Gene Regulatory Network Reconstruction and Pathway Inference from High Throughput Gene Expression Data

by

Weijun Luo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biomedical Engineering)
in The University of Michigan
2008

Doctoral Committee:

- **Assistant Professor Peter J. Woolf, Chair**
- **Professor Steven A. Goldstein**
- **Associate Professor Kerby A. Shedden**
- **Assistant Professor Yongqun He**

To my family and friends

Acknowledgements

I would like to thank my advisor Prof. Peter Woolf for his consistent guidance in doing research, thinking and writing. The understanding, support and freedom he gave me made my graduate school time enjoyable. My life and career benefit much from his sharing of wisdom, knowledge and happiness.

My other committee members have also been very supportive. Prof. Steven Goldstein provided extra mentorship, directed me to opportunities and watched the right direction for me in my graduate education and research. Prof. Kerby Shedden taught me statistical computing methods in his classes and helped me on technical issues in my research through inspiring discussions. Prof. Yongqun “Oliver” He shared with me his insightful perspective on bioinformatics research and career through collaboration and suggestions.

Special thanks to Prof. Kurt Hankenson and Dr. Michael Friedman at University of Pennsylvania. The collaboration project with them funded me for years and motivated a big part of my thesis. Their microarray experiments provided primary data to apply and test my computational methods.

I would also like to thank the faculty and staff at the Biomedical Engineering Department and the Bioinformatics Program for their service and help, particularly Ms. Maria Steel, Ms. Julia Eussen and Prof. Matthew O'Donnell. Thanks to my colleagues at Systems Biology Lab, the Orthopedics Research Labs, and all my friends in Ann Arbor for their help in my life and my research.

Finally, I would like to express my deep gratitude to my family and my wife for their long lasting love, patience and support to me.

Table of Contents

List of Figures

List of Tables

Abstract

Two basic motivating questions in biomedical research are: What genes regulate what other genes? What genes or groups of genes regulate a specific phenotype? Gene regulatory network (GRN) reconstruction and pathway inference are the two computational strategies addressing these two questions respectively. GRN reconstruction is to infer the components and topology of an unknown pathway, while pathway inference is to infer association between known pathways and a phenotype.

This thesis focuses on gene regulatory network reconstruction and pathway inference from high throughput biological data.

In the first part of this work, I developed a novel method, MI3, for de novo GRN reconstruction using continuous three-way mutual information. MI3 addresses three major issues in previous probabilistic methods simultaneously: (1) to handle continuous variables, (2) to detect high order relationships, (3) to differentiate causal vs. confounding relationships. MI3 consistently and significantly outperformed frequently used control methods and faithfully capture mechanistic relationships from gene expression data.

In the second part of this work, I proposed another novel method, GAGE, Generally Applicable Gene Set Enrichment for pathway inference. I successfully apply GAGE to multiple microarray data sets with different sample sizes, experimental designs and profiling techniques. GAGE shows significantly better performance when compared to two other commonly used GSA methods of GSEA and PAGE. GAGE reveals novel and relevant regulatory mechanisms from both published and previously unpublished microarray studies.

In the third part of this work, we conducted a microarray study on transcriptional

programs during BMP6 induced osteoblast differentiation and mineralization, and applied GAGE to recover the regulatory pathways and transcriptional signaling networks in the process. I not only showed which pathways or gene sets are significant, but also when and how they are involved in the osteoblast differentiation and mineralization. Different from common pathway analyses, our work further captures the interconnections among individual pathways or functional groups and integrate them into a whole system.

Chapter I

Introduction

1.1 Motivation

Modern biomedical research focuses on the molecular-level regulatory mechanisms underneath the development of normal functional phenotypes (like differentiation) or abnormal phenotypes (like cancer) as to find strategies for better health and cures for diseases. Two basic motivating questions in biomedical research are (Figure 1.1): (1) What genes regulate what other genes? (2)What genes or groups of genes regulate a specific phenotype? Answers to these questions help us understand the biological systems and effectively interfere with them for ideal biomedical effects.

These two questions are both directly related to gene expression. First question concerns the regulation of gene expression (gene regulation), the second one concerns regulation of phenotype expression by gene expression (phenotype regulation). These two processes together form a causal chain: gene regulatory mechanism causes gene expression, which in turn causes phenotype expression. Two statistical learning problems are defined along this chain in the reverse direction: (1) we infer the gene regulatory mechanisms from gene expression data, which is called reverse engineering; (2) we infer the specific gene expression events associated with a phenotype too, which is called feature selection.

Reverse engineering and feature selection problems are further defined with gene expression data. With traditional molecular/cellular biology techniques such as qPCR, we can only measure the expression of limited number of genes in a study. Correspondingly, we can only learn local regulatory models for individual genes or identify individual gene markers for particular phenotypes. With the development of high throughput technology

such as microarrays, we can profile the expression of the whole genome at a time. This brings the potential to study biology as whole systems. Correspondingly, we can reconstruct gene regulatory network (GRN) involving large number of genes or infer the connections between functional groups or signaling pathways with up to hundreds of genes and particular phenotypes.

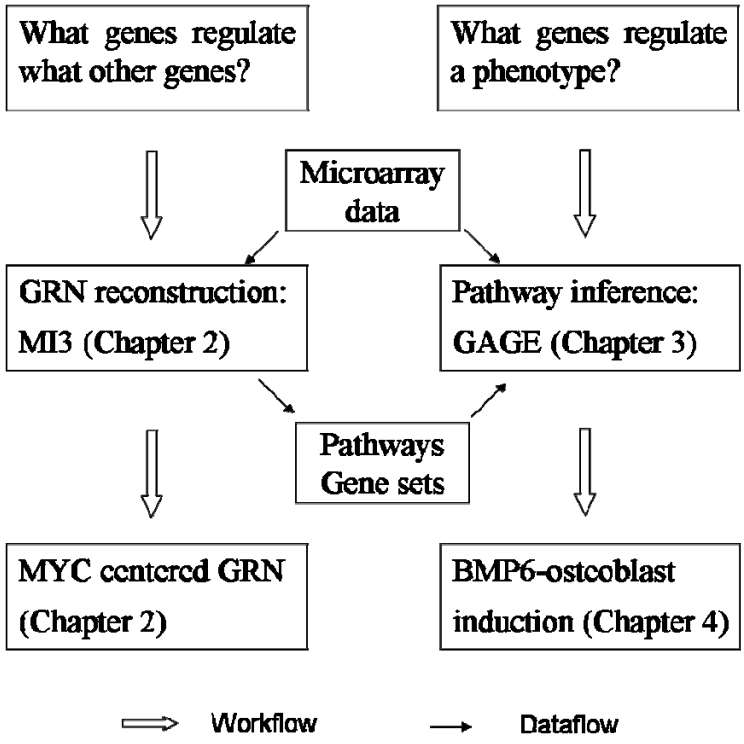


Figure 1.1 Outline of the thesis. The thesis focuses on two problems: GRN reconstruction and pathway inference, which are motivated by two basic questions in biomedical research: what genes regulate what other genes; what genes or groups of genes regulate a specific phenotype. Two novel methods, MI3 and GAGE, were developed to address these two problems, and were applied to two systems biology studies respectively: MYC centered GRN and BMP6 induced osteoblast differentiation and mineralization.

This thesis focuses on gene regulatory network reconstruction and pathway inference from high throughput gene expression data (Figure 1.1 and 1.3). To address these problems, I developed two new computational methods, MI3 and GAGE. These methods are applied to solve two representative biological problems: (1) the gene regulatory network centered at MYC, a transcription factor that regulates the expression of up to

15% of all human genes [1] and a strong oncogene that involves in a variety of cancers and other cellular processes [2]; (2) the regulatory pathways and functional groups involved in BMP6 induced osteoblast differentiation and mineralization, a phenotypic change responsible for skeleton development and multiple bone diseases [3], including osteoporosis, osteogenesis imperfecta, osteosarcoma etc. These applications plus validation experiments suggest that our methods are generally applicable to a broad variety of problems derived from the two motivating questions at the beginning of this section.

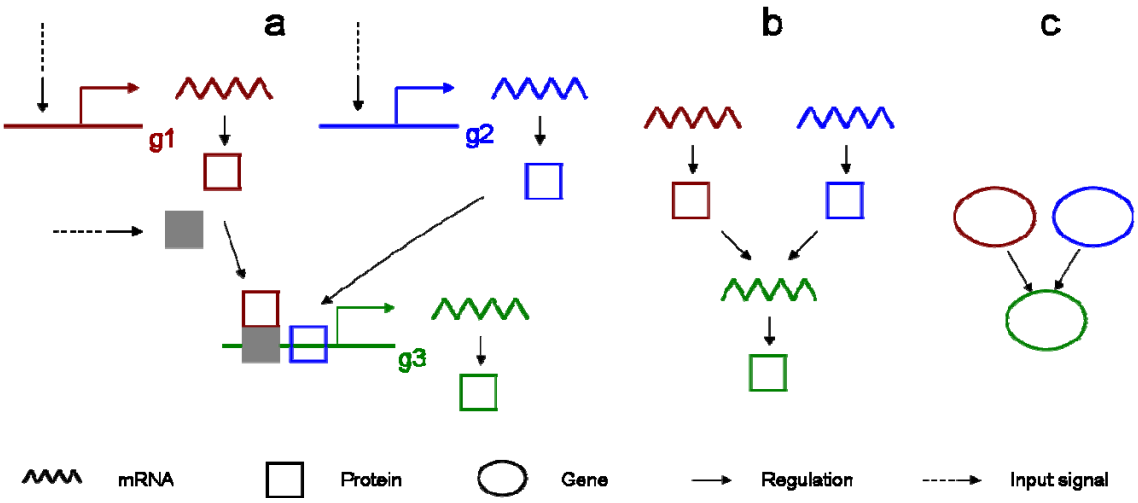


Figure 1.2 An example GRN defined at different detail and abstraction levels. (a) a more complete functional description; (b) a model with both mRNAs and proteins; (c) an abstract model with conceptual nodes the for three genes g1-g3 and their interactions. The dark solid square represents an unknown/unobserved transcription factor protein that binds to g3 promoter. In this thesis, I focus on learning (c) the abstract GRN models from high throughput gene expression data.

1.2 Gene Regulatory Network and reconstruction

A gene regulatory network (GRN) includes genes and interactions that control the transcription or mRNA expression levels of genes (Figure 1.2). GRN may be defined with different levels of detail, each focusing on different aspects of the process. From a functional perspective, a simple GRN would consist of input signaling pathways,

regulatory proteins that integrate the input signals, target genes, and the RNA and proteins produced from those target genes [4] (Figure 1.2a). In addition, such networks often include dynamic feedback loops that provide for further regulation of network architecture and output [4]. There are other slightly different functional descriptions of GRN too. However, all GRN consider the transcriptional regulation of genes.

A GRN may also be considered an abstract model for regulatory interactions among genes (Figure 1.2b-c). In such models, genes are represented as nodes, regulation relationships as edges with arrows standing for the direction. In a detailed GRN model (Figure 1.2b), regulator nodes are the transcriptional regulator proteins, target nodes are the mRNA levels for the target genes. For instance, Gene g1 regulates g3, actually means that the protein of g1 regulates the mRNA transcription of g3 (Figure 1.2b-c). It is not necessary to tell protein vs mRNA explicitly for a gene in abstract GRN models (Figure 1.2c), and a conceptual node for each gene is enough. Frequently we only have expression data at mRNA level (as in microarray datasets) or protein level (as in proteomics datasets), an abstract GRN model such as in Figure 1.2c becomes necessary. In such abstract models, edges represent conceptual regulation relationships rather than a real physical interaction (Figure 1.2c). Hence the statement that g1 regulates g3 does not necessarily means g1 binds to the regulatory elements of g3 DNA sequence, but often g1 may affect the g3 transcription through indirect actions such as being a transporter or modulator of another transcription factor that binds and controls g3 promoter directly (Figure 1.2a).

Schlitt et al proposed to categorize GRN models in four classes according to increasing level of detail in the models [5]. Each class has its own advantages and limitations. The four classes are [5]:

i. parts lists – a collection, description and systematization of network elements in a particular organism or a particular biological system (e.g., transcription factors, promoters, and transcription factor binding sites);

ii. topology models – a description of the connections between the parts; this can be viewed as wiring diagrams where directed or undirected connections between genes represent different types of interactions;

iii. control logic models – a description of combinatorial (synergetic or interfering) effects of regulatory signals – e.g., which transcription factor combinations activate and which repress the transcription of the gene;

iv. dynamic models – the simulation of the real-time behavior of the network and the prediction of its response to various environmental changes, external, or internal stimuli.

Schlitt et al [5] pointed out that for a fixed number of network elements each next level is more detailed and complex. However, the size of the networks that we are able to model at each particular level is constrained by both experimental data and computational power. This said, larger networks can be described on topological level than on the dynamic level [5].

GRN reconstruction or reverse engineering is to infer GRN models from data. A complete GRN model should integrate features of all four classes mentioned above: parts, topology, control logic, and dynamics. However, currently learning GRN models of classes i-iii is the most feasible due to the lack of data and the complexity of the models. Generally speaking, models at higher levels take models at lower level as prerequisite implicitly. Hence, topology models are built on top of part list, and control logic models on topology models (and part list) in turn. Dynamic models are built on all three previous levels, hence are the most complex and complete models. Learning accurate dynamics models requires time series data at all relevant levels, mRNA, protein and metabolites, their active and inactive forms, their distribution and local concentration, as well as kinetic measurements such as transportation rates and reaction rates. Many of these data are difficult or impossible to obtain currently. Even all these data were available, reconstructing fully parameterized dynamic models from these data remains a significant challenge.

Currently, the two most relevant model learning or inference tasks are structure learning and parameter learning, which corresponds to models at the levels of classes ii-iii respectively, i.e. topology and control logics [5]. I will next focus on these models and two major categories of learning methods: linear correlation based methods and probability based methods.

Linear correlation based methods, such as clustering [6, 7], correlation networks [8, 9] and graphical Gaussian models [10], have been frequently used to learn GRN. All linear methods are computationally fast and relatively easy to interpret.

Clustering finds groups of coregulated genes with similar expression pattern [6, 7]. Clustering can be used to infer part list, including coregulated target genes, common cis-regulatory elements such as transcription factor binding sites. Additional data are needed to determine the regulator hence topology model based on clustering results. For example, the common regulator for a coregulated cluster is likely a known TF in that cluster [11, 12], or the common regulator for several member genes of that cluster, or the known or unknown TF with binding motif derived from multiple promoter alignment [13, 14]. There are methods taking clusters as modules or sub-networks of gene regulatory system, and try to learn the gene order or topology of these sub-networks [15, 16]. Assuming each cluster as a gene regulatory network is problematic because coexpression or correlation among cluster genes suggests coregulation rather than a causal relationship. The gene regulatory model learned this way is most likely the highest scoring confounding model.

Correlation networks or relevance networks are networks of highly correlated genes [8, 9]. Edges connect pairs of genes with correlation coefficient over a certain threshold. Correlation networks cluster genes naturally without pre-assigned cluster number. Different from the classic clustering methods, correlation networks keep the strongest pair-wise association between genes, which contain relevant information for functional interpretation of genes and their relationships. However, like clustering, the relationships

between genes in a correlation network are mostly coregulation not causal relationship.

Graphical Gaussian Models (GGMs) [10], also known as “covariance selection” or “concentration graph” models, are a class of graphical models related to correlation networks. The key idea behind GGMs is to use partial correlations as a measure of independence of any two genes conditioned on all other genes. Note that partial correlations are related to the inverse of the correlation matrix. Edges in GGMs represent high conditional dependency, i.e. direct rather than indirect relationships. In contrast, correlation networks define relationships between genes through standard correlation coefficients. Edges in correlation networks only represent high marginal dependency without telling direct vs indirect relationships. Therefore, GGMs are considered a more accurate model over correlation networks for gene regulatory network reconstruction [17]. However, GGMs assume multivariate normality, which is frequently not the case for real biological systems.

Linear correlation methods as a whole suffer from a few important limitations. First among these is that linear methods assume linear relationships between variables, and hence are unable to model non-linear relationships in transcriptional regulatory systems. Furthermore, correlation can only capture association relationships, which are commonly not causal or regulatory interactions.

Probability based methods are a second class of methods commonly used for reconstructing GRNs from biological data. Representative probability methods include Bayesian networks [18-21] and mutual information networks [22, 23]. Probability based methods can capture both linear and nonlinear regulatory relationships and are noise tolerant. Probabilistic graphical models use directed edges to represent causal relationship rather than correlative relationships. However, probabilistic methods require significant more data than correlation based methods. They can be computationally slow. A GRN with several nodes could become intractable using exhaustive search, hence heuristics procedures or local GRN learning are frequently used.

A Bayesian network is a directed graphical probabilistic model that represents the joint probability distribution among variables in a decomposed form [24]. A Bayesian network has two components: a directed acyclic graph (DAG), G, which encodes conditional independent relationships among nodes (variables); and parameters, θ, which is specific a conditional distribution for each variable given its parents. Bayesian networks have been widely used in modeling gene regulatory system [18-21]. These tools have several major advantages: (1) The ability to handle imperfect (incomplete and noisy) data sets; (2) a greater ability to identify causal relationships; (3) direct method to combine domain knowledge and data. A less obvious issue with Bayesian networks is that the joint probability score decomposes into local conditional probability terms. Conditional probability is still a generalized correlative metric for the two-way dependency between the target and the parent set, hence cannot effectively tell the real causal relationships from confounding ones based on observational data. This is a severe issue when there are a large number of correlative variables, like coregulated genes in microarray data.

Mutual information is a probabilistic quantity defined in information theory to measure the similarity or dependency between two variables. Mutual information has been widely used to model gene networks [22, 23, 25]. Mutual information captures both linear and non-linear relationships between two genes, hence can replace correlation to learn more robust relevance or association networks. Mutual information may also work in place of conditional probability to capture dependency between target gene and parent gene set and learn directed causal networks [26]. In both cases, mutual information measure two-way dependency either between two parties (genes or gene sets). Real biological systems frequently involve more complicated, higher order relationships, such as transcriptional regulation coordinated by multiple transcription factors, binding interaction in protein complexes etc. Definition of mutual information has been extended to higher dimensional spaces to measure such high order relationships among multiple variables or variable groups [27-29]. In these senses, mutual information is potentially

more versatile than other probabilistic metrics such as conditional probability used in Bayesian networks.

These are purely computational methods to learn gene regulatory networks from high throughput expression data alone. The resulting models are limited to mRNA or protein levels, and do not include additional information about gene regulation such as promoter sequence data, TF binding data (CHIP-chip), and literature data that contain relevant and complementary information to expression data in GRN learning. When available and ready to integrate, these extra data bring more reliable, relevant and informative results. As such, data integration has been explored and established as an effective analysis strategy in literature works [14, 30, 31] and commercial applications [32-34]. Unfortunately, in most studies, other types of data are either unavailable or not amenable to incorporate. Methods learning from expression data alone or exhaust the maximal potential of expression data are undoubtedly highly valuable. Therefore, the first part of this thesis focuses on learning GRN with expression data as the only experimental data. In the remaining part of the thesis, I also explore pathway/gene set analysis, as data integration based strategy to infer gene regulatory mechanisms.

1.3 Pathway inference and gene set analysis

Differential expression analysis is a well established strategy to screen genes or sets of genes associated with specific phenotypes or sample conditions. There are two categories of differential expression analyses: individual gene analysis (IGA) and gene set analysis (GSA, also called category analysis, pathway inference or analysis) [35]. IGA evaluates the differential expression of individual genes between two sample groups. A list of significantly altered genes is then selected based a certain threshold and downstream biological interpretation is focused on this short list of genes. GSA on the other hand evaluates the coordinate differential expression of predefined gene sets between two sample groups. Downstream biological interpretation is based on the definition or annotation of most significant gene sets.

GSA differs from IGA in two key aspects. First, the zoom-in level of the expression pattern analysis: GSA screens differential expression signals at whole gene set level, whereas IGA screens signals at individual gene level. It is common sense that genes usually work in groups as regulatory pathways, functional groups or target sets in biological systems. Therefore, GSA tends to focus on the right level and see the bigger picture or more sensible patterns, yet IGA frequently zooms-in too much for the finer but less sensible and replicable patterns. Second, the use of prior knowledge: GSA incorporates functionally related gene sets derived from literature works or public databases [35, 36], which bring extra information unavailable in the expression data and place the data analysis in a more relevant context. Therefore, GSA is considered as a knowledge based analysis method.

Gene sets are collected from public databases and literature [35, 36]. Diverse biological knowledge and functional genomics data can be sources for gene set definitions, including signaling pathway (KEGG [37], GenMAPP [38], BioCarta [39] and Reactome [40]), Gene Ontology [41] (molecular function, biological processes, cellular components), genomic location or cytogenetic bands (EntrezGene [42]), cis-acting regulatory motifs for transcription factors or microRNAs (MsigDB [43]), coregulated/coexpressed gene groups (MsigDB [43]), and co-citation in literature (Entrez-PubMed [44]) for example.

A variety of methods have been proposed to test differential expression of gene sets. These methods can be divided into different categories based on whether a cutoff value is imposed on the sorted gene list, statistical inference procedures being used, and whether the detail of gene-gene relationships (topology) is considered.

Cutoff based methods [45] are the earliest form of GSA evolving directly from IGA [35]. Similar to IGA, all genes are ranked based on some differential expression score, and then a short list genes are selected and labeled as differentially expressed based on a cutoff value, with the rest genes as non-differentially expressed. We examine whether

predefined gene sets are overrepresented in this differentially expressed list, using a test for independence in a 2×2 (contingency) table. This approach has been widely used with many minor variations [43]. The major issues with this approach [45, 46]: (1) the choice of cutoff is arbitrary and testing results depended strongly on the choice of cutoff; (2) the ranking information is lost since all differentially expressed or non-differentially expressed genes are treated equally, which will lower the inference power.

Cutoff free methods [35, 36, 46-53] were proposed to address the issues with cutoff based methods. These methods rank all genes based on differential test, aggregate of per gene statistics or scores as the score for the whole gene set, and test whether such gene set scores are significant compared to random control scores. Commonly used per gene statistics or scores include Kolmogorov-Smirnov score, P-values, t statistics, fold changes etc [35]. Cutoff free methods can be further divided into two groups based on the statistical tests used for the significance of gene set scores: sample randomization and gene randomization [35, 51]. Sample randomization methods test significance of gene sets based on permutation of sample labels, with GSEA [36, 47], SAFE [53] and SAM-GS [52] as representatives. In contrast, gene randomization methods test the significance of gene sets based on permutations of gene labels or a parametric distribution over genes, with PAGE [48], T-Profiler [50] and Random-set [49] as representatives. Sample randomization keeps the correlation structure among genes but only applies to large expression data sets with multiple samples per experimental condition. Gene randomization has no limitation on sample size, but may break the correlation structure among genes.

Two newly developed methods incorporate network topology in gene set analysis. First, the impact analysis [54] considers not only the magnitudes of expression changes, but also more pathway-specific factors such as gene type and position in the given pathways, their interactions, etc. however, the impact factor (IF) score used in the impact analysis mixes log based P value and the sum of normalized absolute differential expression

(perturbation factor), hence is ad-hoc and hard to interpret. A more sensible scoring metric should count for gene position and interactions (topology), as well as magnitude and direction of the overall impact of pathway perturbations. Second, Chuang et al [55] extended gene set analysis to infer significantly perturbed protein interaction sub-network significant associated with particular phenotypes. However, network topology was only used to define candidate sub-network boundaries incrementally, and not considered in the sub-network score or statistical inference.

1.4 MYC centered GRN and BMP6 induced osteoblast differentiation and mineralization

From a biology-driven perspective, I am particularly interested in solving the regulatory mechanisms for two representative biological problems.

The first problem is MYC dependent transcription. MYC transcription factor has been established as a universal transcriptional regulator that affects the expression of significant portion of the whole genome [1]. By modifying the expression of its target genes, MYC involves in a variety of cellular processes, including cell cycle, apoptosis, differentiation and stem cell self-renewal, as well as multiple cancers [2]. Therefore, MYC centered GRN is universally important at both molecular level and phenotypic level. However, our current understanding of the system is both incomplete and inaccurate with key questions not answered: what are MYC target genes, and what are the cofactors MYC interact, and how is the specificity of MYC dependent transcription controlled?

Construction of a MYC centered GRN would give answer to these questions and better understanding of MYC affected physiological/pathological processes. Basso et al [22] published a big microarray study with 336 chips on human B cells under various experimental/physiological conditions. Throughout this study, MYC showed significant transcriptional regulatory activity and numerous known MYC target genes significantly correlate with MYC at gene expression level [22]. Theoretically, reconstruction of a

realistic MYC centered GRN is plausible from this microarray study. However, such GRN reconstruction is demanding and requires new effective method since no previous method is appropriate (more in Chapter 2).

The second problem is BMP6 induced osteoblast differentiation and mineralization [56]. This is a phenotypic change responsible for skeleton development and function, its deregulation is associated with multiple common bone diseases [3]. Our previous work showed BMP6 is a primary endogenous factor for human osteoblast differentiation and mineralization [56]. But little is known on BMP6 induced transcriptional programs, particular the regulatory signaling pathways or functional groups responsible for the induction of osteoblast phenotype and function.

High throughput microarray profiling experiments and pathway inference would dissect the regulatory mechanisms throughout this process. We designed and carried out a microarray study with BMP6 addition and withdrawal at different stages of the osteogenic induction. As in most other microarray studies, the sample size for this study is small with two replicates for each state. But different from many other studies, this study has a time series design. No previous pathway inference method handles such dataset effectively. I develop a new pathway inference method (Chapter 3) and apply it for a special temporal pathway inference study (Chapter 4).

1.5 Overview

1.5.1 Problem statement

GRN reconstruction and pathway inference are two basic and related biological problems (Figure 1.1 and 1.3). The former considers what genes regulate what other genes, the latter considers what genes or groups of genes regulate a specific phenotype. Pathways are networks of genes and proteins and their interactions involved in a biochemical process or signal transduction. GRNs are a subset of biochemical pathways, where regulatory signals are transmitted at gene expression level. GRN reconstruction, pathway inference and another problem, gene marker selection are all closely related (Figure 1.3).

Gene marker selection identifies the two-way association between individual genes and a specific phenotype without considering the interactions between these genes (part list, Figure 1.3a). GRN reconstruction instead solves the regulatory interactions among such functionally related genes conditioned on specific phenotype(s) (wiring, Figure 1.3b). Upon constructed, GRNs can be taken as functional modules or regulatory pathways (circuits, Figure 1.3c). Pathway inference examines whether such functional modules are significantly associated with a phenotype. Pathway inference accounts for the aggregate behavior of all genes in a module with or without considering the network topology.

Further connections between GRN reconstruction and pathway inference can be described from a statistical learning perspective (Table 1.1).

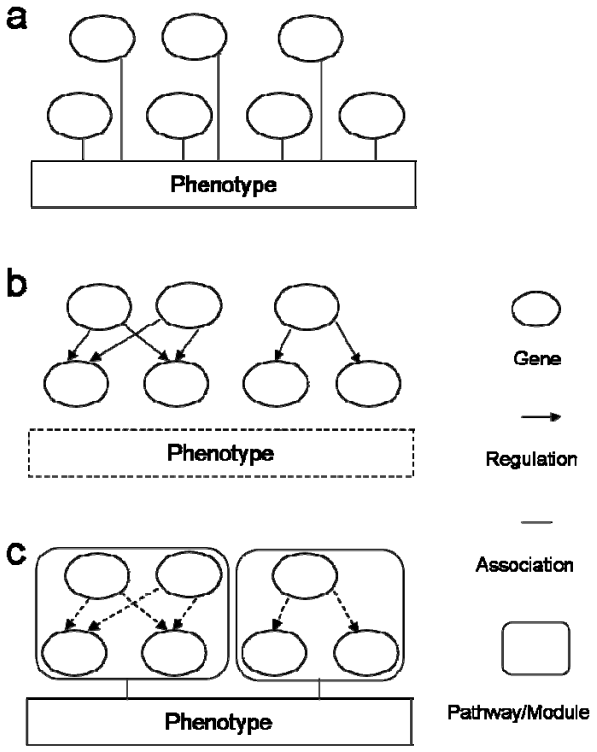
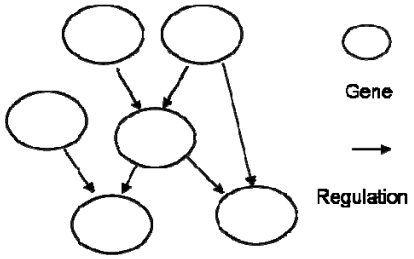
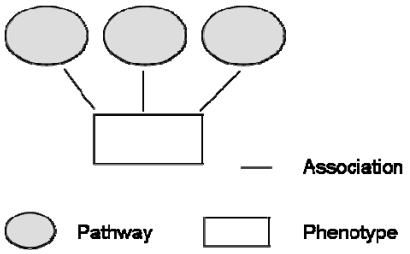


Figure 1.3 A schematic description and comparison of three statistical learning problems using gene expression data. (a) gene marker selection; (b) GRN reconstruction; (c) pathway inference. Parts plotted in dashed lines are dispensable for the problem. In this thesis, I focus on (b) GRN reconstruction and (c) pathway inference.

Table 1.1 Comparison between the two basic problems studied in the thesis, GRN

reconstruction and pathway inference.

Sub-problem	GRN reconstruction	Pathway inference
Super-problem	Reverse engineering	Feature selection
Graph		
Nodes	Gene (or pathway)	Pathway (or gene), discrete phenotype
Edges	Causal relationships between genes or pathways	Naïve two-way association with phenotype, often interpreted as causal.
Summary of Available Methods	Correlation based: clustering [6, 7], correlation networks [8, 9], graphical Gaussian models [10] Probability based: Bayesian networks [18-21], mutual information networks [22, 23]	Cutoff based: 2×2 contingency table [46] Cutoff free: Sample randomization methods such as GSEA [36, 47], SAFE [53] and SAM-GS [52]; Gene randomization methods such as PAGE [48], T-Profiler [50] and Random-set [49] Topology based: the impact analysis [54], protein interaction sub-network analysis [55]
Sample size requirement	Correlation based: $\sim 10^1$ data points; Probability based: $\sim 10^2$ data points	$\sim 10^0$ data points
New Method Proposed	MI3: continuous three-way mutual information network	GAGE: Generally Applicable Gene Set Enrichment (cutoff free, gene randomization)

1.5.2 Thesis outline

This work focuses on gene regulatory network (GRN) reconstruction and pathway inference from high throughput gene expression data and their applications (Figure 1.1). In chapter 2, I present a new method, MI3, for de novo GRN construction using continuous three-way mutual information. I validate the method systematically using synthetic data and applied the method to infer a regulatory network centered at the MYC transcription factor from a published microarray dataset. In chapter 3, I present another

novel method, GAGE, Generally Applicable Gene Set Enrichment for Pathway Inference. I validate GAGE and compare it to established pathway inference methods using published and new microarray datasets of different sample size, experimental design, and by using different profiling techniques. In chapter 4, we conducted a microarray study on transcriptional programs during BMP6 induced osteoblast differentiation and mineralization, and applied GAGE to infer the regulatory pathways and transcriptional signaling networks in the process. I will introduce each part of our work, their relevance and results briefly below.

Chapter 2 Gene Regulatory Network Reconstruction from High Throughput Gene Expression Data Using Continuous Three-Way Mutual Information

Probability based statistical learning methods such as mutual information and Bayesian networks have emerged as a major category of tools for GRN reconstruction from quantitative biological data. In this work I introduce a new statistical learning strategy, MI3 that addresses three common issues in previous methods simultaneously: (1) handling of continuous variables, (2) detection of more complex three-way relationships and (3) better differentiation of real causal versus correlative but confounding relationships. With these improvements, I provide a more realistic representation of the underlying biological system.

I test the MI3 algorithm using both synthetic and experimental data. In the synthetic data experiment, MI3 significantly outperformed the control methods, including Bayesian networks, classical two-way mutual information and a discrete version of MI3. I then used MI3 and control methods to infer a regulatory network centered at the MYC transcription factor from a published microarray dataset. Unlike control methods, MI3 effectively differentiated true causal models from confounding models. MI3 recovered major MYC cofactors, and revealed major mechanisms involved in MYC dependent transcriptional regulation, which are strongly supported by literature. The MI3 network showed that limited sets of regulatory mechanisms are employed repeatedly to control the

expression of large number of genes.

Chapter 3 GAGE: Generally Applicable Gene Set Enrichment for Pathway Inference

Gene set analysis (GSA), also called pathway inference, is a widely used strategy for gene expression data analysis based on pathway knowledge. GSA focuses on sets of related genes and has established major advantages over individual gene analyses, including greater robustness, sensitivity and biological relevance. However, previous GSA methods suffer from limitations in the sample size and experiment design of the data sets they apply to.

To address these limitations, I present a new GSA method called Generally Applicable Gene-set Enrichment (GAGE). I successfully apply GAGE to multiple microarray data sets with different sample sizes, experimental designs and profiling techniques. GAGE shows significantly better results when compared to two other commonly used GSA methods of GSEA and PAGE. I demonstrate this improvement in the following three aspects: (1) consistency across repeated studies/experiments; (2) sensitivity and specificity; (3) biological relevance of the regulatory mechanisms inferred.

GAGE reveals novel and relevant regulatory mechanisms from both published and previously unpublished microarray studies. From two published lung cancer data sets, GAGE derived a more cohesive and predictive mechanistic scheme underlying lung cancer progress and metastasis. For a previously unpublished BMP6 study, GAGE predicted novel yet biologically plausible regulatory mechanisms for BMP6 induced osteoblast differentiation, including the canonical BMP-TGF beta signaling, JAK-STAT signaling, Wnt signaling, and estrogen signaling pathways.

Chapter 4 Time Series Microarray Gene Expression Profiling and Temporal Regulatory Pathway Analysis of BMP6 Induced Osteoblast Differentiation and Mineralization

Osteoblast differentiation and function are implicated directly in skeletal development and bone diseases. Our pervious studies established that BMP6 as a primary endogenous regulator of human osteoblast differentiation and function. Although functionally critical,

BMP6 signaling largely remains uncharacterized. Key problems that still remain unsolved include: what pathways and gene groups are responsible for MSC differentiation to bone in response to BMP6 stimulation? How and when these pathways are altered (induced or repressed) by BMP6 during the process?

To answer these questions, we designed and conducted a time series microarray study on BMP6 induced osteoblast differentiation and mineralization. I conducted a comprehensive temporal pathway analysis using GAGE and predefined gene sets collected from KEGG, GO databases, and literature sources. I inferred novel and coherent sets of regulatory mechanisms and functional groups downstream BMP6 signal during osteoblast differentiation and mineralization. Different from other pathway analyses, our work captures the interconnections between individual pathways or functional groups and integrates them into a whole system. Besides the systems approach, this work also has a dynamic perspective. I not only inferred which pathways or gene sets are significant, but also determined when and how they are involved in the osteoblast differentiation and mineralization.

1.6 References

1. Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ, Ren B: **A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells**. *Proc Natl Acad Sci U S A* 2003, **100**:8164-9.

2. Dang CV, Resar LM, Emison E, Kim S, Li Q, Prescott JE, Wonsey D, Zeller K: **Function of the c-Myc oncogenic transcription factor**. *Exp Cell Res* 1999, **253**:63-77.

3. Favus MJ, American Society for Bone and Mineral Research.: **Primer on the metabolic bone diseases and disorders of mineral metabolism**, 6th edn. Washington, DC: American Society for Bone and Mineral Research; 2006.

4. Genomics:GTL: **Gene Regulatory Networks** [http://genomicsgtl.energy.gov/science/generegulatorynetwork.shtml] 2005

5. Schlitt T, Brazma A: **Current approaches to gene regulatory network modelling**. *BMC Bioinformatics* 2007, **8 Suppl 6**:S9.

6. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**:14863-8.

7. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization**. *Mol Biol Cell* 1998, **9**:3273-97.

8. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS: **Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks**. *Proc Natl Acad Sci U S A* 2000, **97**:12182-6.

9. Moriyama M, Hoshida Y, Otsuka M, Nishimura S, Kato N, Goto T, Taniguchi H, Shiratori Y, Seki N, Omata M: **Relevance network between chemosensitivity and transcriptome in human hepatoma cells**. *Mol Cancer Ther* 2003, **2**:199-205.

10. Schafer J, Strimmer K: **An empirical Bayes approach to inferring large-scale gene association networks**. *Bioinformatics* 2005, **21**:754-64.

11. Zhu Z, Pilpel Y, Church GM: **Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm**. *J Mol Biol* 2002, **318**:71-81.

12. Haverty PM, Hansen U, Weng Z: **Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification**. *Nucleic Acids Res* 2004, **32**:179-88.

13. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture**. *Nat Genet* 1999, **22**:281-5.

14. Clements M, van Someren EP, Knijnenburg TA, Reinders MJ: **Integration of known transcription factor binding site information and gene expression data to advance from co-expression to co-regulation**. *Genomics Proteomics Bioinformatics* 2007, **5**:86-101.

15. Mjolsness E, Mann T, Castano R, Wold B: **From Co-expression to Co-regulation: An Approach to Inferring Transcriptional Regulation among Gene Classes from Large-Scale Expression Data**. *Technical Report JPL-ICTR-99-4* 1999.

16. Wahde M, Hertz J: **Coarse-grained reverse engineering of genetic regulatory networks**. *Biosystems* 2000, **55**:129-136.

17. Strimmer K: **Notes: Graphical Gaussian Models for Genome Data** [http://strimmerlab.org/notes/ggm.html] 2006

18. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA: **Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks**. *Pac Symp Biocomput* 2001:422-33.

19. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data**. *J Comput Biol* 2000, **7**:601-20.

20. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP: **Causal protein-signaling networks derived from multiparameter single-cell data**. *Science* 2005, **308**:523-529.

21. Friedman N: **Inferring cellular networks using probabilistic graphical models**. *Science* 2004, **303**:799-805.

22. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells**. *Nat Genet* 2005, **37**:382-90.

23. Butte AJ, Kohane IS: **Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements**. *Pac Symp Biocomput* 2000:418-29.

24. Heckerman D: **A Tutorial on Learning with Bayesian Networks**. In: *Book A Tutorial on Learning with Bayesian Networks* (Editor ed.^eds.). City: Microsoft Research; 1995.

25. Steuer R, Kurths J, Daub CO, Weise J, Selbig J: **The mutual information: detecting and evaluating dependencies between variables**. *Bioinformatics* 2002, **18 Suppl 2**:S231-40.

26. Friedman N, Nachman I, Pe'er D: **Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm**. In: *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*; 1999; San Francisco, CA. 206-215.

27. Mcgill WJ: **Multivariate Information Transmission**. *Psychometrika* 1954, **19**:97-116.

28. Jakulin A, Bratko I: **Quantifying and Visualizing Attribute Interactions: An Approach Based on Entropy**. *arXiv:cs.AI/0308002* 2004.

29. Nemenman I: **Information theory, multivariate dependence, and genetic network inference**. *arXiv:q-bio/0406015* 2004.

30. Tanay A, Sharan R, Kupiec M, Shamir R: **Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data**. *Proc Natl Acad Sci U S A* 2004, **101**:2981-6.

31. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**:166-76.

32. Seifert M, Scherf M, Epple A, Werner T: **Multievidence microarray mining**. *Trends Genet* 2005, **21**:553-8.

33. Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T: **Pathway mapping tools for analysis of high content data**. *Methods Mol Biol* 2007, **356**:319-50.

34. Ganter B, Zidek N, Hewitt PR, Muller D, Vladimirova A: **Pathway analysis tools and toxicogenomics reference databases for risk assessment**. *Pharmacogenomics* 2008, **9**:35-54.

35. Nam D, Kim SY: **Gene-set approach for expression pattern analysis**. *Brief Bioinform* 2008.

36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**:15545-50.

37. **KEGG: Kyoto Encyclopedia of Genes and Genomes** [http://www.genome.jp/kegg/]

38. **GenMAPP (Gene Map Annotator and Pathway Profiler)** [http://www.genmapp.org/]

39. **BioCarta - Charting Pathways of Life** [www.biocarta.com/]

40. **Reactome** [www.reactome.org/]

41. **Gene Ontology** [www.geneontology.org/]

42. **Entrez Gene** [www.ncbi.nlm.nih.gov/sites/entrez?db=gene]

43. **the Molecular Signature Database** [http://www.broad.mit.edu/GSEA/msigdb/index.jsp]

44. **Entrez-PubMed** [www.ncbi.nlm.nih.gov/pubmed/]

45. Breslin T, Eden P, Krogh M: **Comparing functional annotation analyses with Catmap**. *BMC Bioinformatics* 2004, **5**:193.

46. Goeman JJ, Buhlmann P: **Analyzing gene expression data in terms of gene sets: methodological issues**. *Bioinformatics* 2007, **23**:980-7.

47. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, et al: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nat Genet* 2003, **34**:267-73.

48. Kim SY, Volsky DJ: **PAGE: parametric analysis of gene set enrichment**. *BMC Bioinformatics* 2005, **6**:144.

49. Newton MA, Quintana FA, Den Boon JA, SENGUPTA S, AHLQUIST P: **Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis**. *Ann Appl Stat* 2007, **1**:85-106.

50. Boorsma A, Foat BC, Vis D, Klis F, Bussemaker HJ: **T-profiler: scoring the activity of predefined groups of genes using gene expression data**. *Nucleic Acids Res* 2005, **33**:W592-5.

51. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies**. *Proc Natl Acad Sci U S A* 2005, **102**:13544-9.

52. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P, Yasui Y: **Improving gene set analysis of microarray data by SAM-GS**. *BMC Bioinformatics* 2007, **8**:242.

53. Barry WT, Nobel AB, Wright FA: **Significance analysis of functional categories in gene expression studies: a structured permutation approach**. *Bioinformatics* 2005, **21**:1943-9.

54. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R: **A systems biology approach for pathway level analysis**. *Genome Res* 2007, **17**:1537-45.

55. Chuang HY, Lee E, Liu YT, Lee D, Ideker T: **Network-based classification of breast cancer metastasis**. *Mol Syst Biol* 2007, **3**:140.

56. Friedman MS, Long MW, Hankenson KD: **Osteogenic differentiation of human mesenchymal stem cells is regulated by bone morphogenetic protein-6**. *J Cell Biochem* 2006, **98**:538-54.

Chapter II

Gene Regulatory Network Reconstruction from High Throughput Gene Expression Data Using Continuous Three-Way Mutual Information

2.1 Introduction

A major challenge in systems biology is to infer mechanistic gene interactions from high throughput microarray data [1, 2]. Underlying this challenge is the problem to find causal regulatory relationships among genes, or gene regulatory network (GRN) reconstruction. Robust solutions to this problem would provide us with a *transcriptomic map* of a genome that allows us to accurately predict the effect of gene perturbations.

Previous efforts to detect mechanistic relationships from gene expression data can be broadly divided into linear correlation and probability based methods. Linear correlation based methods, such as clustering [3, 4], correlation networks [5, 6] and graphical Gaussian models [7], have a long and fruitful history in statistical modeling and bioinformatics. These linear methods are computationally fast and relatively easy to interpret. However, a key limitation with these methods is that they assume linear relationships between variables. While some components of any transcriptional regulatory network are linear, nonlinear events such as OR, AND, and XOR type transcriptional regulation are relatively commonplace [8]. These nonlinear interactions would not be captured with a linear model, leading to spurious relationships between variables.

Probability based methods have also been used to detect relationships between genes. These probability methods include Probabilistic Boolean Networks (PBN) [9, 10],

Bayesian networks [11-14] and mutual information networks [15, 16]. Probability based methods can capture both linear and nonlinear regulatory relationships and are noise tolerant. However, many of the current probability based tools used in systems biology suffer from the following three limitations: (1) data discretization [9-14, 16], (2) pairwise testing [15, 16], (3) emphasis on correlation over causality [11, 12, 14, 17]. To transform continuous data into a more easily computable form, most probabilistic methods require the data to first be discretized into a finite number of bins, such as *high, medium,* and *low* [9-14, 16]: The number of bins used in discretization is difficult to choose, and is generally selected at some consistent yet arbitrary point. Unfortunately, different binning procedures can produce different analysis results [12], suggesting that the act of binning alone introduces errors into the analysis. Methods that search for pairwise associations only focus on a single relationship between regulator and target at a time. Pairwise association networks have been created using classical mutual information [15, 16]. However, simple pairwise relationships are likely less common than multivariate relationships in real biological systems, as the expression of most genes is regulated not by a single gene but more likely by multiple genes. Methods that allow multivariate interactions such as Bayesian networks or some fuzzy logic approaches [18] are inherently superior in this respect.

A final challenge in creating mechanistically predictive transcriptional models is the ability to identify not just correlative but also causal models. Although difficult, causal relationships have been learned properly from non-sequential observational data [19, 20]. Probabilistic graphical modeling methods like Bayesian networks have been used to infer causal models from gene expression data [12, 14]. However, many probabilistic approaches are able to make correlative networks but not necessarily causal networks [11, 12, 14, 17]. Their multivariate scoring metrics such as conditional probability and mutual information are still generalized two-way correlation between the target and the parent set. Similar to the classical two-way metrics, these generalized correlations alone cannot

differentiate between a causal versus confounding parent set. True causal relationships like genetic regulation feature positive higher order interaction [21, 22], the non-additive effect above the sum of the lower order interactions [22]. For instance, for regulation involved two regulators such as OR, AND, XOR type relationships, two regulators together account for much more in the target than they individually can (Table 2.1). Intuitively such non-additive effect can be described as coordination or synergy between parents (with respect to the target, more description in Methods). On the other hand, confounding models commonly have no or negative higher order interaction (redundant parents, see the results). We propose that with such high order interaction considered, we can better differentiate true causal model versus confounding models.

In this work, we demonstrate a novel algorithm that attempts to overcome all three limitations using a continuous high order mutual information based scoring metric we call MI3 (Mutual Information 3). Note that continuous two-way mutual information has been described previously [23]. High order interaction information (an extension of mutual information) has been employed to model complex interactions [21, 22, 24]. However, both two-way mutual information and high order interaction information are symmetric and as such unable to make causal statements. MI3 combines 3rd order interaction information with the asymmetric mutual information between target and regulator set to account for the direction of regulation. MI3 is novel as a combinatorial probabilistic metric and an integrated statistical learning method.

In this work, we compare MI3 to other probability based methods quantitatively and qualitatively using synthetic data where the true model is known. Next we apply MI3 and control methods to reconstruct regulatory networks centered at the transcription factor MYC from a published high throughput microarray dataset [15]. The learning results are then evaluated numerically and biologically.

Learning MYC centered transcriptional regulatory network represents an ideal test case for MI3 as MYC is a well characterized transcriptional regulator that acts in tandem with

a finite set of co-effectors and regulates the expression of a large group of genes [25-27]. MYC has been well investigated [26, 28, 29] and online databases of MYC targets [30] are available for validation purpose. Despite these efforts, many cofactors and targets remain unidentified, and corresponding regulatory mechanisms unknown [15, 25, 26, 28]. As a result, an integrated understanding of MYC dependent transcriptional regulation has remained out of reach [15, 25, 26, 28, 29]. In this study, we use MI3 to derive an accurate transcriptomic map surrounding MYC from the same gene expression dataset used to identify MYC targets [15]. The approaches used here are general and can be directly used for any transcriptional regulator given sufficient gene expression data.

Table 2.1 Examples of the non-additive property of high order interactions. The non-additive property of high order interactions, i.e. I(T;R1,R2)-I(T;R1)-I(T;R2) = I(T;R1;R2) >0, is shown by common types of regulatory relationships involving two independent parents (R1 and R2) and a target (T). Entropies (H's) and mutual information (I's) are calculated according to definitions in 2.5.1 Mutual information definition, extension and calculation. These are ideal cases. In reality, we don't always get positive high order interactions due to the data quality and absence of real regulators in the data. Hence we don't impose any threshold on high order interaction alone.

Relationship	OR	AND	XOR
Contigency Table	p R1 R2 T 1/4 0 0 0 1/4 1 0 1 1/4 0 1 1 1/4 1 1 1	p R1 R2 T 1/4 0 0 0 1/4 1 0 0 1/4 0 1 0 1/4 1 1 1	p R1 R2 T 1/4 0 0 0 1/4 1 0 1 1/4 0 1 1 1/4 1 1 0
H(T)	$2-0.75*\log_2 3$	$2-0.75*\log_2 3$	1
H(R1)=H(R2)	1	1	1
H(T,R1)=H(T,R2)	1.5	1.5	2
H(R1,R2)	2	2	2
H(T,R1,R2)	2	2	2
I(T;R1)=I(T;R2)= H(T)+ H(R1)- H(T,R1)	$1.5-0.75*\log_2 3$	$1.5-0.75*\log_2 3$	0
I(T;R1,R2)= H(T)+ H(R1,R2) -H(T,R1,R2)	$2-0.75*\log_2 3$	$2-0.75*\log_2 3$	1
I(T;R1,R2)- I(T;R1)-I(T;R2)	$0.75*\log_2 3-1$ =0.189	$0.75*\log_2 3-1$ =0.189	1

2.2 Results

2.2.1 MI3 validation with synthetic data

We validated MI3 against other commonly used methods listed in Table 2.2, including a discrete version of MI3 (dMI3), two-way mutual information (MI2) and a log conditional probability score used in Bayesian network (BN) learning. Learning was carried out using data sampled from a synthetic regulatory network, described in Figure 2.7 and Table 2.6, where the true network structure is known. We learned the best two-parent regulatory model (Figure 2.1) for each dependent node (u1-u6) by exhaustively searching through each possible parent set and scoring with each metric.

Table 2.2 MI3 and control methods evaluated and compared using the synthetic data. All scores are calculated based on continuous nonparametric probability density estimation, except dMI3 based on discretization using 5 bins of equal size.

Method	Metric	Description	Performance Rank[#]	
			Syn	Real
MI3	2*I(T;R1,R2)-I(T;R1)-I(T; R2) = I(T; R1| R2)+ I(T; R2| R1)	The sum of Correlative and Coordinative Criteria, which equals to the conditional mutual information between the target gene and the each regulator given the other regulator	1	1
dMI3	2*I(T;R1,R2)-I(T;R1)-I(T; R2)	Discrete version of MI3, control score to show the strength of continuous mutual information	3	2
Bayesian network (BN)	logP(T | R1,R2)[†]	Log conditional probability, control score which maximize correlation of the parent set to the target , while ignores the interaction between R1 and R2	2	3
Two-way MI (MI2)	I(T;R1)+I(T;R2)	Control two-way mutual information score to show the strength of three-way metric	4	4

† In this paper, log conditional probability and BN are used interchangeably.

Performance rank for real data experiment is based on qualitative comparison.

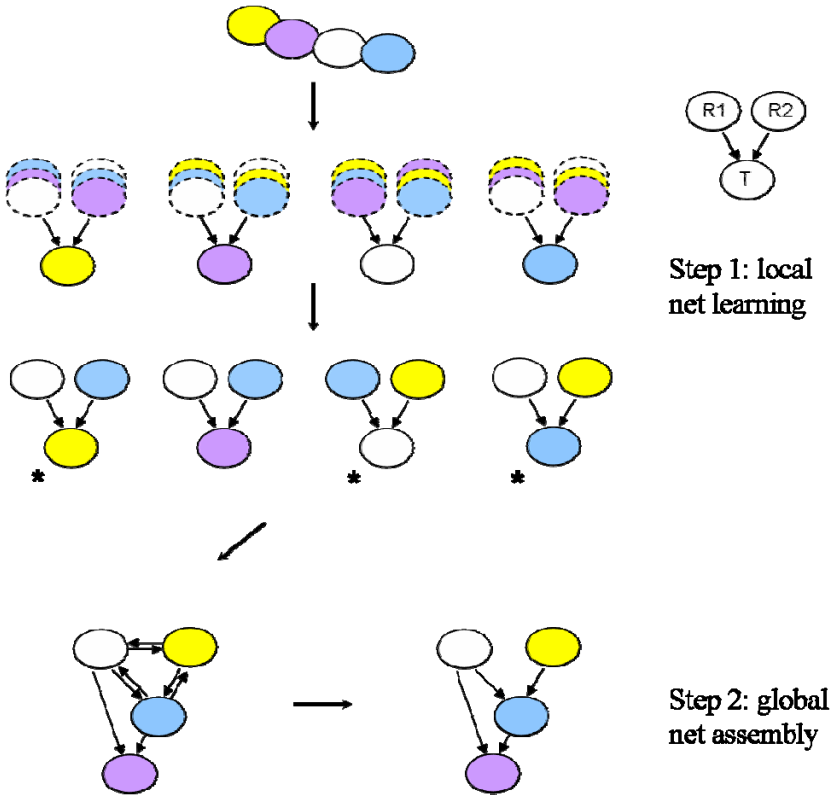


Figure 2.1 A schematic view of the network inference procedure for MI3 and control methods. We learn gene regulatory networks in two steps: (1) learn local regulatory network for each of the interesting nodes through an exhaustive search; (2) assemble local networks up into a unified network if needed. When there is no list of interesting nodes, all nodes becomes interesting in step (1). In the step (2), we may need to reconcile the conflicting local structures (labeled by *) if there are any, mainly the two way edges and cycles. In this work, the key difference between different methods is the score metric being used rather than the network inference procedure. For a fair comparison between scoring metrics, we simple assemble the local networks up without the reconciliation of conflicts in step (2).

The resulting best scoring network from a representative experiment is shown in Figure 2.2. Using the MI3 score, we recovered the true models for all dependent variables with exactly two parents, including u2, u3 and u5. For variables with fewer or more than two parents, i.e. u1, u4 and u6, MI3 detected the best two-parent representative of the true models. Continuous MI3 outperformed dMI3 as dMI3 identified poor models for u1, u4, and u5. The BN tended to select confounding yet correlative models with low or negative

coordination (parents overlapping in their correlation with the target) between the two parents. For example, the BN score selected u2+u3 and x3+u2 over x1+x2 as the top 2 models for u4. Therefore, the coordinative component in MI3 is necessary to differentiate the true parent set from the confounding set. Compared to MI2, MI3 as well as log conditional probability consistently gave more accurate models whenever there was a difference, demonstrating their advantage in capturing higher order relationships. The existence of two way edges or edges with reversed direction showed that MI2 could not

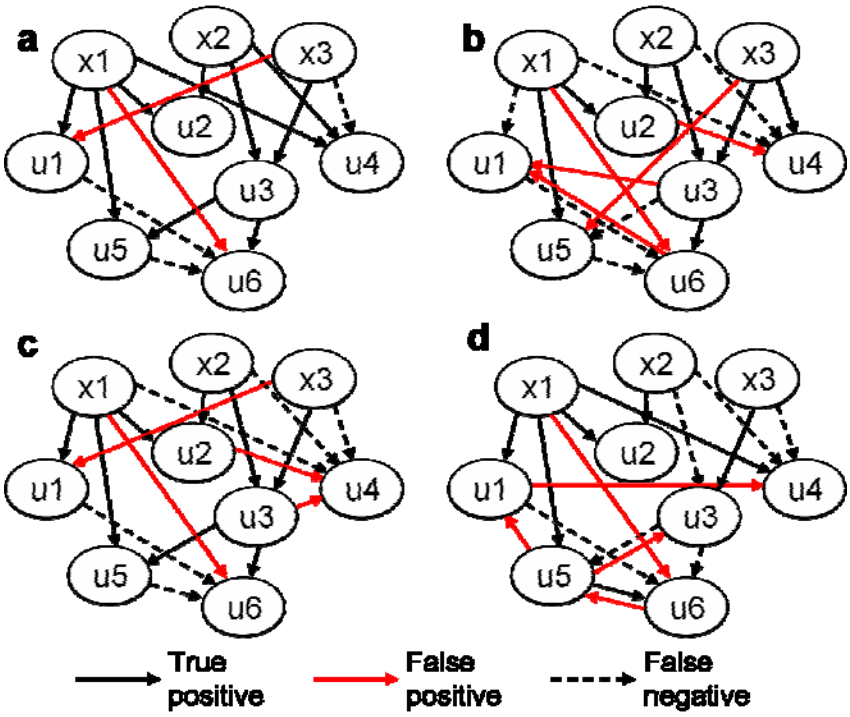


Figure 2.2 Networks inferred by MI3 or control methods from a 350-sample synthetic dataset. (a) MI3, (b) dMI3, (c) BN (log conditional probability) and (d) MI2. The best two parent model for each target gene was selected by using different methods and compared to true models. Here our interesting nodes are all the dependent nodes, u1-u6. Local regulatory networks are learned on these nodes and then assembled. When there is no information on dependent versus independent nodes, local networks are learned for all nodes including x1-x3. Conflicting local structures can be resolved in step (2) of Figure 2.1. For instance, the best two parents for x1 are u3 and u5, which conflicts with the local model for u5 whose parents are x1 and u3. Such conflicts were solved easily based on MI3 score, u3+u5->x1 scores 1.07 while x1+u3 ->u5 scores 1.49; hence the latter is the true model. The results remained essentially the same for MI3, BN and dMI3, but not for

MI2.

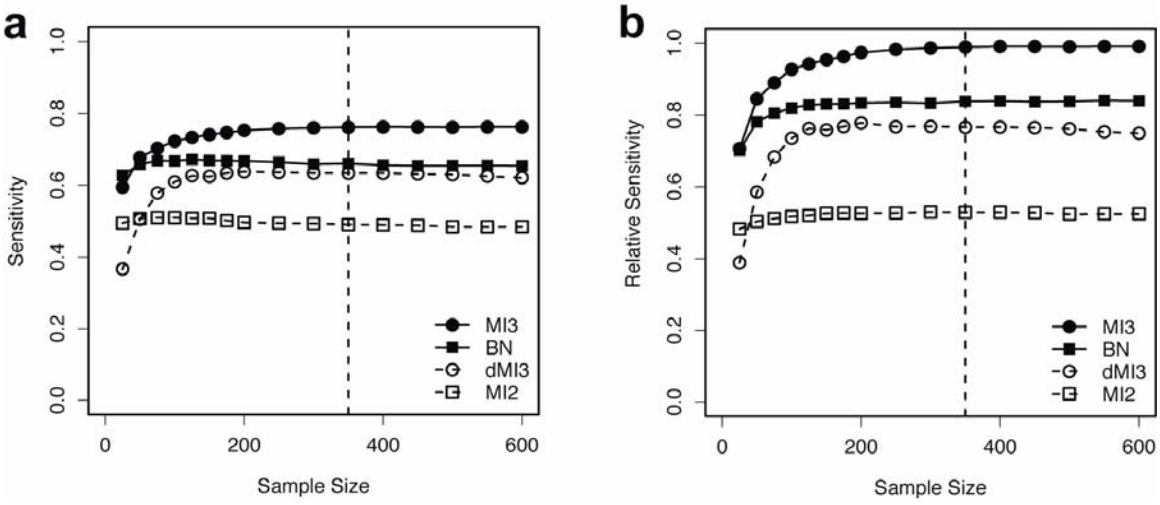


Figure 2.3 Sensitivity curves for MI3 versus control methods in learning two-parent models from the synthetic dataset. (a) Average absolute sensitivity of the 4 methods to recover the known network. (b) Average relative sensitivity of the 4 methods to recover the known network given that only two parents are possible for each dependent node. Vertical dashed lines marked sample size of 350 used in Figure 2.2, which is similar to the experimental sample size used for the MYC study.

identify direction of causality between variables. In addition, the two parents for nodes u1, u4, u5 and u6 picked by MI2 have highly negative coordination with each other. These results demonstrate that, among the methods tested, MI3 most accurately identified the underlying regulatory network for both linear and nonlinear relationships between variables (Table 2.6).

Next we quantitatively compared the performance of MI3 to other commonly used methods in terms of both sensitivity (ratio of correctly inferred interactions among all true interactions) and precision (ratio of correct interactions among all inferred interactions) [15]. In Figure 2.3, only sensitivity curves are shown because the precision curves are essentially the same but shifted. Figure 2.3a provides the absolute performance, while 2b shows the relative performance. The relative performance is a more meaningful comparison, given that the number of parents was fixed, although both results are quite similar. The absolute sensitivity and precision MI3 algorithm achieved were 0.77 and

0.83 respectively (Figure 2.3a), and the relative levels are both 0.99 (Figure 2.3b). In this comparison, MI3 consistently outperformed dMI3 across all different sample sizes. Also MI3 was more robust than dMI3 in that the sensitivity and precision curves have smaller error bars (standard deviation not shown for better plot view). In addition MI3 always outperformed the correlative BN. MI2 consistently demonstrated the lowest performance by a large margin as long as the sample size was greater than 25. All methods reached a plateau at ~250 samples, indicating that the 350 (or 336 for real data) sample default used in this paper is appropriate for all 4 methods to learn two parent regulatory models (3 nodes). Finally, all four methods were ranked in terms of performance in Table 2.2. Overall, MI3 always gave the highest true positive and the lowest false positive rate, and significantly outperformed all control methods (p-value=4.45×10^{-11}).

2.2.2 MI3 applied to high throughput microarray data

We used MI3 and control methods to infer regulatory network centered at MYC transcription factor from a human B cell microarray dataset. Note that the same dataset had been generated and used for identifying MYC target genes by another group using the mutual information tool ARACNE [15]. Instead of doing an exhaustive search of co-regulator pairs for each target as in the synthetic data, we fixed one of the regulators to be MYC and the target to be a known MYC target, and searched for the second regulator. This constraint imposed by our specific biological focus made the analysis more tractable and our results more testable, because we only need to select and test the second regulator (more details given in Footnote 2). Notice that this simplified problem is a sub-case of the synthetic problem. We are still using the same scoring metrics (Table 2.2) and following the same procedure (Figure 2.1), except that one parent node is fixed by introducing extra literature data. In this sense, all methods are still comparable. Experiments with synthetic data showed that such simplification does not change the final results as long as we are introducing a real parent of the target with enough marginal dependency, i.e.

I(T;R1)>0.3, for MI3, dMI3 and BN. For MI2, fixing R1=MYC does change the results, but it makes sense when taken as prior knowledge introduction. We pre-filtered MYC targets, T, with I(T; MYC) ≥0.3 to prevent bias upon fixing R1=MYC, and to speed up analysis similar to candidate parent set selection in the sparse candidate algorithm [31].

The verified targets were retrieved from the MYC Target Gene Database [30] available online [32]. After pre-filtering using the constraint I(T; MYC) ≥0.3, 368 MYC targets remained as shown in Table 2.3. For each filtered target of MYC we selected top 5 cofactor (R2) models using MI3 or control methods. Because for each target gene, there are usually multiple models which score almost the same and are equally interesting. For example, several coregulated cofactors are involved, or multiple genes in a pathway/complex represent the same regulatory action equally well. This is slightly different from the synthetic experiment, where only there is 1 true or best model for each target and the number of regulators is known. Nonetheless, keeping only top 1 model led to almost the same lists of most frequently selected cofactor (Table 2.5) as the list based on top 5 models (Table 2.4), except that the number of targets mapped to individual cofactors was smaller. All other comparisons between MI3 and control methods led to the same results when top 1 models were used (not shown).

Table 2.3 MYC target pre-selection based on two-way mutual information. Genes are selected to be potential MYC targets (T) based on criterion I(MYC; T) > specific cutoff value: cutoff value vs total number of targets, number of targets verified against the MYC target database (http://www.myccancergene.org/), and the verified ratio.

Cutoff	Selected†	Verified	Verified Ratio
<0.1	8358	1156	0.138
0.1	4042	733	0.181
0.2	2226	513	0.230
0.3	1303	368	0.282
0.4	634	231	0.364
0.5	249	107	0.430
0.6	58	34	0.586
0.7	3	3	1.000

† MYC gene itself was pre-excluded from the target selection

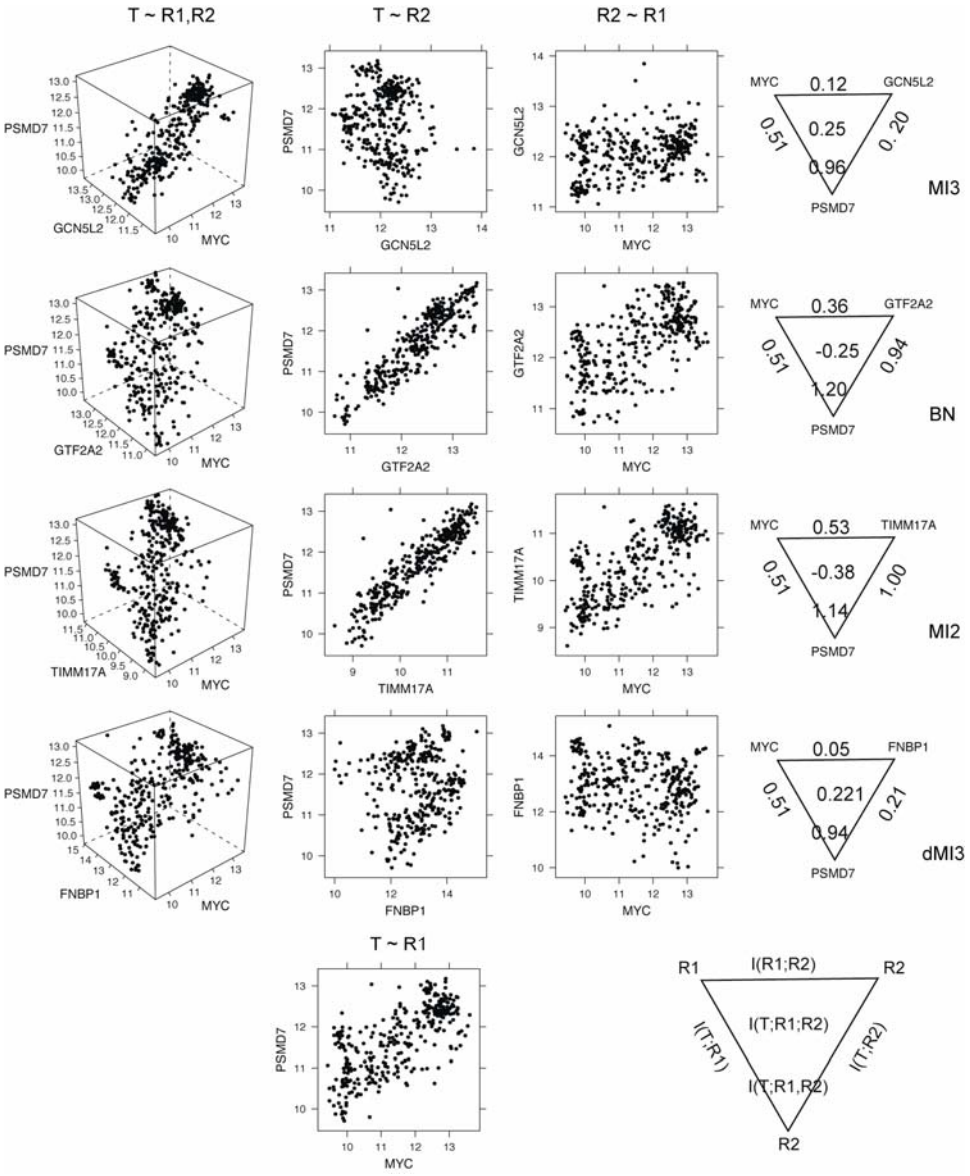


Figure 2.4 Two-way and three-way gene expression patterns and mutual information for representative top two-parent models inferred by MI3 and control methods. For all models, T= PSMD7, R1=MYC. The first three columns show the three-way and two-way gene expression patterns, and the fourth column the mutual information triangles. The bottom row shows the two-way expression pattern for PSMD7-MYC and the legend for mutual information triangle. This Figure 2.gives a concrete example for the difference between MI3 and control scores, echoing the results in Figure 2.5. For high throughput gene expression data, the BN and MI2 metrics both pick up models with high mutual information between parents and between either parent and the target. MI3 selected

relationships with slightly lower I(T;R1,R2) but I(T;R1;R2) much higher than the BN and MI2 metrics.

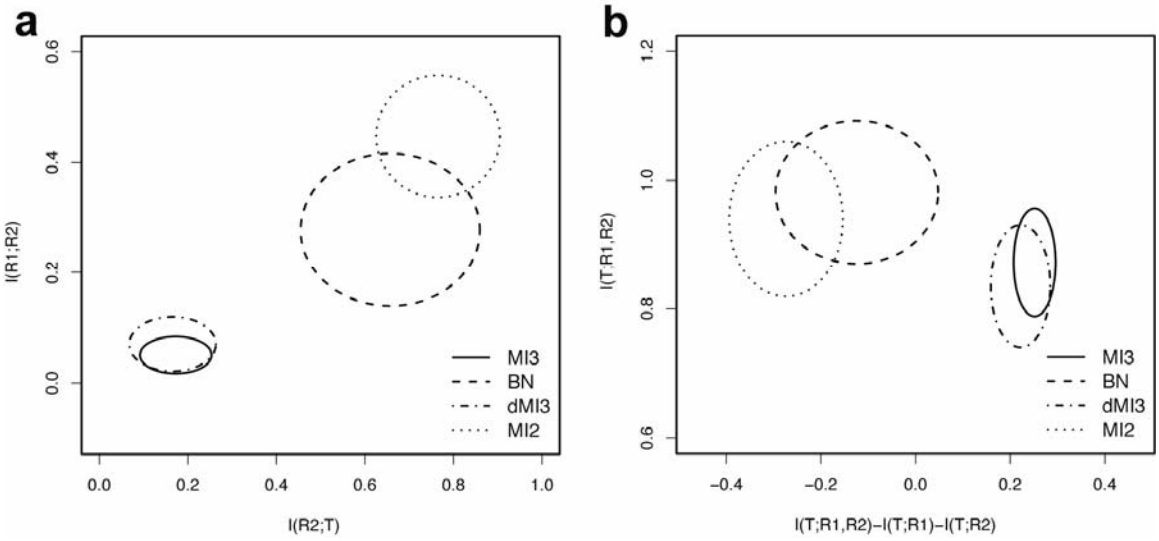


Figure 2.5 Two-way and three-way mutual information distributions for top models selected by MI3 and control methods. For each MYC target gene, the top 5 R2 or MYC cofactors were selected by applying different scoring metrics to the microarray dataset generated by Basso et al [15]. (a). I(R1;R2) vs I(R2;T), i.e. two way mutual information between R2 and R1 or T, (b). I(T;R1,R2) vs I(T;R1,R2)−I(T;R1)−I(T;R2), i.e. the correlative and coordinative components of MI3 score for the top 5 models selected by MI3 or control methods. Each ellipse represents the distribution of top 5 models in the specified mutual information coordinates, with mean as center and standard deviations as width and height. Note that I(R1;T) scores are the same for all methods hence not shown in (a).

MI3 and dMI3 selected models with significant coordination I(T;R1;R2), whereas the BN and MI2 selected models with high two-way dependency or I(T;R2) (note that I(T;R1) is constant because R1 is fixed to MYC) shown by Figure 2.4-5. Models inferred by all methods showed distinct patterns when plotted in three dimensions (T~R1,R2 in Figure 2.4), which means two parents together explain the target expression well. The difference is that BN and MI2 models showed distinct two dimensional patterns as well (T~R1 and R1~R2 in Figure 2.4), while the MI3 and dMI3 models did not. What MI3 and dMI3 captured are 3-way interactions in that neither of the two parents alone can describe the target well enough. In contrast, the relationships BN and MI2 captured are essentially two-way, and as such do not require both parents. This outcome is not surprising in that

the MI3 metric favors strong three way interactions, while the BN and MI2 methods have no such favor and as such would be expected to include confounding two-way models more frequently.

Table 2.4 Top 10 most frequently selected coregulators for the 368 verified MYC targets using different methods. Top 5 highest scoring cofactors are counted for each target. Cofactors in bold font are involved in MYC dependent or general transcriptional regulation, those in italics are in the list of 368 verified MYC targets with I(T; MYC) ≥ 0.3.

Method	MI3		dMI3		BN		MI2	
Rank\R2	Symbol	Targets	Symbol	Targets	Symbol	Targets	Symbol	Targets
1	ARPC1B	45	**PSIP1**	46	**HAT1**	23	*CTPS*	29
2	TRIP12	45	FNBP1	42	**GTF2A2**	15	*JTV1*	24
3	**ASH2L**	41	MRPL28	28	*PSMD14*	14	MRPL3	23
4	**GCN5L2**	35	RAB33A	23	PSMA4	13	**SSRP1**	21
5	**SHOC2**	25	HSPB1	22	*SFRS1*	13	TPX2	20
6	**CSK**	23	TPP2	21	PSMA3	12	*PSMB7*	19
7	**ZNF143**	23	ANKMY2	18	ADRM1	11	*RFC4*	19
8	FNBP1	22	CD59	18	**DNMT1**	10	***MCM7***	18
9	**MIZF**	22	KIAA0922	17	*CCT5*	10	**HAT1**	18
10	**CBX1**	19	**SIAH2**	17	WDR62	10	*HSPC111*	17

Table 2.5 Top 10 most frequently selected coregulators for the 368 verified MYC targets using different methods. Top 1 highest scoring cofactor is counted for each target. Cofactors in bold are involved in MYC dependent or general transcriptional regulation, those in italics are in the list of 368 verified MYC targets with I(T; MYC) ≥ 0.3. This table based on top 1 MYC cofactors is directly comparable to Table 2.4 based on top 5 MYC cofactors.

Method	MI3		dMI3		BN		MI2	
Rank\R2	Symbol	Targets	Symbol	Targets	Symbol	Targets	Symbol	Targets
1	**ASH2L**	18	**PSIP1**	19	*PSMD14*	4	MRPL3	6
2	TRIP12	14	FNBP1	19	*SFRS1*	4	*PES1*	6
3	**ZNF143**	13	MRPL28	14	TXNDC9	3	*HSPC111*	6
4	ARPC1B	11	NIPSNAP1	7	PCID1	3	SSBP1	6
5	**CSK**	9	CD59	7	**GTF2A2**	3	**SSRP1**	6
6	**SIAH2**	7	RAB27A	6	*MSH2*	3	***MCM7***	5
7	FNBP1	6	ACOT8	5	NDUFAB1	3	TMEM53	5
8	**MIZF**	6	ARPC5	5	PSMA3	3	*JTV1*	5

9	**GCN5L2**	6	KIAA0922	5	KIF23	3	TPX2	4
10	PRPSAP1	5	**SIAH2**	5	*CHERP*	2	*MAD2L1*	4

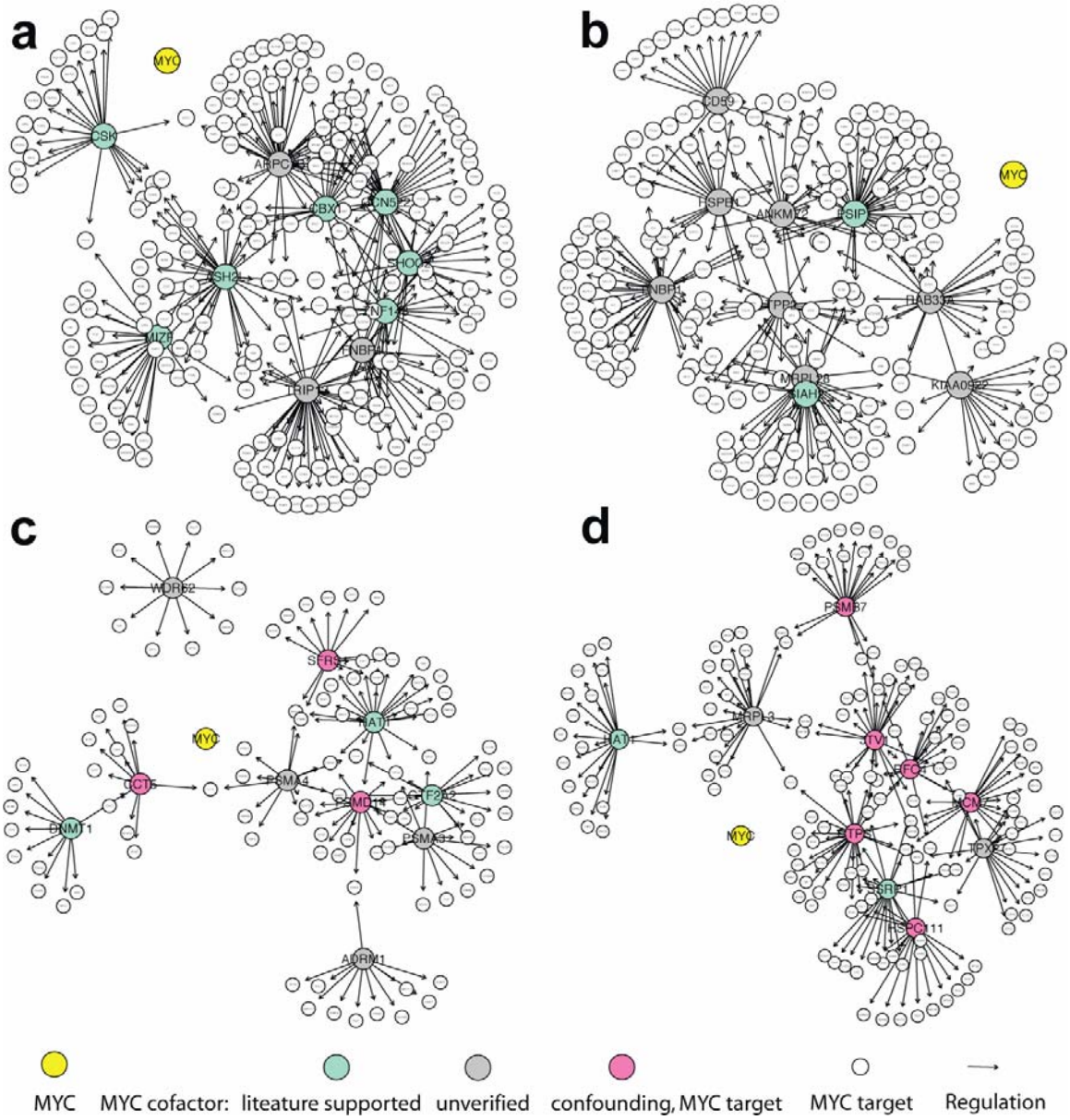


Figure 2.6 The transcriptional regulatory networks centered at MYC transcription factor. Networks included the top 10 most frequently selected MYC cofactors by using MI3 or control methods and the corresponding target genes (transparent). (a-d) are networks inferred by MI3, dMI3, BN and MI2 respectively. Regulators are large nodes and targets are small transparent nodes. Node colors indicate the identity where yellow is MYC, aquamarine are the cofactors involved in MYC dependent or general transcriptional regulation according to literature, gray are unverified cofactors, pink are confounding

cofactors that are actually verified MYC targets. Edges represent transcription regulation. Note that all edges from MYC to targets are hidden for clarity.

BN and MI2 models had low or negative 3-way coordination, and are likely confounding models. The relationship R2~R1 is similar to T~R1 and T~R2 follows a nearly perfect linear pattern (Figure 2.4). Such high similarity between R2 and T is less likely true regulation but rather coregulation relationship when considering various other factors affecting the target gene expression that are not counted by transcription level of the regulator(s), such as mRNA to protein translation, protein modification, and localization of the regulator. We expect that the R2 factors predicted by the BN and MI2 methods is most often another MYC target tightly coregulated with T instead of a coregulator, and indeed many top R2 are MYC targets (Table 2.4 and Figure 2.6, more description next).

Next we collected the top 5 cofactors and ranked each cofactor according to its frequency of being selected. Table 2.4 lists the top 10 most frequently selected cofactors using the four methods. Transcriptional regulatory networks centered at MYC were constructed based on the top 10 cofactors and corresponding targets, as shown in Figure 2.6. Literature validation was focused on these top 10 cofactor lists (Table 2.4).

The top 10 cofactors captured by MI3 and dMI3 were more informative and inclusive regulatory mechanisms than those captured by BN and MI2 (Table 2.4). Correspondingly, top 10 cofactor transcriptional regulatory networks constructed by MI3 and dMI3 were larger than the networks created by BN and MI2 (Figure 2.6). Out of 368 MYC targets, MI3 places 56.3% of these targets while dMI3 places 51.6%, BN places 26.9%, and MI2 places 41.8% of the targets. In other words, more MYC target genes are regulated by the top 10 mechanisms inferred by MI3 or dMI3, which is more consistent with the current mechanistic understanding of MYC dependent transcription that MYC regulates a large number of targets (>1000 verified) [25, 30] as a global transcriptional regulator yet only interacts with a small set of cofactors (13 listed) [26, 28].

Biologically, top 10 MYC cofactor list selected by MI3 was more consistent with the literature than the lists created by the control methods (Table 2.4). Seven out of ten MI3

top MYC cofactors are involved in MYC dependent or general transcriptional regulation. GCN5L2 (known as human GCN5), ASH2L, MIZF, CBX1 (HP1 beta homolog Drosophila) are chromatin structure modifiers, which change chromatin structure around target genes through chemical modification hence activate or repress their transcription. Chromatin structure modification by GCN5L2 and similar enzymes is a well documented mechanism for MYC dependent transcriptional regulation [26, 27, 33, 34]. ZNF143 [35] and MIZF [36] are transcriptional factors. CSK phosphorylates and activates GSK-3beta directly [37] and indirectly [38], while GSK-3beta phosphorylates, deactivates MYC and promotes its degradation [26]. SHOC2 complexes with Ras and Raf and enhances MAP kinase activation [39, 40], which in turn positively regulates MYC stability/activity by phosphorylation [26]. In contrast, only 2 (PSIP1, SIAH2), 3 (HAT1, GTF2A2, DNMT1) and 3 (SSRP1, MCM7, HAT1) top 10 MYC cofactors selected by dMI3, BN and MI2 respectively are transcriptional regulators based on Gene Ontology and literature. Moreover, 3 (SFRS, CCT5, PSMD14) and 6 (CTPS, JTV, PSMB7, RFC4, MCM7, HSPC111) top 10 MYC cofactors selected by BN and MI2 respectively are actually from the 368 verified MYC targets. Other top 10 cofactors selected by BN and MI2 are likely 'unverified' MYC targets given that they either share function annotations have similar expression profile with these questionable cofactors. In other words, BN and MI2 frequently produced confounding models where target genes were mistaken as MYC cofactors, while MI3 and dMI3 produced no confounding models. In Figure 2.6d, the two-way edges between red nodes suggest that MI2 not only confounded coregulators with targets, but also failed to tell the causal direction of the relationships. Combined with numerical comparison in Figure 2.4-5, these biological results show that unlike BN and MI2 scores, MI3 score effectively differentiates true causal models from confounding models because it takes the interaction between regulators into account.

2.3 Discussion

In this study, we have used MI3 to identify mechanistically plausible relationships from gene expression data. For synthetic data, MI3 recovered all true two-parent models, or the best representatives of the true models, and showed superior performance over the commonly used probability based methods including Bayesian networks and classical two-way mutual information. For experimental data, MYC cofactors identified by MI3 are either true or strongly supported by literature, while cofactors identified by control methods make little sense. Notably, the same microarray dataset has been used to identify MYC targets based on two-way mutual information [15].

MI3 uses three strategies to improve its predictions. First, MI3 does not require data discretization, and as such retains more of the information in the data. This continuous method enhanced the learning quality significantly, as shown by the synthetic example in Figure 2.2-3. Second, we extended classical two-way mutual information to three-way, which allows MI3 to capture more complex relationships between regulators and targets. Third, the MI3 score considers high order interaction or coordination and better differentiates causal relationships from confounding relationships as was shown by both the synthetic and MYC problem (Figure 2.2 and 2.6).

MYC cofactors predicted by MI3 details agree with the established literature. Notably, four of the top 10 cofactors selected by MI3 are chromatin structure modifier genes, suggesting that chromatin structure modification is the primary mechanism for MYC dependent transcriptional regulation. This inference is directly supported by the independent experimental results of Knoepfler et al (21), which provides further evidence of the role of MYC on chromatin structure modification via histone acetylation and methylation. Among the top MYC cofactors identified by MI3, GCN5L2 [26, 27, 29, 41], CSK [26, 37, 38], and SHOC2 [39, 40] are known or presumed coregulators for MYC transcriptional activity. All other seven MYC cofactors selected by MI3 are novel, although their connections to MYC or transcription are well documented. All these results demonstrate that MI3 is an accurate and powerful method to infer regulatory

models from microarray data. In contrast, top MYC cofactors inferred using control methods make much less sense biologically. Fewer of them are known transcriptional regulators and none of them is directly connected to MYC function. The fact that multiple MYC targets were mistaken as top MYC coregulators suggests that BN and MI2 methods have difficulty inferring true causal relationships from high throughput gene expression data. It is not likely that these MYC targets taken as co-regulators are real co-regulators because of feedback loops, since almost all of them are not functionally related to transcriptional regulation or MYC regulation activity. Similar confounding regulators were selected by control methods in the synthetic example (Figure 2.2). Figure 2.4-5 show why such confounding models occurred. There are likely feedback loops in MYC regulation, however these feedback relationships could only be identified with knockout data or time series data would be needed for the inference. In this work we only consider the general case where non-sequential observational gene expression data are available.

Learning from high throughput microarray data was different from learning from the small synthetic dataset. Differences between methods were larger for the microarray data (Figure 2.6 and Table 2.4), compared to the synthetic experiment (Figure 2.2). For the microarray data, MI3 and dMI3 were closer, whereas for the synthetic data BN and MI3 were closer (Table 2.2). This change in ranking suggests that the coordinative component was more significant than the difference made by using continuous versus discrete metric (MI3 vs dMI3) or 3-way versus 2-way metric (BN vs MI2) for microarray data, but not for synthetic data. These differences between microarray data and synthetic data can be ascribed to the fact that large numbers of highly correlative confounding models exist for the microarray data due to the large number of variables (genes), especially coexpressed genes, while the synthetic data contained relatively fewer possible confounding models.

The high order mutual information framework presented here is generally applicable, although we have only described and used three-way mutual information. The same set of strategies can be used to model arbitrarily high order relationships. To learn a regulatory

model with d dimensions or nodes (1 child with d-1 parents) by exhaustive searching through a system with v variables, we need ~$10*5^d$ data samples for nonparametric probability density estimation [42-44], and computation time is $O(v^d)$. Although $10*5^d$ is conservative compared to sufficient sample size indicated in the performance curve, ~250 for d=3 (Figure 2.3), undoubtedly, both the required dataset size and computational time exponentially increase with d. Therefore, 4-way or 5-way relationships require more samples than currently available microarray chips.

Through the use of MI3 we have demonstrated that tailored probability based metrics can outperform more standard methods used in systems biology for identifying mechanistic regulatory relationships. We expect that future enhancements to these scoring metrics are possible to identify larger sets of regulators while making fewer assumptions during the analysis.

2.4 Methods

2.4.1 MI3 algorithm

The MI3 algorithm is a novel three-way mutual information engine for local causal model inference. The algorithm is limited to three-way mutual information (two regulators and one target) (Figure 2.5), but the same method can be easily extended to higher order mutual information to model more complicated regulation mechanisms. Note that we call all types of mutual information involving 3 variables 3-way mutual information (2.5.1 Mutual information definition, extension and calculation), while three-way interaction information refers to I(T;R1;R2) only.

The MI3 scoring function has two parts, including correlative and coordinative information components. The correlative component measures the correlation between the target and the parent set, similar to other correlative probabilistic metrics such as log conditional probability for Bayesian networks.

Correlative component: I(T; R1,R2)

Here I() is the mutual information function, T is the target gene, and R1 and R2 are the

regulators as illustrated in Figure 2.1. Mutual information definition and high order extensions are describe in detail in the 2.5.1 Mutual information definition, extension and calculation. Pairs of regulators accurately describing the expression of the target gene will score well by the correlative component.

The coordinative component measures the coordination effect between the regulators with respect to the target. Note this component is actually the third order interaction information between T, R1 and R2, i.e. I(T; R1; R2) [22], and is three-way symmetric.

Coordinative component: I(T; R1,R2)- I(T; R1)-I(T; R2)

The coordinative component of the score identifies how well pairs of regulators versus individual regulators predict the target (examples in Table 2.1). Confounding models commonly have a negative coordinative score because parents overlap in their correlation with the target. The coordinative component can be rearranged to I(T; R1|R2)- I(T; R1), suggesting that this component measures how much better R1 predicts T given R2 versus not given R2. The coordinative component provides a quantitative measurement for the well-known 'selection bias' (also called Berkson's paradox) [45] in statistics or the 'explaining-away phenomenon' in Bayesian network theory [46].

The MI3 score is the sum of the correlative and coordinative component.

MI3 score: 2*I(T; R1,R2) - I(T; R1)-I(T; R2)= I(T; R1| R2)+ I(T; R2| R1)

The symmetric coordinative component captures higher order interactions and differentiates causal relationships from confounding ones without telling the causal direction. The asymmetric correlative component determines the direction of the causal relationship. By merging these two components, the MI3 score considers connections between the regulators as well as dependency between child and regulators. The MI3 score can be rearranged and simplified to I(T;R1|R2)+ I(T;R2|R1). This rearrangement can be interpreted as the conditional mutual information between the target gene and the each regulator given the other regulator, which better shows the three-way nature of this score. The MI3 score is structurally different from yet related to other probability scoring

metrics such as log based conditional probability used in Bayesian network learning logP(T|R1,R2)[11, 12] and two-way mutual information I(T;R1)+I(T;R2) [15, 16] (described in Table 2.2 and 2.5.2 Comparison between MI and log-based local conditional probability).

Network inference procedure

Regulatory network inference procedure based on MI3 is shown in Figure 2.1. Note that the key difference between MI3 and control methods is the scoring metrics, less in the network construction procedure. For a fair comparison between methods, we keep the procedure for all methods the same as in Figure 2.1. For more details on how the local network was selected see Footnote 1.

MI3 is implemented in the statistical computing language R, and codes are available online [47].

2.4.2 Nonparametric probability density estimation for continuous variables

To avoid discretizing our data to calculate mutual information, we have adopted a continuous method for mutual information calculation based on a classical nonparametric Gaussian kernel method in probability density estimation [42, 43]. To estimate the probability density at a specific location, we used all our data points. First we calculate the probability density at an interesting location based on a Gaussian distribution centered at each data point (kernel), and then take the average of all these densities using the following expression:

$$f(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(2\pi\sigma^2)^{d/2}}e\left(-\frac{|x-xi|^2}{2\sigma^2}\right) \qquad (1.1)$$

Here x is the position where probability density is to be estimated, and xi ($i=1,2,,N$) is the ith data point, both x and xi are d-dimension vectors, σ is the standard deviation of the kernel Gaussian distribution. We used optimal bandwidth described by Scott [43]. As others have noted, the choice of kernel distribution makes little difference in probability

estimation [42]. The reason we chose to use a Gaussian kernel is that it is intuitive and the result probability density distribution is continuous and infinitely differentiable [42]. Data may be transformed into a uniform distribution before the kernel density estimation to eliminate the potential effect of specific distributions. We found uniform transformation does help but the improvement is limited when the gene expression data are log transformed.

Following our description above, to calculate entropy and mutual information for continuous variables, we calculated a probability density estimate at the positions of sample data points, then took the sample mean of log probability density [23], to approximate the full integration. The probability density estimation was the most computationally intensive step for this work.

Nonparametric probability density estimation for continuous variables effectively eliminates the inaccuracies introduced by discretizing data. However, this method is computationally demanding, and requires a large sample size [42, 43]. Due to these limitations, we limited our MI calculation to 3 variables. Notice that the sufficient sample only depends on the number of relevant dimensions of the local models (3 nodes, Figure 2.1), and has nothing to do with the size of the total number of variables.

To compare our continuous approach to more commonly used discretization approaches, we used 5 bins of equal size.

2.4.3 Generation of synthetic testing data

Synthetic data was used to validate our MI3 method as an example of a completely known gene regulatory network. We created a synthetic network structure with algebraic relationships between variables found in Figure 2.7 and Table 2.6 online. We sampled 25 to 1000 samples from this network to generate a set. At each sample size, the sampling-learning procedure was repeated 500 times to determine the average sensitivity and precision of MI3 and control methods. This model structure is designed to mimic a miniature gene regulatory system, with regard to the network size, overall and local

structure, and dependency relationships.

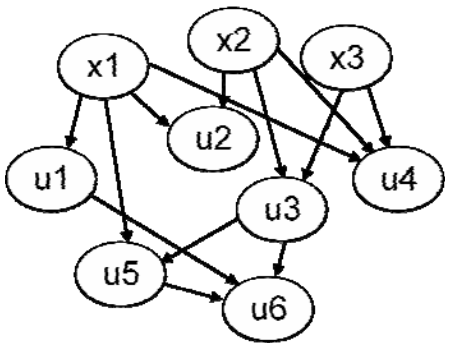


Figure 2.7 The synthetic gene regulatory network. This synthetic model structure is designed to mimic a miniature gene regulatory network, with several major features. First the network contains a number of variables, 9 variables in total, 3 of which are independent and 6 dependent. Second, the variables are assembled into a hierarchy of regulatory relationships, with independent variable mimicking regulators and cofactors, and dependent variables mimicking target genes. Third, the complexity of the network is controlled in that dependent variables have 1-3 parents, mostly 2 or 3, and each regulator/cofactor controls a set of targets. Targets may share regulators and thus may have different levels of coregulation/coexpression, which can lead to confounding models. Fourth, a diverse set of continuous non-linear and logical relationships among variables were encoded by the algebraic formulas in Table 2.6 to describe a realistic, yet complicated regulatory network.

Table 2.6 Relationships encoded into the true models for the synthetic dataset. N(μ, σ) is a normal random distribution with mean of μ and standard deviation of σ.

Variable	Algebraic formula	True parent set
x1	N(0,1)	
x2	N(10,5)	
x3	N(0,10)	
u1	$(x1)^3$ + N(0,0.1)	x1
u2	x1 + N(0,0.1), x1+10≥x2 x2/10 + N(0,0.1), x1+10<x2	x1, x2
u3	(x2-x3)/(x2+10) + N(0,0.05)	x2, x3
u4	x1+sin(x3) + N(0,0.1), x1+10≥x2 x2/10+sin(x3) + N(0,0.1), x1+10<x2	x1, x2, x3
u5	log(exp(x1)+exp(u3)) + N(0,0.1)	x1, u3
u6	(u1+u5)*u3/2+ N(0,0.05)	u1, u3, u5

2.4.4 Gene expression data processing and annotation

A gene expression dataset of human B cells with 336 samples was used for our study. These data were collected on the Affymetrix HG-U95Av2 platform and published by another group [15]. The raw data in .CEL format was collected from Gene Expression Omnibus (GEO) and processed by using RMA [48] method implemented in Bioconductor [49] Affy package [50]. A up-to-date probe set definition (.CDF file) based on Entrez Gene sequence, Hs95Av2_Hs_ENTREZG_7, created by the Microarray Lab at University of Michigan [51, 52], is used in place of the Affymetrix original probe set definition provided by Bioconductor [53]. The corresponding annotation data was generated with AnnBuilder package based on the latest release of public databases, including Entrez Gene, UniGene, PubMed of NCBI, Gene Ontology (GO) and KEGG.

For downstream analysis, all genes are included without discriminative filtering process based on magnitude of changes. The expression level for each gene is standard normalized before use.

2.5 Appendices

2.5.1 Mutual information definition, extension and calculation

Here we describe entropy and mutual information definition for discrete variables. The corresponding definition for continuous variables remained the same [23], except that the summation becomes integration in the following formulas.

In information theory, for a discrete variable, X, Shannon entropy H(X) is defined to be [54]:

$$H(X) = -\sum_{i=1}^{Mx} P(x_i) \log_2 P(x_i) \tag{1.2}$$

Where $X=x_i$ (i=1,2, , M_x), corresponding to M_x different states of variable X, notice that M_x may be different from total number of data points. Shannon entropy is a measurement for the randomness of variable distribution, i.e. how unpredictable the value or state of a variable is. The higher the Shannon entropy is, the harder to predict the value or state of this variable. Similarly, the entropy of joint distribution of two discrete variables X and Y is defined to be [54]:

$$H(X,Y) = -\sum_{i=1}^{Mx}\sum_{j=1}^{My} P(x_i, y_j) \log_2 P(x_i, y_j) \tag{1.3}$$

Where $Y=y_j$ (j=1,2, , M_y), corresponding to M_y different states of variable Y.

Mutual information between two variable X and Y, I(X;Y), is defined based on Shannon entropy, it equals the difference between the sum of entropy of X and Y individually vs the entropy of them jointly [54, 55]:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{1.4}$$

Mutual information measures the difference in predictability when considering two variables together versus considering them independently. Said another way, mutual information is a measurement of dependency between variables. High dependency or mutual information usually occurs when there is causal relationship between variables, or

common causal factors exit. Therefore, mutual information can be used to identify best predictors, or even causal factors and target/dependent factors of variables.

One specific problem addressed in this work is the mutual information among multiple variables. We extended entropy and mutual information definitions in Formula 1.2-1.4 correspondingly. For 3 variables X, Y, Z, we can define three types of three-way mutual information: total correlation C(X;Y;Z) [56], generalized two-way I(X;Y,Z), and three-way interaction information I(X;Y;Z) [21, 22]:

$$C(X;Y;Z) = H(X) + H(Y) + H(Z) - H(X,Y,Z) \tag{1.5}$$

$$I(X;Y,Z) = H(X) + H(Y,Z) - H(X,Y,Z) \tag{1.6}$$

$$\begin{aligned} I(X;Y;Z) = {} & H(X,Y) + H(Y,Z) + H(X,Z) \\ & - H(X) - H(Y) - H(Z) - H(X,Y,Z) \end{aligned} \tag{1.7}$$

These are all generalized mutual information of order 3, different in lower order terms:

$$I(X;Y,Z) = C(X;Y;Z) - I(Y;Z) \tag{1.8}$$

$$I(X;Y;Z) = I(X;Y,Z) - I(X;Y) - I(X;Z) \tag{1.9}$$

Table 2.1 show common examples, where the relationships are high order and can only be fully captured by high order mutual information.

Conditional entropy and mutual information can also be defined based on conditional probability. A rearranged version of conditional mutual information can derived by starting with the definition of conditional probability given Z:

$$I(X;Y \mid Z) = \frac{1}{N}\sum_{k=1}^{N} \log_2 \frac{P(x_k, y_k \mid z_k)}{P(x_k \mid z_k)P(y_k \mid z_k)} \tag{1.10}$$

Next, apply Bayes' rule and rearrange to yield:

$$I(X;Y \mid Z) = \frac{1}{N}\sum_{k=1}^{N} \log_2 \left[\frac{P(x_k, y_k, z_k)}{P(x_k)P(y_k, z_k)} \frac{P(y_k, z_k)}{P(y_k)P(z_k)} \right] \tag{1.11}$$

Re-write into mutual information:

$$I(X;Y \mid Z) = I(X;Y,Z) - I(X;Z) \tag{1.12}$$

Apparently, this conditional mutual information is of order 3 and is closely related to all other types of three-way mutual information. So far, we have been focusing on three-way mutual information and entropy. Similarly, the conception of entropy and mutual information can be directly extended to arbitrary higher order to capture even complicated relationships among multiple variables or multiple sets of variables.

2.5.2 Comparison between MI and log-based local conditional probability

Plug entropy definitions Formula 1.2 and 1.3 into Formula 1.4, we get the expanded Formula for mutual information based on probability:

$$I(X;Y) = \sum_{i=1}^{Mx}\sum_{j=1}^{My} P(x_i, y_j)\log_2\frac{P(x_i, y_j)}{P(x_i)P(y_j)} = \frac{1}{N}\sum_{k=1}^{N}\log_2\frac{P(x_k, y_k)}{P(x_k)P(y_k)} \tag{1.13}$$

Where $X=x_k$ (j=1,2, , N) $Y=y_k$ (j=1,2, , N), corresponding to N data points of variable X or Y.

The counterpart to mutual information in Bayesian network (BN) is log-based local conditional probability, or log likelihood (LL) can be expanded as:

$$LL(X \mid Y) = \log\prod_{k=1}^{N} P(x_k \mid y_k) = \sum_{k=1}^{N}\log\frac{P(x_k, y_k)}{P(y_k)} \tag{1.14}$$

It can be seen that mutual information is close to log likelihood. However mutual information is more standardized, with a weighted-averaging term 1/N and normalizing term $P(x_k)$, which minimize the effects of sample size and specific distribution of individual variables.

2.6 Footnotes

1. In MI3, model learning was focused locally, i.e. we scored and compared all possible local regulatory models for specific target T. This target centered model learning applied to both synthetic data and experimental data, even though biologically we are interested in constructing models centered at particular R1=MYC in the latter case. It would be less appropriate to compare models across different T's because they are not mutually exclusive. Similarly, in Bayesian network, logP(T|R1,R2) is only comparable for fixed T, where all other terms including P(R1)P(R2) in the full product form of joint probability [11, 12] cancelled out. Therefore, we only searched for best R1-R2 pairs given T, but not best R2-T pairs given R1 when learning probabilistic models based on MI3 score or log conditional probability or any other established score. This local approach makes it affordable for MI3 to conduct exhaustive search, which leads to globally optimized models. Heuristic search can be taken when computing time is limited.

2. When MI3 is applied to an experimental gene expression dataset, two key differences between experimental data and synthetic data need to be considered. First, in our gene expression data there are 8359 genes, which is significantly larger system than the 9-variable synthetic network. For an exhaustive search for the best two-parent set for each gene, this problem size would require searching $\sim 10^{11}$ (83593) combinations—a scale that is currently out of reach computationally. In this work, we focus on the construction transcription regulatory networks centered to MYC. Therefore, we can fix one regulator, R1, to MYC, and only search across cofactors (R2s) and targets (T). This reduced problem requires the search of $\sim 10^{7}$ (83592) combinations for our gene expression data. This scale of problem is computationally tractable. For both scenarios, we constrain MYC targets (T) with I(T; MYC)≥0.3, i.e. targets that have enough marginal dependency on MYC to ensure that MYC does likely regulate the target based on the microarray dataset. Second, there are frequently multiple equally interesting and closely scoring regulatory models learned from experimental data for each target. For example, several regulators are equally important, or multiple genes in a pathway/complex represent the same regulatory action equally well. Correspondingly, we kept the top 5 highest scoring 2-parent models for each target gene, rather than the top 1 as in the synthetic data. Keeping top 1 model only led to almost the same list of top 10 MYC cofactors (Table 2.4-5), except that the number of targets mapped to individual cofactors was too small for quantitative evaluation.

2.7 References

1. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, et al: **Expression monitoring by hybridization to high-density oligonucleotide arrays**. *Nat Biotechnol* 1996, **14**:1675-80.

2. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**:467-70.

3. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**:14863-8.

4. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization**. *Mol Biol Cell* 1998, **9**:3273-97.

5. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS: **Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks**. *Proc Natl Acad Sci U S A* 2000, **97**:12182-6.

6. Moriyama M, Hoshida Y, Otsuka M, Nishimura S, Kato N, Goto T, Taniguchi H, Shiratori Y, Seki N, Omata M: **Relevance network between chemosensitivity and transcriptome in human hepatoma cells**. *Mol Cancer Ther* 2003, **2**:199-205.

7. Schafer J, Strimmer K: **An empirical Bayes approach to inferring large-scale gene association networks**. *Bioinformatics* 2005, **21**:754-64.

8. Alon U: **An introduction to systems biology : design principles of biological circuits**. Boca Raton, FL: Chapman & Hall/CRC; 2007.

9. Shmulevich I, Dougherty ER, Kim S, Zhang W: **Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks**. *Bioinformatics* 2002, **18**:261-74.

10. Shmulevich I, Zhang W: **Binary analysis and optimization-based normalization of gene expression data**. *Bioinformatics* 2002, **18**:555-65.

11. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA: **Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks**. *Pac Symp Biocomput* 2001:422-33.

12. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data**. *J Comput Biol* 2000, **7**:601-20.

13. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP: **Causal**

protein-signaling networks derived from multiparameter single-cell data. *Science* 2005, **308**:523-529.

14. Friedman N: **Inferring cellular networks using probabilistic graphical models**. *Science* 2004, **303**:799-805.

15. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells**. *Nat Genet* 2005, **37**:382-90.

16. Butte AJ, Kohane IS: **Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements**. *Pac Symp Biocomput* 2000:418-29.

17. Pearl J: **Causality : models, reasoning, and inference**. Cambridge, U.K. ; New York: Cambridge University Press; 2000.

18. Woolf PJ, Wang Y: **A fuzzy logic approach to analyzing gene expression data**. *Physiol Genomics* 2000, **3**:9-15.

19. Dupont WD: **Making causal inferences from observational data**. *Biometrics* 1978, **34**:713-4.

20. Winship C, Morgan SL: **The Estimation of Causal Effects from Observational Data**. *Annual Review of Sociology* 1999, **25**:659-706.

21. Mcgill WJ: **Multivariate Information Transmission**. *Psychometrika* 1954, **19**:97-116.

22. Jakulin A, Bratko I: **Quantifying and Visualizing Attribute Interactions: An Approach Based on Entropy**. *arXiv:cs.AI/0308002* 2004.

23. Steuer R, Kurths J, Daub CO, Weise J, Selbig J: **The mutual information: detecting and evaluating dependencies between variables**. *Bioinformatics* 2002, **18 Suppl 2**:S231-40.

24. Nemenman I: **Information theory, multivariate dependence, and genetic network inference**. *arXiv:q-bio/0406015* 2004.

25. Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ, Ren B: **A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells**. *Proc Natl Acad Sci U S A* 2003, **100**:8164-9.

26. Adhikary S, Eilers M: **Transcriptional regulation and transformation by Myc proteins**. *Nat Rev Mol Cell Biol* 2005, **6**:635-45.

27. Knoepfler PS, Zhang XY, Cheng PF, Gafken PR, McMahon SB, Eisenman RN: **Myc influences global chromatin structure**. *Embo Journal* 2006, **25**:2723-2734.

28. Eisenman RN: **Deconstructing myc**. *Genes Dev* 2001, **15**:2023-30.

29. Cowling VH, Cole MD: **Mechanism of transcriptional activation by the Myc oncoproteins**. *Semin Cancer Biol* 2006, **16**:242-52.

30. Zeller KI, Jegga AG, Aronow BJ, O'Donnell KA, Dang CV: **An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets**. *Genome Biol* 2003, **4**:R69.

31. Friedman N, Nachman I, Pe'er D: **Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm**. In: *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*; 1999; San Francisco, CA. 206-215.

32. **The MYC Target Gene Database** [http://www.myccancergene.org/site/mycTargetDB.asp]

33. Pal S, Yun R, Datta A, Lacomis L, Erdjument-Bromage H, Kumar J, Tempst P, Sif S: **mSin3A/histone deacetylase 2- and PRMT5-containing Brg1 complex is involved in transcriptional repression of the Myc target gene cad**. *Mol Cell Biol* 2003, **23**:7475-87.

34. Ogawa H, Ishiguro K, Gaubatz S, Livingston DM, Nakatani Y: **A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells**. *Science* 2002, **296**:1132-6.

35. Schuster C, Krol A, Carbon P: **Two distinct domains in Staf to selectively activate small nuclear RNA-type and mRNA promoters**. *Mol Cell Biol* 1998, **18**:2650-8.

36. Mitra P, Xie RL, Medina R, Hovhannisyan H, Zaidi SK, Wei Y, Harper JW, Stein JL, van Wijnen AJ, Stein GS: **Identification of HiNF-P, a key activator of cell cycle-controlled histone H4 genes at the onset of S phase**. *Mol Cell Biol* 2003, **23**:8110-23.

37. Fan G, Ballou LM, Lin RZ: **Phospholipase C-independent activation of glycogen synthase kinase-3beta and C-terminal Src kinase by Galphaq**. *J Biol Chem* 2003, **278**:52432-6.

38. Dominguez-Caceres MA, Garcia-Martinez JM, Calcabrini A, Gonzalez L, Porque PG, Leon J, Martin-Perez J: **Prolactin induces c-Myc expression and cell survival through activation of Src/Akt pathway in lymphoid cells**. *Oncogene* 2004, **23**:7378-90.

39. Rodriguez-Viciana P, Oses-Prieto J, Burlingame A, Fried M, McCormick F: **A phosphatase holoenzyme comprised of Shoc2/Sur8 and the catalytic subunit of PP1 functions as an M-Ras effector to modulate Raf activity**. *Mol Cell* 2006, **22**:217-30.

40. Li W, Han M, Guan KL: **The leucine-rich repeat protein SUR-8 enhances**

MAP kinase activation and forms a complex with Ras and Raf. *Genes Dev* 2000, **14**:895-900.

41. Liu X, Tesfai J, Evrard YA, Dent SY, Martinez E: **c-Myc transformation domain recruits the human STAGA complex and requires TRRAP and GCN5 acetylase activity for transcription activation**. *J Biol Chem* 2003, **278**:20405-12.

42. Silverman BW: **Density estimation for statistics and data analysis**. London ; New York: Chapman and Hall; 1986.

43. Scott DW: **Multivariate density estimation : theory, practice, and visualization**. New York: Wiley; 1992.

44. Scott DW, Wand MP: **Feasibility of Multivariate Density Estimates**. *Biometrika* 1991, **78**:197-205.

45. Grimes DA, Schulz KF: **Bias and causal associations in observational research**. *Lancet* 2002, **359**:248-52.

46. Pearl J: **Probabilistic reasoning in intelligent systems: networks of plausible inference**: Morgan Kaufmann Publishers Inc.; 1988.

47. **The MI3 Algorithm R packages** [http://sysbio.engin.uimich.edu/~luow/downloads]

48. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**:249-64.

49. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**:R80.

50. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level**. *Bioinformatics* 2004, **20**:307-15.

51. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, et al: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucleic Acids Res* 2005, **33**:e175.

52. **The Microarray Lab at the University of Michigan** [http://brainarray.mhri.med.umich.edu]

53. **The BioConductor Project** [http://bioconductor.org/]

54. Shannon CE: **A Mathematical Theory of Communication**. *Bell System Technical Journal* 1948, **27**:379-423.

55. Kolmogor.An: **Logical Basis for Information Theory and Probability Theory**.

Ieee Transactions on Information Theory 1968, **IT14**:662-&.

56. Watanabe S: **Information Theoretical Analysis of Multivariate Correlation**. *Ibm Journal of Research and Development* 1960, **4**:66-82.

Chapter III

GAGE: Generally Applicable Gene Set Enrichment for Pathway Inference

3.1 Introduction

A central goal of biomedical research is to define mechanistic causes for cellular behavior and disease. High throughput technologies such as gene expression profiling provide a rich starting point to identify mechanistic causes. Ideally we would like to contextualize gene expression patterns with the known biochemical processes and regulatory signaling pathways. This way we gain a more systems level and informative view of the biological states that have been perturbed, which in turn allows us to identify points where we could intervene to change cellular behavior.

Gene set analysis (GSA) , also called pathway inference, is a widely used strategy for gene expression data analysis based on pathway knowledge [1-10]. Unlike previous strategies which focus on individual or a limited number of genes, GSA focuses on sets of related genes and has demonstrated three major advantages. First, GSA methods are better able to detect biologically relevant signals and give more coherent results across different studies [2, 4]. Second, GSA uses all of the available gene expression data instead of prefiltering the data for a short list of strongly differentially expressed genes. Indeed, small coordinated gene expression changes in a pathway can have a major biological effect even if these changes are not statistically significant for any individual gene [2]. Third, GSA incorporates prior knowledge of biological pathways and other experimental results in the form of gene sets [2, 3]. These gene sets are constantly

updated in the literature and represent a significant repository of useful biological knowledge.

There are two categories of GSA based on the statistical tests used: sample randomization and gene randomization [1, 7]. Sample randomization methods test significance of gene sets based on permutation of sample labels, with GSEA [2, 3], SAFE [9] and SAM-GS [8] as representatives. In contrast, gene randomization methods test the significance of gene sets based on permutations of gene labels or a parametric distribution over genes, with PAGE [4], T-Profiler [6] and Random-set [5] as representatives. Sample randomization keeps the correlation structure among genes but only applies to large expression data sets with multiple samples per experimental condition. For a two-state comparison, a minimum of 8 chips for each state is required for 1000 balanced (presence of the two sample states) permutation or 6 chips for 1000 unbalanced permutation. Gene randomization has no limitation on sample size, but may break the correlation structure among genes [10]—an issue that may not be a problem (detailed in discussion) [4, 5]. Sample randomization and gene randomization test different but related null hypotheses , hence combinatory procedures [1, 7] were proposed to achieve more robust results.

In spite of its advantages, GSA as a whole strategy still suffers from three major limitations.

First, no GSA method currently available is appropriate for small data sets, yet most gene expression data sets fall into this category. As mentioned above, the sample randomization strategy used by methods such as GSEA is not appropriate for studies with under 8 gene chips per state, thus gene randomization remains to be the only feasible option [1, 2]. Gene randomization methods such as PAGE have been applied to small data set [4], but these methods tend to make large number of (false) positive calls with extremely small P-values [11, 12] (also see the results). T-profiler targets data sets with one sample pair [6], however, it can’t combine results from multiple paired experiments nor can it be applied to studies with non-paired studies [6].

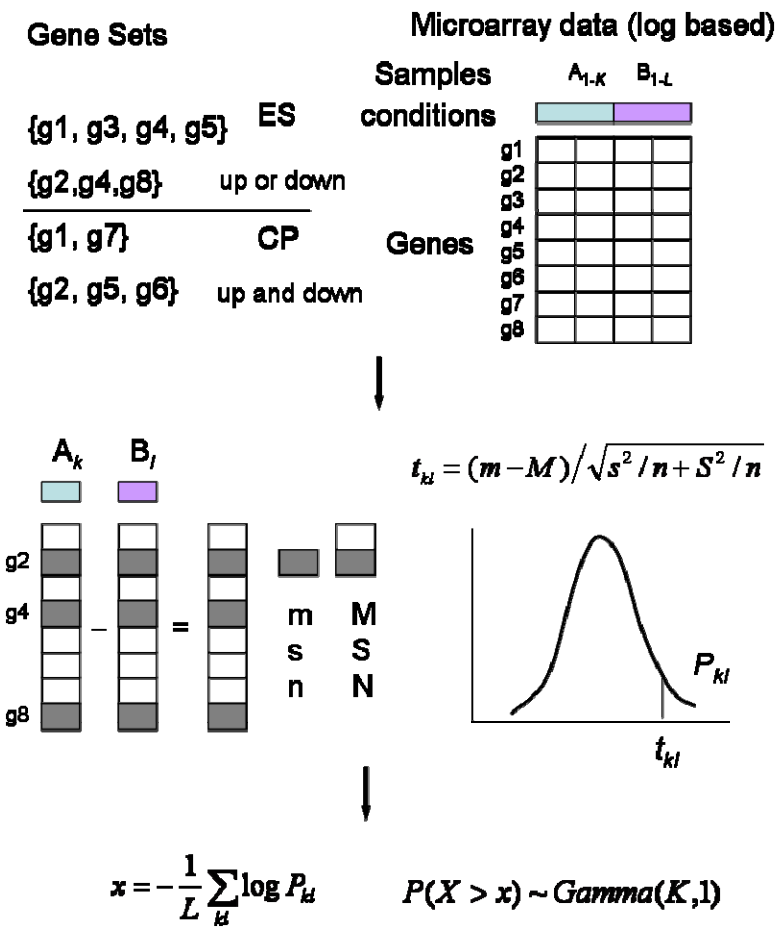


Figure 3.1 A schematic overview of the GAGE algorithm. GAGE has three major steps. Step 1: input preparation. Separate gene sets into two categories: experimental sets (ES) and canonical pathways (CP), for differential treatment in significant test. Step 2: one-on-one comparison between samples from the two experimental conditions. For each experiment-control pair, calculate differential expression in log based fold change for all genes. Test whether specific gene sets are significantly differentially expressed relative to the background whole set using two-sample t-test. Step 3: summarization. For each gene set, derive a global P value based on a meta test on the P-values from all one-on-one comparisons. More details of GAGE are given in the Methods. Variables m, s and n are the mean fold change, standard deviation and number of genes in a gene set, M, S and N are those for the whole set.

Second, no GSA method currently available handles data sets with different sample sizes and experiment designs consistently. For data sets with few or no replicates, t-test statistics, signal noise ratios, or their corresponding P-values are not robust estimates of differential expression for genes or simply not applicable. Therefore, fold change (log

based) is frequently used as more versatile per gene statistics [2, 4-6, 13]. This gives rise to two issues that have been largely neglected so far. First, average fold change does not account for different experimental designs, i.e. pair-matched samples or non-paired samples. The per gene statistics such as t-test statistics may vary significantly depending on if the samples are paired or not, yet there is no difference in fold change. Second, average fold change does not contain any information for the sample size. Sample size largely determines the confidence or significance level of our inference, yet is dropped when using fold change. Fold change makes sense in one-on-one paired comparison, as in T-profiler [6]. However for data sets with replicate samples, the test power or the significance of relevant gene sets would be underestimated.

Third, most GSA methods only consider transcriptional regulation towards one direction in a gene set. This directional bias makes sense for experimentally derived gene sets, but not for gene sets based on canonical signaling pathways, which frequently show reciprocal gene regulation in both directions upon perturbation [14, 15]. Thus it is advisable to consider both cases for an inclusive analysis for regulatory mechanisms.

To address these issues, we have developed a novel method called Generally Applicable Gene-set Enrichment (GAGE, Figure 3.1). GAGE applies to data sets with any number of samples and is based on a parametric gene randomization procedure. Similar to Parametric Analysis of Gene Set Enrichment (PAGE) and T-profiler, GAGE uses log-based fold changes as per gene statistics. However, GAGE differs from PAGE and T-profiler in three significant ways (Figure 3.1). First, GAGE assumes a gene set comes from a different distribution than the background and uses two-sample t-test to account for the gene set specific variance as well as the background variance. In contrast, PAGE assumes gene sets comes from the same distribution as the background and uses one-sample z-test that only considers the background variance. T-profiler also employs two-sample t-test, but it is essentially a one-sample z-test since the sample size of a gene set is not comparable to its complementary set [6] (Footnote 1 and Methods). Second,

GAGE adjusts for different microarray experimental designs (paired or non-paired) and sample sizes by decomposing group-on-group comparisons into one-on-one comparisons between samples from different groups. GAGE derives a global P-value using a meta test on the P-values from these comparisons for each gene set. Third, GAGE separates experimentally perturbed gene sets (from literature) and canonical pathways (from pathway databases). Experimental sets are taken as genes coregulated towards single direction, whereas canonical pathways allowed changes towards both directions. This gene set separation strategy give GAGE more test power in detecting relevant biological signals.

In this work, we show that GAGE is generally applicable to data sets with different sample sizes and experimental designs (Footnote 2). We first apply GAGE to two lung cancer data sets [16, 17] and one type 2 diabetes data set [3], which as been analyzed by GSEA [2, 3] and PAGE [4] as example cases. These are representatives for large data sets with tens of samples per condition frequently seen in large clinical or experimental studies. We then analyze a smaller dataset describing mesenchymal stem cell response to BMP6 treatment. This is a typical small data set with as few as two samples per condition like in most experimental studies. BMP6 treated samples and controls are one-on-one matched, which is a frequently used experiment design particularly for all the two-channel microarray studies. In each case, we compare GAGE to GSEA and PAGE. Finally, we also detail the major strategies employed by GAGE.

3.2 Results

3.2.1 Application to large data sets with the GSEA and PAGE as control methods

As a test case, we applied GAGE, PAGE and GSEA to two lung cancer data sets [16, 17] which were originally analyzed and compared by GSEA [2]. These two data sets were generated by two independent studies done in Boston [17] and Michigan [16], containing gene expression profiles of lung adenocarcinomas samples from patients. Patients were classified as having “good” or “poor” clinical outcomes. For each data set, we defined the

control set as patient profiles with good clinic outcomes, and selected the most differentially regulated gene sets associated with poor outcomes. Note that we used the updated curated gene set collection c2 from MSigDB [2, 18]. For fair comparison, experimental sets and the canonical pathways were separated for all methods.

Table 3.1 GAGE applied to the two lung cancer datasets of large sample sizes. Top 10 most significantly enriched experimental sets and canonical pathways in poor clinical outcomes vs good outcomes were inferred by GAGE from two published lung adenocarcinoma data sets used in the GSEA paper [2]. Both positively and negatively regulated gene sets were collected and ranking by the P-value, and by absolute value of average t-statistics (data not shown) for ties. Consistencies between the two data sets are shown in bold font. Notes show the connections of the gene sets to cancer related topics: t for tumor related; bt for tumor metastasis and bad outcome; c for cell growth and proliferation related; blank represents no evident connection. These annotations came from the original studies for experimental sets, and from relevant literature for the canonical pathway.

Boston study			Michigan study		
Experimental Sets	P-value	Notes	Experimental Sets	P-value	Notes
Tarte_Plasma_Blastic	<1.0E-16	c	**Tarte_Plasma_Blastic**	<1.0E-16	c
Uvb_Nhek3_All	<1.0E-16	t	**Cancer_Undifferentiat***	<1.0E-16	bt
Peng_Glutamine_Dn	<1.0E-16	c	**Brca_Er_Neg**	<1.0E-16	bt
Lei_Myb_Regulated_Ge*	<1.0E-16	bt,c	Serum_Fibroblast_Cell*	<1.0E-16	bt,c
Peng_Leucine_Dn	<1.0E-16	c	**Uvb_Nhek3_All**	<1.0E-16	t
Cancer_Undifferentiat*	<1.0E-16	bt	Zhan_Mm_Cd138_Pr_*	1.8E-15	bt
Brca_Er_Neg	<1.0E-16	bt	Li_Fetal_Vs_Wt_Kidne*	3.6E-14	t
Peng_Rapamycin_Dn	<1.0E-16	c	Dox_Resist_Gastric_Up	9.5E-14	bt
Cancer_Neoplastic_Me*	<1.0E-16	t	Idx_Tsa_Up_Cluster3	2.3E-13	c
Rcc_Nl_Up	<1.0E-16	t	Tarte_Mature_Pc	6.0E-13	c
Canonical Pathways	P-value	Notes	Canonical Pathways	P-value	Notes
Gpcrs_Class_A_Rhod*	<1.0E-16	bt	**Gpcrs_Class_A_Rhod***	3.0E-10	bt
Gpcrdb_Class_A_Rho*	<1.0E-16	bt	**Gpcrdb_Class_A_Rho***	1.1E-09	bt
Blood_Clotting_Casca*	3.4E-15	bt	Androgen_Genes	5.0E-08	bt
Intrinsicpathway	5.2E-15	bt	Cytokinepathway	1.8E-07	bt
Fibrinolysispathway	5.9E-13	bt	Prostaglandin_And_Leu*	2.8E-05	bt
Peptide_Gpcrs	1.9E-12	bt	Proliferation_Genes	5.1E-05	c
Tyrosine_Metabolism	7.9E-09	bt	**Peptide_Gpcrs**	5.8E-05	bt
Extrinsicpathway	4.2E-07	bt	**Intrinsicpathway**	8.4E-05	bt
Gpcrdb_Other	5.1E-06	bt	Androgen_And_Estroge*	4.1E-04	bt
Small_Ligand_Gpcrs	6.1E-06	bt	**Blood_Clotting_Casca***	7.0E-04	bt

We compared the top 10 most significant gene sets inferred by the three methods (Table 3.1-3) and identified evident differences in four aspects. First, the top experimental gene sets selected by GAGE and PAGE overlapped significantly, but the canonical pathways identified by GAGE, PAGE, and GSEA did not (Table 3.3). The lack of overlap for the canonical pathways is expected because GAGE allows perturbations in both directions in canonical pathways. Second, GAGE derived more modest P-values and numbers of significant gene sets compared to GSEA and PAGE (Table 3.2). While others have suggested that GSEA suffers from low sensitivity [4, 7, 8], our results suggest that PAGE is overly sensitive (low specificity). Third, the top 10 gene sets inferred by GAGE are more consistent between the two studies: 4 experimental sets and 5 canonical pathways are the same for GAGE results, 4 and 4 for PAGE and 2 and 0 for GSEA respectively (Table 3.2). Fourth, the top 10 gene sets inferred by GAGE better describe poor outcomes of lung cancer mechanistically (Table 3.2). Experimental sets inferred by GAGE and by PAGE are similarly indicative of tumor occurrence and prognostic of metastasis or poor clinical outcomes, and both are better than those inferred by GSEA. Canonical pathways inferred by GAGE are by far the most indicative of tumor occurrence and metastasis.

Table 3.2 Comparison between GAGE, PAGE and GSEA results from the two lung cancer datasets. The top 10 most significantly enriched experimental sets and canonical pathways in poor clinical outcomes vs good outcomes were inferred by GAGE, PAGE, and GSEA from two published lung adenocarcinoma data sets used in the GSEA paper [2].Data columns are overlap between top 10 gene sets for the two studies, top 10 P-values, number of top 10 gene sets related to metastasis (bt) and tumor (t and bt), and numbers of significant gene sets with P-values ≤ 0.001.

Gene Sets & Methods		Overlap	Top 10 P-values	Metastasis	Tumor	Sign. Sets
Experimental Sets	GAGE	4	<1.0E-16, 5.9E-13	3, 5	6, 7	245, 122
	PAGE	4	2.1E-170, 4.1E-85	6, 4	8, 6	647, 563
	GSEA	2	1.6E-2, 6.4E-3	1, 2	6, 3	2, 5
Canonical Pathways	GAGE	5	6.1E-6, 7.0E-4	10, 9	10, 9	23, 10
	PAGE	4	8.5E-26, 7.5E-27	2, 3	4, 3	160, 146
	GSEA	0	7.5E-2, 1.4E-2	2, 1	6, 4	2, 4

Table 3.3 Overlaps between GAGE, PAGE and GSEA results from the two lung cancer datasets. The top 10 most significantly enriched experimental sets and canonical pathways in poor clinic outcomes vs good outcomes were inferred by GAGE, PAGE and GSEA from two published lung adenocarcinoma data sets used in the GSEA paper [2].

Boston

Gene Sets & Methods		GAGE	PAGE	GSEA
Experiment Sets	GAGE	NA	5	0
	PAGE	5	NA	0
	GSEA	0	0	NA
Canonical Pathways	GAGE	NA	1	0
	PAGE	1	NA	2
	GSEA	0	2	NA

Michigan

Gene Sets & Methods		GAGE	PAGE	GSEA
Experiment Sets	GAGE	NA	5	0
	PAGE	5	NA	0
	GSEA	0	0	NA
Canonical Pathways	GAGE	NA	0	0
	PAGE	0	NA	3
	GSEA	0	3	NA

Several major mechanistic themes predictive of poor clinical outcomes emerged from the list of top gene sets inferred by GAGE. These themes included G-protein coupled receptors (GPCRS) associated signals (sets 1, 2, 6, 9, 10 of Boston and sets 1, 2, 7 of Michigan in Table 3.1), thrombosis or blood coagulation activation (sets 3, 4, 5, 8 of Boston and set 8, 10 of Michigan in Table 3.1), and hormone and cytokine (sets ranking >10 of Boston not shown, and set 3, 4, 9 of Michigan in Table 3.1). Indeed, G-protein-coupled receptors, the largest family of cell-surface molecules involved in signal transmission, have recently emerged as crucial players in the growth and metastasis of multiple human cancers [19, 20]. Thrombosis or blood coagulation activation has been implicated in the disease and is an predictor for poor survival rates for lung cancer patients [21, 22]. Androgen level and cytokine profiles influence clinic outcomes of non-small cell lung cancer [23, 24]. All these factors are likely the major

causal or contributing mechanisms for non-small cell lung cancer progress and metastasis.

We also applied GAGE, PAGE and GSEA to another large dataset describing type 2 diabetes progression that was analyzed by GSEA [3] and PAGE [4] previously (Table 3.4). This comparison performed similarly to the cancer study mentioned above. In particular, GAGE pinpointed multiple experimental sets and canonical pathways which are directly involved in type 2 diabetes or closely related metabolism processes.

Table 3.4 Comparison between GAGE, PAGE and GSEA results from the type 2 diabetes dataset. The most significantly enriched experimental sets and canonical pathways in type 2 diabetes patients vs healthy controls were inferred by GAGE, PAGE and GSEA from published data set generated by Mootha et al [3]. Data columns are top 10 P values, number of top 10 gene sets related to type 2 diabetes and metabolism, and numbers of significant gene sets with cutoff P value = 0.001.

Gene Sets & Methods		Top 10 P	Diabetes	Metab.	Sign. Calls
Experiment Sets	GAGE	<1.0E-16	3	3	169
	PAGE	<1.0E-307	0	0	911
	GSEA	3.6E-2	1	0	1
Canonical Pathways	GAGE	<1.0E-16	5	2	39
	PAGE	<1.0E-307	3	4	325
	GSEA	6.1E-2	0	3	0

3.2.2 Application to small data sets with PAGE and GSEA-g (GSEA with gene permutation option) as control methods

We applied GAGE and PAGE to a microarray data set generated by our group to select the most differentially expressed gene sets in human mesenchymal stem cells (MSC) upon BMP6 treatment (Table 3.5-7). The data set contains a total of 4 gene chip measurements from duplicate experiments each with paired measurements of human MSC with or without 8 hours BMP6 treatment. Notice that GSEA by default is not applicable to this data set because the sample size is too small for permutation based inference. However, GSEA with gene labels permutation option (GSEA-g) works. Since GSEA-g does not implement the sample randomization strategy recommended by the

authors [2], we mainly compared GAGE to PAGE here (Table 3.7-8). GAGE conducts one-on-one comparisons, hence was applied to each of the two BMP6 experiments individually (Table 3.5). For exact comparison, PAGE was slightly modified to enable one-on-one comparison (Table 3.6). The GSEA software took multiple samples per condition hence not applicable to the experiments individually.

Table 3.5 GAGE applied to the BMP6-MSC dataset of small sample size. Top 10 most significantly differentially expressed experimental sets and canonical pathways were inferred by GAGE from human MSCs following an 8 hour BMP6 treatment. Two replicate experiments were done, each with BMP6 treated sample and control. Therefore GAGE was applied to each experiment and derived corresponding P-values (P.exp1-2). Gene sets were ranked based on global P-values from both experiments.

Experimental Sets	t-statistic	P-value	P.exp1	P.exp2
Ifna_Hcmv_6hrs_Up	-3.80	2.8E-07	1.9E-04	8.0E-05
Der_Ifnb_Up	-3.47	1.7E-06	1.7E-03	5.5E-05
Baf57_Bt549_Dn	-3.09	1.4E-05	3.6E-03	2.6E-04
Ifn_Beta_Up	-2.92	5.4E-05	6.1E-03	6.7E-04
Sana_Ifng_Endothelial_Up	-2.88	6.7E-05	6.1E-03	8.2E-04
Ifn_Any_Up	-2.76	1.2E-04	1.2E-02	7.0E-04
Dac_Bladder_Up	-2.65	2.9E-04	1.2E-03	2.0E-02
Grandvaux_Ifn_Not_Irf3_Up	-2.76	2.9E-04	1.9E-02	1.3E-03
Ifna_Uv-Cmv_Common_Hcmv_6hrs_Up	-2.55	5.0E-04	8.2E-03	5.6E-03
Bennett_Sle_Up	-2.48	7.2E-04	6.8E-03	1.0E-02
Canonical Pathways	t-statistic	P-value	P.exp1	P.exp2
Tgf_Beta_Signaling_Pathway	3.15	1.8E-05	1.1E-03	1.1E-03
Wnt_Signaling	2.47	5.6E-04	3.0E-03	1.7E-02
Alkpathway	2.46	7.3E-04	8.8E-03	7.8E-03
Proliferation_Genes	2.27	1.3E-03	6.8E-03	1.9E-02
Cell_Proliferation	2.24	1.5E-03	2.1E-02	7.5E-03
Hematopoesis_Related_Transcription_Factors	2.05	4.0E-03	1.8E-02	2.5E-02
Erythpathway	1.98	7.0E-03	2.5E-02	3.4E-02
Smooth_Muscle_Contraction	1.79	1.0E-02	2.6E-02	5.2E-02
Apoptosis	1.73	1.4E-02	7.1E-02	2.5E-02
Breast_Cancer_Estrogen_Signaling	1.61	2.1E-02	8.1E-02	3.5E-02

Using a P-value cutoff of <0.01, GAGE identified fewer gene sets than PAGE (Table

3.5-7). GAGE identified 39 significant experimental sets and 7 canonical pathways (Table 3.9-10). There were only 17 experimental sets and 4 canonical pathways significant (Table 3.9-10) after removing the redundancy among gene sets, which is reasonable number of pathways triggered by a single perturbation in a single cell line. In contrast, PAGE gave 745 significant experimental sets and 187 significant canonical pathways. Most significant genes sets selected by PAGE were not significant according to GAGE using the same cutoff P-value (full result tables not shown). After removing the redundancy in these sets, there were more than 200 and 40 non-redundant experimental sets and canonical pathways respectively (not shown, Footnote 3). Presumably, PAGE made a large number of false positive calls. Similar differences between GAGE and PAGE were observed for the two lung cancer data sets and the type 2 diabetes data set (Table 3.2 and 3.4). This difference came from the different statistical tests used by GAGE and PAGE, i.e. two-sample t-test vs one-sample z-test (detailed in the subsection of 'Dissection of major strategies employed by GAGE'). GSEA-g gave P-values and a predicted number of significant gene sets comparable to GAGE when nominal P-values were used (Table 3.7, source data table not shown).

Biologically, GAGE gene sets were mechanistically more relevant for BMP6 effects compared to those sets selected by PAGE. 9 out of 10 experimental sets inferred by GAGE (Table 3.5) are directly related to interferon or STAT pathway [25], which is a target of BMP signaling [26, 27].The experimental sets selected by PAGE alone have less connection to BMP (Table 3.6). GAGE and PAGE differed in 8 entries of the top 10 canonical pathways. Of GAGE predictions (Table 3.5), Wnt signaling [28, 29], proliferation [30, 31] are all known pathways or process regulated by BMP treatment in MSC or osteoblastic cell lineages. BMPs regulate hematopoiesis and erythrocyte differentiation [32, 33]. Breast cancer estrogen signaling interacts with BMP signal [34, 35]. None of these pathways were significant according to PAGE (Table 3.6, full result table not shown). The GSEA-g top experimental sets overlapped with GAGE, but the

canonical pathways were more similar to PAGE (Table 3.8).

Table 3.6 PAGE applied to the BMP6-MSC dataset of small sample size. Top 10 most significantly differentially expressed experimental sets and canonical pathways in human MSC following 8 hour BMP6 treatment were inferred by PAGE. PAGE by default apples to whole data set with duplicate samples and gave the global P value. Upon small modification to enable one-on-one comparison, PAGE was applied to each of the two BMP6 experiments individually the same way as GAGE in Table 3.5.

Experiment Sets	t-statistic	P-value	P.exp1	P.exp2
Rett_Dn	37.2	5.2E-291	2.2E-170	2.5E-233
Gh_Hypophysectomy_Rat_Up	30.7	4.4E-202	1.2E-60	1.3E-280
Uvc_High_D2_Dn	29.1	6.4E-182	8.3E-79	1.0E-194
Gh_Igf_Chondrocytes_Up	29.1	2.2E-181	2.5E-77	9.5E-197
Passerini_Growth	-28.3	1.4E-172	1.4E-95	2.8E-146
Ifna_Hcmv_6hrs_Up	-26.6	6.2E-153	1.4E-85	3.6E-128
Lvad_Heartfailure_Dn	-25.8	1.3E-143	6.3E-88	1.4E-108
Uvc_Low_C1_Dn	25.1	4.0E-136	1.0E-88	3.0E-95
Baf57_Bt549_Dn	-25.0	1.0E-135	4.9E-66	4.6E-131
Der_Ifnb_Up	-24.6	2.6E-131	4.6E-64	1.6E-126
Canonical Pathways	t-statistic	P-value	P.exp1	P.exp2
Apoptosis	-14.9	7.2E-50	1.9E-31	6.4E-37
Tgf_Beta_Signaling_Pathway	13.6	8.2E-42	2.5E-26	3.5E-31
Valine_Leucine_And_Isoleucine_Degradation	-13.0	3.0E-38	7.2E-32	4.7E-19
Striated_Muscle_Contraction	12.7	1.4E-36	5.0E-20	3.6E-32
Tob1pathway	-12.5	1.1E-35	4.0E-25	3.9E-23
Gpcrdb_Other	12.4	2.0E-35	9.4E-12	7.5E-48
Badpathway	11.8	3.3E-32	1.54E-19	2.3E-25
Mitochondria	-11.6	5.7E-31	2.0E-08	9.1E-48
Eicosanoid_Synthesis	11.5	1.1E-30	1.5E-12	1.1E-35
Apoptosis_Genmapp	-10.9	1.7E-27	4.4E-12	3.1E-30

Significant gene sets inferred by GAGE were consistent across replicate experiments and within the top 10 lists. The top 10 gene sets are almost the same if we used either one of the two experiments only (Table 3.5). And the difference between the P-values from the two experiments almost never exceeded one order of magnitude. On the other hand, the top 10 gene set lists inferred by the PAGE and corresponding P-values are more different cross the two experiments (Table 3.6, not all top sets for individual experiments included).

There was also high level of internal consistency in the top 10 gene sets inferred by GAGE (Table 3.5). For example, 9 out of 10 experimental sets were directly related to interferon signal. Among the canonical pathways, there were two proliferation and two hematopoietic differentiation related pathways, and Alk pathway overlapped with TGF beta and Wnt signaling pathways. In contrast, the PAGE (Table 3.6) and GSEA-g (not shown) top gene sets had lower internal consistencies. These results indicate that GAGE is a method robust against the heterogeneity in experiments or gene set definition. Notice that redundant gene sets representative of the same effect or pathway were kept here for exact comparison between methods, but they can be differentiated and combined by GAGE program if needed (Table 3.9-10).

Table 3.7 Comparison between GAGE, PAGE and GSEA-g results from the BMP6-MSC dataset. The significantly enriched experimental sets and canonical pathways in human MSC following 8 hour BMP6 treatment were inferred by GAGE, PAGE and GSEA-g (permutation of gene labels). Top 10 t- or z-statistics and P-values and the numbers of significant gene sets were shown (P-value < 0.01). Note that GSEA-g results shown were based on nominal P-values.

Gene Sets & Methods		Top 10 abs(T/Z)	Top 10 P-values	Sign. Sets
Experiment Sets	GAGE	2.48	7.22E-4	39
	PAGE	24.6	2.62E-131	745
	GSEA-g	1.97	<1.0E-3	86
Canonical Pathways	GAGE	1.61	1.96E-2	7
	PAGE	10.9	1.77E-27	187
	GSEA-g	1.55	3.70E-2	6

Table 3.8 Overlaps between GAGE, PAGE and GSEA-g results from the BMP6-MSC dataset. The top 10 most significantly differentially expressed experimental sets experimental sets and canonical pathways in human MSC following 8 hour BMP6 treatment were inferred by GAGE, PAGE and GSEA-g.

Gene Sets & Methods		GAGE	PAGE	GSEA
Experiment Sets	GAGE	NA	3	6
	PAGE	3	NA	3
	GSEA-g	6	3	NA
Canonical	GAGE	NA	2	1

Pathways	PAGE	2	NA	5
	GSEA-g	1	5	NA

Table 3.9 Full and non-redundant list of experimental sets inferred by GAGE. The list of significantly (P<0.01) differentially expressed experimental sets were inferred by GAGE from human MSCs following an 8 hour BMP6 treatment. Two replicate experiments were done, each with BMP6 treated sample and control. Therefore GAGE was applied to each experiment and derived corresponding P-values (P.exp1-2). Gene sets were ranked based on global P-values from both experiments. The 17 gene sets in bold typeface are the non-redundant sub-list, all other 22 gene sets are removed during the redundant test.

Experimental Sets	t-statistic	P-value	P.exp1	P.exp2
Ifna_Hcmv_6hrs_Up	**-3.80**	**2.8E-07**	**1.9E-04**	**8.0E-05**
Der_Ifnb_Up	-3.47	1.6E-06	1.7E-03	5.5E-05
Baf57_Bt549_Dn	**-3.09**	**1.4E-05**	**3.6E-03**	**2.6E-04**
Ifn_Beta_Up	-2.92	5.4E-05	6.1E-03	6.7E-04
Sana_Ifng_Endothelial_Up	-2.88	6.6E-05	6.1E-03	8.2E-04
Ifn_Any_Up	-2.76	1.1E-04	1.2E-02	7.0E-04
Dac_Bladder_Up	-2.65	2.8E-04	1.2E-03	2.0E-02
Grandvaux_Ifn_Not_Irf3_Up	**-2.76**	**2.8E-04**	**1.9E-02**	**1.3E-03**
Ifna_Uv-Cmv_Common_Hcmv_6hrs_Up	-2.55	5.0E-04	8.2E-03	5.6E-03
Grandvaux_Irf3_Up	-2.56	7.6E-04	1.5E-02	4.7E-03
Dac_Ifn_Bladder_Up	-2.45	9.9E-04	6.4E-03	1.5E-02
Serum_Fibroblast_Core_Dn	**-2.32**	**1.1E-03**	**1.0E-02**	**1.1E-02**
Der_Ifna_Up	-2.21	1.5E-03	5.6E-02	2.8E-03
Der_Ifng_Up	-2.22	1.8E-03	2.8E-02	6.5E-03
Radaeva_Ifna_Up	-2.18	2.1E-03	3.6E-02	6.3E-03
Chang_Serum_Response_Dn	-2.12	2.5E-03	7.0E-03	3.8E-02
Lee_Myc_Tgfa_Up	**-2.14**	**2.7E-03**	**1.9E-02**	**1.6E-02**
Lvad_Heartfailure_Dn	**-2.11**	**3.2E-03**	**1.1E-02**	**3.3E-02**
Vegf_Huvec_30min_Up	**-1.98**	**3.7E-03**	**2.8E-03**	**1.5E-01**
Uvc_High_D4_Dn	**1.92**	**3.8E-03**	**2.7E-03**	**1.6E-01**
Ifn_Gamma_Up	-2.05	4.0E-03	4.0E-02	1.2E-02
Hif1_Targets	**-2.02**	**4.7E-03**	**1.8E-02**	**3.0E-02**
Cmv_Hcmv_Timecourse_All_Up	-1.82	5.2E-03	1.7E-01	3.5E-03
Roth_Htert_Up	-2.04	5.3E-03	1.2E-02	5.2E-02
Nf90_Up	-2.00	5.4E-03	5.1E-02	1.3E-02
Lei_Myb_Regulated_Genes	**-1.90**	**5.4E-03**	**9.1E-02**	**7.2E-03**
Ifnalpha_Hcc_Up	-1.97	5.8E-03	4.7E-02	1.5E-02
Zhan_Multiple_Myeloma_Vs_Normal*	**-1.91**	**5.9E-03**	**7.6E-02**	**9.3E-03**
Ifnalpha_Nl_Hcc_Up	-1.98	6.4E-03	2.8E-02	2.8E-02
Human_Cd34_Enriched_Transcriptio*	**1.89**	**6.5E-03**	**1.3E-02**	**6.1E-02**

Ifnalpha_Nl_Up	-1.94	6.7E-03	3.6E-02	2.3E-02
Brca_Er_Pos	**-1.85**	**6.7E-03**	**9.5E-03**	**8.7E-02**
Shipp_Dlbcl_Cured_Up	**1.85**	**7.8E-03**	**9.3E-03**	**1.1E-01**
Htert_Up	**-1.83**	**8.8E-03**	**5.8E-02**	**1.9E-02**
Cmv_Hcmv_Timecourse_18hrs_Dn	**-1.88**	**9.1E-03**	**3.7E-02**	**3.1E-02**
Verhaak_Aml_Npm1_Mut_Vs_Wt_Dn	**-1.82**	**9.1E-03**	**2.1E-02**	**5.6E-02**
Takeda_Nup8_Hoxa9_3d_Up	-1.82	9.3E-03	5.1E-02	2.3E-02
Takeda_Nup8_Hoxa9_16d_Up	-1.81	9.4E-03	2.0E-02	6.1E-02

Table 3.10 Full and non-redundant list of canonical pathways inferred by GAGE. The list of significantly (P<0.01) differentially expressed canonical pathways were inferred by GAGE from human MSCs following an 8 hour BMP6 treatment. Two replicate experiments were done, each with BMP6 treated sample and control. Therefore GAGE was applied to each experiment and derived corresponding P-values (P.exp1-2). Gene sets were ranked based on global P-values from both experiments. The 4 gene sets in bold typeface are the non-redundant sub-list, all other 3 gene sets are removed during the redundant test.

Canonical Pathways	t-statistic	P-value	P.exp1	P.exp2
Tgf_Beta_Signaling_Pathway	**3.15**	**1.7E-05**	**1.1E-03**	**1.1E-03**
Wnt_Signaling	**2.47**	**5.6E-04**	**3.0E-03**	**1.7E-02**
Alkpathway	2.46	7.3E-04	8.8E-03	7.8E-03
Proliferation_Genes	**2.27**	**1.3E-03**	**6.8E-03**	**1.9E-02**
Cell_Proliferation	2.24	1.5E-03	2.1E-02	7.5E-03
Hematopoesis_Related_Transcription*	**2.05**	**3.9E-03**	**1.8E-02**	**2.5E-02**
Erythpathway	1.98	6.9E-03	2.5E-02	3.4E-02

3.2.3 Impact of GAGE strategies: gene set separation, two-sample t-test, and one-on-one comparisons

Compared to PAGE (and GSEA), GAGE employs three different strategies: (1) gene set separation, (2) two-sample t-test, and (3) one-on-one comparisons between experiment and control samples. In this section, we show the results of each of these three strategies. We compare GAGE to PAGE on these aspects if possible, or to GAGE variants which ensembles PAGE in each one of these three aspects for exact comparison. GSEA is either not or less comparable in these aspects.

3.2.3.1 Gene set separation

In contrast to PAGE and GSEA, GAGE separates canonical pathways from experimental sets and considers potential perturbations towards both directions (i.e. up and down regulation simultaneously) in canonical pathways. Expression data directly showed that genes in the most relevant canonical pathways were regulated towards both directions (Figure 3.2). Figure 3.2a shows the gene expression level changes following BMP6 treatment in top 3 different significant canonical pathways inferred by GAGE and PAGE (Table 3.5-7). These canonical pathways inferred by GAGE are directly related to BMP induced osteoblast differentiation [29, 30] (Alk pathway is essentially TGF Beta signaling + Wnt signaling). Figure 3.2b shows the gene expression level changes in the TGF beta-BMP signaling pathway following BMP6 treatment. This pathway is a presumable gold standard as it is the primary signal triggered directly by BMPs (KEGG). The changes of gene expression were not uniform. TGF-beta pathway includes both positive effectors such as BMPs, BMPR1-2, SMAD1/5/8, ID1-4, and THBS, and negative effectors such as NOG, SMAD2/3, and SMAD6/7. Clearly, both types of effectors were regulated up and down. Genes were regulated in both directions not only for the whole pathway but also within the sub-pathways like BMP or TGF-beta signaling branches. These results demonstrate that genes in canonical pathways are frequently up- and down-regulated simultaneously because (1) they play positive or negative roles [15] and (2) homeostatic mechanisms tend to bring a certain level of balance back to the system when it is perturbed [14]. Therefore, it is necessary to treat canonical pathways differently from experimental sets and count both up and down regulation when doing gene set analyses.

Compared to the top 10 canonical pathways assuming one-way changes, the top 10 canonical pathways allowing two-way changes better described BMP induced osteoblast differentiation mechanistically (Table 3.5 and 3.11). TGF beta signaling, Wnt signaling and cell proliferation are all known essential signals or processes for osteoblast differentiation [29, 30], yet they were were not significant in the one-way changing list

(Table 3.11, full Table 3.not shown). One-way assumption tended to select metabolism pathways (6 out of 10 canonical pathways in Table 3.11), which are likely to be tightly coregulated as relative simple functional group. In other words, top canonical pathways with one-way changes are still interesting if they are not complicated regulatory pathways.

Table 3.11 GAGE with the opposite assumption on canonical pathways vs experimental sets. Top 10 most significantly differentially expressed experimental sets and canonical pathways in human MSC following 8 hour BMP6 treatment were inferred by GAGE with the exact opposite assumption that all genes in a canonical pathways are regulated towards the same direction, either up or down, whereas genes in an experimental set can be regulated towards both directions at the same time. This analysis is the same as that for Table 3.5 otherwise.

Experiment Sets	t-stat	P	P.exp1	P.exp2
Cmv_Hcmv_Timecourse_All_Dn	5.81	7.8E-16	1.6E-09	1.2E-08
Uvc_High_All_Dn	4.21	4.2E-09	1.9E-06	9.5E-05
Baf57_Bt549_Dn	4.23	4.2E-09	3.0E-05	6.0E-06
Baf57_Bt549_Up	4.09	1.4E-08	8.4E-05	7.4E-06
Cmv_Hcmv_6hrs_Dn	3.98	7.9E-08	3.7E-05	1.1E-04
Takeda_Nup8_Hoxa9_3d_Up	3.71	2.5E-07	5.0E-04	2.7E-05
Cmv_Hcmv_Timecourse_All_Up	3.61	4.8E-07	2.6E-04	1.0E-04
Li_Fetal_Vs_Wt_Kidney_Up	3.61	5.4E-07	1.7E-04	1.8E-04
Cmv-Uv_Hcmv_6hrs_Up	3.63	5.5E-07	3.0E-04	9.9E-05
Boquest_Cd31plus_Vs_Cd31minus_Dn	3.48	1.2E-06	6.4E-04	1.1E-04
Canonical Pathways	t-stat	P	P.exp1	P.exp2
Valine_Leucine_And_Isoleucine_Degra*	-2.32	1.3E-03	4.2E-03	3.0E-02
Mitochondria	-2.15	1.3E-03	9.7E-02	1.4E-03
Apoptosis	-2.07	3.6E-03	1.8E-02	2.3E-02
Propanoate_Metabolism	-1.69	1.3E-02	1.2E-02	1.5E-01
Gpcrdb_Other	1.67	1.4E-02	1.3E-01	1.4E-02
Human_Mitodb_6_2002	-1.40	2.3E-02	3.1E-01	1.1E-02
Apoptosis_Genmapp	-1.53	2.6E-02	1.0E-01	3.9E-02
Limonene_And_Pinene_Degradation	-1.56	2.7E-02	3.3E-02	1.3E-01
Beta_Alanine_Metabolism	-1.46	3.0E-02	2.6E-02	1.8E-01
Raspathway	-1.46	3.5E-02	7.1E-02	8.0E-02

3.2.3.2 Two-sample t-test

GAGE uses a two-sample t-test to compare expression level changes of a gene sets to the

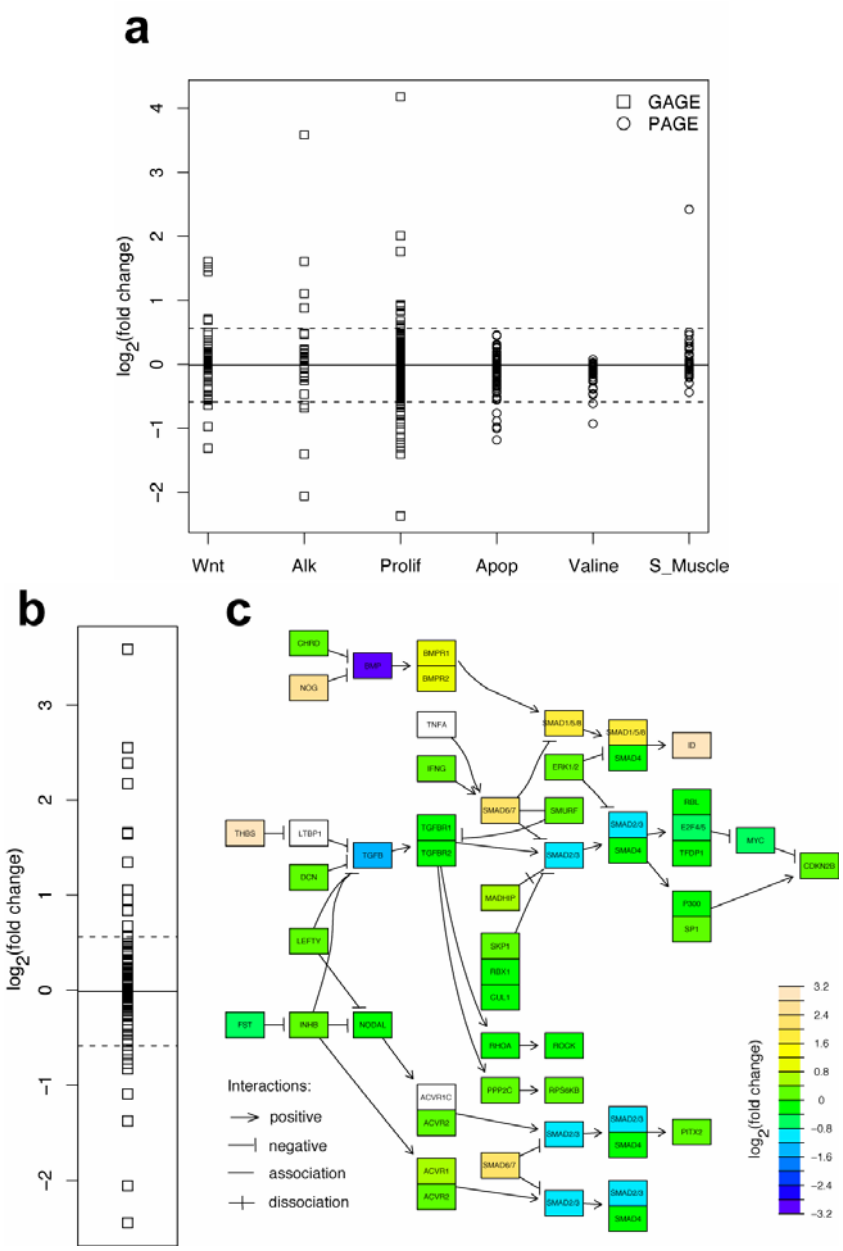


Figure 3.2 GAGE captured canonical pathways which are significantly perturbed towards both directions following 8h BMP6 treatment in human MSC. (a) Gene expression level changes in the top 3 different significant canonical pathways inferred by GAGE and PAGE. (b) Gene expression level changes in the canonical TGF beta signaling pathway and (c) plotted in pseudo-color on the pathway topology derived from KEGG database. The solid horizontal line and dashed lines in (a-b) mark the mean fold changes of all genes and the positive/negative two times standard deviation from the mean respectively. Note that in (c), one KEGG node may correspond to multiple closely related genes with

the same function, and the maximum fold changes among these genes are plotted as the color of the node.

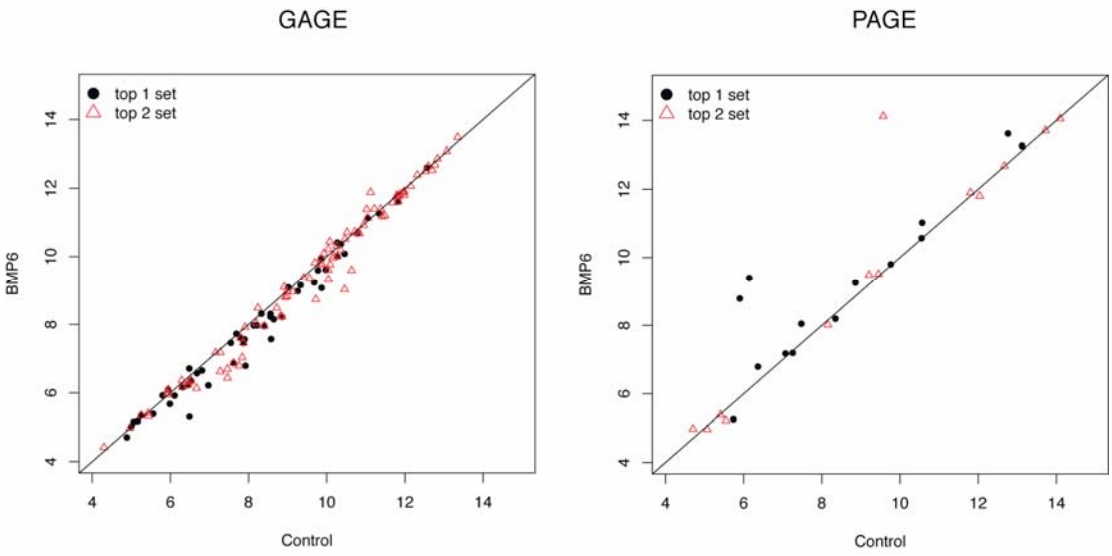


Figure 3.3 Differential gene expression in the top 2 significant experimental sets inferred by GAGE or PAGE. Gene expression levels are log 2 based, and compared between human MSC with 8 hour BMP6 treatment vs control. Results for the first experiment are shown, and the second replicate experiment is similar.

whole set background, whereas PAGE uses a one-sample z-test. GAGE's use of a two-sample t-test has three effects. First, two-sample t-test considers the variance for both the target gene set distribution as well as the background distribution (Formula 2.1), while a one-sample z-test only considers the variance for the background distribution and ignores the effect of specific target gene set distribution (Formula 2.2). The background variance is very small and often negligible compared to the within gene set variance, hence PAGE can produce unrealistically large z-scores and small P-values (Table 3.6) in contrast to GAGE (Table 3.5). Second, the two-sample t-test used by GAGE identifies gene sets with modest but consistent changes in gene expression level, whereas PAGE tends to identify gene sets with a few extremely changed outliers (Figure 3.3, more comments in Footnote 4). In other words, GAGE is more robust to experimental noise or variations in gene set definitions than PAGE. Many top gene sets selected by PAGE were not significant according to GAGE (Table 3.5, Table 3.6, full tables not shown) because

the within gene set variance is too large (Figure 3.4). On the other hand, significant gene sets inferred by GAGE are almost always selected as significant by PAGE (Table 3.5, Table 3.6, full tables not shown). Said another way, GAGE is likely as sensitive (high true positive calls) as PAGE, but more specific (low false positive calls) than PAGE. Third, there is higher level of consistency within the top 10 gene sets inferred by GAGE (Table 3.5) than by PAGE (Table 3.6), and between the top 10 gene sets cross experiments (Table 3.5 vs Table 3.6). This consistency is because the two-sample t-test is more robust than one-sample z-test for gene set analysis. All these observations for PAGE also apply to GAGE-z (GAGE variant doing one-sample z-test, data not shown).

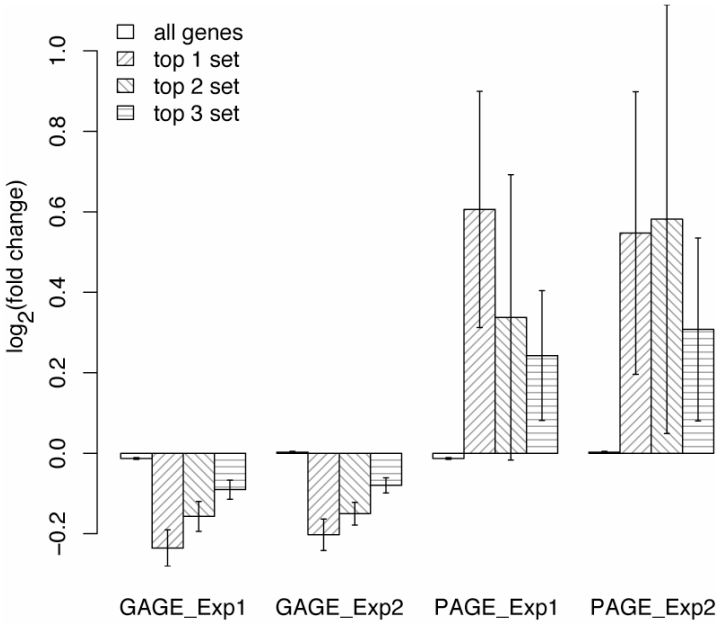


Figure 3.4 Gene expression fold changes (log 2 based) in the top 3 significant experimental sets inferred by GAGE or PAGE. For each gene set, the bar height represents mean and error bar represent standard error of gene expression fold changes induced by 8 hour BMP6 treatment in human MSC. GAGE uses two-sample t-test and PAGE does one-sample z-test. PAGE frequently selected gene sets with extreme up or down regulation in a few genes and almost no changes in the rest. Such gene sets have too large within-group variances to be called significantly different from the background based on two-sample t-test, even though their mean fold changes are big.

3.2.3.3 One-on-one comparisons

GAGE carries out one-on-one comparisons between experiments and controls, whereas

PAGE compares experiments and controls as two groups together. One-on-one comparisons are natural when the experiment samples and controls are paired. This one-on-one pairing is still preferred over group-on-group comparison even though experiments are not pair-matched for two reasons. First, multiple tests on all experiment-control pairs are more statistically powerful than single test on group averages, as the P-values would be orders of magnitude smaller for the one-on-one comparisons versus the group comparisons (Table 3.12). Second, comparisons between two specific samples makes sense but not between two sample groups when the net effect of the whole gene set is non-additive, for instance, being expressed as mean of the absolute fold changes for canonical pathways (Footnote 5). As expected, a one-on-one comparison approach produced more consistent and biologically meaningful results across independent studies (Table 3.12). The enumeration of all one-on-one comparisons is not always advantageous as it can be slow for data sets with large number of replicates. To circumvent this problem for larger datasets, we can take the average gene expression levels for all controls as a single reference state and do gene set analysis on each experiment sample vs this reference state, because controls are often more homogenous than experiments. Correspondingly, GAGE has the options for three-way comparison schemes specified as 1-on-1, 1-on-grp and grp-on-grp. The option 1-on-grp produces similar results to 1-on-1 but different results to grp-on-grp (Table 3.12). The difference between these three options is better shown when the sample conditions are complicated as in the large clinical data sets above.

3.3 Discussion

In this work we have presented a new software tool GAGE that is generally applicable to gene expression data sets of all sample sizes and experimental designs and in general performs better than two most frequently used gene set analysis packages. We have demonstrated GAGE's performance by comparing it to GSEA and PAGE in the following

three aspects: (1) consistency across parallel studies or experiments; (2) sensitivity and specificity of the pathway inference; (3) biological relevance of the pathways identified. Our results show a significant impact of separating gene sets into pathway and experimentally derived gene sets as is shown in Figure 3.2. We showed that two-way perturbations commonly occur in regulatory pathways (Figure 3.2 and Table 3.1, also in Table 3.5), which would otherwise be overlooked (Table 3.11). However, pathway derived gene sets do not always show regulation in both directions. For example, we see that metabolic pathways or functional groups such as GO term categories tend to be coregulated toward one direction (Table 3.11). Strictly speaking, these kinds of pathway gene sets are not regulatory canonical pathways and could be further separated from canonical signaling pathways (such as in MSigDB collection c2). In response to this observation, GAGE provides the option for two rounds of screening on MSigDB pathway sets. The first round assumes two-way regulation for regulatory signaling pathways while the second round assumes one-way for coregulated functional groups.

Table 3.12 The three comparison schemes of GAGE, 1-on-1, 1-on-grp and grp-on-grp. The top 10 significantly enriched experimental sets and canonical pathways in poor clinical outcomes vs good outcomes were inferred by GAGE using these three different comparison schemes from two published lung adenocarcinoma data sets [2]. Data columns are overlap between top 10 gene sets for the two studies, top 10 P-values, number of top 10 gene sets related to metastasis (bt) and tumor (t and bt), and numbers of significant gene sets with P-values ≤ 0.001.

Gene Sets & Methods		Overlap	Top 10 P-values	Metastasis	Tumor	Sign. Sets
Experiment Sets	1-on-1	3	<1.0E-16, 1.1E-9	2, 3	5, 4	203, 55
	1-on-grp	4	<1.0E-16, 5.9E-13	3, 5	6, 7	245, 122
	grp-on-grp	3	5.1E-8, 1.8E-4	3, 4	6, 8	52, 17
Canonical Pathways	1-on-1	6	5.41E-5, 3.5E-4	9, 9	9, 9	18, 9
	1-on-grp	5	6.1E-6, 7.0E-4	10, 9	10, 9	23, 10
	grp-on-grp	1	1.1E-1, 6.0E-2	4, 5	6, 5	0, 0

GAGE made two assumptions in conducting two sample t-tests on the log based fold changes of target gene set and control sets. The first assumption is approximate normal

distribution for the mean fold change of the two sets. The central limit theorem states that the distribution of an average of sampled observations is normal regardless of the nature of parent distribution when sampling size is large enough. Indeed, the mean of fold change values for gene sets with ≥10 genes are close to normal distribution as shown by q-q plot previously [4]. The second assumption is that the fold changes of genes are independent and identically distributed (IID). Dependency between genes has been a concern for all gene randomization methods. However, Netwon et al [5] argued that dependency is not necessarily an issue when GSA was conditioned on the differential expression analysis results (like fold changes). Moreover, we think dependency (coregulation) is rare for randomly sampled control gene sets. For most curated gene sets there is no coregulation under the specific condition of the microarray study (even though they might be under certain other condition), and the null hypothesis holds. For the few interesting gene sets where genes are coregulated, there will be a significant difference in expression between these sets and random control sets, hence the null hypothesis gets rejected. Therefore, gene sets which violate the IID assumption are the few significant sets and will be captured this way [4, 5]. GAGE results clearly showed that our arguments work. The same logic has also been quite successful in well established gene randomization methods [4-6].

The one-on-one comparison scheme is generally applicable to data sets of all sample sizes and experiment designs. We used a meta test to infer a global P-value for all the individual comparison P-values. The global P-values and the number of significant gene sets we derived are sensible. As in common statistical tests, these P-values tend to decrease when the sample size increases, and can become small for large data sets like the lung cancer data sets (Table 3.1), hence the number of significant gene sets can be large especially when all the redundant gene sets are kept (Table 3.2). This is still sensible because large clinical data sets (like the lung cancer studies) are much more heterogeneous than small experimental data sets (like the BMP6 study). Large data sets

study complicated problems, like cancer, and a variety of different mechanisms cause and affect tumor and metastasis.

There are frequently multiple significant gene sets which share multiple genes or represent the same regulatory mechanism, especially for experimental gene sets. This redundant gene sets problem has been discussed elsewhere in detail [36]. In response to this issue, GAGE has the option to combine redundant gene sets and give more concise significant gene set lists (Table 3.9-10). In this work, we chose not to combine these redundant gene sets for exact comparison between methods. As a benefit of not merging these sets, we took these overlapping sets as an internal control to validate the internal consistency of the predictions. There could also be multiple test issue, i.e. gene sets may become significant when gene set number is very large. We did not address this issue because the number of significant call is limited after removal of redundant gene sets and adjustments on P-values based on FDR are usually conservative. Furthermore, such adjustment is complicated when gene sets are not strictly independent and of different number of genes.

3.4 Methods

A schematic overview of GAGE procedure is shown in Figure 3.1. Here we describe the major steps of GAGE.

3.4.1 Gene sets separation

GAGE uses curated gene sets [2] collected from individual studies or pathway databases for regulatory mechanisms inference. In contrast to other gene set analysis approaches, GAGE requires that each curated gene set be identified as either a pathway set (canonical pathways) or an experimentally derived differential expression set (experiment sets). GAGE treats these two categories differently. Genes in an experimental set are assumed to be regulated in the same direction, either all up or all down, as they were in the original study. In contrast, genes associated with a pathway gene set may be heterogeneously regulated in either direction. This separation better reflects the origin of the gene set and

is therefore expected to produce better results.

For an experimental set the test statistic (score) used in GAGE is the average of the per-gene test statistics—similar to the scoring scheme used by other gene set analysis methods. However, for canonical pathways GAGE uses the average of the absolute values of the per gene test statistics to account for both up- and down-regulation.

3.4.2 Significance test

To test whether a gene set is significantly correlated with a phenotype or an experiment condition, we exam the fold changes of gene expression level in the experiment condition (or phenotype) vs control condition. Correspondingly, we want to test whether the mean fold changes of a target gene set is significantly different from that of the background set (the whole gene set of the microarray). This is a prototype two-sample t-test, as shown in Formula 2.1, in contrast to the one-sample z-test used in PAGE [4] shown in Formula 2.2.

$$t = (m - M) \Big/ \sqrt{s^2 / n + S^2 / n} \tag{2.1}$$

$$z = (m - M) \Big/ \sqrt{S^2 / n} \tag{2.2}$$

Where m, s and n are the mean fold change (log ratio of expression levels), standard deviation, and number of genes in a particular gene set, and M and S are the mean fold change and standard deviation for all of the genes in the dataset. Notice that this is a two sample t-test between interesting gene set of n genes and a virtual random set of the same size derived from the background (comparable to the one-sample z-test control set in Formula 2.2). The degree of freedom (df) for this two-sample t-test with unequal variance is given in Formula 2.3. Two sample t-test would be inaccurate when the two sample sizes are not comparable [37], for instance, comparison between a gene set and the whole set or all other genes as in T-profiler [6] (Footnote 1). The assumptions we made for the two-sample t-test are described in Discussion in detail.

$$df = \frac{(s^2 / n + S^2 / n)^2}{(s^2 / n)^2 / (n-1) + (S^2 / n)^2 / (n-1)} = (n-1)\frac{(s^2 + S^2)^2}{s^4 + S^4} \tag{2.3}$$

3.4.3 One-on-one comparison between microarray experiment and control samples

For microarray studies with one-on-one paired experiment and control samples, we calculate fold changes and carried out gene set significance tests for each experiment vs control sample pair. For microarray studies with multiple unpaired experimental and control samples, GAGE has two options: 1-on-1 and 1-on-grp. In 1-on-1 we enumerate all pairs of experiment-control and do gene set significance tests. In the 1-on-grp option we take the average gene expression level for all control samples as the sole reference, compare each experimental sample against this reference and do gene set significance tests. 1-on-1 is more rigorous theoretically. Our experiment showed that 1-on-grp gives very close results and is much faster when the sample size is large. We take 1-on-1as our standard, and leave 1-on-grp as a computationally fast option (default for unpaired experiments in this paper). We also implemented the commonly used comparison between experiment group and control group as the grp-on-grp option.

3.4.4 Combination of multiple comparisons or experiments

GAGE derived multiple t-statistics and P-values from Formula 2.1 when doing 1-on-1 or 1-on-grp comparison for data sets with replicate samples. We derive a global P-value by combining these individual P-values. Individual P-value follows a Uniform(0,1) distribution under the null hypothesis of the two-sample t-test and the negative log sum of K independent P-values follows a Gamma(K,1) distribution. Hence we can do a meta-test for all the P-values of a gene set cross multiple samples (Formula 2.4-5).

$$x = -\sum_k \log P_k \tag{2.4}$$

$$P(X > x) \sim Gamma(K,1) \tag{2.5}$$

Note that this analysis assumes that individual P-values come from independent comparisons. However, the 1-on-1 comparisons are not all independent for unpaired studies (with k=1,.,K experiments and l=1,.,L controls), thus we need to take the average of the P-values for all L comparisons of a experiment to different controls as the P-value

for that experiment (Formula 2.6) and then apply Formula 2.5 to these K independent P-values.

$$x = -\frac{1}{L}\sum_{kl}\log P_{kl} \tag{2.6}$$

3.4.5 Implementation of GAGE

GAGE is implemented in the statistical computing language R and is freely available online [38]. The gene sets used in this paper are from the Molecular Signature Database of GSEA website [18]. From this site, we used the curated gene sets (collection c2), and treat the two sub-collections experimental sets (CGP: chemical and genetic perturbations) and canonical pathways differently. There are 16966 unique gene symbols in c2, 3834 of them are nonstandard. Among these nonstandard symbols, 1190 were converted standard symbols automatically by using GAIQ database [39]. Database access and scripts for the gene symbol standardization is available upon request.

3.4.6 Comparison software

GAGE was compared to two widely used gene set analysis software packages: PAGE and GSEA. GSEA-P-R.1.0 was downloaded form GSEA website [40], and PAGE is implemented in R as part of GAGE package based on description of the authors [4] and source codes in PGSEA package [41].

3.4.7 Data sets

The gene set analysis software were compared using three datasets including two large studies and one small one.

The two large studies included a lung cancer set was provided with GSEA-R package [40] and a type 2 diabetes data set comes from ChipperDB [42]. These datasets were chosen because they were originally used to validate and/or compare GSEA [2, 3] and PAGE [4]

The small dataset is a gene expression study from our group describing human MSC response to 8 hours of exposure to the signaling molecule BMP6. This dataset includes

two experimental groups each with paired treatment and control samples, resulting in a total of 4 gene chips. The raw data were processed by using RMA implemented in the Bioconductor Affy package [43] with up-to-date probe set definition (.CDF file) based on Entrez Gene sequence, Hs133P_Hs_ENTREZG_8 [44]. Annotation data were retrieved from the GAIQ website [39]. The type 2 diabetes data set was processed similarly from raw data files.

3.5 Footnotes

3. T-profiler employs two-sample t-test, but it compares a gene set to the complementary set of all other genes, and assumes equal variance between the two set, which made it similar to a one-sample z-test in PAGE. In the formulas of the T-profiler [6], given $N_{G'}$ is much greater than N_G, the pooled standard error s approximately equal $s_{G'}$, and t statistics essentially equals z statistics.

4. I did not use synthetic data for the comparison between different methods. Real microarray data experiments accurately/consistently showed the differences between methods, which are more convincing and revealing than synthetic data experiments. The differences between methods were well explained by theoretical description in 3.2.3. Similarly, most other gene set analysis methods, including GSEA and PAGE, are evaluated based on real microarray data alone.

5. Note that we used GAGE program to remove PAGE redundant gene sets since PAGE doesn't offer such function. This remover program has been optimized for GAGE, where there were no or very few false positive calls. When applied to PAGE results, the large number of false positive calls may result in excessive redundancy removal, hence the non-redundant list could be shorter than it should be. Nonetheless, the non-redundant list is a good reference for the comparison between GAGE and PAGE.

6. GAGE more stresses the overall expression changes of the whole set, whereas PAGE is more sensitive to big changes of individual genes. GAGE can be considered more competitive (Q1) and PAGE more like self-contained (Q2) according to the classification described by Geoman [10] and Nam [1], although both are assigned to the big competitive (Q1) category.

7. First, for two groups each with n samples, we can do n independent-tests on sample pairs, yet only one test on the group averages. Obviously, the former is more powerful than the latter. Second, for gene set based analyses, what really matters is the change for the whole gene set not that for individual genes. Hence big fluctuation for single gene expression level is considered common for the same experiment condition (or within group variance) as long as the whole set net effect is zero. Taking average of such fluctuated gene expression levels within the group as the representative expression level would be misleading when the net effect of the gene set is non-additive as seen in many canonical pathways (where we take set mean of the absolute fold changes). Take a simplified example, we have a gene set of two genes, the expression level for the control condition is (2, 2). This set is perturbed for the two samples under experiment condition and becomes (4, 0) and (0, 4), both are able to achieve certain effect because the two genes are functionally related (like A OR B but not A AND B). But the average over the experimental condition is (2, 2), no different than the control at all.

3.6 References

1. Nam D, Kim SY: **Gene-set approach for expression pattern analysis**. *Brief Bioinform* 2008.

2. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**:15545-50.

3. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, et al: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nat Genet* 2003, **34**:267-73.

4. Kim SY, Volsky DJ: **PAGE: parametric analysis of gene set enrichment**. *BMC Bioinformatics* 2005, **6**:144.

5. Newton MA, Quintana FA, Den Boon JA, SENGUPTA S, AHLQUIST P: **Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis**. *Ann Appl Stat* 2007, **1**:85-106.

6. Boorsma A, Foat BC, Vis D, Klis F, Bussemaker HJ: **T-profiler: scoring the activity of predefined groups of genes using gene expression data**. *Nucleic Acids Res* 2005, **33**:W592-5.

7. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies**. *Proc Natl Acad Sci U S A* 2005, **102**:13544-9.

8. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P, Yasui Y: **Improving gene set analysis of microarray data by SAM-GS**. *BMC Bioinformatics* 2007, **8**:242.

9. Barry WT, Nobel AB, Wright FA: **Significance analysis of functional categories in gene expression studies: a structured permutation approach**. *Bioinformatics* 2005, **21**:1943-9.

10. Goeman JJ, Buhlmann P: **Analyzing gene expression data in terms of gene sets: methodological issues**. *Bioinformatics* 2007, **23**:980-7.

11. Bussemaker HJ, Ward LD, Boorsma A: **Dissecting complex transcriptional responses using pathway-level scores based on prior information**. *BMC Bioinformatics* 2007, **8 Suppl 6**:S6.

12. Baur JA, Pearson KJ, Price NL, Jamieson HA, Lerin C, Kalra A, Prabhu VV, Allard JS, Lopez-Lluch G, Lewis K, et al: **Resveratrol improves health and survival of mice on a high-calorie diet**. *Nature* 2006, **444**:337-42.

13. Smid M, Dorssers LC: **GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms**. *Bioinformatics* 2004, **20**:2618-25.

14. Saxena V, Orgill D, Kohane I: **Absolute enrichment: gene set enrichment analysis for homeostatic systems**. *Nucleic Acids Res* 2006, **34**:e151.

15. Kemp DM, Nirmala NR, Szustakowski JD: **Extending the pathway analysis framework with a test for transcriptional variance implicates novel pathway modulation during myogenic differentiation**. *Bioinformatics* 2007, **23**:1356-62.

16. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, et al: **Gene-expression profiles predict survival of patients with lung adenocarcinoma**. *Nat Med* 2002, **8**:816-24.

17. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, et al: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses**. *Proc Natl Acad Sci U S A* 2001, **98**:13790-5.

18. **the Molecular Signature Database** [http://www.broad.mit.edu/GSEA/msigdb/index.jsp]

19. Dorsam RT, Gutkind JS: **G-protein-coupled receptors and cancer**. *Nat Rev Cancer* 2007, **7**:79-94.

20. Li S, Huang S, Peng SB: **Overexpression of G protein-coupled receptors in cancer cells: involvement in tumor progression**. *Int J Oncol* 2005, **27**:1329-39.

21. Altiay G, Ciftci A, Demir M, Kocak Z, Sut N, Tabakoglu E, Hatipoglu ON, Caglar T: **High plasma D-dimer level is associated with decreased survival in patients with lung cancer**. *Clin Oncol (R Coll Radiol)* 2007, **19**:494-8.

22. Antoniou D, Pavlakou G, Stathopoulos GP, Karydis I, Chondrou E, Papageorgiou C, Dariotaki F, Chaimala D, Veslemes M: **Predictive value of D-dimer plasma levels in response and progressive disease in patients with lung cancer**. *Lung Cancer* 2006, **53**:205-10.

23. Montgrain PR, Quintana R, Rascon Y, Burton DW, Deftos LJ, Casillas A, Hastings RH: **Parathyroid hormone-related protein varies with sex and androgen status in nonsmall cell lung cancer**. *Cancer* 2007, **110**:1313-20.

24. Hidalgo GE, Zhong L, Doherty DE, Hirschowitz EA: **Plasma PGE-2 levels and altered cytokine profiles in adherent peripheral blood mononuclear cells in non-small cell lung cancer (NSCLC)**. *Mol Cancer* 2002, **1**:5.

25. Darnell JE, Jr., Kerr IM, Stark GR: **Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins**. *Science* 1994, **264**:1415-21.

26. Kristof AS, Marks-Konczalik J, Billings E, Moss J: **Stimulation of signal transducer and activator of transcription-1 (STAT1)-dependent gene transcription by lipopolysaccharide and interferon-gamma is regulated by mammalian target of rapamycin**. *J Biol Chem* 2003, **278**:33637-44.

27. Rajan P, Panchision DM, Newell LF, McKay RD: **BMPs signal alternately through a SMAD or FRAP-STAT pathway to regulate fate choice in CNS stem cells**. *J Cell Biol* 2003, **161**:911-21.

28. Fujita K, Janz S: **Attenuation of WNT signaling by DKK-1 and -2 regulates BMP2-induced osteoblast differentiation and expression of OPG, RANKL and M-CSF**. *Mol Cancer* 2007, **6**:71.

29. Rawadi G, Vayssiere B, Dunn F, Baron R, Roman-Roman S: **BMP-2 controls alkaline phosphatase expression and osteoblast mineralization by a Wnt autocrine loop**. *J Bone Miner Res* 2003, **18**:1842-53.

30. Kulterer B, Friedl G, Jandrositz A, Sanchez-Cabo F, Prokesch A, Paar C, Scheideler M, Windhager R, Preisegger KH, Trajanoski Z: **Gene expression profiling of human mesenchymal stem cells derived from bone marrow during expansion and osteoblast differentiation**. *BMC Genomics* 2007, **8**:70.

31. Balint E, Lapointe D, Drissi H, van der Meijden C, Young DW, van Wijnen AJ, Stein JL, Stein GS, Lian JB: **Phenotype discovery by gene expression profiling: mapping of biological processes linked to BMP-2-mediated osteoblast differentiation**. *J Cell Biochem* 2003, **89**:401-26.

32. Larsson J, Karlsson S: **The role of Smad signaling in hematopoiesis**. *Oncogene* 2005, **24**:5676-92.

33. Maguer-Satta V, Bartholin L, Jeanpierre S, Ffrench M, Martel S, Magaud JP, Rimokh R: **Regulation of human erythropoiesis by activin A, BMP2, and BMP4, members of the TGFbeta family**. *Exp Cell Res* 2003, **282**:110-20.

34. Helms MW, Packeisen J, August C, Schittek B, Boecker W, Brandt BH, Buerger H: **First evidence supporting a potential role for the BMP/SMAD pathway in the progression of oestrogen receptor-positive breast cancer**. *J Pathol* 2005, **206**:366-76.

35. Ong DB, Colley SM, Norman MR, Kitazawa S, Tobias JH: **Transcriptional regulation of a BMP-6 promoter by estrogen receptor alpha**. *J Bone Miner Res* 2004, **19**:447-54.

36. Jiang Z, Gentleman R: **Extensions to gene set enrichment**. *Bioinformatics* 2007, **23**:306-13.

37. Stonehouse JM, Forrester GJ: **Robustness of the t and U tests under combined assumption violations**. *Journal of Applied Statistics* 1998, **25**:63-74.

38. **GAGE R package** [http://sysbio.engin.umich.edu/~luow/downloads.php]

39. **Gene Annotation & Information Query (GAIQ)** [http://sysbio.engin.umich.edu/~luow/project/geneInfo.php]

40. **GSEA software** [http://www.broad.mit.edu/gsea/downloads.jsp]

41. Furge KA, Chen J, Koeman J, Swiatek P, Dykema K, Lucin K, Kahnoski R, Yang XJ, Teh BT: **Detection of DNA copy number changes and oncogenic signaling abnormalities from gene expression data reveals MYC activation in high-grade papillary renal cell carcinoma**. *Cancer Res* 2007, **67**:3171-6.

42. **ChipperDB type 2 diabetes data set**.

43. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level**. *Bioinformatics* 2004, **20**:307-15.

44. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, et al: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucleic Acids Res* 2005, **33**:e175.

Chapter IV

Time Series Microarray Gene Expression Profiling and Temporal Regulatory Pathway Analysis of BMP6 Induced Osteoblast Differentiation and Mineralization

4.1 Introduction

This is a systematic experimental and computational study on the regulatory mechanisms involved in BMP6 induced osteoblast differentiation and mineralization.

Osteoblasts, the bone forming cells, are responsible for bone matrix production and mineralization [1]. In concert with osteoclasts, osteoblasts coordinate bone remodeling, a physiologic process by which bone mass is maintained constant throughout adult life in vertebrates [1]. Osteoblasts arise from osteoprogenitor cells or mesenchymal stem cells (MSC) residing in the periosteum and the bone marrow [1]. Osteoblast differentiation and function are implicated directly in skeletal development and bone diseases. Identifying the endogenous factors controlling osteoblast differentiation and function and in turn the signaling and transcriptional networks activated is essential for understanding bone related physiological and pathological processes.

One set of soluble factors, the bone morphogenetic protein family (BMP) [2], induce osteoblast differentiation when delivered to cells in vitro and in vivo [3]. Among the BMPs, BMP2, 4, 6 and 7 are the best known and characterized osetogenic factors [4]. Our previous study [5] show that: (1) human MSC produce BMP6 in defined, serum-free conditions, but not BMP2, 4, or 7, (2) BMP6 is upregulated under mild osteogenic stimulus (dexamethasone), (3) exogenous BMP6 potently induces osteoblast

differentiation, but responses to BMP2, 4, or 7 are inconsistent and require higher doses, (4) exogenous BMP-6 induces the expression or upregulation of a repertoire of osteoblast-related genes in human MSC. These results establish that BMP6 is an important endogenous regulator of human osteoblast differentiation [5].

Although functionally critical, BMP6 signaling largely remains uncharacterized. Molecular descriptions of osteogenic BMP signals in a whole are still incomplete. Significant efforts, particularly a series of high throughput microarray studies [6-12] have been undertaken to uncover BMP (including BMP6) responsive genes, transcriptional programs and their roles in osteoblast development. However, an integrated understanding of the regulatory mechanisms for osteoblast differentiation and mineralization has not been achieved. Two key problems still remains unsolved include: (1) what pathways and gene groups are responsible for MSC differentiation to bone in response to BMP6 stimulation? (2) How and when these pathways are altered (induced or repressed) by BMP6 during the osteogenic induction?

To answer these two questions, we conducted a time series microarray study on BMP6 osteogenic induction and a comprehensive pathway analysis on the temporal data. We met special challenges at both experiment and data analysis levels.

At experiment level, we considered two major issues: (1) Given our limited experimental resources, what rate to sample the time series? The time intervals should be short enough to capture the dynamics and continuity, but long enough to show phenotypically significant changes. We determined time intervals based on the minimal BMP6 treatment durations needed for significant phenotypic changes, including the expression of osteoblast markers and formation of mineralized extracellular matrix (Figure 4.1). (2) Osteoblast induction is a temporal process with accumulative effects and gradual changes, how to dissect out the net effect of BMP6 at different phenotypic stages? We employed a BMP6 addition and withdraw scheme (Figure 4.1).

At data analysis level, we employed the effective method and procedure for a novel

temporal pathway analysis. Gene set analysis (GSA) is a well established strategy to identify pathways or gene sets associated with particular phenotypes or conditions [13-17]. However, no previous method is capable of inferring a dynamic list of pathways or gene sets continuously changing over a time course, or temporal pathway analysis. For temporal pathway analysis, a method needs to be: (1) applicable to time series datasets with small sample size at each time point or condition. (2) both sensitive and selective to capture subtle yet real regulatory signals over time. Our newly developed GAGE (Generally Applicable Gene-set Enrichment) method meets these technical challenges well [18]. To further handle the unevenly distributed short time series (a few time points)

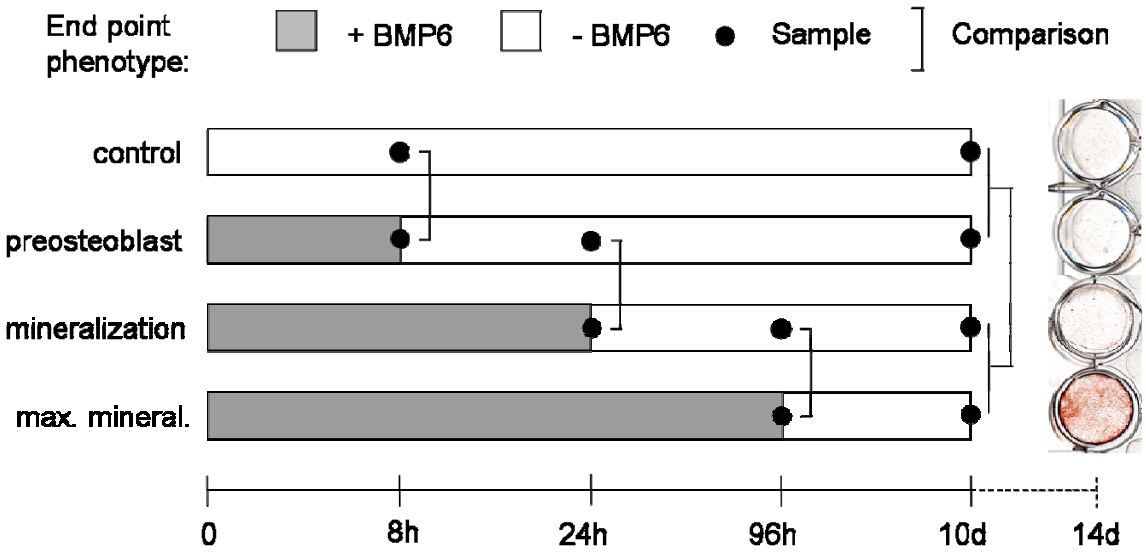


Figure 4.1 Design for the microarray study on BMP6 induced osteoblast differentiation. Human MSC cells were pre-cultured for 4 days and subsequently treated with BMP6 for 0 hours, 8 hours, 24 hours, and 96 hours. These four time points correspond to four phenotypic groups, of control, preosteoblast (no mineralization), (sub-maximal) mineralization, and maximal mineralization at 14 days (18 days in total). Cells were harvested at 8 hours, 24 hours, 96 hours and 10 days for microarray profiling. Mineralization level was quantified at 14 days by Alizarin Red S staining (right column). GAGE was applied to infer the most differentially expressed pathways or gene sets between the matched samples with or without BMP6 at different time points. For 8, 24 and 96 hours, GAGE compares between two sample conditions for the net BMP6 effect at that time, for 10 days, GAGE compares between two mineralized conditions versus two non-mineralized conditions.

dataset, we designed a special and flexible analysis procedure (Figure 4.1): GAGE was applied to compare BMP6 addition or withdrawal samples at different time slices for net temporal effect of BMP6 treatment.

In this systematic and dynamic microarray study and pathway analysis, we discovered a novel and coherent sets of regulatory mechanisms and functional groups downstream of BMP6 signaling during osteoblast differentiation and mineralization. We not only inferred which pathways or gene sets are significant, but also determined when and how they are involved in the osteoblast differentiation and mineralization.

4.2 Results

Following our previous study [5], we explored BMP6 induced human MSC osteoblast gene expression and function. Our preliminary experiments showed that 8 hours BMP6 treatment was sufficient to induce early osteoblast differentiation marker in human MSC. At least 24 hours BMP6 treatment was required to form mineralized matrix at 14 days after the initiation of BMP treatment. A maximal mineralization response was observed upon 96 hours of BMP6 treatment.

We designed a high throughput microarray study to explore the regulatory mechanisms underneath these phenotypic changes at different stages of human MSC osteoblast differentiation (Figure 4.1). Our newly developed GAGE method [18] was applied to infer the significantly perturbed KEGG pathways, GO term processes or functional groups, and experimentally derived gene sets (experimental sets for short) by BMP6 treatment at different times along the induction process (Figure 4.1). We examined these significant gene sets in details below.

4.2.1 Significantly perturbed KEGG pathways during BMP6 osteogenic induction

Essentially the same set of KEGG pathways are significantly perturbed at gene expression level throughout BMP6 induction process (Figure 4.2 and Table 4.1). In other words, these regulatory mechanisms are constantly involved at different stages of BMP6 induced osteoblast differentiation and mineralization. To test whether a KEGG pathway

are significantly associated with a phenotype or a sample condition, we account for gene expression level perturbation in both directions (both up and down regulation), since a pathway commonly includes both positive and negative effectors and local feedback loops to keep the system balanced and fine tuned. Therefore, we frequently call a KEGG pathway significantly perturbed rather than activated or inhibited as a whole.

These significant KEGG pathways show different perturbation patterns (Figure 4.2). TGF-beta signaling pathway, Cytokine-cytokine receptor interaction and Wnt signaling pathway are most perturbed at 8h; Jak-STAT signaling pathway, MAPK signaling pathway and p53 signaling pathway most perturbed at 24h, whereas Focal adhesion and ECM-receptor interaction at 10d. We checked differential gene expression induced by BMP6 treatment in three representative pathways in detail (Figure 4.3 a-c, animation not shown). (1) TGF-beta signaling pathway is the top 1 significant pathway (compared to other pathways) at 8h and is most perturbed (compared to itself at other time points) also at 8h. Therefore, this is the pathway triggered directly by BMP6 treatment, or the signal nitiating the osteoblast differentiation. This observation is supported by previous work [19] and the common sense: BMP6 as a canonical BMP triggers canonical BMP signal, which is one major branch of TGF-beta signaling pathway. (2) Focal adhesion is top 1 significant pathway (compared to other pathways) after 24 h and is most perturbed (compared to itself at other time points) at 10 days. This pathway is the convergence point of the regulatory signals, and it is associated with late stage osteoblast differentiation and mineralization. These results are consistent with the role of Focal adhesion inferred from experimental works [20, 21]. (3) MAPK signaling pathway is the most perturbed at 24 h or the middle stage, which suggests that this pathway is likely an intermediate step during the BMP6 induced signal relay process at gene expression level. Indeed, MAPKs have been reported to mediate BMP effect during osteoblast differentiation [22, 23]. The temporal perturbation patterns in other pathways suggested similar roles in the BMP6 induction process, which are supported by literature works and

Table 4.1 Interpretation and validation of the significant KEGG pathways inferred by GAGE. Temporal perturbation patterns, literature findings and experimental evidence suggest regulatory roles of these KEGG pathways in BMP6 induced osteoblast differentiation and mineralization.

KEGG pathways	Most perturbed	Literature (Pubmed ID)	Other evidences
TGF-beta signaling pathway	8h	7750491, 16317727	BMP signal targets IDs, SMAD6-7, DLX5 extremely up
Cytokine-cytokine receptor interaction	8h	16313349, 16551243	IFN target gene sets down (Table 4.4)
Wnt signaling pathway	8h	17971207, 14584895	
Jak-STAT signaling pathway	24h	16313349, 12796477	STAT1 target genes down
MAPK signaling pathway	24h	18056716, 16000303	--
p53 signaling pathway	24h	16380437, 16533949	--
Focal adhesion	10d	17081517, 17459803	--
ECM-receptor interaction	10d	11771655, 9830051	--

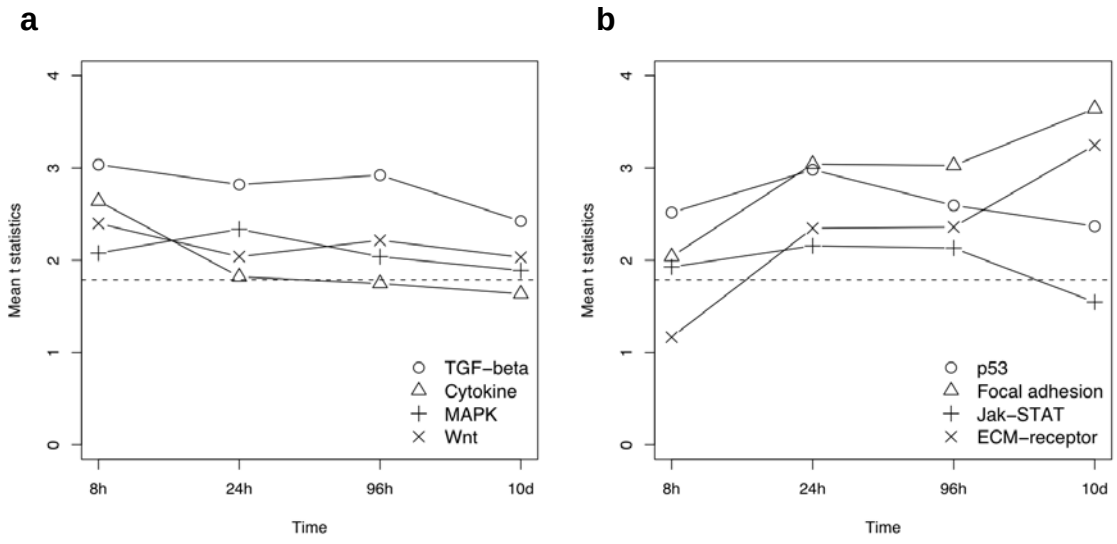


Figure 4.2 The expression perturbation patterns induced by BMP6 treatment in eight significant KEGG pathways. These pathways are consistently significantly differentially expressed or near so based on GAGE. The mean t-statistics from multiple one-on-one comparisons between the two sample conditions is used as overall perturbation magnitude for each pathway. Perturbation magnitude here measures absolute gene expression change without considering the direction. The dashed line indicates a t-statistics of 1.79, which roughly corresponds to p=0.01 for 8-96 h or 0.001 for 10 d. Two panels are used only for better view here, and similarly in Figure 4.5 and 4.7.

extra evidences from expression data (Table 4.1). The temporal roles assigned to the pathways are just relatively, because all these pathways are important throughout BMP6 induced osteoblast differentiation and mineralization (Figure 4.2). Notice that the observations are all based on gene expression data, perturbations in these pathways occurred at transcriptional level and the regulatory signal in this level is much slower or longer lasting than that in post-transcriptional level (i.e. protein phosphorylation level).

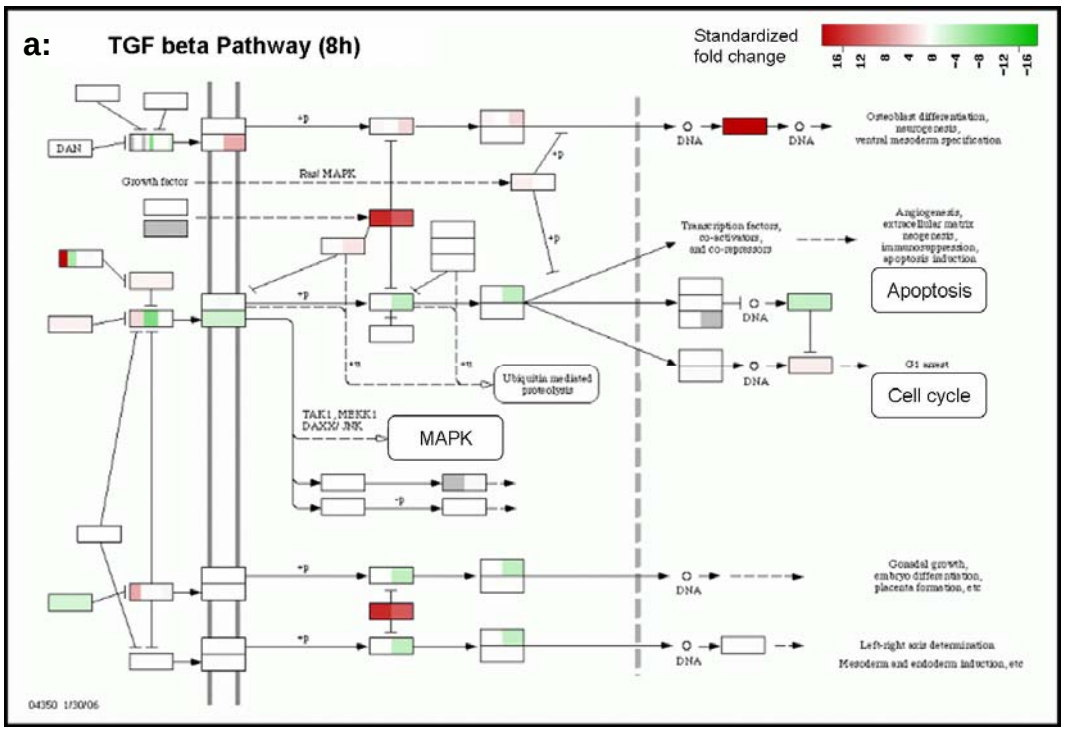
All these significant KEGG pathways are not separated but rather they work as a super regulatory system. They interconnected to each other as shown on KEGG pathway graphs. For instance, TGF-beta signal triggers MAPK signaling pathway (Figure 4.3a), whereas MAPK signaling pathway connects to Focal adhesion (Figure 4.3b). They also share common downstream response processes: including apoptosis, cell cycle etc (Figure 4.3a-c). We collected top KEGG pathways inferred by GAGE [18], gene expression data and connections between pathways from LinkDB module of KEGG databases [25], and present an integrated dynamic network of pathways in Figure 4.4. The upstream nodes including TGF-beta signaling pathway and Cytokine-cytokine receptor interaction etc are most perturbed at 8h or early stage, downstream nodes including Focal adhesion and ECM-receptor interaction are most perturbed at 10d or late stage, whereas midstream nodes including MAPK signaling pathway, p53 signaling pathway etc are most perturbed at 24h or middle stage. These sequential perturbation patterns cross interconnected pathways clearly suggest a dynamic transmission process of the regulatory signals induced by BMP6 treatment at gene expression level. Meanwhile, there are no real boundaries between these KEGG pathways. They frequently share multiple component genes that are evidently perturbed in expression level, and these overlap are statistically significant (Table 4.2). Particularly, overlaps between significant KEGG pathways (Table 4.2) are consistent with the connections between pathways (Figure 4.4): high overlaps (Table 4.2) almost always suggest direct connections between pathways (Figure 4.4), and vise versa. These overlap component genes serve as bridges cross these relative

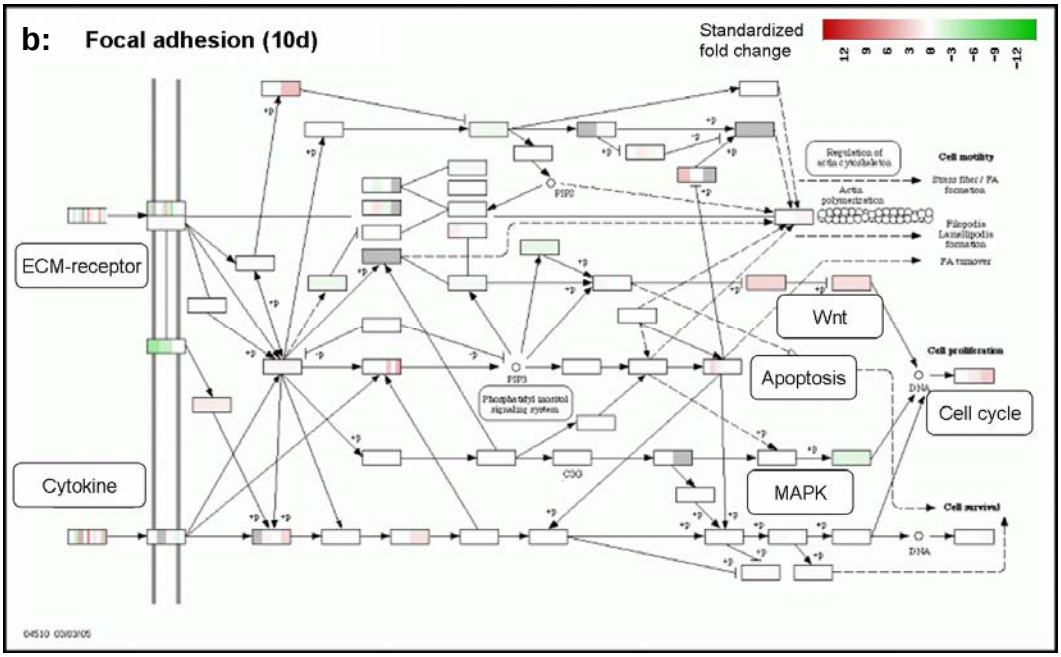
independent functional modules or pathways, hence perturbation in one pathway such as BMP-TGF beta signaling pathway) can be robustly propagated throughout other relevant pathways. All these data suggest that these significant KEGG pathways work as an integrated super regulatory system.

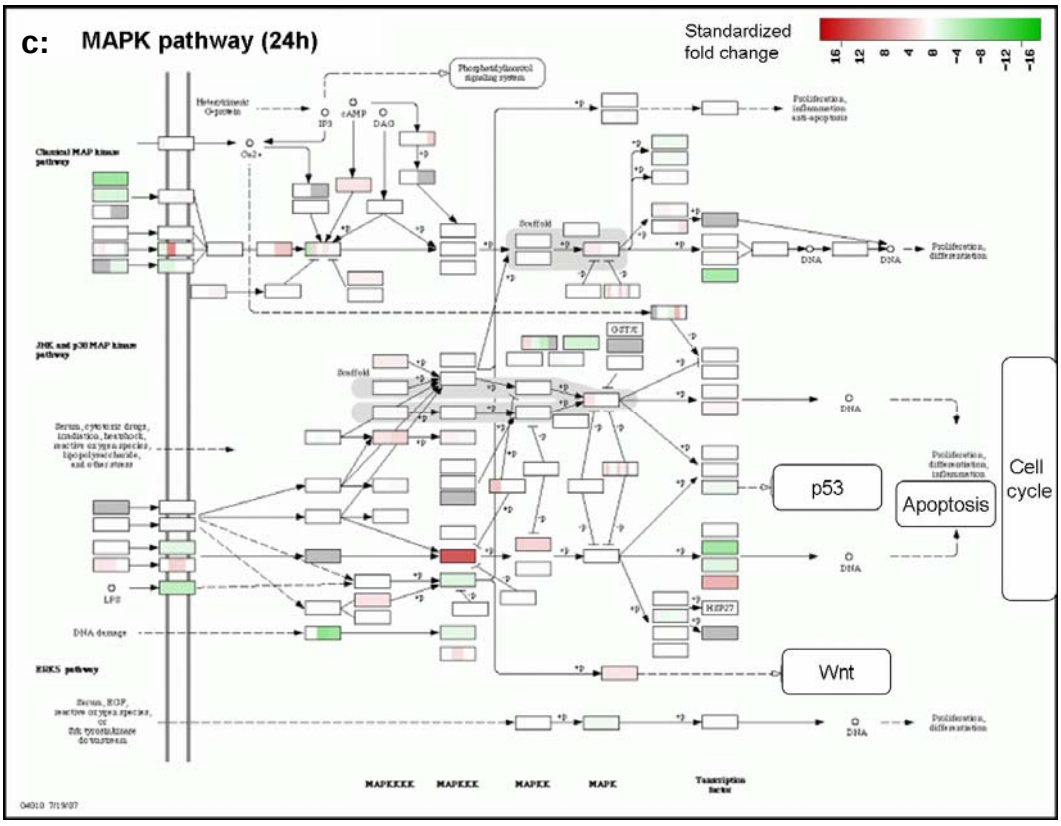
4.2.2 Significantly perturbed GO term gene sets during BMP6 osteogenic induction

Different from top KEGG pathways, top significant GO term gene sets change with time (Figure 4.5). Most relevant GO term gene sets are significantly up or down regulated only for part of the induction process, except Notch signaling pathway and insulin-like growth factor receptor binding. Different from complex KEGG pathways, genes in a GO term set are assumed to be coregulated towards one direction, either all up or all down regulated.

Figure 4.3 Gene expression fold changes induced by BMP6 in three representative significant KEGG pathways. Each pathway shown is at its most perturbed time point. Gene expression level fold changes are standardized over the standard deviation of fold changes for all genes and visualized using KEGGanim [24]. Note that one KEGG node may correspond to multiple closely related genes with the same function. Other relevant pathways are magnified locally for better view.




b:
Focal adhesion (10d)
Standardized fold change
12 9 6 3 0 -3 -6 -9 -12
ECM-receptor
Cytokine
Wnt
Apoptosis
Cell cycle
MAPK
Cell motility
Cell proliferation
Cell survival


c:
MAPK pathway (24h)
Standardized fold change
16 12 8 4 0 -4 -8 -12 -16
Classical MAP kinase pathway
JNK and p38 MAP kinase pathway
ERK5 pathway
p53
Apoptosis
Cell cycle
Wnt
MAPKKK
MAPKK
MAPK

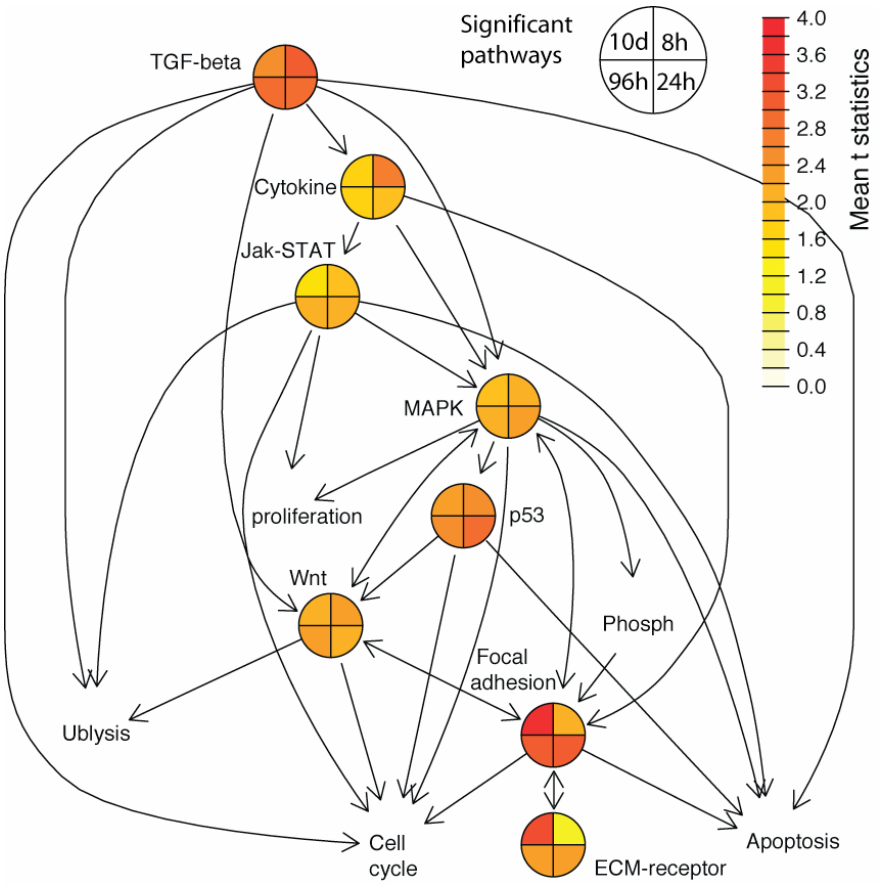


Figure 4.4 An integrated network of the significant KEGG pathways with their temporal perturbation patterns. Significant KEGG pathways (color pies) or closely related other pathways (text only) connected to them are presented by nodes. Connections between these pathways collected from KEGG database/graphs are represented by arrows plus edges. The mean t-statistics from multiple one-on-one comparisons at four different time points are plotted in pseudo heat color in the pie nodes as average perturbation magnitude for significant KEGG pathways. Full names for some abbreviated KEGG pathways are: Cytokine-cytokine receptor interaction (Cytokine), Ubiquitin mediated proteolysis (Ublysis), Phosphatidylinositol signaling system (phosph).

The top GO term gene sets are novel yet plausible regulatory processes or functional group involved in BMP6 induced osteoblast differentiation and mineralization. Some are self-explaining, including skeletal development and ossification. Other less evident top GO term gene sets are supported by literature works. Notch signaling pathway is directly involved in BMP2 induced osteoblast differentiation [11, 26]. Both insulin-like growth factor receptor binding and insulin-like growth factor binding suggest that IGF signal is

critical for osteoblast differentiation and mineralization. Indeed IGF1 [27] and IGF1R [28] promote min and bone formation, whereas IGFBP 3-6 [29-32] sequester IGF1 and inhibit osteoblast differentiation and mineralization. Direct connection between immune response and bone formation and metabolism has been well appreciated [33, 34], with a new research area, osteoimmunology [34], dedicated to this connection. Genes from homophilic cell adhesion [35] and oxidoreductase activity [36] sets play a role in ob. Selenium deficiency is associated with osteoporosis [37], a disease in bone formation and metabolism. Our GAGE predictions (Table 4.3) are novel in that (1) these GO term gene

Table 4.2 The overlaps in perturbed member genes between the significant KEGG pathways inferred by GAGE. For each cell in the table, in the parenthesis is the number of perturbed member genes overlapping between the significant KEGG pathways; outside is the significance of this overlap between KEGG pathways inferred by a statistical test based on hypergeometric distribution. A gene is counted as perturbed when its absolute fold change is at least one standard deviation higher than the mean of all gene perturbations. Full names for these KEGG pathways are: TGF-beta signaling pathway (TGFb), Cytokine-cytokine receptor interaction (Cyto), Wnt signaling pathway (Wnt), MAPK signaling pathway (MAPK), Jak-STAT signaling pathway (Jak), p53 signaling pathway (p53), Focal adhesion (Focal), ECM-receptor interaction (ECM).

	TGFb	Cyto	Wnt	MAPK	Jak	p53	Focal	ECM
TGFb	0 (22)	5.6E-20 (8)	3.3E-06 (2)	4.3E-10 (4)	3.1E-04 (1)	3.1E-04 (1)	2.9E-06 (2)	4.1E-05 (1)
Cyto	5.6E-20 (8)	0 (34)	1 (0)	3.4E-11 (5)	8.3E-29 (11)	1 (0)	6.4E-10 (4)	1 (0)
Wnt	3.3E-06 (2)	1 (0)	0 (24)	6.9E-10 (4)	2.9E-06 (2)	3.5E-09 (3)	2.3E-08 (3)	1 (0)
MAPK	4.3E-10 (4)	3.4E-11 (5)	6.9E-10 (4)	0 (33)	7.0E-04 (1)	2.7E-06 (2)	5.5E-10 (4)	1 (0)
Jak	3.1E-04 (1)	8.3E-29 (11)	2.9E-06 (2)	7.0E-04 (1)	0 (21)	1.4E-04 (1)	3.4E-04 (1)	1 (0)
p53	3.1E-04 (1)	1 (0)	3.5E-09 (3)	2.7E-06 (2)	1.4E-04 (1)	0 (15)	3.0E-09 (3)	1.9E-05 (1)
Focal	2.9E-06 (2)	6.4E-10 (4)	2.3E-08 (3)	5.5E-10 (4)	3.4E-04 (1)	3.0E-09 (3)	0 (23)	2.1E-24 (7)
ECM	4.1E-05 (1)	1 (0)	1 (0)	1 (0)	1 (0)	1.9E-05 (1)	2.1E-24 (7)	0 (8)

Table 4.3 Interpretation and validation information of the significant GO term gene sets inferred by GAGE. Temporal perturbation patterns (significantly perturbed time period and the direction), supportive literature and experimental evidence all suggest the regulatory roles of these GO terms in BMP6 induced osteoblast differentiation and mineralization.

GO terms	Perturbation	Literature (Pubmed ID)	Other evidences
Notch signaling pathway	8h-10d up	15695512, 15207708	Figure 4.6
insulin-like growth factor receptor binding	8h-10d up	12215457, 10875273	--
insulin-like growth factor binding	10d down	12733722 ,14584894, 18292241, 18395833	--
homophilic cell adhesion	24h-96h up	12070283	--
immune response	96h-10d down	16551243, 11117729	--
skeletal development	8h-96h up	--	--
ossification	96h-10d up	--	--
oxidoreductase activity	8h-24h, 10d down	17949499	--
selenium binding	96h-10d up	--	--

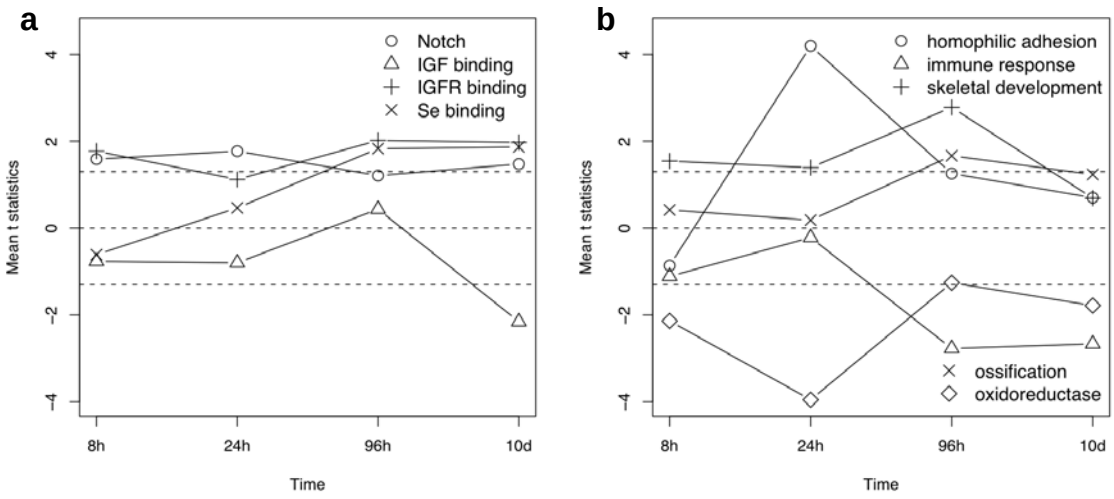


Figure 4.5 The expression perturbation patterns induced by BMP6 treatment in nine significant GO term gene sets. Each of these GO term gene sets is significant in at least one time point based on GAGE. The mean t-statistics from multiple one-on-one comparisons between the two sample conditions is used as overall perturbation for each GO term. Different from KEGG pathways, perturbation direction is considered here. Dashed line marks t=+/-1.30, roughly corresponds to p=0.05 for 8-96 h or 0.01 for 10 d.

sets as whole processes or functional groups instead of individual genes are involved in regulation of osteoblast differentiation and mineralization, (2) these are regulatory mechanisms triggered by BMP6 induction, (3) they are effective for specific time periods during the induction process.

We take Notch signal as an example prediction for direct experiment validation (Figure 4.6). We block Notch signal using the specific gamma secretase inhibitor (GSI, L-685458) [38, 39]. GSI inhibits BMP6 induced osteoblast differentiation and mineralization, and this inhibition is effective at all stages of the induction process (Figure 4.6). These results confirmed that Notch signaling pathway is a critical throughout BMP6 osteogenic induction. Downstream effect of Notch signal is likely mediated by HEY1 and HEY2, two transcription factors which are significantly unregulated (Figure 4.8c).

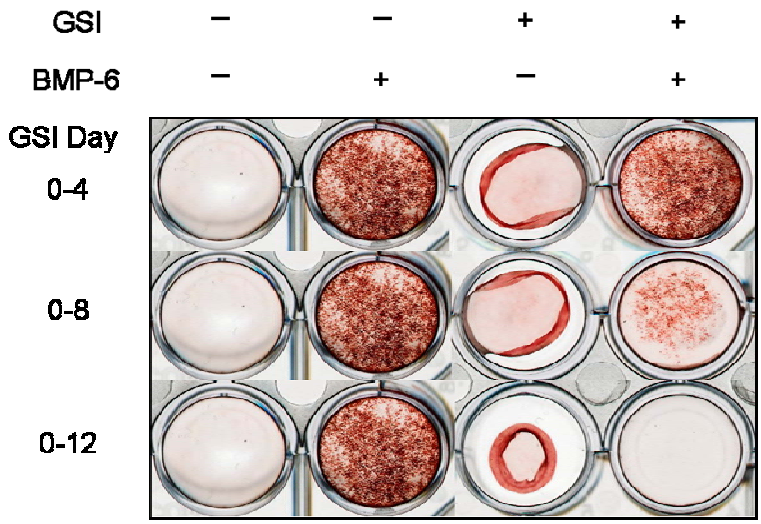


Figure 4.6 Effect of blocking Notch signal using GSI on BMP6 induced osteoblast differentiation and mineralization. Human MSC cells were treated with BMP for 4 days or untreated (control). GSI (the gamma secretase inhibitor, L-685458) was added for the indicated days. The cells were harvested and stained with Alizarin Red S at 16 days after the initiation of BMP treatment. Note that cells in the + GSI - BMP6 column peeled off from the culture surface. This is presumably a direct effect of notch signal inhibition as GSI is a specific notch signal inhibitor [37, 38] and integrin-mediated cell adhesion is directly regulated by gamma-secretase-mediated intramembranous cleavage of membrane-bound Notch [54].

4.2.3 Significantly perturbed experimentally derived gene sets during BMP6 osteogenic induction

The most significantly up or down-regulated experimental sets during BMP6 induction are selected using GAGE [18]. There are commonly multiple experimental sets describing the same perturbation or mapping to the same regulatory mechanism. A non-redundant subset of the top experimental sets were collected in Table 4.4 with their perturbation patterns shown in Figure 4.7. Similar to gene sets based on GO terms but not to those on KEGG pathways, experimental sets are significantly perturbed towards one direction, either up or down regulated.

MYB transcription factor is selected as a novel regulator for osteoblast differentiation. MYB target gene set was down regulated at 8 h, 24 h and 10 d with no change in MYB expression level. MYB transcriptional activity at protein level can be inhibited through two potential mechanisms following BMP6 treatment: the activation of Wnt signaling pathways (Table 4.1 and Figure 4.2) phosphorylates and degrades of MYB protein [40], and BMP/TGF beta and Wnt responsive OVOL1 antagonizes transcriptional activation of MYB by competing for target promoter binding [41].

Another novel transcriptional regulator for osteoblast differentiation we predicted is BAF57, which is the regulatory subunit SWI/SNF chromatin remodeling complex [42]. Multiple BAF57 target genes are directly related to osteoblast differentiation and function (Table 4.5). Indeed, SWI/SNF regulates osteoblast-specific transcription through chromatin structure modification [43]. BMP6 treatment may target SWI/SNF to nucleus though SMAD1 signal [44] or p38 MAPK pathway targets SWI-SNF chromatin-remodeling complex [45]. Interestingly, BAF57 positive target genes are up-regulated and negative targets down-regulated during BMP6 induction, which further confirms the involvement of BAF57 activity. The different timing of the positive and negative regulation likely suggests different dynamics of these actions.

Other interesting regulatory mechanisms are inferred based on the top ranking experimental sets. Interferon beta (interferon alpha and gamma too, but not shown)

positive target gene sets are down-regulated, which is consistent with Jak-STAT pathway from KEGG (Table 4.2). VEGF positive targets are down-regulated, likely because VEGF gene expression is down-regulated due to MYB inhibition [46]. IRS negative targets are down-regulated is consistent with activation of IGF signal, particularly up-regulation of IGF receptor binding proteins (Table 4.4 and Figure 4.5).

4.3 Discussion

This is the first high throughput microarray study on BMP6 induced transcriptional program in human MSCs. It covers the whole process from early to late stage osteoblast differentiation and mineralization. We conducted a comprehensive gene set analysis to identify relevant regulatory mechanisms and functional groups. We inferred a series of significant KEGG pathways, GO terms and experimental sets at different stages of BMP6 induction process. We not only showed which pathways or gene sets are significant, but also when and how they are involved in the osteoblast differentiation and mineralization. Different from common pathway analyses [14-17], our work further captures the interconnections among individual pathways or functional groups and integrate them into

Table 4.4 Interpretation and validation information on the significant experimental sets inferred by GAGE. Temporal perturbation patterns, supportive literature works and extra experimental evidences all suggest regulatory roles in BMP6 osteoblast induction. Experimental sets came from MSigDB [52] c2 collection, where the original names for these significant gene sets are: Lei_MYB_Regulated_Genes, BAF57_BT549_Dn, BAF57_BT549_Up; (B) Der_IFNB_Up, VEGF_Mmmec_6hrs_Up, IRS_Ko_Adip_Up.

Experimental sets	**Perturbation**	**Literature (Pubmed ID)**	**Other evidences**
MYB targets	8h-24h,10d down	15082531, 17311813	Wnt signal (KEGG)
BAF57 down	8h-24h,10d down	16769725, 16772287, 15231748, 15208625	BAF57 up
BAF57 up	24h-96h up	16769725, 16772287, 15231748, 15208625	BAF57 down
IFNb up	8h,10d down	--	Jak-STAT signal (KEGG)
VEGF up	10d down	16650815	MYB targets
IRS down	8h-10d down	--	IGF signal (GO)

Table 4.5 Eleven BAF57 positive target genes (the 'BAF57 up' gene set) evidently induced by 8 hours BMP6 treatment. A gene is counted as evidently induced with its fold change is one standard deviation higher than the mean for all genes. Eight of these genes are likely regulators for osteoblast differentiation based on their functional annotation.

Gene	Relevant function
FZD1	Wnt signaling
PRRX1	Transcription factor
DIO2	--
AGC1	ECM, osteoblast function
PRRX2	Transcription factor
GHR	Osteoblast prolif./diff.
JAG1	Notch signaling; mineralization
LBH	--
BMPR2	BMP signaling
LMCD1	--
SOCS2	Insulin signaling, osteoblast function

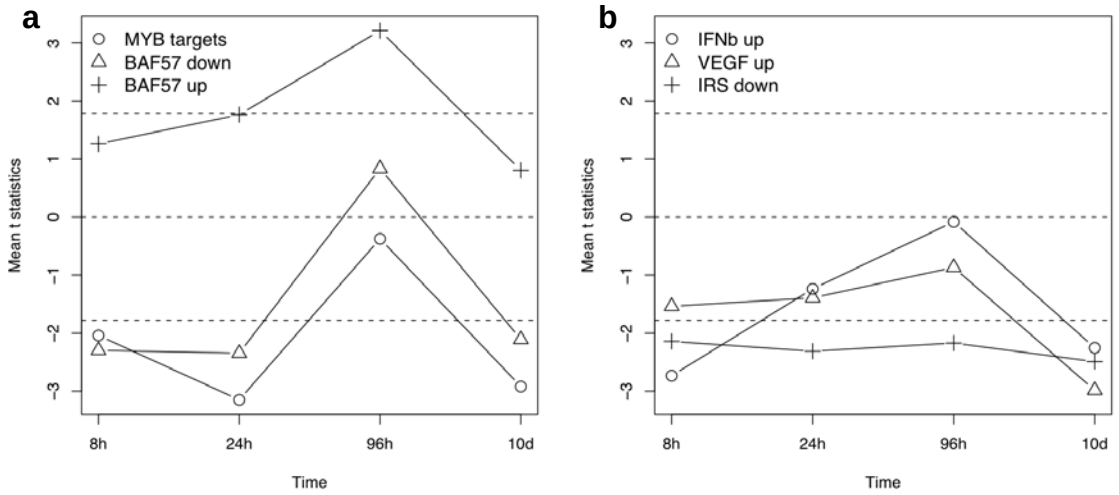


Figure 4.7 The expression perturbation patterns induced by BMP6 treatment in six significant experimental sets. Each of these experimental sets is significantly up or down regulated in at least one time point based on GAGE. The mean t-statistics from multiple one-on-one comparisons is used as overall perturbation for each gene set. Different from KEGG pathways, perturbation direction is considered here. Dashed line marks t=+/-1.78, which roughly corresponds to p=0.01 for 8-96 h or 0.001 for 10 d.

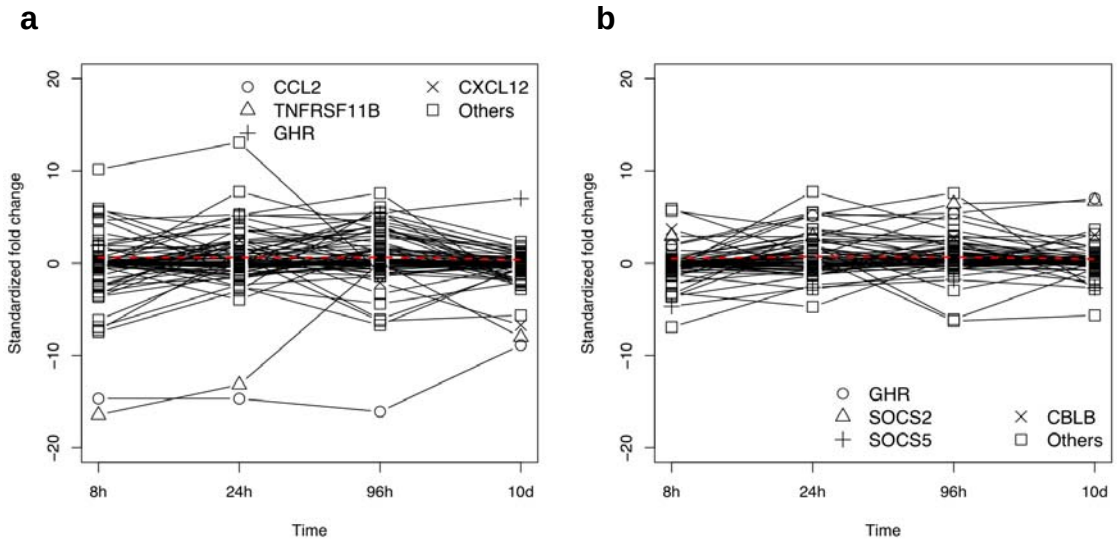
a whole system. Taken together, this work provides clearer mechanistic picture of osteoblast differentiation and function. We inferred novel and coherent sets of regulatory mechanisms downstream of BMP6 signaling during osteoblast differentiation and mineralization. First, the same set of KEGG pathways are constantly involved in BMP6 induction. Their roles in osteogenic induction are clarified based on their perturbation patterns and connected to relevant discoveries in literature. These significant KEGG pathways are not separated but rather they work as a unified super regulatory system, and the pathway perturbation patterns we derived reflect a dynamic transmission process of the regulatory signal at transcriptional level along the super system. Second, a varying set of GO processes and functional groups are involved at different stage of BMP6 induced osteoblast differentiation and mineralization. These suggest novel yet plausible regulatory mechanisms, which are validated using experiment and/or connected to literature works. Third, the most significant experimental sets suggest novel transcriptional regulators including MYB and BAF57, and regulatory pathways consistent with predictions based on KEGG and GO gene sets above.

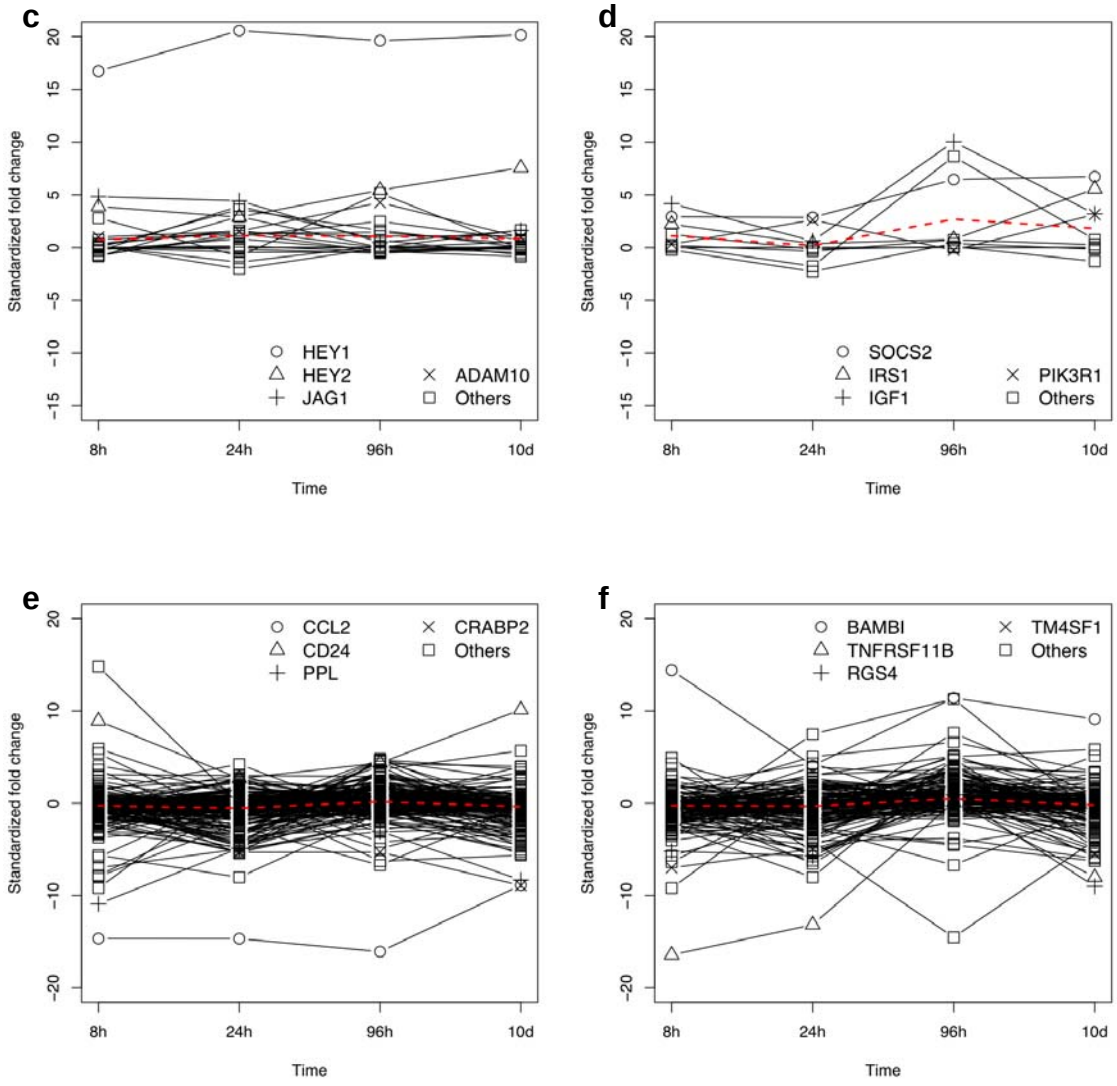
Connections between KEGG pathways are evident as shown in the super regulatory network of pathways (Figure 4.4 and Table 4.2). Perturbation or signal propagates along the super network at two levels: at protein level, the phosphorylation, binding, activation/inhibition events relay along pathways and transmit into interconnected pathways, as stated by KEGG graphs (Figure 4.3); at transcriptional level, gene expression perturbation propagates through auto-regulatory (feedback and feed forward) loops within pathways, and bridges into its neighbor pathways through the multiple component genes in common, as suggested by our pathway analysis results (Figure 4.4 and Table 4.2). Protein level transmission is much faster, but transcriptional level transmission lasts much longer hence ensure the long term biological effects. The latter echoes the former process with a long lag. The hard wired KEGG pathways and interconnections between them define how BMP6 signal triggers downstream programs

(Figure 4.4).

Presumably, there should be connections between BMP6 signal and the processes/groups represented by GO terms and experimental sets too. For example, Notch signal and IGF signal are involved in the whole induction process (Figure 4.5a) like all significant KEGG pathways (Figure 4.2). It follows that these two signals should be also part of the super regulatory system and interconnected with multiple significant KEGG pathways of the network (Figure 4.4). This hypothesis is well supported by our data and literature: (1) Notch signal directly interacts with BMP signal. SMAD1 and NIC synergize to induce expression of HEY1 and other Notch targets [47-49]. Indeed up-regulation of HEY1 (and HEY2) requires continued BMP6 treatment by 96 hours (Figure 4.8c). Besides this binding-synergy at protein level, Notch ligand JAG1 expression is also up-regulated by BMP6 (Figure 4.8c). Notch interacts with Wnt signal too [50]. (2) GH (Growth hormone)

Figure 4.8 Individual gene expression perturbation patterns induced by BMP6 treatment in the representative significant gene sets. (a) Cytokine-cytokine receptor interaction (KEGG); (b) Jak-STAT signaling pathway (KEGG); (c) Notch signaling pathway (GO); (d) insulin-like growth factor receptor binding (GO); (e) MYB targets, (f) BAF57 down. Log 2 based expression level fold change induced by BMP6 treatment were standardized over the standard deviation for all genes. Top 4 most perturbed genes were labeled out differently from other genes. The red dashed line marks the mean for each gene set.





signal [51] as part of cytokine-cytokine receptor interaction (KEGG pathway 04060) and Jak-STAT pathways (KEGG pathway 04630) are activated and by BMP signal (Figure 4.4) with GHR up-regulated (Figure 4.8a-b), GH signal up-regulates (IGF1 and IRS1, Figure 4.8d) and activates IGF signal in turn [51]. Connections between BMP signal and the two predicted transcriptional regulators, MYB and BAF57, are described above in the Results.

We find strong consistency among significant KEGG pathways, GO terms and experimental sets. For example, KEGG focal adhesion and ECM receptor interaction

pathways (Table 4.1), GO homophilic cell adhesion (Table 4.3) and extracellular matrix structural constituent (not shown) groups consistently show the relevance of cell adhesion and extracellular matrix in osteoblast differentiation and mineralization. GO immune response groups (Table 4.3) echoes KEGG cytokine-cytokine receptor interaction pathway (Table 4.1). Similarly, significant experimental target genes sets closely reflected changes in the regulatory KEGG pathways or GO processes/groups (Table 4.4).

Discrepancies among significant KEGG pathways, GO terms, and experimental sets exist too. For example, Notch signaling is defined both as KEGG pathway and GO process. This KEGG pathway is not significant (not shown) but this GO process is (Figure 4.5). This discrepancy between arises from two sources: (1) different definitions, i.e. KEGG pathways contain partially different set of genes from corresponding GO processes. While KEGG pathways tend to cover the whole haemostatic signal transmission systems even cross multiple transcriptional cycles, GO usually covers one or multiple discrete steps or functional groups for a process. KEGG and GO definitions can be considered complementary and both provide valuable gene sets for our analysis. (2) GAGE [18] treats KEGG pathways and GO term gene sets differently: genes under a GO term are taken as a group coregulated towards a single direction, either all up or all down regulated, whereas genes in a KEGG pathway are frequently not coregulated and expression changes in both directions are counted. Timing discrepancies exist between experimental sets and corresponding KEGG pathways. For example, IFN positive target sets (only IFN beta shown) are not significant at 24-96 hours and MYB target set not at 96 hours (Figure 4.7) while Jak-STAT pathway and Wnt signaling pathway are significantly perturbed all the time (Figure 4.2). This can be explained by the fact that two-directional perturbation treatment for KEGG pathway does not account for direction or net effect of the perturbation, whether inhibited, activated or no overall effect. In the other hand, GO term analysis has no such issue, and IRS negative set and corresponding GO IGF receptor binding group are both significant all the time.

In this work, we not only took a systems approach in studying MSC differentiation, but also followed a systems biology research procedure: rational experimental design, whole-system or genome-wide expression profiling, high throughput data process and analysis, results interpretation and experimental validation at pathway and system level. This study combines experimental and computational work to synthesize a unified picture of BMP6 signaling. Because the methods in this paper are generally applicable, the same set of experimental design and computational approach could be used to study other physiological and pathological processes, including cell differentiation of other cell types and tumorogeneis for example.

4.4 Methods

4.4.1 Cell culture and BMP6 osteogenic induction

Passage 5 human MSC (5×10^5) were plated in 24-well dishes and cultured for 3 days. The cells were subsequently placed in serum free media supplemented with ITS for 24 hours. BMP was then added for the pre-designed time periods and then removed. To remove BMP6, cells were rinsed 2 times and fresh media without BMP was added. Ascorbate and b-glycerolphosphate were added 4 days after the initiation of BMP treatment. Cells were harvested 14 days after the initiation of BMP treatment. To quantifying mineralization, the plates were stained with Alizarin Red S [5].

4.4.2 Microarray experiment and analysis

Human MSC cells underwent osteogenic induction with BMP6 treatment for 0 hours, 8 hours, 24 hours, and 96 hours, which correspond to four phenotypic groups, i.e. control, preosteoblast (no mineralization), (sub-maximal) mineralization, and maximal mineralization at 14 days after the initiation of BMP treatment (18 days in total). Cells were harvested at 8 hours, 24 hours, 96 hours and 10 days for microarray profiling using Affymetrix U133 plus Genechip® platform. The raw data (.CEL file) were processed by using RMA implemented in the Bioconductor Affy package [43] with up-to-date probe

set definition (.CDF file) based on Entrez Gene sequence, Hs133P_Hs_ENTREZG_8 [44]. Annotation data were retrieved from the GAIQ website [39]. GAGE [18] was applied to infer the most differentially expressed pathways or gene sets between the BMP6 added or withdrawn samples under comparison at different time points (Figure 4.1). Note that comparisons at 8, 24 and 96 hours were between two sample conditions, whereas comparison at 10 days was between the 24 and 96 hours BMP6 groups versus 0 and 8 hours BMP6 groups.

4.4.3 Notch signal inhibition experiment

Human MSC were treated with a gamma secretase inhibitor (GSI, L-685458) to antagonize Notch signaling. The cells were treated with BMP for 4 days or untreated (control). GSI was added for the first 4 days of induction, for 8 days, or for 12 days. The cells were harvested and stained at 16 days after the initiation of BMP treatment.

4.4.4 Pathway analysis using GAGE

We use GAGE [18], Generally Applicable Gene set Enrichment, a novel method we developed for gene set/pathway analysis. GAGE procedure has been described in detail in the original paper [18]. Here is the brief procedure.

Step 1 Gene sets separation. Gene sets are derived or collected from KEGG pathways [25], GO [52] and MSigDB [53] databases, as KEGG pathways, GO terms and experimental sets respectively. GAGE treats KEGG pathways differently from GO terms and experimental sets: member genes for a GO term or experimental set are taken as a group coregulated towards a single direction, either all up or all down regulated, whereas genes in a KEGG pathway are frequently not coregulated and expression changes in both directions are counted.

Step 2 One-on-one comparison. Instead of comparing BMP6 treated samples vs controls as two groups, GAGE does one-on-one comparison between samples from the two groups at a time. For each one-on-one comparison, log based fold changes are calculated for all genes. GAGE conducts two-sample t-test on the average fold change in specific

gene sets against that for the background whole set. Repeat such one-on-one comparison procedure for all potential experiment-control pairs.

Step 3 Summarization. For each gene set, GAGE derives a global P value based on a meta test on the negative log sum of multiple P-values for this set from all one-on-one comparisons between experiments and controls.

4.4.5 Perturbation pattern visualization

For significant KEGG pathways, we generated graphs to visualize the dynamic expression perturbation at two levels: individual genes (Figure 4.3, animation not shown) and whole pathways (Figure 4.4). Gene expression level fold changes are standardized over the standard deviation of fold changes for all genes. The standardized fold changes for individual genes in KEGG pathways are visualized by using KEGGanim web tool [24] in Figure 4.3 (animation not shown). We present an integrated network to show the connections between pathways and average expression perturbations for them in Figure 4.4. Significant KEGG pathways or closely related other pathways connected to them are presented by nodes. Connections between these pathways collected from KEGG database/graphs [25] are represented by arrows plus edges. The mean t-statistics from two-sample t-test from multiple one-on-one comparisons are plotted in pseudo heat color as average perturbation magnitude for significant KEGG pathways.

4.5 References

1. Harada S, Rodan GA: **Control of osteoblast function and regulation of bone mass**. *Nature* 2003, **423**:349-55.

2. Chen D, Zhao M, Mundy GR: **Bone morphogenetic proteins**. *Growth Factors* 2004, **22**:233-41.

3. Li X, Cao X: **BMP signaling and skeletogenesis**. *Ann N Y Acad Sci* 2006, **1068**:26-40.

4. Lavery KS, Swain PM, Falb D, Alaoui-Ismaili MH: **BMP-2/4 and BMP-6/7 differentially utilize cell surface receptors to induce osteoblastic differentiation of human bone marrow derived mesenchymal stem cells**. *J Biol Chem* 2008.

5. Friedman MS, Long MW, Hankenson KD: **Osteogenic differentiation of human mesenchymal stem cells is regulated by bone morphogenetic protein-6**. *J Cell Biochem* 2006, **98**:538-54.

6. de Jong DS, Vaes BL, Dechering KJ, Feijen A, Hendriks JM, Wehrens R, Mummery CL, van Zoelen EJ, Olijve W, Steegenga WT: **Identification of novel regulators associated with early-phase osteoblast differentiation**. *J Bone Miner Res* 2004, **19**:947-58.

7. Peng Y, Kang Q, Cheng H, Li X, Sun MH, Jiang W, Luu HH, Park JY, Haydon RC, He TC: **Transcriptional characterization of bone morphogenetic proteins (BMPs)-mediated osteogenic signaling**. *J Cell Biochem* 2003, **90**:1149-65.

8. Harris SE, Guo D, Harris MA, Krishnaswamy A, Lichtler A: **Transcriptional regulation of BMP-2 activated genes in osteoblasts using gene expression microarray analysis: role of Dlx2 and Dlx5 transcription factors**. *Front Biosci* 2003, **8**:s1249-65.

9. Korchynskyi O, Dechering KJ, Sijbers AM, Olijve W, ten Dijke P: **Gene array analysis of bone morphogenetic protein type I receptor-induced osteoblast differentiation**. *J Bone Miner Res* 2003, **18**:1177-85.

10. Kalajzic I, Staal A, Yang WP, Wu Y, Johnson SE, Feyen JH, Krueger W, Maye P, Yu F, Zhao Y, et al: **Expression profile of osteoblast lineage at defined stages of differentiation**. *J Biol Chem* 2005, **280**:24618-26.

11. de Jong DS, Steegenga WT, Hendriks JM, van Zoelen EJ, Olijve W, Dechering KJ: **Regulation of Notch signaling genes during BMP2-induced differentiation of osteoblast precursor cells**. *Biochem Biophys Res Commun* 2004, **320**:100-7.

12. Balint E, Lapointe D, Drissi H, van der Meijden C, Young DW, van Wijnen AJ, Stein JL, Stein GS, Lian JB: **Phenotype discovery by gene expression profiling:**

mapping of biological processes linked to BMP-2-mediated osteoblast differentiation. *J Cell Biochem* 2003, **89**:401-26.

13. Nam D, Kim SY: **Gene-set approach for expression pattern analysis**. *Brief Bioinform* 2008.

14. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**:15545-50.

15. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, et al: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nat Genet* 2003, **34**:267-73.

16. Kim SY, Volsky DJ: **PAGE: parametric analysis of gene set enrichment**. *BMC Bioinformatics* 2005, **6**:144.

17. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies**. *Proc Natl Acad Sci U S A* 2005, **102**:13544-9.

18. Luo W, Friedman M, Shedden K, Hankenson KD, Woolf P: **GAGE: Generally Applicable Gene Set Enrichment for Pathway Inference**. *Submitted to BMC Bioinformatics* 2008.

19. Gosselet FP, Magnaldo T, Culerrier RM, Sarasin A, Ehrhart JC: **BMP2 and BMP6 control p57(Kip2) expression and cell growth arrest/terminal differentiation in normal primary human epidermal keratinocytes**. *Cell Signal* 2007, **19**:731-9.

20. Kim JB, Leucht P, Luppen CA, Park YJ, Beggs HE, Damsky CH, Helms JA: **Reconciling the roles of FAK in osteoblast differentiation, osteoclast remodeling, and bone regeneration**. *Bone* 2007, **41**:39-51.

21. Salasznyk RM, Klees RF, Williams WA, Boskey A, Plopper GE: **Focal adhesion kinase signaling pathways regulate the osteogenic differentiation of human mesenchymal stem cells**. *Exp Cell Res* 2007, **313**:22-37.

22. Ulsamer A, Ortuno MJ, Ruiz S, Susperregui AR, Osses N, Rosa JL, Ventura F: **BMP-2 induces Osterix expression through up-regulation of Dlx5 and its phosphorylation by p38**. *J Biol Chem* 2008, **283**:3816-26.

23. Celil AB, Campbell PG: **BMP-2 and insulin-like growth factor-I mediate Osterix (Osx) expression in human mesenchymal stem cells via the MAPK and protein kinase D signaling pathways**. *J Biol Chem* 2005, **280**:31353-9.

24. Adler P, Reimand J, Janes J, Kolde R, Peterson H, Vilo J: **KEGGanim: pathway animations for high-throughput data**. *Bioinformatics* 2008, **24**:588-90.

25. **KEGG: Kyoto Encyclopedia of Genes and Genomes** [http://www.genome.jp/kegg/]

26. Nobta M, Tsukazaki T, Shibata Y, Xin C, Moriishi T, Sakano S, Shindo H, Yamaguchi A: **Critical regulation of bone morphogenetic protein-induced osteoblastic differentiation by Delta1/Jagged1-activated Notch1 signaling**. *J Biol Chem* 2005, **280**:15842-8.

27. Zhao G, Monier-Faugere MC, Langub MC, Geng Z, Nakayama T, Pike JW, Chernausek SD, Rosen CJ, Donahue LR, Malluche HH, et al: **Targeted overexpression of insulin-like growth factor I to osteoblasts of transgenic mice: increased trabecular bone volume without increased osteoblast proliferation**. *Endocrinology* 2000, **141**:2674-82.

28. Zhang M, Xuan S, Bouxsein ML, von Stechow D, Akeno N, Faugere MC, Malluche H, Zhao G, Rosen CJ, Efstratiadis A, et al: **Osteoblast-specific knockout of the insulin-like growth factor (IGF) receptor gene reveals an essential role of IGF signaling in bone matrix mineralization**. *J Biol Chem* 2002, **277**:44005-12.

29. Strohbach C, Kleinman S, Linkhart T, Amaar Y, Chen ST, Mohan S, Strong D: **Potential involvement of the interaction between insulin-like growth factor binding protein (IGFBP)-6 and LIM mineralization protein (LMP)-1 in regulating osteoblast differentiation**. *J Cell Biochem* 2008.

30. Mukherjee A, Rotwein P: **Insulin-like growth factor-binding protein-5 inhibits osteoblast differentiation and skeletal growth by blocking insulin-like growth factor actions**. *Mol Endocrinol* 2008, **22**:1238-50.

31. Silha JV, Mishra S, Rosen CJ, Beamer WG, Turner RT, Powell DR, Murphy LJ: **Perturbations in bone formation and resorption in insulin-like growth factor binding protein-3 transgenic mice**. *J Bone Miner Res* 2003, **18**:1834-41.

32. Zhang M, Faugere MC, Malluche H, Rosen CJ, Chernausek SD, Clemens TL: **Paracrine overexpression of IGFBP-4 in osteoblasts of transgenic mice decreases bone turnover and causes global growth retardation**. *J Bone Miner Res* 2003, **18**:836-43.

33. Arron JR, Choi Y: **Bone versus immune system**. *Nature* 2000, **408**:535-6.

34. Walsh MC, Kim N, Kadono Y, Rho J, Lee SY, Lorenzo J, Choi Y: **Osteoimmunology: interplay between the immune system and bone metabolism**. *Annu Rev Immunol* 2006, **24**:33-63.

35. Arai F, Ohneda O, Miyamoto T, Zhang XQ, Suda T: **Mesenchymal stem cells in**

perichondrium express activated leukocyte cell adhesion molecule and participate in bone marrow formation. *J Exp Med* 2002, **195**:1549-63.

36. Bennett KP, Bergeron C, Acar E, Klees RF, Vandenberg SL, Yener B, Plopper GE: **Proteomics reveals multiple routes to the osteogenic phenotype in mesenchymal stem cells**. *BMC Genomics* 2007, **8**:380.

37. Ebert R, Jakob F: **Selenium deficiency as a putative risk factor for osteoporosis**. *International Congress Series: Nutritional Aspects of Osteoporosis 2006. Proceedings of the 6th International Symposium on Nutritional Aspects of Osteoporosis, 4-6 May 2006, Lausanne, Switzerland* 2007, **1297**:158-164.

38. Curry CL, Reed LL, Golde TE, Miele L, Nickoloff BJ, Foreman KE: **Gamma secretase inhibitor blocks Notch activation and induces apoptosis in Kaposi's sarcoma tumor cells**. *Oncogene* 2005, **24**:6333-44.

39. Kang JH, Lee DH, Lee JS, Kim HJ, Shin JW, Lee YH, Lee YS, Park CS, Chung IY: **Eosinophilic differentiation is promoted by blockage of Notch signaling with a gamma-secretase inhibitor**. *Eur J Immunol* 2005, **35**:2982-90.

40. Kanei-Ishii C, Ninomiya-Tsuji J, Tanikawa J, Nomura T, Ishitani T, Kishida S, Kokura K, Kurahashi T, Ichikawa-Iwata E, Kim Y, et al: **Wnt-1 signal induces phosphorylation and degradation of c-Myb protein via TAK1, HIPK2, and NLK**. *Genes Dev* 2004, **18**:816-29.

41. Nair M, Bilanchone V, Ortt K, Sinha S, Dai X: **Ovol1 represses its own transcription by competing with transcription activator c-Myb and by recruiting histone deacetylase activity**. *Nucleic Acids Res* 2007, **35**:1687-97.

42. Garcia-Pedrero JM, Kiskinis E, Parker MG, Belandia B: **The SWI/SNF chromatin remodeling subunit BAF57 is a critical regulator of estrogen receptor function in breast cancer cells**. *J Biol Chem* 2006, **281**:22656-64.

43. Villagra A, Cruzat F, Carvallo L, Paredes R, Olate J, van Wijnen AJ, Stein GS, Lian JB, Stein JL, Imbalzano AN, et al: **Chromatin remodeling and transcriptional activity of the bone-specific osteocalcin gene require CCAAT/enhancer-binding protein beta-dependent recruitment of SWI/SNF activity**. *J Biol Chem* 2006, **281**:22695-706.

44. Colland F, Jacq X, Trouplin V, Mougin C, Groizeleau C, Hamburger A, Meil A, Wojcik J, Legrain P, Gauthier JM: **Functional proteomics mapping of a human signaling pathway**. *Genome Res* 2004, **14**:1324-32.

45. Simone C, Forcales SV, Hill DA, Imbalzano AN, Latella L, Puri PL: **p38 pathway targets SWI-SNF chromatin-remodeling complex to muscle-specific loci**. *Nat Genet* 2004, **36**:738-43.

46. Lutwyche JK, Keough RA, Hunter J, Coles LS, Gonda TJ: **DNA**

binding-independent transcriptional activation of the vascular endothelial growth factor gene (VEGF) by the Myb oncoprotein. *Biochem Biophys Res Commun* 2006, **344**:1300-7.

47. Itoh F, Itoh S, Goumans MJ, Valdimarsdottir G, Iso T, Dotto GP, Hamamori Y, Kedes L, Kato M, ten Dijke Pt P: **Synergy and antagonism between Notch and BMP receptor signaling pathways in endothelial cells**. *Embo J* 2004, **23**:541-51.

48. Dahlqvist C, Blokzijl A, Chapman G, Falk A, Dannaeus K, Ibanez CF, Lendahl U: **Functional Notch signaling is required for BMP4-induced inhibition of myogenic differentiation**. *Development* 2003, **130**:6089-99.

49. Takizawa T, Ochiai W, Nakashima K, Taga T: **Enhanced gene activation by Notch and BMP signaling cross-talk**. *Nucleic Acids Res* 2003, **31**:5723-31.

50. Deregowski V, Gazzerro E, Priest L, Rydziel S, Canalis E: **Notch 1 overexpression inhibits osteoblastogenesis by suppressing Wnt/beta-catenin but not bone morphogenetic protein signaling**. *J Biol Chem* 2006, **281**:6203-10.

51. Leroith D, Nissley P: **Knock your SOCS off!** *J Clin Invest* 2005, **115**:233-6.

52. **Gene Ontology** [www.geneontology.org/]

53. **The Molecular Signature Database** [http://www.broad.mit.edu/GSEA/msigdb/index.jsp]

Chapter V

Conclusions and future work

This work focuses on gene regulatory network reconstruction and pathway inference from High Throughput Biological Data. The data we used are high throughput mRNA expression data. But the methods are generally applicable to other types of quantitative biological data, for instance protein expression data, or even non-biological data.

In chapter 2, I developed a new GRN reconstruction strategy, MI3 that addresses three major issues simultaneously: (1) to handle continuous variables, (2) to detect high order relationships, (3) to differentiate causal vs. confounding relationships. MI3 consistently and significantly outperformed frequently used control methods and faithfully capture mechanistic relationships from gene expression data. I used MI3 to infer a regulatory network centered at the MYC transcription factor from a published microarray dataset. This MYC centered GRN not only include MYC target genes but also corresponding MYC cofactors. This network reveals that MYC regulates the transcription of a large number of target genes through limited number of mechanisms, which detail agrees with experimental evidences from literature. This is the first time any comparably complex and realistic gene regulatory network is constructed on mammalian systems from microarray data alone.

In this work, I only applied MI3 to learn static GRNs. When we have time series data, the exact same method can learn dynamic GRN the same way as dynamic Bayesian network. The high order mutual information framework resented here is generally applicable, although I have only described and used three-way mutual information. The same set of

strategies can be used to model arbitrarily high order relationships given enough data.

In chapter 3, I present another novel method GAGE for pathway inference. GAGE is generally applicable to gene expression data sets with different sample sizes and experimental designs. GAGE consistently outperformed two most frequently used GSA methods and inferred statistically and biologically more relevant regulatory pathways.

GAGE reveals novel and relevant regulatory mechanisms from both published and previously unpublished microarray studies. From two published lung cancer data sets, GAGE derived a more cohesive and predictive mechanistic scheme underlying lung cancer progress and metastasis. For a previously unpublished BMP6 study, GAGE predicted novel yet biologically plausible regulatory mechanisms for BMP6 induced osteoblast differentiation.

One major strategy employed by GAGE is differential treatment of different types of gene sets, i.e. allows expression perturbation towards both directions in a canonical pathway yet only in one direction in a experimental sets. Further improvement can be made by special treatment of canonical pathways or functional groups. For instance, genes can be assign different weights based on their roles and positions in a pathway, high weight for hubs genes or genes with critical and indispensable roles, low weight to less important or redundant genes. Similarly, other pathway specific information like network topology can also be accounted for in gene set analysis.

Pathway inference methods can be improved by adopting ideas from GRN reconstruction, since GRN reconstruction and pathway inference are two related problems (detailed in Chapter 1). While topology is the major focus of network reconstruction, pathway inference commonly does not consider the pathway or network topology (Figure 1.3c). It has been proposed that topology and other pathway specific information would enable more sensitive and specific pathway inference. Pathway inference commonly identifies pathways differentially expressed between two discrete phenotypes. In GRN reconstruction, genes are treated as continuous or multi-state discrete variables. Similar

procedures dealing with continuous or multiple phenotypes would make pathway inference more generally applicable and more powerful. Feature selection takes features (either genes or pathways) as independent variables. GRN suggests that it is frequently an oversimplified assumption, since individual regulators may have very low dependency with the target yet the whole regulator set together predict the target very well. Similarly, procedures without this independent assumption would capture more realistic high order interaction or the combinatory effects between features, hence allow identifying the real causal genes or pathways more efficiently.

In chapter 4, I present the first high throughput microarray study on BMP6 induced transcriptional program in human MSC. It covers the whole process from early to late stage osteoblast differentiation and mineralization. I conducted a comprehensive pathway inference using GAGE method to identify relevant regulatory mechanisms and functional groups. I inferred a series of significant KEGG pathways, GO terms and experimental sets at different stages of BMP6 induction process. I not only showed which pathways or gene sets are significant, but also when and how they are involved in the osteoblast differentiation and mineralization. Different from common pathway analyses, our work further captures the interconnections among individual pathways or functional groups and integrate them into a whole system. Taken together, this work provides clearer mechanistic picture of osteoblast differentiation and function. I followed a systems biology study procedure in this work: rational experimental design, whole-system or genome-wide expression profiling, high throughput data process and analysis, results interpretation and experimental validation at pathway and system level, and further high throughput experiment and analysis design. This study combines experimental and computational work seamlessly and reaches novel and robust conclusions.

To better define the transcriptional programs involved in BMP6 induced osteoblast differentiation and mineralization, we can expand the microarray dataset with more time points and shorter intervals. More time points with smaller interval at early stage (<8

hours) can tell the temporal order of the KEGG pathways directly in greater detail. More time points not only allow better timing of predefined pathways/gene sets, but also better differentiate waves of undefined transcriptional program with individual genes. We may better define the phenotypic stages from MSC to fully mature osteoblast, identify novel and informative marker genes or transcriptional events. With more time points, we may also compare the transcriptional programs triggered by BMP6 to those by BMP2 along temporal axis, which would allow us to trace the specificities of different osteogenic BMPs.

In this work, I apply GAGE to an unevenly distributed short time series (a few time points) dataset for a temporal pathway inference. The success in such demanding analysis highlights two advantages of GAGE: (1) applicable to time series datasets with small sample size at each time point or condition. (2) both sensitive and selective to capture subtle yet real regulatory signals over time. GAGE was not designed for time series data analysis but rather applicable for robust two-state comparison at each time slice due to these two advantages. For evenly distributed time-series data, GAGE can be revised to test whether gene sets are significantly correlated with the phenotype, which is taken as a continuously changing variable similar to the gene expression levels then (unlike the discrete phenotype in Table 1.1). To our knowledge, there is no method dedicated to such time series pathway inference, nor method for pathway inference other than GAGE sharing above-mentioned two advantages.

With adequate evenly distributed time-series data, other advanced statistical learning such as time series analysis and dynamic mutual information networks or dynamic Bayesian networks would be plausible too, either for GRN reconstruction or temporal pathway inference when phenotype is taken as a continuous node.