT procedure can be applied to association analyses with various types of endpoints. To compare the statistical power of LOCITO procedure to that of other approaches, simulations were performed in a simple setting involving two treatment groups, where the other approaches could be applied. Twenty disjoint loci were generated as in table 5.5 with sample size 100 in each of the two treatment groups. Two types of genomic features were simulated in each locus, both of which were assumed to be

Table 5.3: Top genes associated with OS in AML with orthogonal projection

Gene	Nprobe	Ndim	Stat	Pvalue
SOCS3	4	4	7.48	0.015
RRM2	6	6	6.55	0.021
DCK	2	2	6.74	0.035
ABCB1	3	3	5.78	0.05
DCTD	20	14	4.3	0.064
CDA	10	8	4.51	0.076
XRCC1	5	4	4.1	0.083
ABCG2	3	3	4.34	0.094

Table 5.4: Top genes associated with day 22 MRD in AML with orthogonal projection

Gene	Nprobe	Ndim	Stat	Pvalue
CDA	10	8	6.82	0.019
CMPK	7	7	4.75	0.071

multivariate normal. These genomic features could be gene expression and DNA methylation. In DNA methylation, the logit transformation of the fraction of the methylated signal over total signal(M values) can be used.

In the 20 hypothetical loci, the first six loci(A to F) were small loci with 10 members each(5 from Type A and 5 from Type B); Loc G to M had moderate size with 30 members (10 from Type A and 20 from Type B); and Loc N to T had large size with 100 members (20 from Type A and 80 from Type B). There was no difference between the two groups in Loc A, F, G, N and T. In Loc B, 2 out of 5 members from type A had moderate increase (1) of mean in group 2 compared to group 1 and members from type B had no difference in mean between group 1 and 2. In Loc C, all 5 member from type A had moderate increase (0.5) of mean in group 2; 2 members of type B had big increase of mean (2) in group 2. Loc D to S were set up similarly with various combination of number of differential mean within type A and/or type B. Variance and covariance for each locus or sub-locus was drawn from Wishart distribution with toeplitz matrix and

Table 5.5: Locus size and group means of twenty loci in a simulation study.

GeneSet	Type A				Type B			
	Size	Sub Size	Group 1	Group 2	Size	Sub Size	Group 1	Group 2
Loc A	5	5	5	5	5	5	0	0
Loc B	5	2	5	6	5	5	0	0
		3	5	5				
Loc C	5	5	5	5.5	5	2	0	2
						3	0	0
Loc D	5	5	5	5	5	5	0	1
Loc E	5	3	5	6	5	5	0	0
		2	5	5				
Loc F	5	5	5	5	5	5	0	0
Loc G	10	10	5	5	20	20	0	0
Loc H	10	10	5	4	20	20	0	1
Loc I	10	2	5	7	20	5	0	-2
		8	5	5		15	0	0
Loc J	10	10	5	5	20	5	0	2
						15	0	0
Loc K	10	3	5	7	20	20	0	0.5
		7	5	5				
Loc L	10	10	5	5	20	5	0	1
						15	0	0
Loc M	10	3	5	7	20	20	0	0
		7	5	5				
Loc N	20	20	5	5	80	80	0	0
Loc O	20	20	5	6	80	30	0	1
						50	0	0
Loc P	20	5	5	7	80	40	0	1
		15	5	5		40		
Loc Q	20	20	5	5	80	10	0	2
						70	0	0
Loc R	20	20	5	5	80	50	0	0.5
						30	0	0
Loc S	20	20	5	6	80	80	0	0
Loc T	20	20	5	5	80	80	0	0

number of members plus 10 as parameter. The Variance and covariance with Toeplitz structure instead of Identity matrix was used to introduce correlation structure among features within a locus/sub-locus and a form of genomic data. The variance and covariance matrixes were scaled with diagonal elements around 1.

For each locus and sub-locus, random samples were drawn from multivariate normal distribution with mean in table 5.5 and variance-covariance generated as above. One thousand random data sets were drawn. The LOCIT procedure with orthogonal projection (LOCITO), SAFE (Barry *et al.*, 2005), MRPP test (Nettleton *et al.*, 2008) and GSA (Efron and Tibshirani, 2007) were applied to each of the 1000 data sets. The power and type 1 error of the four methods are summarized in table 5.4.

In the five loci without differential expression and methylation, LOCITO method maintained the nominal alpha level of 5% (2.2% to 5%) for all the three selected $\delta : 0.8, 0.9$ and 1; the rest three methods also maintained the nominal alpha level of 5% well (SAFE: 0% to 0.3%; MRPP: 3.7% to 5.3%; GSA: 0%). SAFE and GSA are too conservative in alpha level control. In the 15 loci with differential expression and/or methylation, both LOCITO and MRPP had high power detect difference between the two treatment groups. SAFE had negligible power to detect difference except for Loc I (91.6%) and Q (36.9%). GSA lacked power to detect difference except for Loc H (78%), Loc I (100%) and Q (97.6%). In the simulation setting, 15 out 20 loci had differential expression and/or methylation. It demonstrated that the SAFE and GSA procedure loses power if most of loci are differentially expressed or methylated due to the nature of the two tests: not selfcontained. MRPP had power to detect locus with any difference, especially with large difference in part of the loci. This result is consistent with the fact that MRPP test is designed to detect any differential expression in multivariate spaces. Unfortunately, the MRPP test is hard to extend to complicated models such as adjusting for known factors. LOCITO, GSA and SAFE methods can be applied to complicated statistical modeling with various phenotypes. From the simulation studies, LOCITO method performed better in most of the simulated locus settings.

5.5 Summary and Discussion

In this chapter we have shown that the LOCIT procedure for locus-based association studies is a very flexible method that can be adapted to many types of endpoints. In the applications to real datasets, we illustrated the implementation of the LOCIT test for binary endpoint (logistic regression) and time to event endpoint (Cox proportional hazards model). The application of the LOCIT to continuous normal endpoint is trivially amounts to a linear regression. In application of the LOCIT to time to event endpoint with competing events, the procedures Fine and Gray (1999)[56] for competing risk regression can be applied to model event of interest with competing events. In general, LOCIT test can be applied to any parametric and non-parametric models with asymptotically normal test statistics. Thus the LOCIT is reasonably versatile and adaptable to most types of endpoints of practical interest.

Moreover, LOCIT is self-contained. LOCIT test for one locus does not influence the test on another locus. Geoman and Buhlmann 2007 pointed out that self-contained null hypothesis testing in gene set analysis gives advantage of valid *p*-values and easy interpretability. This is also true for locus-based test. The LOCIT test is model based and does not require permutation to determine *p*-values, although it does require resampling technique to estimate one parameter. The LOCIT test is still computationally efficient and can be easily applied to genomic level association testing.

LOCIT test statistic is currently defined as a quadratic form with the corresponding scaled eigenvalues as diagonal elements. This choice of weight assumes that the more variation among features in each feature type in a locus carries the more information of association with endpoint of interest.It also assumes equal contribution of feature types to the overall association of the locus with the endpoint of interest. In most cases, these are

Table 5.6: Summary of simulation results on twenty loci with sample size 100 in both groups.

Loci	Differential	LOCITO			SAFE	MRPP	GSA
		$\delta = 0.8$	$\delta = 0.95$	$\delta = 1$			
Loc A	0	4.9	4.6	4.3	0	5	0
Loc F	0	3.3	2.8	2.3	0	3.7	0
Loc G	0	3.8	3.5	2.2	0.3	4.4	0
Loc N	0	5	4.8	2.7	0	4.1	0
Loc T	0	5	4.3	2.6	0	5.3	0
Loc B	1	100	100	100	0	100	0
Loc C	1	100	100	100	3.9	100	0.1
Loc D	1	100	100	100	0	100	21.3
Loc E	1	100	100	100	0	100	0
Loc H	1	100	100	100	0	100	78
Loc I	1	100	100	100	91.6	100	100
Loc J	1	100	100	100	2.3	100	13.4
Loc K	1	100	100	100	0.5	100	0.8
Loc L	1	100	100	100	0	100	0
Loc M	1	99.3	99.3	99.3	0	100	0
Loc O	1	100	100	100	3.8	100	0
Loc P	1	100	100	100	1.3	100	0
Loc Q	1	100	100	100	36.9	100	97.6
Loc R	1	76.3	75.3	61.9	0	94.2	0
Loc S	1	100	100	100	0	100	0

Notes: Differential: 0: no difference, 1: difference between the two groups; For loci without difference, the false positive percentages (type 1 error) are shown; For loci with difference between the two groups, the true positive percentages (power) are shown.

reasonable assumptions. However, these assumptions need to be further investigated in some circumstances.

In current implementation of LOCIT, we used orthogonal projection in each feature type to reduce dimension. For genomic study with two types of high throughput data in which one type regulates the other, we also could perform feature selection for large loci. One potential strategy is to apply sparse cannocial correlation by Karkhomenko *et al.* (2009), in which a subset of features are selected to maximize first-order approximation of correlation matrix. This will be also useful to extend LOCIT to big set level association testing in genetic studies.

Chapter 6

SUMMARY AND FUTURE RESEARCH

New revolutionary technologies have provided scientists with tools to study thousands of genomic features simultaneously. These studies have shed light on mechanisms underlying the biology of complex diseases through accumulation of knowledge of gene-gene relationships and identification of gene sets and graphical pathways. This knowledge has provided vital information for testing global hypotheses about gene sets and consequently interpreting results from high dimension genetic studies. In the preceding chapters, we proposed new procedures for utilizing pathway information in high throughput genetic data.

6.1 Pathways and applications in Bayesian framework

Using the information from graphical pathways and gene networks that capture the gene relationships in cells, we have described a procedure for incorporating prior knowledge of gene relationships using directed graphs into genomic testing in a Bayesian framework. This extends the work of Pan (2006) and Wei and Pan (2008) that used random Markov field in a mixture model by capturing the actual geometric direction of gene relationships. The utility of our method is demonstrated by an application to a real data set with MAPK pathway derived from KEGG in adult AML.

Many public and commercial databases have been developed to structure and store biological knowledge at various genetic levels and in various organisms. However, pathways defined by these databases are from multiple tissues and various resources. Some of the relationships inferred were from high throughput experiments such as gene profiling, proteomics, Chip-Chip experiment. However, some of these relationships in pathways may not be applicable to specific experiment units from a different organism. Moreover, a gene is sometimes represented by multiple names in various databases. Matching gene names in an experimental platform to gene names in a pathway is a non-trivial and uncertain task, especially because current pathway databases are not statistically or computationally friendly. There is an urgent and practical need for pathways or gene networks to be expressed in numeric form, such as an adjacency matrix. This will be more accessible to statistical and computational tools.

In the interest of biological interpretation, pathways or gene networks should be tailored unambiguously to the specific organism/tissue type of interest. This will allow researchers to use data from similar organism and tissue/cell type for correlation analysis and to correctly (biologically) trim pathways derived from pathway databases. Combining pathway information of relationships from similar biological entities will yield more biologically correct prior information of genes. Efforts to incorporate such pathways, with directed graphs, into the construction of posterior information in Bayesian analysis is one of our future research plans.

6.2 POST in genetic studies

In the construction of POST (Projection onto Orthogonal Space Test), we designed a test of association of gene sets with diverse types of endpoints of interest. This procedure has several desirable features: it is flexible, self-contained and amenable to subject permutation for parameter estimation. POST assumes that more variation among probes in a set carries more information of association with an endpoint of interest. This is a critical assumption for the validity of POST. In most applications, this assumption should hold. POST captures the correlation structure among genomic features within a set, but this is different from jointly modeling the genomic features in a statistical multivariate analysis. As shown in a simulation study, it is less powerful than multivariate analysis such as MRPP test (Nettleton *et al.* 2008).

POST procedure tests association of a set with endpoint of interest. However it is not able to pinpoint the most important genomic features or subset of features driving the association. Therefore, a subsequent analysis is needed to identify the features driving the association.

POST was developed for gene profiling data. One of our future research plans will be to extend POST in application to other types of data such as methylation data, microRNA, or even SNP after proper transformation. In methylation data, M values, the logit transformation of beta values are assumed normally distributed. Probes in CpG island/shore in promoter region tends to be hypo/hyper methylated coordinately. We plan future research to use POST in this context based on the fact that it is biologically and statistically feasible to perform gene set association test of methylation data with endpoints of interest.

In addition, we will extend POST to multiple endpoints, specifically, to clinical trials where multiple presenting features and outcome variables are collected. The relationship of these variables are known under specific treatment model. Pounds *et al.* (2009)[34] have proposed PROMISE procedure for association of gene expression with multiple endpoints and then extended it to SNP data (2011)[57]. POST can be extended to test for gene set association with multiple endpoints of interest under the PROMISE framework. We plan to develop a POST-PROMISE procedure for gene set level testing with multiple endpoints.

6.3 LOCIT in genetic studies

Along the same line of thinking, we will extend our work to extend the LOCIT and LOCITO procedures. In genetic studies with multiple forms of high throughput data, it is traditional to test association within high throughput data and look for overlap at certain FDR control. LOCIT was proposed to perform integrated association test and to alleviate the difficulties encountered due to overlap.

Since LOCIT provides one P-value for multiple data, it is especially valuable for prioritizing gene/locus for follow-up study. The method was originally applied to small locus such as gene with equal weights. As noted in the Chapter 5, simulation studies and real applications indicated that the procedure incurs too much penalty on locus with moderate to large number of noise features. As an extension, LOCITO was then developed to overcome the difficulty by performing orthogonal projection within each form data. LOCITO reduces dimensionality of locus and puts equal overall weight across data forms. We plan to apply LOCITO to small loci and big loci sets. We will exam two forms of genetic data with one form regulating the other. We will also investigate possible extensions for performing feature selection by sparse canonical correlation similar to Karkhomenko *et al.* (2009)[30].

REFERENCES

[1] Consortium, S et al. (1998) {Genome sequence of the nematode $C.elegans$: A platform for investigating biology}. *Science* **282**, 2012–2018.

[2] Adams, M. D, Celniker, S. E, Holt, R. A, Evans, C. A, Gocayne, J. D, Amanatides, P. G, Scherer, S. E, Li, P. W, Hoskins, R. A, Galle, R. F, et al. (2000) The genome sequence of drosophila melanogaster. *Science* **287**, 2185–2195.

[3] Chinwalla, A. T, Cook, L. L, Delehaunty, K. D, Fewell, G. A, Fulton, L. A, Fulton, R. S, Graves, T. A, Hillier, L. W, Mardis, E. R, McPherson, J. D, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.

[4] Initiative, A. G et al. (2000) Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature* **408**, 796.

[5] Goff, S. A, Ricke, D, Lan, T.-H, Presting, G, Wang, R, Dunn, M, Glazebrook, J, Sessions, A, Oeller, P, Varma, H, et al. (2002) A draft sequence of the rice genome (oryza sativa l. ssp. japonica). *Science* **296**, 92–100.

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

[7] Khatri, P & Drăghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **21**, 3587–3595.

[8] Goeman, J. J & Bühlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**, 980–987.

[9] Nam, D & Kim, S.-Y. (2008) Gene-set approach for expression pattern analysis. *Briefings in bioinformatics* **9**, 189–197.

[10] Bentwich, I, Avniel, A, Karov, Y, Aharonov, R, Gilad, S, Barad, O, Barzilai, A, Einat, P, Einav, U, Meiri, E, et al. (2005) Identification of hundreds of conserved and nonconserved human micrornas. *Nature genetics* **37**, 766–770.

[11] Lieb, J. D, Liu, X, Botstein, D, & Brown, P. O. (2001) Promoter-specific binding of rap1 revealed by genome-wide maps of protein–dna association. *Nature genetics* **28**, 327–334.

[12] Johnson, D. S, Mortazavi, A, Myers, R. M, & Wold, B. (2007) Genome-wide mapping of in vivo protein-dna interactions. *Science* **316**, 1497–1502.

[13] Ashburner, M, Ball, C. A, Blake, J. A, Botstein, D, Butler, H, Cherry, J. M, Davis, A. P, Dolinski, K, Dwight, S. S, Eppig, J. T, et al. (2000) Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25–29.

[14] Kandasamy, K, Mohan, S. S, Raju, R, Keerthikumar, S, Kumar, G. S, Venugopal, A. K, Telikicherla, D, Navarro, J. D, Mathivanan, S, Pecquet, C, et al. (2010) Netpath: a public resource of curated signal transduction pathways. *Genome biology* **11**, R3.

[15] Croft, D, OKelly, G, Wu, G, Haw, R, Gillespie, M, Matthews, L, Caudy, M, Garapati, P, Gopinath, G, Jassal, B, et al. (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research* **39**, D691–D697.

[16] Kanehisa, M & Goto, S. (2000) Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30.

[17] Kanehisa, M, Goto, S, Hattori, M, Aoki-Kinoshita, K. F, Itoh, M, Kawashima, S, Katayama, T, Araki, M, & Hirakawa, M. (2006) From genomics to chemical genomics: new developments in kegg. *Nucleic acids research* **34**, D354–D357.

[18] Yu, N, Seo, J, Rho, K, Jang, Y, Park, J, Kim, W. K, & Lee, S. (2012) hipathdb: a human-integrated pathway database with facile visualization. *Nucleic acids research* **40**, D797–D802.

[19] Allison, D. B, Cui, X, Page, G. P, & Sabripour, M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* **7**, 55–65.

[20] Mootha, V. K, Bunkenborg, J, Olsen, J. V, Hjerrild, M, Wisniewski, J. R, Stahl, E, Bolouri, M. S, Ray, H. N, Sihag, S, Kamal, M, et al. (2003) Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115**, 629–640.

[21] Subramanian, A, Tamayo, P, Mootha, V. K, Mukherjee, S, Ebert, B. L, Gillette, M. A, Paulovich, A, Pomeroy, S. L, Golub, T. R, Lander, E. S, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–15550.

[22] Tian, L, Greenberg, S. A, Kong, S. W, Altschuler, J, Kohane, I. S, & Park, P. J. (2005) Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 13544–13549.

[23] Kim, S.-Y & Volsky, D. J. (2005) Page: parametric analysis of gene set enrichment. *BMC bioinformatics* **6**, 144.

[24] Efron, B & Tibshirani, R. (2007) On testing the significance of sets of genes. *The annals of applied statistics* pp. 107–129.

[25] Jiang, Z & Gentleman, R. (2007) Extensions to gene set enrichment. *Bioinformatics* **23**, 306–313.

[26] Barry, W. T, Nobel, A. B, & Wright, F. A. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* **21**, 1943–1949.

[27] Lu, Y, Liu, P.-Y, Xiao, P, & Deng, H.-W. (2005) Hotelling's t2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics* **21**, 3105–3113.

[28] Srivastava, D. K, Boyett, J. M, Jackson, C. W, Tong, X, & Rai, S. N. (2007) A comparison of permutation hotelling's t 2 test and log-ratio test for analyzing compositional data. *Communications in StatisticsTheory and Methods* **36**, 415–431.

[29] Nettleton, D, Recknor, J, & Reecy, J. M. (2008) Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics* **24**, 192–201.

[30] Parkhomenko, E, Tritchler, D, & Beyene, J. (2009) Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology* **8**, 1–34.

[31] Witten, D. M, Tibshirani, R, & Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* p. kxp008.

[32] Wu, M. C, Lee, S, Cai, T, Li, Y, Boehnke, M, & Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82–93.

[33] Lee, S, Wu, M. C, & Lin, X. (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775.

[34] Pounds, S, Cheng, C, Cao, X, Crews, K. R, Plunkett, W, Gandhi, V, Rubnitz, J, Ribeiro, R. C, Downing, J. R, & Lamba, J. (2009) Promise: a tool to identify genomic features with a specific biologically interesting pattern of associations with multiple endpoint variables. *Bioinformatics* **25**, 2013–2019.

[35] Lamba, J. K, Crews, K. R, Pounds, S. B, Cao, X, Gandhi, V, Plunkett, W, Razzouk, B. I, Lamba, V, Baker, S. D, Raimondi, S. C, et al. (2011) Identification of predictive markers of cytarabine response in aml by integrative analysis of gene-expression profiles with multiple phenotypes. *Pharmacogenomics* **12**, 327–339.

[36] Pan, W. (2006) Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics* **22**, 795–801.

[37] Wei, P & Pan, W. (2008) Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics* **24**, 404–411.

[38] Albert, J. H & Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* **88**, 669–679.

[39] Pounds, S & Cheng, C. (2006) Robust estimation of the false discovery rate. *Bioinformatics* **22**, 1979–1987.

[40] Benjamini, Y & Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* pp. 1165–1188.

[41] Storey, J. D. (2003) The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of statistics* pp. 2013–2035.

[42] Tomfohr, J, Lu, J, & Kepler, T. B. (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC bioinformatics* **6**, 225.

[43] Duchesne, P & Lafaye De Micheaux, P. (2010) Computing the distribution of quadratic forms: Further comparisons between the liu–tang–zhang approximation and exact methods. *Computational Statistics & Data Analysis* **54**, 858–862.

[44] Farebrother, R. (1984) Algorithm as 204: The distribution of a positive linear combination of χ 2 random variables. *Applied Statistics* pp. 332–339.

[45] Yekutieli, D & Benjamini, Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* **82**, 171–196.

[46] Benjamini, Y & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300.

[47] Ross, M. E, Mahfouz, R, Onciu, M, Liu, H.-C, Zhou, X, Song, G, Shurtleff, S. A, Pounds, S, Cheng, C, Ma, J, et al. (2004) Gene expression profiling of pediatric acute myelogenous leukemia. *Blood* **104**, 3679–3687.

[48] Rubnitz, J. E, Inaba, H, Dahl, G, Ribeiro, R. C, Bowman, W. P, Taub, J, Pounds, S, Razzouk, B. I, Lacayo, N. J, Cao, X, et al. (2010) Minimal residual disease-directed therapy for childhood acute myeloid leukaemia: results of the aml02 multicentre trial. *The lancet oncology* **11**, 543–552.

[49] Mikesch, J, Steffen, B, Berdel, W, Serve, H, & Müller-Tidow, C. (2007) The emerging role of wnt signaling in the pathogenesis of acute myeloid leukemia. *Leukemia* **21**, 1638–1647.

[50] Anastas, J. N & Moon, R. T. (2012) Wnt signalling pathways as therapeutic targets in cancer. *Nature Reviews Cancer* **13**, 11–26.

[51] Wagner, E. F & Nebreda, Á. R. (2009) Signal integration by jnk and p38 mapk pathways in cancer development. *Nature Reviews Cancer* **9**, 537–549.

[52] Hussain, S. P, Hofseth, L. J, & Harris, C. C. (2003) Radical causes of cancer. *Nature Reviews Cancer* **3**, 276–285.

[53] Nambiar, M, Kari, V, & Raghavan, S. C. (2008) Chromosomal translocations in cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1786**, 139–152.

[54] Mitra, A, Crews, K, Pounds, S, Cao, X, Downing, J, Raimondi, S, Campana, D, Ribeiro, R, Rubnitz, J, & Lamba, J. (2011) Impact of genetic variation in fkbp5 on clinical response in pediatric acute myeloid leukemia patients: a pilot study. *Leukemia* **25**, 1354–1356.

[55] Cao, X, Mitra, A. K, Pounds, S, Crews, K. R, Gandhi, V, Plunkett, W, Dolan, M. E, Hartford, C, Raimondi, S, Campana, D, et al. (2013) Rrm1 and rrm2 pharmacogenetics: association with phenotypes in hapmap cell lines and acute myeloid leukemia patients. *Pharmacogenomics* **14**, 1449–1466.

[56] Fine, J. P & Gray, R. J. (1999) A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**, 496–509.

[57] Pounds, S, Cao, X, Cheng, C, Yang, J. J, Campana, D, Pui, C.-H, Evans, W, & Relling, M. (2011) Integrated analysis of pharmacologic, clinical and snp microarray data using projection onto the most interesting statistical evidence with adaptive permutation testing. *International journal of data mining and bioinformatics* **5**, 143–157.