A CAUSAL GENE NETWORK WITH GENETIC VARIATIONS INCORPORATING BIOLOGICAL KNOWLEDGE AND LATENT VARIABLES

By

Jee Young Moon

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2013

Date of final oral examination: 12/21/2012

The dissertation is approved by the following members of the Final Oral Committee:

Brian S. Yandell. Professor, Statistics, Horticulture

Alan D. Attie. Professor, Biochemistry

Karl W. Broman. Professor, Biostatistics and Medical Informatics

Christina Kendziorski. Associate Professor, Biostatistics and Medical Informatics

Sushmita Roy. Assistant Professor, Biostatistics and Medical Informatics, Computer Science, Systems Biology in Wisconsin Institute of Discovery (WID)

To my parents and brother,

ACKNOWLEDGMENTS

I greatly appreciate my adviser, Prof. Brian S. Yandell, who has always encouraged, inspired and supported me. I am grateful to him for introducing me to the exciting research areas of statistical genetics and causal gene network analysis. He also allowed me to explore various statistical and biological problems on my own and guided me to see the problems in a bigger picture. Most importantly, he waited patiently as I progressed at my own pace. I would also like to thank Dr. Elias Chaibub Neto and Prof. Xinwei Deng who my adviser arranged for me to work together. These three improved my rigorous writing and thinking a lot when we prepared the second chapter of this dissertation for publication.

It was such a nice opportunity for me to join the group of Prof. Alan D. Attie, Dr. Mark P. Keller, Prof. Karl W. Broman and Prof. Christina Kendziorski. I appreciate Prof. Attie and Dr. Keller for sharing their enormous dataset from experimental cross studies and insightful biological opinions. I also learned a lot by observing how Profs. Broman and Kendziorski interacted with Prof. Attie and Dr. Keller. It was valuable experience for me to see how they formulated statistical problems, explained their statistical models to biologists and made sure the collaboration proceeded.

I feel grateful to my dear friends in Madison and in Korea. I was lucky to spend valuable time with my friends in Madison. They were great company and we laughed, relaxed and helped each other mentally and academically. I especially thank Kiyoung Park whom I met again in Madison after high school and it became regular for us to go to the gym together and talk about the excitements of science. She gave me the most honest, realistic and sincere advice. I am also greatly thankful to Yujin Chung who spent lots of time with me discussing our research, futures and even difficulties we encountered. I relied a lot on our conversations to make wise decisions. I'd like

to express my sincere gratitude to Jungwon Mun, Hyuna Yang, Jihoon Kim, Lisa Mijung Chung, Youngdeok Hwang, Dongjun Chung, Jee Yeon Kim, Heejung Shim, Seho Park and the TV show Infinite Challenge. In addition, I owe a lot to my friends in Korea — Jiyoung Kim, Sunhee Lee, Yoojung Lee, Myungshin Lee and class of 3-6. I am thankful to them for keeping our friendship constant even though I could only see them for a short time each year.

Most importantly, I cannot thank my parents more, Hanjoong Moon and Hyunsook Park, and my brother, Seounghyun Moon, who always believed in me and gave me unconditional love and support in every possible way. They were one big influence that kept me moving forward. Also, I'd like to thank my relatives who warmly welcomed me whenever I visited them. I happily dedicate this dissertation to my family. Love you!

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES vii

ABSTRACT x

1 Introduction 1

1.1 Biological networks 2
1.2 Causal networks 4
1.3 Experimental cross study 6
1.4 A causal gene network on experimental cross study 12
1.5 Motivation of my work 15
1.6 Contribution of this dissertation 16

2 Bayesian causal phenotype network incorporating genetic variation and biological knowledge 18

2.1 Introduction 18
2.2 Joint inference of causal phenotype network and causal QTLs 21
2.2.1 Standard Bayesian network model 21
2.2.2 HCGR model 22
2.2.3 Systems genetics and causal inference 24
2.2.4 QTL mapping conditional on phenotype network structure 28
2.2.5 Joint inference of phenotype network and causal QTLs 31
2.3 Causal phenotype network incorporating biological knowledge 32
2.3.1 Model 32
2.3.2 Sketch of MCMC 37
2.3.3 Summary of encoding of biological knowledge 40
2.4 Simulations 43
2.5 Analysis of yeast cell cycle genes 48
2.6 Conclusion 53

Page

3 Causal network incorporating genetic variations with latent phenotypes 57

3.1 Motivation 57
3.2 Ancestral graph 59
3.3 Ancestral graph for phenotypes and genotypes 66
3.3.1 Model 67
3.3.2 Graphical properties of extended network 70
3.3.3 The parametric family of the extended network 74
3.3.4 Conjecture: Distribution equivalence is the same as the Markov equivalence for Gaussian family and HCG family 76
3.4 Algorithms for ancestral graph inference of phenotypes and QTLs 84
3.4.1 Score-based search – MCMC 84
3.4.2 Summarizing MCMC samples 88
3.5 Simulation 88
3.6 Real data analysis 90
3.7 Conclusion 96

4 Summary 99

4.1 Summary of my work 99
4.2 Future work 100

APPENDIX Additional Figures 103

A.1 Inferred yeast cell cycle network with causal QTLs integrating TF information by QTLnet-prior 103
A.2 Convergence diagnostics of yeast cell cycle network 105

LIST OF TABLES

Table Page

1.1 Configuration of genotypes of pseudomarker p on a gamete conditional on markers 1 and 2. Let $r = r_1 + r_2 - 2r_1r_2$. 10

1.2 Development of models using genotypes and phenotypes in an experimental cross study 15

2.1 Models G^1_Y and G^3_Y are distribution/likelihood equivalent. 27

2.2 Extended models G^1 and G^3 are no longer distribution/likelihood equivalent. 27

2.3 Four methods which differ in the use of genetic variation information and biological knowledge. 44

3.1 Decomposition of ancestral graph G of genotypes and phenotypes into $G_Q, G_{Q\to Y}$ and G_Y. Nodes and edges consisting of each subgraph are summarized and the corresponding values in the statistical model are summarized. 69

3.2 A DMAG that the reparameterization complies with the transformed DMAG by a legitimate mark change from $t \to v$ to $t \leftrightarrow v$. 83

3.3 DMAGs violating assumptions in a legitimate mark change 84

LIST OF FIGURES

Figure Page

1.1 An example DAG 5

1.2 Experimental cross 8

1.3 Causal gene network decompositions 14

2.1 Example network with five phenotypes ($Y_1, \ldots, Y_5$) and four QTLs ($Q_1, \ldots, Q_4$). . . . 23

2.2 Output of the unconditional QTL mapping analysis for the phenotypes in Fig. 2.1. Dashed and pointed arrows represent direct and indirect QTL/phenotype causal relationships, respectively. 28

2.3 QTL mapping tailored to the network structure. Dashed, pointed and wiggled arrows represent, respectively, direct, indirect and incorrect QTL/phenotype causal relationships. (a) Mapping analysis of Y_5 conditional on Y_3 and Y_4 still detects Q_1 and Q_2 as QTLs for Y_5, since failing to condition on Y_2 leaves the paths $Q_1 \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_5$ and $Q_2 \rightarrow Y_2 \rightarrow Y_5$ in Fig. 2.1 open. In other words, (Y_3, Y_4) cannot d-separate (Q_1, Q_2) from Y_5 in the true causal graph. (b) Mapping analysis of Y_4 conditional on Y_1, Y_3 and Y_5 incorrectly detects Q_5 as a QTL for Y_4 because in the true network the paths $Y_4 \rightarrow Y_5 \leftarrow Q_5$ and $Y_4 \leftarrow Y_3 \rightarrow Y_5 \leftarrow Q_5$ in Fig. 2.1 are open when we condition on Y_5. 30

2.4 The comparison of four methods by the area under the ROC curves with respect to the accuracy of biological knowledge. 46

2.5 Precision-recall curves show the trade-off between the probability that an inferred edge is the true edge and the probability that a true edge is included. As the biological knowledge gets positively informative, the precision-recall curve moves to the upper right which means good performance. The performance is better when the network is estimated by QTLnet-prior compared to the network by WH-prior. 47

2.6 The distribution of median weight W of posterior sample by QTLnet-prior inference. Each panel shows the median W distribution when biological knowledge is defective ($\delta = -0.1$), non informative ($\delta = 0$), and informative ($\delta = 0.1$). 49

Figure Page

2.7 Yeast cell cycle phenotype network by QTLnet-prior, integrating transcription factor binding information. A solid edge is the inferred edge with its posterior probability over 0.5 and the darkness of the edge is in proportion to the posterior probability. Dark nodes are transcription factors. The edge consistent with transcription factor binding information is marked with a star, $*$. The TF binding relation recovered by an indirect path in the inferred network is represented by a dashed edge. 51

2.8 The posterior distribution of weight W of transcription factor information in reconstructing a yeast cell cycle network by QTLnet-prior. 52

2.9 Comparison of the posterior probability of every possible directed edge between the network inferred by QTLnet-prior and the network inferred by QTLnet. 52

3.1 Schematic view of an ancestral graph. Reproduced from from Fig. 4 in Richardson and Spirtes [2002]. 61

3.2 Schematic view of an ancestral graph of phenotypes Y and pseudomarkers Q. 69

3.3 A network for simulation 89

3.4 A network after marginalizing c out 89

3.5 Frequency of detection of each edge in the true skeleton 90

3.6 Correlation between Y_1 and Y_2 and correlation between Y_2 and Y_4. 91

3.7 Frequency of edge types. Thick arrows correspond to true directions. 91

3.8 497 genes out of 593 genes are connected to each other in the estimated undirected graph by glasso. The gene of interest (Nfatc2) is numbered to be 428 in red and directly connected genes are colored in blue. 93

3.9 6 genes are directly connected to Nfatc2 94

3.10 104 genes are connected to Nfatc2 on the top in red by at most 2 steps. 6 genes that are directly connected to Nfatc2 is in blue. 95

3.11 A causal network of Nfatc2, Iqsec1 and Pcnt allowing latent variables. chrK@x are identified QTLs at x cM in chromosome K conditional on the phenotype network. . . . 96

3.12 Scatter plots of gene expression after adjusting for sex, batch and QTL effects 97

Appendix
Figure Page

A.1 Yeast cell cycle network integrating transcription factor binding information inferred by QTLnet-prior. The edge darkness is in proportion to the posterior probability. . . . 104

A.2 The top two figures are the trace plot and the autocorrelation plot of BIC scores for sampled causal networks. The bottom two figures are the trace and the autocorrelation plots of the sampled weights (W) on transcription factor binding information. 106

A CAUSAL GENE NETWORK WITH GENETIC VARIATIONS INCORPORATING BIOLOGICAL KNOWLEDGE AND LATENT VARIABLES

Jee Young Moon

Under the supervision of Professor Brian S. Yandell

At the University of Wisconsin-Madison

A Bayesian network has often been modeled to infer a gene regulatory network from expression data. Genotypes along with gene expression can further reveal the regulatory relations and genetic architectures. Biological knowledge can also be incorporated to improve the reconstruction of a gene network. We propose a Bayesian framework to jointly infer a gene network and weights of prior knowledge by integrating expression data, genetic variations, and prior biological knowledge. The proposed method encodes biological knowledge such as transcription factor and DNA binding, gene ontology annotation, and protein-protein interaction into a prior distribution of the network structures. A simulation study shows that the incorporation of genetic variation information and biological knowledge improves the reconstruction of gene network as long as biological knowledge is consistent with expression data.

There are some assumptions that the Bayesian network inference is making. One assumption is that all variables are included, in other words, there are no latent variables. An ancestral graph properly represents the conditional independence relations on observed variables when there are latent variables in an underlying Bayesian network. Without needing to introduce latent variables, we modeled a set of recursive equations for an ancestral graph of phenotypes extended with genetic variations. We showed that 1) its parametric family is a homogeneous conditional Gaussian (HCG) family and 2) the extended graph satisfies the ancestral graph conditions.

Brian S. Yandell

ABSTRACT

A Bayesian network has often been modeled to infer a gene regulatory network from expression data. Genotypes along with gene expression can further reveal the regulatory relations and genetic architectures. Biological knowledge can also be incorporated to improve the reconstruction of a gene network. We propose a Bayesian framework to jointly infer a gene network and weights of prior knowledge by integrating expression data, genetic variations, and prior biological knowledge. The proposed method encodes biological knowledge such as transcription factor and DNA binding, gene ontology annotation, and protein-protein interaction into a prior distribution of the network structures. A simulation study shows that the incorporation of genetic variation information and biological knowledge improves the reconstruction of gene network as long as biological knowledge is consistent with expression data.

There are some assumptions that the Bayesian network inference is making. One assumption is that all variables are included, in other words, there are no latent variables. An ancestral graph properly represents the conditional independence relations on observed variables when there are latent variables in an underlying Bayesian network. Without needing to introduce latent variables, we modeled a set of recursive equations for an ancestral graph of phenotypes extended with genetic variations. We showed that 1) its parametric family is a homogeneous conditional Gaussian (HCG) family and 2) the extended graph satisfies the ancestral graph conditions.

Chapter 1

Introduction

One of the ultimate goals of biology is to understand how genes, proteins, metabolites and other molecules regulate each other to maintain, activate or inhibit biological processes. With the development of various types of high-throughput experiments and the accumulation of experimental data, different types of evidence for regulations and interactions could be collected on a large-scale. The large-scale collection could yield us a balanced viewpoint on how they are regulating each other and could be used to reconstruct large-scale networks.

We can classify the inferred networks from collection of data largely into two classes. One class is a physical network, which is produced by experiments designed to detect the very physical relations such as transcription factor binding, protein-protein interaction and chemical reactions. The other class is a statistical model for biological networks, which is statistically estimated from mostly gene expression data to connotatively reveal the connections among genes. Since gene expression is the most frequently available data so far, it is often assumed to represent the activity of the corresponding gene and protein and statistical methods are applied to identify the relationships. This is why the connections are connotative. For example, a high correlation of gene expression in a coexpression network does not necessarily mean that they are related, however, it suggests that they could be related and generates hypotheses on whether they are regulated by common regulators and which gene is a candidate master regulator. Besides coexpression networks, causal networks can be statistically inferred based on Bayesian network models. The causal networks are more meaningful than coexpression networks because they infer causal relationships, which are the main interest to understand biological processes. In this dissertation, I focus on causal gene network inference, which can generate compelling hypotheses on gene regulations. Before fully

developing causal networks and the system of experiments that I consider to infer causal gene networks, I will briefly go over several biological networks mentioned above in Section 1.1.

1.1 Biological networks

Biological networks can be classified into two groups by whether the network is constructed with physical relationships or not. Physical networks include protein-protein interaction networks, metabolic networks and transcription networks. The other class of networks, statistical networks, include coexpression networks and causal networks.

A protein-protein interaction network shows the direct physical contact between proteins [De Las Rivas and Fontanillo, 2010, Stelzl et al., 2005, Rual et al., 2005]. Two types of interaction data are often produced — one is pairwise direct interaction from yeast two hybrid (Y2H) experiments [Suter et al., 2008, Stelzl et al., 2005] and the other is direct or indirect interaction by being in the same protein complex from tandem affinity purification with mass spectrometry (TAP-MS) [Bauer and Kuster, 2003, Gavin et al., 2011]. The latter type of interactions is processed to distinguish direct interactions from indirect interactions by using methods such as spoke model and matrix model [Gavin et al., 2011]. The evidence of interactions is reposited in databases such as BioGRID (thebiogrid.org), MIPS (mips.helmholtz-muenchen.de/proj/ppi/), BIND [Bader et al., 2003] and DIP (dip.doe-mbi.ucla.edu/dip).

A metabolic network consists of metabolites and its biochemical reactions catalyzed by enzymes [Junker and Schreiber, 2008]. Enzymes mostly made of proteins take substrates and accelerate the conversion into products by lowering activation energy of the reaction. The reactions catalyzed by enzymes are stored in the reaction databases such as Brenda (www.brenda-enzymes.org), ENZYME (expasy.org/enzyme) and LIGAND (www.genome.jp/ligand). With the sequencing of a genome, a metabolic network of a specific species can be reconstructed [Francke et al., 2005]. A gene can be annotated with predicted molecular functions by finding genes in other organisms with similar sequences. The catalytic function of the gene represented by enzyme number can then be searched for its reactions in the reaction databases. These procedures reconstruct an automated metabolic network of a species and it is further curated for better accuracy. KEGG

(www.genome.ad.jp/kegg) and BioCyc (www.biocyc.org) are examples of databases of metabolic networks.

A transcription network is a network of transcription factors, its targets and targets of target transcription factors [Lee et al., 2002, Neph et al., 2012]. Transcription factors have a special property that it binds to a specific DNA sequence. Chromatin-immunoprecipitation (ChIP) experiments can identify the pair of transcription factor and its target genes by sequencing the DNA regions bound by a transcription factor where the complex of DNA and transcription factor is captured by the antibodies to the transcription factor. Lee et al. [2002] constructed a transcriptional regulatory network by assembling transcription factor binding configurations in addition to coexpression of genes. DNase I footprinting provides information about the accessibility of chromatins, and most transcription factors can bind to a DNA when the chromatin is open (accessible). DNAse I footprinting and motif of transcription factors can be used to construct a transcription factor network [Neph et al., 2012].

As a statistical model for biological networks, a coexpression network shows co-expressed genes in expression data and identifies module structures and hub genes [Ruan et al., 2010, Allen et al., 2012, Segal et al., 2003]. Genes in the same module could also share functional similarities, such as gene ontology terms. Hub genes are genes that are connected to many other genes and might be master regulators. Thus, the lack of a properly functioning hub gene may have a significant effect on the system. The connections in a coexpression network can be inferred from pairwise correlations or mutual information [Stuart et al., 2003, Zhang and Horvath, 2005], or partial correlations to distinguish direct and indirect connections [Schäfer and Strimmer, 2005, Meinshausen and Bühlmann, 2006, Peng et al., 2009].

Another statistical model for biological networks is a causal network. A causal network infers the causal relations, often using Bayesian networks. Bayesian networks consist of nodes and directed edges without directed cycles, and the directed edges could imply causal relations. Gene expression data at a steady state is used to infer the structure of a Bayesian network [Friedman et al., 2000, Pe'er et al., 2001, Hartemink et al., 2001]. Time-series gene expression measurements can easily infer directed cycles such as feedback loops in biological pathways because time moves

in one direction. It is often modeled by dynamic Bayesian networks [Murphy and Mian, 1999, Friedman et al., 1998, Husmeier, 2003, Geier et al., 2007, Grzegorczyk and Husmeier, 2011] or ordinary differential equations [de Jong, 2002, Gardner et al., 2003, Bansal et al., 2006]. Logsdon and Mezey [2010] proposed a method to infer directed cyclic networks with pre-defined cis-acting genotypes.

The physical networks and statistical networks described above would not necessarily share common features between each other because each network characterizes different gene activities – physical interactions and statistical associations — and gene products are in different states — proteins and mRNAs [Penfold and Wild, 2011]. For example, mRNA levels are not necessarily correlated with protein levels or activities [Rogers et al., 2008] even though the mRNA level tells a lot about the biology and could be assumed to represent protein levels. By integrating various evidence of physical interactions with statistical networks, a more comprehensive network could be reconstructed.

In the next Section 1.2, I will now elucidate causal networks, which are the main interest of the dissertation.

1.2 Causal networks

When we think of an underlying genetic mechanism that probably would be composed of activations and inhibitions of gene activities, it is natural to think that the relationships between genes are generated by causal relations, not associations. In statistics, it is regarded that a causal effect can be figured out only when in an experiment there is a manipulation or a temporal ordering, or the experiment is a randomized design such that individuals are assigned to treatments randomly [Pearl, 2000]. Correlations between variables or coefficients in a linear regression generally show associations, not causal relations. However, Pearl [2000] presented that some causal relations can be identified by examining the patterns of conditional independencies. For example, X and Z are dependent, Y and Z are dependent, and X and Y are independent. Then, the possible causal structure is that $X \rightarrow Z \leftarrow Y$ where $\rightarrow$ means a causal relation. Therefore, by examining conditional independencies, we can find the causal structure that explains how the data is generated. Most

importantly, finding the causal structure is more meaningful and closer to the scientific goals than figuring out associations.

One popular class of graphs to represent causal structures is a directed acyclic graph (DAG), which consists of nodes and at most one directed edge between two nodes and contains no directed cycles. The corresponding probabilistic model for a DAG is called a Bayesian network. A directed edge in a DAG could practically mean a causal relation. What a DAG actually represents is a set of conditional independence relations between nodes. A characteristic Markov property of a DAG regarding conditional independencies is that conditional on parent nodes, a node is independent of other nodes except for descendants. By this Markov property, the joint distribution of a Bayesian network can be factored by probabilities on parent and child relationships. For example in Figure 1.1, let Y_i represent the gene expression level of gene i. Then, the joint distribution can be written to be

$$P(Y_1, Y_2, Y_3, Y_4) = P(Y_4|Y_3)P(Y_3|Y_1, Y_2)P(Y_2)P(Y_1).$$

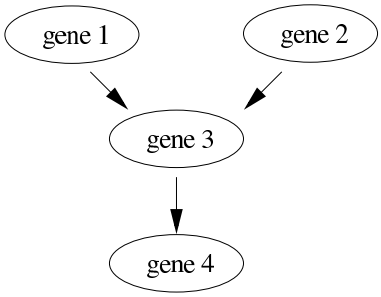


Figure 1.1: An example DAG

If the joint distribution is assumed to be Gaussian, each factored probability can be modeled by a linear regression and hence, the joint distribution can be represented by a set of linear regressions.

In the example, it can be written to be

$$
\begin{aligned}
Y_4 &= \mu_4 + \beta_{43} Y_3 + \epsilon_4, \quad \epsilon_4 \sim N(0, \sigma_4^2) \\
Y_3 &= \mu_3 + \beta_{31} Y_1 + \beta_{32} Y_2 + \epsilon_3, \quad \epsilon_3 \sim N(0, \sigma_3^2) \\
Y_2 &= \mu_2 + \epsilon_2, \quad \epsilon_2 \sim N(0, \sigma_2^2) \\
Y_1 &= \mu_1 + \epsilon_1, \quad \epsilon_1 \sim N(0, \sigma_1^2),
\end{aligned}
$$

where β_{tv} is a causal effect of Y_v on Y_t.

There are three approaches to learn a Bayesian network — examining conditional independence relations [Spirtes et al., 2000], scoring structures [Cooper and Herskovits, 1992], or hybridizing the two methods [Tsamardinos et al., 2006]. The first approach tests the independence between two variables conditional on a subset of the remaining variables, removes the edge if they are independent, and configures the orientation from the set of conditional independence relations [Spirtes et al., 2000]. The second approach searches the space of networks, scores the networks, and finds the network of the maximum score [Cooper and Herskovits, 1992]. The third approach learns the skeleton of the network by conditional independence relations and scores the network by restricting a set of parents to be connected to the variable [Tsamardinos et al., 2006].

More details of Bayesian networks will be further explained in Section 2.2.1. The system I consider to reconstruct a causal gene network is from an experimental cross study with genotypes and gene expression data. Genotypes on gene expression have a special feature that is advantageous to infer causal relations. First of all, I will describe an experimental cross study and classical analysis to associate genotypes and expression of a single gene in the next Section 1.3. Later in Section 1.4 I will expand to the case when there are multiple genes in experimental cross study and present the model of a causal gene network of genotypes and gene expressions.

1.3 Experimental cross study

An experimental cross study is conducted to identify the association between genotypes and phenotypes. A population of experimental cross is generated by a series of crosses, beginning from two inbred lines, diagramed in Figure 1.2. An inbred line of an experimental organism such

as mice, rat, Arabidopsis and maize, is a sub-population of a specific species that is genetically identical to each other. The inbred line is made by many generations of inbreedings so that every pair of homologous chromosomes is homogeneous. Hence, every gamete it produces has exactly the same genetic backgrounds. Let's denote that the inbred line *A* has a homozygous genotype *AA*. If we cross line *A* and line *B*, the offspring will have genotypes *AB* where *A* lies on only one of homologous chromosomes, coming from the gamete of line *A*, and *B* lies on the other. Individuals in this F1 population have identical genetic backgrounds to each other. We can cross F1 individuals to get an F2 population. When an F1 population produces a gamete during meiosis, recombinations occur between homologous chromosomes by aligning homologous chromosomes together and crossing over them. Hence, a gamete can be a mosaic of *A* and *B* along a chromosome. By mating these gametes generated in the F1 population, the F2 population can have every combination of genotypes AA, AB, BB along the homologous chromosomes in the ratio of 1:2:1. Hence, the F2 population individuals have diverse genetic backgrounds. Another type of experimental cross is a backcross, which crosses F1 population with one of parental inbred lines. In a backcross, genotypes *AA* and *AB* occur in the ratio of 1:1.

Genetic variations can be inspected through restriction fragment length polymorphisms (RFLPs), repeat variations and single nucleotide polymorphisms (SNPs) [Schlötterer, 2004, Silver, 1995]. Restriction enzymes detect a specific DNA sequence and cleave the sequence. If there is a difference in the DNA sequence, there are differences in the length and number of fragmented sequences, and hence it is called a restriction fragment length polymorphism (RFLP). Repeat variations are the variations in the number of repeats of a short sequence. Microsatellites are the repeats of 1-10 base pairs (bp) and minisatellites are the repeats of 10-60 bp [Olson et al., 1989]. A SNP is a DNA location which is found to have a different nucleotide between individuals of the same species [Syvänen, 2001] and it is deposited in dbSNP [Sherry et al., 2001]. A SNP microarray is designed with probes in the array and the probes hybridize to fragments of DNA sequences differently according to SNP genotypes. Since a SNP has only one nucleotide difference, multiple probes are designed for each SNP and differential hybridization is used for calling a SNP genotype [Yang et al., 2009]. As of 2012, there are approximately 4 SNPs per 1Kbp, 4 RFLPs per 1Mbp, and

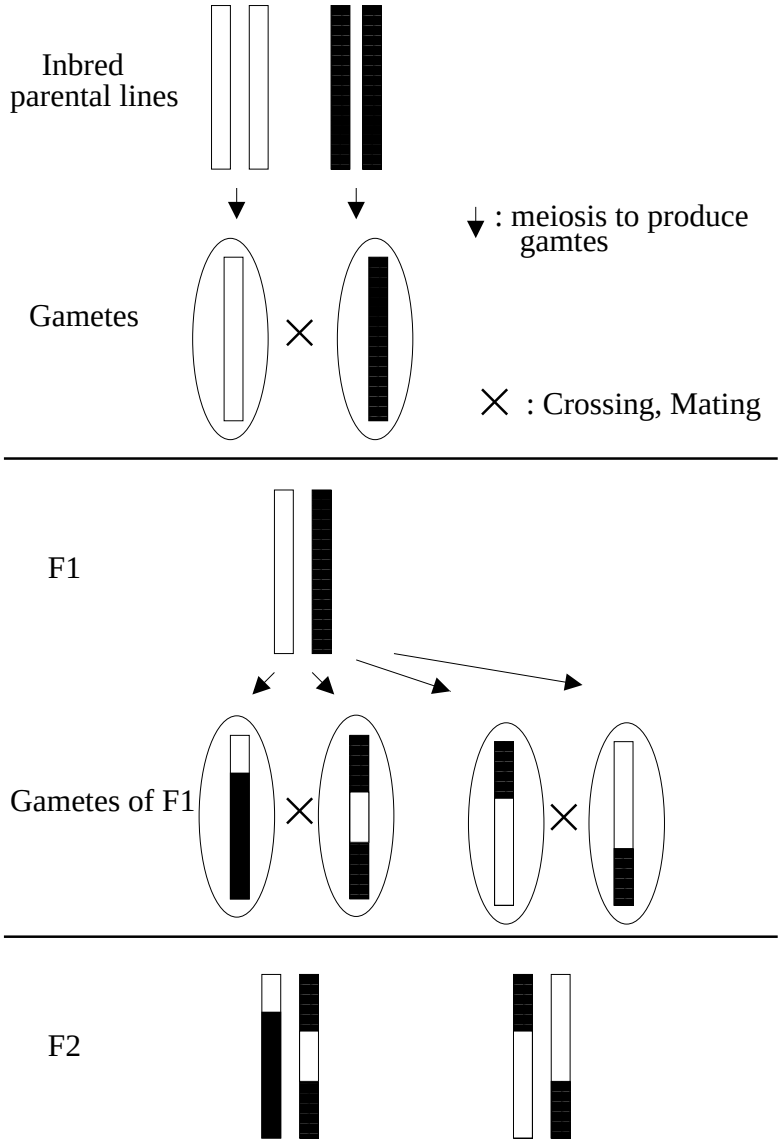


Figure 1.2: Experimental cross

0.6 microsatellites per 1Mbp retrieved from the Mouse Genome Database (MGD), Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine (www.informatics.jax.org) on Oct. 20, 2012 [Eppig et al., 2012]. A next generation sequencing is another good tool to identify SNP since it can read along a sequence for hundreds of bases by current techniques [Quail et al., 2012].

Gene expression can be measured in terms of mRNA levels by DNA microarrays or next generation sequencing techniques such as RNA-seq. Microarrays measure mRNA levels by hybridizing the mRNA sample to probes like cDNAs in a cDNA array, matched and mismatched probe sets of 25-mer oligonucleotides in an Affymetrix geneChip, or 60-mer oligonucleotides in an Agilent array. A sample of mRNAs is prepared by attaching fluorescent dyes to mRNAs and the sample is hybridized to the microarray. Then, the intensity of the fluorescence is measured and brighter signal means there is a higher level of mRNA targeted for the probe. The intensity data of microarray undergoes the normalization procedure to take care of dye labeling efficiency, background adjustment, scanning sensitivity, and so on [Irizarry et al., 2003, Zahurak et al., 2007]. RNA-seq can measure mRNA expression levels by the count of short mRNA fragments by high-throughout sequencing techniques [Wang et al., 2009, Mardis, 2008].

A quantitative trait loci (QTLs) mapping identifies genomic loci that influence phenotypes. A significant genomic location for a phenotype is called a QTL. The simplest model of QTL mapping is a linear model of a phenotype (Y) in respect to a single genetic marker at the location k, (X_k):

$$Y = \mu + \theta_k X_k + \epsilon, \tag{1.1}$$

where X_k is a coding variable for genotype at locus k in the genome, θ_k is the genetic effect of the genotype at locus k. In an intercross, there are three genotypes at a locus — AA, Aa, aa — and in a BC, there are two genotypes — AA, Aa. The effect of the change from a to A, called an additive effect, can be obtained as a coefficient in the model 1.1 after coding the genotype by the number of capital A. Then, AA is coded into 2, Aa is 1 and aa is 0. When the phenotype value of Aa is not the intermediate of AA and aa, there is a dominance effect that is the amount of deviation of Aa from the intermediate of AA and aa. In the case of complete dominance, Aa and AA would have the same phenotype values and we interpret that possession of one A is enough for the increases. Genotypes can be coded for dominance effects, in addition to coding variables for additive effects,

by *Aa* to be $1/2$ and *AA* and *aa* to be $-1/2$. The significance of the genotype on a phenotype is often noted by LOD (logarithm of the odds favoring linkage) score [Broman, 2001]. The LOD score is the logarithm of the likelihood ratio of model with QTL versus model without QTL.

A genomic region with sparse genetic markers can be enhanced with pseudomarkers to refine QTL locations. A genotype at a pseudomarker can be inferred from two flanking markers. Suppose a pseudomarker p is between marker 1 and 2. Let the distance between pseudomarker p and marker 1 be d_1 cM and between pseudomarker p and marker 2 be d_2 cM. One centimorgan (cM) is defined such that the expected average number of crossovers is 0.01. We observe two markers are recombined when there is an odd number of crossovers. Haldane map corresponds the centimorgan distance (d) and recombination frequency (r) in a form of $r = \frac{1-\exp(-2d/100)}{2}$ [Zhao and Speed, 1996]. Or, approximately, $r = d/100$ for small d. Let the recombination frequency corresponding to d_1 and d_2 to be r_1 and r_2, respectively. On a gamete, the probability of a genotype at a pseudomarker p is calculated for each combination of genotypes at marker 1 and marker 2 in Table 1.1.

Table 1.1: Configuration of genotypes of pseudomarker p on a gamete conditional on markers 1 and 2. Let $r = r_1 + r_2 - 2r_1r_2$.

marker 1	marker 2	pseudomarker p: A	pseudomarker p: a
A	A	$(1-r_1)(1-r_2)/(1-r)$	$r_1r_2/(1-r)$
A	a	$(1-r_1)r_2/(1-r)$	$r_1(1-r_2)/(1-r)$
a	A	$r_1(1-r_2)/(1-r)$	$(1-r_1)r_2/(1-r)$
a	a	$r_1r_2/(1-r)$	$(1-r_1)(1-r_2)/(1-r)$

In a backcross, the genotypes at a pseudomarker p on a pair of homologous chromosomes whether it is *AA* or *Aa* is just the conditional probability in Table 1.1 because the other chromosome is only one type like *A* with a probability 1. In an intercross, genotypes at pseudomarker p among AA, Aa, aa can be obtained by multiplying conditional probabilities.

With the calculated probabilities for pseudomarker's genotypes, LOD score can be calculated. Lander and Botstein [1989] presents that the likelihood of a model with a QTL at a pseudomarker

is

$$L(QTL) = \prod_i [p(AA)L_i(AA) + p(Aa)L_i(Aa) + p(aa)L_i(aa)],$$

where $L_i(g)$ is the likelihood when the genotype at the pseudomarker is g for an individual i, and $p(g)$ is the probability that the genotype is g inferred from flanking markers. The maximum likelihood can be obtained by expectation-maximization (EM) algorithm [Lander and Botstein, 1989]. Haley and Knott [1992] simplified the maximization of likelihood such that the QTL is located between two markers and the expected value of phenotype for each configuration of flanking markers will be the mean of theoretical values ($2*additive$ for AA, $additive+dominance$ for Aa and 0 for aa, assuming QTL at the pseudomarker) weighted by the frequency of the pseudomarker's genotype. The expected value will be expressed in terms of additive and dominance variables and recombination frequency, and additive and dominance effects can be obtained by fitting against the observed phenotype values. Sen and Churchill [2001] proposed a method for getting maximized likelihood through imputing genotypes at the pseudomarker. These various methods to get LOD is available at the R package QTL [Broman et al., 2003].

To identify a significant QTL, a p-value of LOD score can be used. A p-value can be obtained by permuting phenotype values and getting the distribution of LOD scores. Since multiple genomic locations are tested in QTL mapping, a p-value is adjusted by multiple testings. This is done by permuting phenotype value and obtaining the distribution of maximum LOD score across loci. 5% cutoff is often used to declare significant QTLs.

For better detection of QTLs influencing a phenotype, eqn 1.1 can be extended with multiple QTLs and epistasis (interaction of QTLs),

$$\begin{aligned} Y &= \mu + \sum_{k=1}^{K} \theta_k X_k + \sum_{k=1} \sum_{l \neq k} \theta_{kl}\, vec(X_k \otimes X_l) + \epsilon \\ &= \mu + \boldsymbol{\theta}\mathbf{X} + \epsilon, \end{aligned} \tag{1.2}$$

where X_k is a column vector of coding variables for additive and dominance and θ_k is a row vector of each effects. The term $vec(X_k \otimes X_l)$ is a set of coding variables for epistasis and θ_{kl} is a set of corresponding effects. The operator $\otimes$ is an outer product, which multiplies all combinations

of each coding variable and is defined to be that $X_k \otimes X_l = X_k X_l'$, where $'$ is the transpose. The operator $vec(\cdot)$ makes the matrix into a column vector by stacking columns. Simply, in eqn (1.2), $\mathbf{X}$ represents the design matrix of every genotype with all the coding variables for additive, dominance and interactions, and $\boldsymbol{\theta}$ represents the corresponding effects. Model selection criteria for multiple QTL have been proposed — simultaneous two QTLs in a model [Jansen and Stam, 1994], BIC score [Broman and Speed, 2002] and Bayesian approach [Yi, 2004]. Epistasis have been taken care of in these papers : Manichaikul et al. [2009], Yi et al. [2007].

The identified QTLs by QTL mapping methods described above affect the phenotype directly or indirectly. Indirect effects can happen when a QTL influences phenotype 1 and phenotype 1 influence phenotype 2. In this situation, the QTL of phenotype 1 could be identified as QTL of phenotype 2.

1.4 A causal gene network on experimental cross study

As genome-wide gene expression (phenotypes) can be measured in an experimental cross study, we want to know how genotypes regulate gene products and how gene products regulate each other. In this perspective, the reconstruction of a causal network of genotypes and phenotypes is desirable to elucidate how genetic variations can affect the cascade of phenotypes and, in the end, trigger some disease. An experimental cross study has a special feature which helps to infer more causal relations compared to expression data alone. Since a genotype affects phenotypes, not the other way around, we can restrict the directionality to be from genotype to phenotype. Also, since genetic recombination occurs randomly and experimental cross study is conducted in a controlled condition, the design of experimental cross is a randomized design. As a result, we can get a causal effect of a genotype to phenotypes without any confounding effects or biases.

If we know the causal network among phenotypes, we can distinguish whether the identified QTLs by QTL mapping has a direct effect on a phenotype or an indirect effect via other phenotypes. I will call QTLs directly affecting a phenotype as “causal QTLs”. As in the example mentioned in the last Section 1.3, let the true underlying pathway be that Q_1 is a causal QTL for phenotype 1 (Y_1) and phenotype 1 causally affects phenotype 2 (Y_2), that is, $Q_1 \rightarrow Y_2 \rightarrow Y_2$. Suppose we

know the causal relations between phenotypes, $Y_1 \rightarrow Y_2$. By QTL mapping, Q_1 will be identified to be a QTL for both Y_1 and Y_2. Taking in the causal relations on phenotypes ($Y_1 \rightarrow Y_2$), the QTL mapping on Y_2 after taking into account Y_1 will not identify Q_1 as a QTL for Y_2. Hence, the causal network among phenotypes can distinguish between direct causal QTL and indirect QTLs.

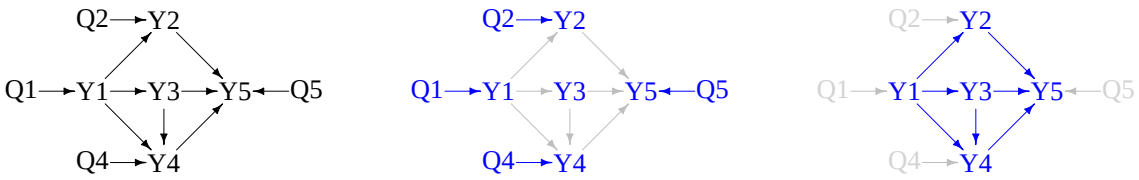
On the other hand, if we know causal QTLs, we could distinguish causal networks among phenotypes. For instance, Q_1 is known to be a causal QTL for Y_1 and Y_2 and Y_2 are correlated. It is possible that the causal relation on phenotypes is either $Y_1 \rightarrow Y_2$ or $Y_1 \leftarrow Y_2$. By incorporating causal QTL, the whole network is either $Q_1 \rightarrow Y_1 \rightarrow Y_2$ or $Q_1 \rightarrow Y_1 \leftarrow Y_2$. The first network ($Q_1 \rightarrow Y_1 \rightarrow Y_2$) will have this conditional independence relation: Q_1 and Y_2 are independent conditional on Y_1. The second network ($Q_1 \rightarrow Y_1 \leftarrow Y_2$) will have the relation: Q_1 and Y_2 are dependent conditional on Y_1. Hence, causal QTL Q_1 helps to distinguish two network structures on phenotypes.

Therefore, the joint inference of casual network among phenotypes (G_Y) and causal QTLs ($G_{Q \rightarrow Y}$) will be the main interest to decode the regulatory network of genotypes and phenotypes. As in Figure 1.3, I will denote the joint network of genotypes and phenotypes to be G and G is taken apart into a phenotype network (G_Y, causal network among phenotypes) and causal QTLs ($G_{Q \rightarrow Y}$). The joint inference of phenotype network (G_Y) and causal QTLs ($G_{Q \rightarrow Y}$) have been researched in an experimental cross study in these papers: gene network via structural equation modeling [Li et al., 2006], the joint inference of causal network on phenotypes and QTLs [Chaibub Neto et al., 2010a, Hageman et al., 2011].

The model of a casual network (G) of genotypes and phenotypes can be written by a set of linear equations. Each phenotype ($t = 1, \ldots, T$) can be modeled as follows [Chaibub Neto et al., 2010a]:

$$Y_t = \mu_t + \boldsymbol{\theta_t}\mathbf{X} + \sum_{v \in pa(t) \text{ in } G_Y} Y_v \beta_{tv} + \epsilon_t, \tag{1.3}$$

where $\boldsymbol{\theta_t}\mathbf{X}$ is the causal QTL effect in $G_{Q \rightarrow Y}$ and $\sum_{v \in pa(t) \text{ in } G_Y} Y_v \beta_{tv}$ is the parental phenotype effect in a phenotype network G_Y.



(a) G : Causal network of genotypes and phenotypes

(b) $G_{Q \to Y}$: QTL mapping given G_Y

(c) G_Y : Phenotype network given QTLs ($G_{Q \to Y}$)

Figure 1.3: Causal gene network decompositions

A genome wide association study (GWAS) for human has a similar data structure to an experimental cross study. It has genotype and phenotype data. However, there are challenges for the inference of causal network on GWAS data. First, because GWAS data is collected in uncontrolled conditions, there are unobserved non-genetic factors such as environments and they make other estimates to be confounded and biased. Second, as in association testing in GWAS, the population structure should be stratified for unbiased estimation.

1.5 Motivation of my work

An experimental cross study creates genotypes (X) and phenotypes (Y) data. QTL mapping introduced in Section 1.3 associates genotypes (X) and a phenotype (Y). With the availability of multiple phenotypes, a causal gene network in Section 1.4 finds causal relations among phenotypes in addition to causal QTLs for phenotypes. Table 1.2 conceptually compares these two models and two other models that I am going to develop in Chapters 2 and 3. In an effort to construct a more comprehensive regulatory network, I incorporate biological knowledge through a prior on phenotype networks on top of the causal gene network in Chapter 2. In Chapter 3, I consider the case when there are unmeasured variables and the causal network is extended to take account into the possibility of latent variables.

1. QTL mapping	$Y_t = \mu_t + \boldsymbol{\theta_t}\mathbf{X} + \epsilon_t$. A significant Q is called a QTL.
2. Causal network	$Y_t = \mu_t + \boldsymbol{\theta_t}\mathbf{X} + \sum_{v \in pa(t) \text{ in } G_Y} Y_v \beta_{tv} + \epsilon_t$.
3. Prior by knowledge	$P(G_Y)$ - Biological knowledge sets a prior on phenotype network structures (G_Y).
4. Causal network embedded with latent variables	$\epsilon \sim N(0, \Omega)$ Structured covariance of ϵ due to latent variables

Table 1.2: Development of models using genotypes and phenotypes in an experimental cross study

In Chapter 2 we incorporate biological knowledge when a causal network of genotypes and phenotypes is estimated. Since a causal network of genotypes and phenotypes is a statistical network, it may not reflect biological knowledge. At present there is huge amount of biological evidence in various aspects 1.1 and the incorporation of biological knowledge on a causal network reconstruction would get us a comprehensive biological network. We propose to integrate biological knowledge such as transcription factor binding and gene ontology to the inference of a causal gene network of genotypes and phenotypes from an experimental cross study. Biological knowledge sets a prior on phenotype networks (G_Y) with a scale parameter W to control the contribution of prior biological knowledge on the causal network construction. We introduce the ways to encode biological knowledge into a numeric matrix to be plugged in the prior probability ($P(G_Y)$). A Markov chain Monte Carlo is developed to find a causal gene network that fits genotype and phenotype data and biological knowledge well.

In Chapter 3 we consider the case when there are unmeasured variables in constructing causal networks. One reason for considering latent variables is that we often take a subset of variables to build a network due to the computational complexity for a large network. Another reason is that we are not certain that the data includes all the variables in an underlying network. It has been shown that if a Bayesian network is an underlying network but some variables are not measured, there may not exist a Bayesian network that properly represents the relations on the observed variables. Hence, we consider an extended version of Bayesian networks, ancestral graphs [Richardson and Spirtes, 2002], as a gene network. We model a gene network with genotypes in the framework of ancestral graphs and prove the graphical and statistical properties of our model. A Markov chain Monte Carlo is developed to search over directed ancestral graphs.

1.6 Contribution of this dissertation

Integration of various types of biological knowledge with experimental data is attracting lots of attention as biological knowledge gets accumulated. In the Bayesian perspective, the accumulation of biological knowledge should benefit us. At present, there is no standard method to integrate biological knowledge with experimental data. However, as we model the integration method in

diverse ways and keep comparing them, we anticipate to learn better integration methods and as a result, the accumulated knowledge could be fully utilized.

One of key motivation to study biology comes from variations — people are different to each other, species are different, environments are different, and time points are different. The association between variations and experimental data such as gene expressions could answer some biological questions. What is more interesting is to decipher causal relations between variations and gene expressions and causal relations among gene expressions simultaneously, which is a causal gene network.

In this dissertation, we integrate biological knowledge with gene expressions and genetic variations to infer a causal gene network. We learn the advantages and disadvantages of our integration model when inferring a causal network and learn the key points for improvements in future research.

When inferring a causal gene network with genetic variations, consideration of latent variables has not be attempted yet and this dissertation lays out a rigorous modeling with latent variables. Despite the computational challenge, it could give us as much information as the data have.

Chapter 2

Bayesian causal phenotype network incorporating genetic variation and biological knowledge [1]

2.1 Introduction

A key interest in molecular biology is to understand how DNA, RNA, proteins and metabolic products regulate each other. In this regard, people have considered to construct the regulatory networks composed of candidate regulatory relationships from microarray expression data with time-series measurements or transcriptional perturbations [Friedman et al., 2000, Gardner et al., 2003]. A regulatory network can also be constructed in a segregating population where genotypes perturb the gene expression, protein and metabolite levels. The genetic variation information can decipher genetic effects on traits and help discover causal regulatory relationships between phenotypes. In addition, knowledge of regulatory relationships is available in various biological databases, which can improve the reconstruction of causal networks. This chapter focuses on combining genetic variations in a segregating population and biological knowledge to improve the inference of causal networks.

Given the quantitative nature of a gene expression phenotype, one can perform quantitative trait loci (QTL) mapping to detect the genomic locations affecting the phenotype [Jansen and Nap, 2001]. The genotypes at a location are often coded as *AA*, *Aa*, or *aa*, where allele *A* and *a* are distinct variant forms of a genetic locus. A quantitative phenotype/trait is any observable physical or biochemical quantitative feature of an organism such as weight, blood pressure, gene

[1] This chapter is under review for publication: Moon, J.-Y., Chaibub Neto, E., Deng, X., Yandell, B. S. (2011) Bayesian causal phenotype network incorporating genetic variation and biological knowledge. In *Probabilistic Graphical Models Dedicated to Applications in Genetics*

expression, or protein levels. The basic idea of QTL mapping is to detect genomic regions, or QTLs, where variation in genotype is associated with quantitative variation in phenotype. For example, tall parents tend to have tall children while short parents tend to have short children. Then, it appears that there are genetic factors to be associated with the height and the genetic factors can be identified by QTL mapping. In an experimental population, where genotypes are randomly assigned, the genetic variation at QTLs can be interpreted as causing later changes in the phenotype of interest.

In a segregating population, QTL mapping can identify QTLs with a causal effect on a phenotype. The causal effect can be direct from QTL to phenotype, or indirect via other intermediate phenotypes. We only label the direct QTLs as "causal QTLs", recognizing that they have a more proximal effect on a phenotype than indirect QTLs. We also acknowledge that there may be many other molecular factors in a pathway between the QTL and the phenotype that were not measured in a particular study. Indirect and direct QTLs can be used to help determine the direction of the edges in a causal phenotype network (i.e., a directed graph composed of phenotype nodes, whose edges represent causal relations). Several approaches in the literature take advantage of QTLs identified by QTL mapping to determine causal relations among phenotypes including: structural equation modeling [Li et al., 2006, Liu et al., 2008, Aten et al., 2008]; score-based methods for Bayesian networks [Zhu et al., 2004, 2008, Winrow et al., 2009]; causal algorithms for Bayesian networks based on independence tests [Chaibub Neto et al., 2008, Valente et al., 2010]; and causality tests on pairs of phenotypes [Schadt et al., 2005, Kulp and Jagalur, 2006, Chen et al., 2007, Millstein et al., 2009, Chaibub Neto et al., 2010b]. A common feature of the above approaches is that QTL mapping and phenotype network reconstruction are conducted separately. The QTL mapping without consideration of a phenotype network may find indirect QTLs. As pointed out by Chaibub Neto et al. [2010a], incorrect or indirect QTLs may compromise the inference of causal relationships among phenotypes. To address this issue, several researchers [Chaibub Neto et al., 2010a, Hageman et al., 2011] proposed to jointly infer causal phenotype networks and causal QTLs.

Various sources of biological knowledge have been incorporated with gene expression in the reconstruction of phenotype networks because it is difficult to determine the causal direction of

gene regulation using expression data only. Transcription factor binding information was leveraged by [Tamada et al., 2003], whereas [Nariai et al., 2004] used protein-protein interaction knowledge to construct phenotype networks. Methods integrating multiple sorts of biological knowledge were proposed by [Imoto et al., 2004], [Werhli and Husmeier, 2007], and [Christley et al., 2009].

In this chapter, we propose a Bayesian approach to jointly inferring a causal phenotype network and causal QTLs with a prior distribution on phenotype network structures adjusted by biological knowledge. The joint network of causal phenotype relationships and causal QTLs is modeled as a Bayesian network[1] adopted from Chaibub Neto et al. [2010a], QTLnet. Causal QTLs can be inferred by QTL mapping conditional on the phenotype network. Since the phenotype network is unknown, QTLnet traverses the space of phenotype networks and updates causal QTLs using Markov Chain Monte Carlo (MCMC). We extend the framework of QTLnet by incorporating biological knowledge into the prior distribution on phenotype network structures. The incorporation of biological knowledge is expected to increase the accuracy of the model, enhancing the predictive power of the network [Zhu et al., 2008]. The prior probability on phenotype network structures is based on the Gibbs distribution to integrate different sources of biological information allowing for flexible tuning of the analyst's confidence on this knowledge [Werhli and Husmeier, 2007]. The consideration of reliability of biological knowledge is necessary since biological knowledge can be incomplete and inaccurate. While Zhu et al. [2008] proposed a method to incorporate genetic variation and biological knowledge to phenotype networks, their method does not consider the reliability of biological knowledge. Our proposed approach (QTLnet-prior) can integrate phenotype data, genetic variation and several sources of biological knowledge (protein-protein interaction,

[1]Note that Bayesian networks can be inferred in a Bayesian framework or a frequentist framework. Here, we take a Bayesian approach to infer a Bayesian network. The reason that the term "Bayesian" is used in a Bayesian network is described in the book [Pearl, 2000]. An excerpt from page 14 of the book [Pearl, 2000]: *Bayesian networks, a term coined in Pearl (1985) to emphasize three aspects: (1) the subjective nature of the input information; (2) the reliance on Bayes's conditioning as the basis for updating information; and (3) the distinction between causal and evidential modes of reasoning, a distinction that underscores Thomas Bayes's paper of 1763.*

gene ontology annotation, and transcription factor and DNA binding information) with the consideration of the reliability of each source of biological knowledge in the network reconstruction algorithm.

The details of our integrated framework for the joint inference of causal phenotype network and causal QTLs are organized as follows. Section 2.2 describes the QTLnet method for the joint inference of causal network and causal QTLs. Section 2.3 presents the proposed QTLnet-prior, which incorporates biological knowledge into the prior probability distribution of phenotype network structures. A simulation study is conducted in Section 2.4 to compare the proposed method with several existing approaches and it shows that the incorporation of biological knowledge and genetic variation is advantageous in inferring a causal network, especially when biological knowledge is reliable. In Section 2.5, the proposed method is used to reconstruct a network of 26 genes involved in the yeast cell cycle. Finally, in Section 2.6, we discuss the strengths and caveats of our approach and point out future research directions.

2.2 Joint inference of causal phenotype network and causal QTLs

In Section 2.2.1, we first present a standard Bayesian network for modeling phenotype data. Next, in Section 2.2.2, we present an extended model, based on the homogeneous conditional Gaussian regression (HCGR) model, to incorporate QTL nodes into phenotype networks. Directed edges in the standard Bayesian network can be interpreted as causal relationships. By extending the phenotype network with causal QTL nodes we can further claim causal interpretations. In Section 2.2.3, we present a rationale for the joint inference of the causal phenotype network and causal QTLs and in Section 2.2.4, we describe the QTL mapping conditional on the phenotype network. Finally, we give an overview of our joint approach for phenotype network and causal QTL inference in Section 2.2.5.

2.2.1 Standard Bayesian network model

A standard Bayesian network is a probabilistic graphical model whose conditional independence is represented by a directed acyclic graph (DAG). A node t in a DAG G corresponds to a

random variable Y_t in the Bayesian network. A directed edge from node u to node v can supposedly represent that Y_v is causally dependent on Y_u, though an edge truly represents the conditional dependency. The local directed Markov property of Bayesian networks states that each variable is independent of its non-descendant variables conditional on its parent variables:

$$Y_t \perp Y_{V \setminus de(t)} \mid Y_{pa(t)} \quad \text{for all } t \in V,$$

where V is the set of all nodes in a DAG, $de(t)$ is the set of descendants of node t, $pa(t)$ is the set of parents of node t and $Y_{pa(t)}$ is a set of variables indexed by $pa(t)$, that is, $\{Y_i : i \in pa(t)\}$. Assume the node index is ordered such that the index of descendants is always bigger than the index of their parents. Since $\{t-1, \ldots, 1\}$ is a set of non-descendants of node t and $pa(t)$ is included in the non-descendant set $\{t-1, \ldots, 1\}$, Y_t is independent of $Y_{\{t-1,\ldots,1\}}$ conditional on $Y_{pa(t)}$. That is, $P(Y_t \mid Y_{pa(t)})$ is equivalent to $P(Y_t \mid Y_{t-1}, \ldots, Y_1)$. The joint distribution can be written to be

$$\begin{aligned} P(Y_1, \ldots, Y_T) &= \prod_{t=1}^{T} P(Y_t \mid Y_{t-1}, \ldots, Y_1) \\ &= \prod_{t=1}^{T} P(Y_t \mid Y_{pa(t)}), \end{aligned} \tag{2.1}$$

where the first equality is satisfied by the chain rule in probability theory[2].

2.2.2 HCGR model

The parametric family of a Bayesian network that jointly models phenotypes and QTL genotypes corresponds to a homogeneous conditional Gaussian regression (HCGR) model. Conditional

[2]In probability theory, the chain rule permits that the joint probability of two variables X and Y can be written as $P(X, Y) = P(Y \mid X)\,P(X) = P(X \mid Y)\,P(Y)$. This can be extended to the joint probability of multiple variables:

$$\begin{aligned} P(Y_T, \ldots, Y_1) &= P(Y_T \mid Y_{T-1}, \ldots, Y_1) P(Y_{T-1}, \ldots, Y_1) \\ &= P(Y_T \mid Y_{T-1}, \ldots, Y_1) P(Y_{T-1} \mid Y_{T-2}, \ldots, Y_1) P(Y_{T-2} \mid Y_{T-3}, \ldots, Y_1) \\ &= \cdots \\ &= \prod_{t=1}^{T} P(Y_t \mid Y_{t-1}, \ldots, Y_1). \end{aligned}$$

on the QTL genotypes and covariates, the phenotypes are distributed according to a multivariate normal distribution, where QTLs and covariates enter the model via the mean, and the correlation structure among the phenotypes is explicitly modeled according to the DAG representing the phenotype network structure [Chaibub Neto et al., 2010a]. Figure 2.1 depicts one example of a joint Bayesian network of phenotypes and QTL genotypes.

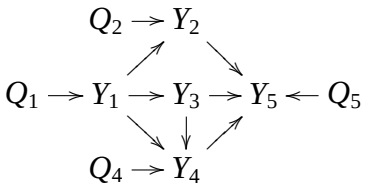


Figure 2.1: Example network with five phenotypes ($Y_1, \ldots, Y_5$) and four QTLs ($Q_1, \ldots, Q_4$).

The HCGR model is derived from a series of linear regression equations. For $i = 1, \ldots, n$ and $t = 1, \ldots, T$, let Y_{ti} be the value of phenotype for individual i and trait t. Then we assume for each phenotype that Y_{ti} can be modeled as follows:

$$Y_{ti} = \mu_{ti}^{\star} + \sum_{v \in pa(t)} \beta_{tv} Y_{vi} + \epsilon_{ti}, \quad \epsilon_{ti} \sim \mathcal{N}(0, \sigma_t^2). \tag{2.2}$$

The model can be decomposed into three parts: a genetic part ($\mu_{ti}^{\star}$), a phenotypic part ($\sum_{v \in pa(t)} \beta_{tv} Y_{vi}$), and an error term ($\epsilon_{ti}$). In the phenotypic part, β_{tv} is the effect of parent phenotype v on phenotype t. The error term, ϵ_{ti}, follows a normal distribution. The genetic part, $\mu_{ti}^{\star}$, corresponds to a model of QTL genotypes and possibly covariates:

$$\mu_{ti}^{\star} = \mu_t + \sum_{k=1}^{C} \vartheta_{tk} Z_{ki} + \sum_{k=1}^{K} \gamma_{tk} \theta_{tk} X_{ki}, \tag{2.3}$$

where μ_t is the overall mean for trait t, Z_{ki} represents a covariate, ϑ_{tk} represents the effect of the covariate on the phenotype, and $\sum_{k=1}^{K} \gamma_{tk} \theta_{tk} X_{ki}$ is the overall effect of QTLs. For the simplicity, we will not consider the covariates Z later on. The parameter γ_{tk} is unknown, and represents the inclusion ($\gamma_{tk} = 1$) or exclusion ($\gamma_{tk} = 0$) of the QTL located at the genomic position k into the model. The genetic effects of QTL can be partitioned into different types of genetic effects, e.g.

additive and dominance effects, and hence the genotype of the QTL is coded into the variables to estimate the different genetic effects. The vector X_{ki} represents a column vector of coded variables of the genotype at the genomic location k for individual i and the vector θ_{tk} is a row vector of several types of genetic effects of QTL at the location k on phenotype t. The coding of a genotype may follow Cockerham's genetic model [Kao and Zeng, 2002]. For example, in an intercross, the segregating genotypes at a locus are denoted by *AA*, *Aa*, or *aa*, and we can code the genotype into an additive variable by the number of *A* alleles in the genotype minus 1 and a dominance variable by 1/2 if it is *Aa* and -1/2 otherwise. In this case, the additive effect is the effect of substituting one allele *a* with another allele *A* and the dominance effect is the deviation of *Aa* from the mean of *AA* and *aa*. Accordingly, in an intercross, X_{ki} is a column vector of additive and dominance coding variables, and θ_{tk} is the row vector of additive and dominance effects on phenotype t. It was shown by Chaibub Neto et al. [2010a] that these linear regression equations in eqn (2.2) set a HCGR model for phenotypes and QTL genotypes.

2.2.3 Systems genetics and causal inference

Systems genetics aims to understand the complex interrelations between genetic variations and phenotypes from large scale genotype and phenotype data [Nadeau and Dudley, 2011]. Here we explain how the systems genetics approach can infer causal networks. Causal relations from QTLs to phenotypes are justified by the unidirectional influence of the genotype on phenotype and the random allocation of genotypes to individuals. In contrast, causal relations among phenotypes are induced from conditional independence. The key idea of systems genetics is that by incorporating QTL nodes into phenotype networks we create new sets of conditional independence relationships for distinguishing network structures that would, otherwise, belong to the same equivalence class (see Tables 2.1 and 2.2).

First, we give a more detailed description for the causal relations between QTLs and phenotypes. As stated in the central dogma of molecular biology, the hereditary DNA information is transferred to phenotypes. Thus a genotype influences phenotypes in general but not the other

way around. A genotype is assumed to be randomized to other environmental factors by independent segregation of chromosomes in meiosis and random mating between gametes. These special characteristics enable us to infer causal effects of QTLs on phenotypes since, by analogy with a randomized experiment, we have that: (1) the treatment (genotype) to an experimental unit precedes the measured outcome (phenotype), and (2) random allocation of treatments to experimental units guarantees that other common causes get averaged out. Two loci on the same chromosome are highly correlated when their distance is small. But crossovers between two loci can still occur randomly in proportion to the distance. One can distinguish the true causal QTL and false nearby QTL with a large sample size. This random allocation is explicit in an experimental cross such as a backcross or an intercross[3]. While this idea can be extended to natural populations, special attention must be paid to admixture, kinship and other forms of relatedness.

Second, the explanation of causal inference among phenotypes requires the concept of conditional independence in DAGs composed of phenotypes and QTL nodes. In the next three paragraphs we present some definitions and results that allow us to infer phenotype-to-phenotype causal relationships.

Here are definitions. In graph theory, a *path* is defined as any unbroken, non-intersecting sequence of edges in a graph, which may go along or against the direction of arrows. We say that a path p is *d-separated* [Pearl, 1988, 2000] by a set of nodes Z if and only if: (1) p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node is in Z, or (2) p contains a collider $i \rightarrow m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z. We say that Z d-separates X from Y if and only if Z blocks every path from a node in X to a node in

[3]An experimental cross is generated by crossing inbred lines. An inbred line is obtained by repeated generations of inbreedings so that any genotype of the inbred line is homozygous, *AA*. Therefore, breeding within the inbred line produces genetically identical offspring to its parents. Both backcross and intercross first produce the first generation of population by mating two different inbred lines, *AA* and *BB*. The first generation is identical to each other with heterozygous genotypes, *AB*. The backcross population is produced by mating the first generation to one of its parental inbred lines such as *AA*. Then, the backcross population has genotypes either *AA* or *AB* in a ratio of 1 : 1. The intercross population is produced by mating the first generation itself so that it has genotypes *AA*, *AB*, or *BB* in a ratio of 1 : 2 : 1.

Y. The *skeleton* of a DAG is the undirected graph obtained by replacing its arrows by undirected edges. A *v-structure* is composed by two converging arrows whose tails are not connected by an arrow.

The equivalence concept plays a key role in learning the structure of networks from the data. Here we present three important equivalence relations for graphs or statistical models of graphs. Two graphs are *Markov equivalent* if they have the same set of d-separation relations [Spirtes et al., 2000]. Two structures m_1 and m_2 for Y are *distribution equivalent* with respect to the distribution family F if they represent the same joint distributions for Y, that is, for every θ_1, there exists a θ_2 such that $p(Y \mid \theta_1, m_1) = p(Y \mid \theta_2, m_2)$ [Heckerman et al., 2006]. In other words, m_1 and m_2 are distribution equivalent if the parameters θ_1 and θ_2 are simple re-parametrizations of each other. If m_1 and m_2 are distribution equivalent, then the invariance principle of maximum likelihood estimates guarantees $p(Y \mid \hat{\theta}_1, m_1) = p(Y \mid \hat{\theta}_2, m_2)$, and m_1 and m_2 cannot be distinguished using the data. In this case we say that m_1 and m_2 are *likelihood equivalent*. In a Bayesian setting we define likelihood equivalence using the prior predictive distribution, $\int p(Y \mid \theta, m_1)\, p(\theta \mid m_1)\, d\theta = \int p(Y \mid \theta, m_2)\, p(\theta \mid m_2)\, d\theta$. If models m_1 and m_2 are distribution equivalent and we adopt a proper prior $p(\theta \mid m)$, it is often reasonable to expect $p(Y \mid m_1) = p(Y \mid m_2)$, so that we cannot distinguish m_1 and m_2 for any data set Y [Heckerman et al., 2006].

Now we state four important results regarding causal inference in systems genetics: (1) Two DAGs are Markov equivalent if and only if they have the same skeletons and the same set of v-structures [Verma and Pearl, 1990]; (2) Distribution equivalence implies Markov equivalence, but the converse is not necessarily true [Spirtes et al., 2000]; (3) For a Gaussian regression model, Markov equivalence implies distribution equivalence [Heckerman and Geiger, 1996]; (4) For the homogeneous conditional Gaussian regression model, Markov equivalence implies distribution equivalence [Chaibub Neto et al., 2010a].

Therefore, for the HCGR parametric family, two DAGs are distribution and likelihood equivalent if and only if they are Markov equivalent. This implies that we can simply check out if any two DAGs have the same skeleton and the same set of v-structures in order to determine if they are likelihood equivalent and hence cannot be distinguished using the data.

Getting back to the idea of causal inference among phenotypes, let G_Y be a phenotype network represented by a standard Bayesian network of phenotypes, Y. Phenotype data alone can distinguish some network structures by its likelihood but may fail to distinguish some other network structures. For example, consider the three network structures in Table 2.1. Models G_Y^1 and G_Y^3 have the same skeleton ($Y_1 - Y_2 - Y_3$) and the same set of v-structures (no v-structure) and, thus, are distribution/likelihood equivalent. Model G_Y^2, on the other hand, has the same skeleton but a different set of v-structures and, hence, is not distribution/likelihood equivalent to models G_Y^1 and G_Y^3. Therefore, phenotype data alone can identify G_Y^2 but cannot distinguish G_Y^1 and G_Y^3.

Table 2.1: Models G_Y^1 and G_Y^3 are distribution/likelihood equivalent.

DAG structures	skeletons	v-structures
$G_Y^1 = Y_1 \rightarrow Y_2 \rightarrow Y_3$	$Y_1 - Y_2 - Y_3$	$\emptyset$
$G_Y^2 = Y_1 \rightarrow Y_2 \leftarrow Y_3$	$Y_1 - Y_2 - Y_3$	$Y_1 \rightarrow Y_2 \leftarrow Y_3$
$G_Y^3 = Y_1 \leftarrow Y_2 \rightarrow Y_3$	$Y_1 - Y_2 - Y_3$	$\emptyset$

Adding causal QTL nodes to a phenotype network allows the inference of causal relationships between phenotypes that could not be distinguishable using phenotype data alone. For example, if we add a causal QTL Q_1 to Y_1 in phenotype networks G_Y^1 and G_Y^3 in the above example, then the corresponding extended network structures G^1 and G^3 have different v-structures as shown in Table 2.2.

Table 2.2: Extended models G^1 and G^3 are no longer distribution/likelihood equivalent.

Extended DAG structures	skeletons	v-structures
$G^1 = Q_1 \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_3$	$Q - Y_1 - Y_2 - Y_3$	$\emptyset$
$G^3 = Q_1 \rightarrow Y_1 \leftarrow Y_2 \rightarrow Y_3$	$Q - Y_1 - Y_2 - Y_3$	$Q \rightarrow Y_1 \leftarrow Y_2$

2.2.4 QTL mapping conditional on phenotype network structure

Now we examine the inference of QTLs conditional on a phenotype network. QTL mapping can be done in a conditional or unconditional fashion. In the unconditional mapping analysis, we measure the association of a trait Y_t and QTL Q using the LOD score (logarithm of odds)

$$LOD(y_t, q) = \log_{10}\left(\frac{f(y_t \mid q)}{f(y_t)}\right),$$

where $f(y_t \mid q)$ represents the predictive density of a linear model with Q as an independent variable and $f(y_t)$ the predictive density of the baseline model. Here a predictive density is given by a maximized likelihood in a frequentist setting, or by the prior predictive density in a Bayesian setting. A high LOD score means that Y_t and Q are associated. Note that unconditional analysis can detect QTLs that directly affect the phenotype under investigation, as well as QTLs with indirect effects [Chaibub Neto et al., 2010a]. For example, if we consider the causal network of phenotypes and QTLs in Fig. 2.1, then the unconditional QTL mapping of Y_2 detects a direct QTL Q_2 as well as an indirect QTL Q_1 that affects Y_2 via Y_1. Figure 2.2 shows the expected results of the unconditional analysis for each phenotype.

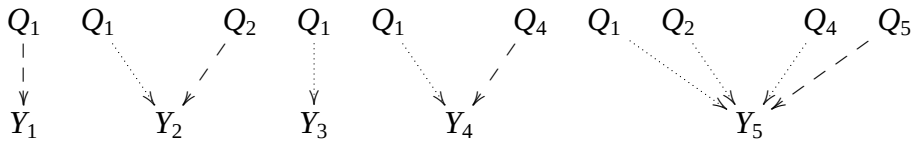


Figure 2.2: Output of the unconditional QTL mapping analysis for the phenotypes in Fig. 2.1. Dashed and pointed arrows represent direct and indirect QTL/phenotype causal relationships, respectively.

The conditional mapping analysis, on the other hand, incorporates other traits as covariates, and measures the association of Y_t and Q conditional on these covariates (say y_z) using the conditional LOD score

$$\begin{aligned} LOD(y_t, q \mid y_z) &= \log_{10}\left(\frac{f(y_t \mid q, y_z)}{f(y_t)}\right) - \log_{10}\left(\frac{f(y_t \mid y_z)}{f(y_t)}\right) \\ &= LOD(y_t, q, y_z) - LOD(y_t, y_z). \end{aligned}$$

Now, consider QTL mapping analysis tailored to a known phenotype network structure. In this situation we can avoid detecting indirect QTLs by simply performing mapping analysis of the phenotypes conditional on their parents. For instance, in Fig. 2.1, if we perform QTL mapping of Y_5 conditional on Y_2, Y_3 and Y_4 we do not detect Q_1, Q_2 and Q_4 because of the following independence relations: $Y_5 \perp Q_1 \mid Y_2, Y_3, Y_4$; $Y_5 \perp Q_2 \mid Y_2, Y_3, Y_4$; and $Y_5 \perp Q_4 \mid Y_2, Y_3, Y_4$. We only detect Q_5 due to the following relation: $Y_5 \not\perp Q_5 \mid Y_2, Y_3, Y_4$.

In practice, however, the structure of the phenotype network is unknown, and performing QTL mapping conditional on a misspecified phenotype network structure can result in the inference of misspecified causal QTLs as shown in Fig. 2.3. The mapping analysis of a phenotype conditional on downstream phenotypes in the true network, induces dependencies between the phenotype and QTLs affecting downstream phenotypes. This leads to the erroneous inference that the phenotype includes downstream QTLs as its QTLs. For example, the mapping analysis of Y_4 conditioning on Y_1, Y_3 and a downstream phenotype Y_5 includes downstream Q_5 as its QTLs in Fig. 2.3(b). However, a model with misspecified phenotype network and QTLs will generally have a lower marginal likelihood score than the model with the correct causal order for the phenotypes and correct QTLs. Since in practice QTLnet adopts a model selection procedure to traverse the space of network structures, it tends to prefer models closer to the true data generating process. Simulation studies presented in Chaibub Neto et al. [2010a] corroborate this point.

Note that, as pointed out in Chaibub Neto et al. [2010a], the conditional LOD score can be adopted as a formal measure of independence between a phenotype and QTLs. Even though we restrict our attention to HCGR models, conditional LOD profiling is a general framework for the detection of conditional independencies between continuous and discrete random variables. Contrary to partial correlations, the conditional LOD score does not require the assumption of multinormality of the data in order to formally test for independence, and it can handle QTL by covariate interactions.

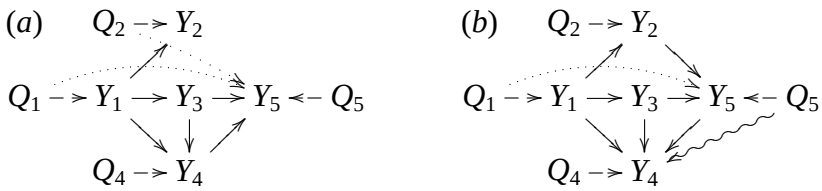


Figure 2.3: QTL mapping tailored to the network structure. Dashed, pointed and wiggled arrows represent, respectively, direct, indirect and incorrect QTL/phenotype causal relationships. (a) Mapping analysis of Y_5 conditional on Y_3 and Y_4 still detects Q_1 and Q_2 as QTLs for Y_5, since failing to condition on Y_2 leaves the paths $Q_1 \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_5$ and $Q_2 \rightarrow Y_2 \rightarrow Y_5$ in Fig. 2.1 open. In other words, (Y_3, Y_4) cannot d-separate (Q_1, Q_2) from Y_5 in the true causal graph. (b) Mapping analysis of Y_4 conditional on Y_1, Y_3 and Y_5 incorrectly detects Q_5 as a QTL for Y_4 because in the true network the paths $Y_4 \rightarrow Y_5 \leftarrow Q_5$ and $Y_4 \leftarrow Y_3 \rightarrow Y_5 \leftarrow Q_5$ in Fig. 2.1 are open when we condition on Y_5.

2.2.5 Joint inference of phenotype network and causal QTLs

Section 2.2.4 describes the QTL mapping conditional on a phenotype network. In practice, the phenotype network is generally unknown and we cannot directly infer the correct causal QTLs. Therefore, we need to perform a joint inference of phenotype network and causal QTLs.

Recall that Y are phenotypes, X are genetic variations as defined in Section 2.2.2, and G is a Bayesian network structure of phenotypes and QTLs. Let G_Y represent a phenotype network and let $G_{Q\to Y}$ represent a graph from causal QTL nodes to phenotype nodes. Note that G_Y and $G_{Q\to Y}$ are subgraphs of the extended network structure G. Conforming to the HCGR model in eqn (2.2), G_Y corresponds to the collection of causal relations from $pa(t)$ to trait t and $G_{Q\to Y}$ corresponds the collection of causal relations from nonzero γ_{tk} to traits. Denote θ_G to be the parameter sets $(\beta_{tv}, \sigma_t^2, \mu_t, \theta_{tk})$. From eqn (2.2), the likelihood of a Bayesian network of phenotypes and causal QTLs can be written as a product of normal densities:

$$\begin{aligned} P(Y \mid G, X, \theta_G) &= P(Y \mid G_Y, G_{Q\to Y}, X, \theta_G) \\ &= \prod_{t=1}^{T} \prod_{i=1}^{n} \mathcal{N}\left(\mu_{ti}^{\star} + \sum_{y_k \in pa(y_t)} \beta_{tk} y_{ki}\,,\ \sigma_t^2\right). \end{aligned}$$

The marginal likelihood of phenotypes and causal QTLs $P(Y \mid G, X)$ is calculated by integrating parameters θ_G out in the Bayesian network

$$P(Y \mid G, X) = \int P(Y \mid G, X, \theta_G)\, P(\theta_G \mid G)\, d\theta_G.$$

The posterior probability of G conditional on the data is given by

$$P(G \mid Y, X) = \frac{P(Y \mid G, X) P(G)}{\sum_G P(Y \mid G, X) P(G)},$$

where $P(G)$ represents the prior probability of the network structure G. In the next section we devote our attention to the specification of $P(G)$ using integrated biological knowledge.

Following Chaibub Neto et al. [2010a], we adopt the QTLnet framework that jointly infers the phenotype network structure and causal QTLs. Most of the current literature in genetical network reconstruction has treated the problems of QTL inference and phenotype network reconstruction

separately, generally performing QTL inference first, and then using QTLs to help determine the phenotype network structure [Zhu et al., 2008, Chaibub Neto et al., 2008]. As indicated in Section 2.2.4, such a strategy can include QTLs with indirect effects into the network.

2.3 Causal phenotype network incorporating biological knowledge

Besides the causal QTLs, biological knowledge is another useful and important information resource to enhance the construction of the phenotype network. Such knowledge can be integrated on top of the causal network to provide a more comprehensive picture of how genes are regulated. This integrated network could generate a new hypothesis on gene regulation, having an overall consistency with biological knowledge.

In this section, we propose a network inference method, QTLnet-prior, from phenotype data with genetic variations, integrating biological knowledge. The QTLnet-prior extends the framework of QTLnet referred to at the end of Section 2.2.5. It specifies the prior probability on phenotype network structures to integrate multiple sources of biological knowledge with flexible tuning parameters on confidence of knowledge [Werhli and Husmeier, 2007]. The weighted integration of biological knowledge could produce a more predictive Bayesian network. The details of our extended framework, QTLnet-prior, are presented in Section 2.3.1. In Section 2.3.2, we sketch a Metropolis-Hastings (M-H) MCMC scheme for QTLnet-prior implementation that integrates the sampling of network structures [Madigan and York, 1995, Grzegorczyk and Husmeier, 2008], the QTL mapping, and the sampling of biological knowledge weights. In Section 2.3.3, we present how to encode biological knowledge into the prior distribution over phenotype network structures.

2.3.1 Model

2.3.1.1 Extended model

Denote by G a Bayesian network structure of phenotypes and QTLs. The graph G consists of a phenotype network (G_Y) and causal QTLs to phenotypes ($G_{Q\to Y}$). Let Y be phenotype data, X be genetic variations, and W represent weights set on various sources of biological knowledge B. The biological knowledge B is considered to be relations between phenotypes such as transcription

factor binding, protein-protein interaction and gene ontology annotation. That is, biological knowledge B can give a prior probability only for the phenotype network G_Y. The QTLnet framework presented in Section 2.2 assumes intrinsically a uniform prior over phenotype network structures. Additionally, we specify a prior distribution on the weights of biological knowledge in order to control the consistency between phenotype data and knowledge. Because the prior information can be inaccurate or incompatible with the phenotype data, it is important to quantify its uncertainty. We write the extended model as follows:

$$\begin{aligned} P(G, W \mid Y, X, B) &\propto P(Y \mid G, W, X, B) P(G, W \mid X, B) \\ &= P(Y \mid G, X) P(G, W \mid X, B) \\ &= P(Y \mid G, X) P(G_Y, W \mid X, B) P(G_{Q \to Y} \mid X, B) \\ &= P(Y \mid G, X) P(G_Y, W \mid B) P(G_{Q \to Y} \mid X) \\ &= P(Y \mid G, X) P(G_Y \mid B, W) P(W \mid B) P(G_{Q \to Y} \mid X). \end{aligned} \tag{2.4}$$

In the first step, the posterior probability of a network G and weights W is calculated by multiplying the marginal likelihood $P(Y \mid G, W, X, B)$ of the traits given the network G and the prior probability $P(G, W \mid X, B)$ of a network and weights given genetic variations and biological knowledge. The marginal likelihood $P(Y \mid G, W, X, B)$ can be simplified to be $P(Y \mid G, X)$ as in the second step. In the third step, the prior probability $P(G, W \mid X, B)$ can be decomposed into $P(G_Y, W \mid X, B)$ and $P(G_{Q \to Y} \mid X, B)$ by assuming the independence between a phenotype network G_Y along with the weights W and causal QTLs $G_{Q \to Y}$ given genetic variations X and biological knowledge B. The fourth step is provided by the fact that $P(G_Y, W \mid X, B)$ is equal to $P(G_Y, W \mid B)$ because the genetic variations are not included in the structure of the phenotype network G_Y, and $P(G_{Q \to Y} \mid X, B)$ is equal to $P(G_{Q \to Y} \mid X)$ because B affects G_Y but not $G_{Q \to Y}$. The extended model in eqn (2.4) shows that prior distributions on phenotype network structure $P(G_Y \mid B, W)$, biological knowledge weights $P(W \mid B)$ and causal QTLs of traits $P(G_{Q \to Y} \mid X)$ must be specified. We will describe how to set $P(G_Y \mid B, W)$, $P(W \mid B)$, and $P(G_{Q \to Y} \mid X)$ in the following.

2.3.1.2 Prior on phenotype network structures $P(G_Y \mid B, W)$

Incorporation of *a priori* biological knowledge into a prior on network structures can lead to discriminate Bayesian networks having the same likelihood [Werhli and Husmeier, 2007, Zhu et al., 2007]. If G^1 and G^2 have the same likelihood ($P(Y \mid G^1) = P(Y \mid G^2)$) but have different prior probabilities ($P(G^1) \neq P(G^2)$), the posterior probabilities would become different ($P(G^1 \mid Y) \neq P(G^2 \mid Y) \propto P(Y \mid G^2)P(G^2)$). For example, consider two graphs for nodes t and v: one is $t \rightarrow v$ and the other is $v \rightarrow t$. Their likelihoods are the same because they are Markov equivalent. If a prior indicates that one direction ($t \rightarrow v$) is more likely than the other direction ($t \leftarrow v$), then the posterior of one direction ($t \rightarrow v$) becomes higher than the other direction ($v \rightarrow t$). The biological knowledge B along with its weight W can therefore give different prior probabilities $P(G_Y \mid B, W)$ for the phenotype network G_Y.

Various types of information can supplement the learning of a phenotype network. We can encode this supplementary information into unequal priors on network structures. A transcription factor binding location can be used to prefer the direction from a transcription factor to the target gene [Bernard and Hartemink, 2005]. Pathway information can also guide to infer directions among phenotypes [Werhli and Husmeier, 2007]. For example, consider a network with three nodes t, v and u, where a path from t to v is known. Then, we can at least distinguish these two relations: $t \rightarrow v \leftarrow u$ and $t \rightarrow v \rightarrow u$. Regulation inference [Peleg et al., 2010, Yeang et al., 2004, Ourfali et al., 2007] from knock-out data and protein-protein interaction [Imoto et al., 2003] can be used as a prior for network structure. We will describe how to encode this information in Section 2.3.3. Since QTLnet is a Bayesian approach, we can flexibly incorporate various sources of biological knowledge by constructing meaningful priors for the network structures.

Now, it remains to set the prior distribution on phenotype network structure G_Y with respect to biological knowledge B. Since a Bayesian network distribution can be factored by its parent-child relations $\prod_t P(Y_t \mid Y_{pa(t)})$, it is natural to assume the prior on DAG structures to be factored by its parent-child relations. Adapting the prior formulation over network structures in Werhli and Husmeier [2007], we will show below that the prior satisfies the parent-child factorization.

Let us define the *energy* of a phenotype network G_Y relative to the biological knowledge B to be

$$\mathcal{E}(G_Y) = \sum_{i,j=1}^{T} |B(i, j) - G_Y(i, j)|, \tag{2.5}$$

where B is an encoding meant to describe biological knowledge ranging from 0 to 1 and G_Y is represented by the adjacency matrix of a network structure. The adjacency matrix is a 0-1 matrix which assigns $G_Y(i, j)$ to be 1 if there is a directed edge from node i to j, and to be 0 otherwise. The energy $\mathcal{E}(G_Y)$ acts as a distance measure between biological knowledge and a network structure G_Y. For a fixed biological knowledge matrix B, network structures will have small energy if they agree with the biological knowledge, and will have large energy if they disagree with the knowledge.

The energy can be decomposed into the sum of local pseudo-energies defined by parent-child relations for each trait:

$$\begin{aligned}\mathcal{E}(G_Y) &= \sum_{j=1}^{T}\left(\sum_{i\in pa(j)}(1 - B(i, j)) + \sum_{i\notin pa(j)} B(i, j)\right) \\ &= \sum_{j=1}^{T}\left(\frac{|B|}{T} + \sum_{i\in pa(j)}(1 - 2B(i, j))\right) = \sum_{j=1}^{T}\left(\frac{|B|}{T} + \mathcal{E}_{j,pa(j)}(G_Y)\right),\end{aligned}$$

where $|B| = \sum_{i,j=1}^{T} B(i, j)$ and $\mathcal{E}_{j,pa(j)}(G_Y) = \sum_{i\in pa(j)}(1 - 2B(i, j))$, which is the local pseudo-energy defined by phenotype j and and its parents. Therefore, the prior distribution on network structures can be constructed in terms of energy and it is shown to be the Gibbs distribution factorized by parent-child relations:

$$\begin{aligned}P(G_Y|B, W) &= \frac{\exp(-W\mathcal{E}(G_Y))}{Z(W)} \\ &= \frac{\prod_{j=1}^{T}\exp(-W\mathcal{E}_{j,pa(j)}(G_Y))}{Z'(W)}, \quad G_Y \in \text{DAG}\end{aligned} \tag{2.6}$$

where $Z(W)$ is a normalizing constant given by $\sum_{G_Y\in\text{DAG}}\exp(-W\mathcal{E}(G_Y))$ and $Z'(W)$ is another normalizing constant given by $Z(W)/\exp(-W|B|)$. For a fixed W, network structures with small energy will have higher prior probabilities than network structures with large energy. The weight W of biological knowledge B is introduced to tune the confidence of biological information which

sometimes can be inaccurate or incompatible with expression data. As W goes toward 0, the influence of *a priori* knowledge gets negligible and the prior distribution of network structure is assumed to be almost uniform. On the contrary, as W goes to the infinity, the prior on network structure peaks at the biological knowledge.

Multiple sources of biological knowledge can be integrated into a prior on network structures with different weights.

$$P(G_Y \mid B, W) = \frac{\exp(-\sum_k W_k \mathcal{E}_k(G_Y))}{Z(W)}, \quad G_Y \in \text{DAG}$$

where B_k is an encoding matrix of biological knowledge from source k, B is the vector of biological knowledge matrices $(B_1, \ldots, B_k)$, W_k is the weight of B_k, W is the weight vector $(W_1, \ldots, W_k)$, and $Z(W)$ is the summation of the numerator over all DAGs.

2.3.1.3 Prior on biological knowledge weights $P(W \mid B)$

The weight parameter is introduced to control the influence of biological knowledge on the phenotype network. A higher value of the weight would increase the influence of the biological knowledge on the posterior distribution of networks. Specifically, a large W puts significant prior probability on the phenotype network structures which consistently agree with biological knowledge B. Conversely, a small W puts fairly equal prior probabilities on all possible networks. If biological knowledge B is similar to the true network from which the expression data are generated, then the posterior probability will peak at high W. On the contrary, if biological knowledge is deviated substantially from the true network, the posterior will peak at small W. This happens because a smaller W leads to a smaller ratio of prior probabilities of the deviated network and the true network. Consequently, the posterior of the true network can be larger than the posterior of the deviated network by the virtue of likelihood ratio overcoming the prior ratio at a small W.

For each biological knowledge B_k, we specify the prior probability distribution of the weight W_k to be an exponential distribution such that

$$P(W_k \mid B_k) = \phi \exp(-\phi W_k), \tag{2.7}$$

with the rate parameter ϕ. Such an exponential prior for W_k has several advantages. First, it does not impose an upper bound on W_k. Second, it would not allow the weight going to infinity too easily since an infinite weight always results in a network closer to the biological knowledge regardless of expression data. Third, when biological knowledge is inaccurate or incompatible with expression data, the exponential distribution can control the contribution of negative biological knowledge more easily than a uniform distribution. The rate parameter ϕ is set to be 1 in our simulation because this rate balances the prior and likelihood well in the empirical study.

2.3.1.4 Prior on causal QTLs $P(G_{Q\to Y} \mid X)$

Without any specific information about the causal QTLs, we set the prior of causal QTLs to be a uniform distribution. Several alternative specifications can be found in Bayesian QTL mapping such as in Yi et al. [2005] and in Yi et al. [2007].

2.3.2 Sketch of MCMC

A main challenge in the reconstruction of networks is that the graph space grows super-exponentially with the number of nodes. An exhaustive search approach over all network structures is impractical even for small networks. Hence, heuristic approaches are needed to efficiently traverse the graph space. We adopt a Metropolis-Hastings (M-H) MCMC scheme that integrates the sampling of network structures [Madigan and York, 1995, Husmeier, 2003], the QTL mapping, and the sampling of biological knowledge weights W. The MCMC scheme iterates between accepting a network structure G and accepting k weights $W_1, \cdots, W_k$ corresponding to k types of biological knowledge.

1. Sample a new phenotype network structure G_Y^{new} from a network structure proposal distribution $R(G_Y^{new} \mid G_Y^{old})$.

2. Given the phenotype network structure G_Y^{new}, sample a new set of causal QTLs $G_{Q\to Y}$ from a QTL proposal distribution $R(G_{Q\to Y}^{new} \mid G_{Q\to Y}^{old})$.

3. Accept the new extended network structure G^{new} composed of G_Y^{new} and $G_{Q\to Y}^{new}$ given the biological knowledge weights W with a probability

$$A_G = \min\{1, \frac{P(Y \mid G^{new}, X)P(G_Y^{new} \mid B, W)P(G_{Q\to Y}^{new} \mid X)}{P(Y \mid G^{old}, X)P(G_Y^{old} \mid B, W)P(G_{Q\to Y}^{old} \mid X)} \times \frac{R(G_Y^{old} \mid G_Y^{new})R(G_{Q\to Y}^{old} \mid G_{Q\to Y}^{new})}{R(G_Y^{new} \mid G_Y^{old})R(G_{Q\to Y}^{new} \mid G_{Q\to Y}^{old})}\}.$$

4. For each biological knowledge k,

   (a) Sample a new weight W_k^{new} for biological knowledge B_k from a weight proposal distribution $R(W_k^{new} \mid W_k^{old})$.

   (b) Accept the new biological weight W_k^{new} given the phenotype network G_Y with a probability

$$A_{W_k} = \min\{1, \frac{P(G_Y \mid W_k^{new}, W_{-k}^{old}, B)}{P(G_Y \mid W^{old}, B)} \frac{P(W_k^{new} \mid B)}{P(W_k^{old} \mid B)} \frac{R(W_k^{old} \mid W_k^{new})}{R(W_k^{new} \mid W_k^{old})}\}.$$

5. Iterate the steps 1-4 until the chain converges.

In step 1, a new phenotype network structure is proposed by a mixture of single edge operations (single edge addition, single edge deletion, single edge reversal) and edge reversal moves with orphaning [Grzegorczyk and Husmeier, 2008]. The edge reversal move with orphaning consists of selecting an edge $i \to j$, removing the parents of each node on the selected edge, sampling new parents of node i (including node j), and sampling new parents of node j, as long as it does not make a cycle. It has been shown that edge reversal moves can significantly improve the convergence of MCMC sampler [Grzegorczyk and Husmeier, 2008]. The proposal distribution puts the same probability, summing to 1, to the graphs that can be reached by a corresponding edge move.

In step 2, causal QTLs can be sampled conditional on the phenotypes' parents. There are several ways to sample causal QTLs. One way is a Bayesian QTL mapping proposed in Yi et al. [2005] for each phenotype. The prior distribution for the indicators of QTLs is $\prod w_k^{\gamma_k}(1 - w_k)^{\gamma_k}$ where $w_k = p(\gamma_k = 1)$ is the prior inclusion probability for the k^{th} QTL. We can use this independent prior for the prior distribution and the proposal distribution for a causal QTL. Another way

is the classical interval mapping of QTL for each phenotype conditional on its phenotypic parents. The classical interval mapping regresses a phenotype on a single QTL and picks every QTL over the significance threshold computed by permutations. Thus, this approach is deterministic as it chooses the same set of QTLs given the same set of parent phenotypes. It is a fast algorithm approximating the Bayesian mapping of QTL though it might fail to satisfy the irreducibility of the Markov Chain. We use the interval mapping for practical reasons.

In step 3, the computation of the ratio of marginal likelihoods, or Bayes factor, $P(Y \mid G^{new}, X)/P(Y \mid G^{old}, X)$, can be approximated by the difference of BIC scores [Kass and Raftery, 1995] when the sample size is large,

$$\frac{P(Y \mid G^{new}, X)}{P(Y \mid G^{old}, X)} \approx \exp(-\frac{1}{2}(BIC_{G^{new}} - BIC_{G^{old}})).$$

The BIC score is defined to be $-2 \log L + k \log n$ where L is the maximized value of the likelihood for the estimated model, k is the number of free parameters estimated, and n is the sample size.

In step 4, a new weight W_k^{new} can be sampled from a moving uniform distribution $U(W_k^{old} - 1, W_k^{old} + 1)$ and if the sampled W_k^{new} is less than 0, we take a negative of the new weight. This proposal distribution makes a ratio of proposal distributions, $R(W_k^{old} \mid W_k^{new})/R(W_k^{new} \mid W_k^{old})$, being 1. In addition, we need to compute

$$\frac{P(G_Y \mid W_k^{new}, W_{-k}^{old}, B)}{P(G_Y \mid W^{old}, B)} = \frac{\frac{\exp(-W_k^{new}\mathcal{E}_k(G_Y) - \sum_{k' \neq k} W_{k'}^{old}\mathcal{E}_{k'}(G_Y))}{Z(W_k^{new}, W_{-k}^{old})}}{\frac{\exp(-\sum_k W_k^{old}\mathcal{E}_k(G_Y))}{Z(W^{old})}},$$

where $Z(W) = \sum_{G_Y \in \mathrm{DAG}} \exp(-\sum_k W_k \mathcal{E}_k(G_Y))$ is a normalizing constant. Note that it is not feasible to compute the exact $Z(W)$ due to the exclusion of cyclic networks. We approximate the normalizing constant by the summation over directed graphs with restriction on the number of parents, e.g. 3 as adopted by Werhli and Husmeier [2007].

After running an MCMC chain, we need to efficiently summarize the chain for the inference of a network structure after discarding a transient burn-in period. The burn-in period can be arbitrary determined by looking at the chain which seems to be away from an equilibrium distribution. The choice by the highest posterior network structure might not produce a convincing model because the graph space grows rapidly with the number of phenotype nodes and the most probable network

structure might still have a very low probability. Therefore, instead of selecting the network structure with the highest posterior probability, we perform Bayesian model averaging [Hoeting et al., 1999] over the causal links between phenotypes, to infer an averaged network. Explicitly, let Δ_{uv} represent a causal link from u to v, that is, $\Delta_{uv} = \{Y_u \rightarrow Y_v\}$. Then

$$\begin{aligned} P(\Delta_{uv} \mid Y, X) &= \sum_G P(\Delta_{uv} \mid G, Y, X)\, P(G \mid Y, X) \\ &= \sum_G \mathbb{1}\{\Delta_{uv} \in G\}\, P(G \mid Y, X). \end{aligned}$$

The averaged network is represented by the causal links with maximum posterior probability or with posterior probability above a predetermined threshold, e.g. 0.5.

2.3.3 Summary of encoding of biological knowledge

In eqn (2.6), we have constructed a prior distribution on a network structure G_Y in terms of energy $\mathcal{E}(G_Y)$ relative to biological knowledge B. Now we describe how to encode a biological knowledge matrix B from several sources of biological information. Recall that B is an encoding meant to describe biological knowledge ranging from 0 to 1, and energy $\mathcal{E}(G_Y)$ is defined to be a distance measure between B and G_Y in eqn (2.5). When there is no available biological knowledge, we would put every element in B as $1/2$. Then all DAGs have the same energy and therefore the probability of a network structure conditional on W is $1/K$ with K as the number of all DAGs. In contrast, when biological knowledge is available, we will look at several ways of encoding biological knowledge into B : transcription factor and DNA binding [Bernard and Hartemink, 2005], protein-protein interaction [Jansen et al., 2003] and gene ontology annotations [Lord et al., 2003].

2.3.3.1 Transcription factor and DNA binding

Chromatin immunoprecipitation with microarray experiments is used to investigate the interaction of proteins and DNA *in vivo*. This technology has been employed to generate putative lists of transcription factor/target gene interactions [Lee et al., 2002]. Bernard and Hartemink [2005] suggested an approach to convert a p-value P_{ij}, quantifying the evidence that a transcription factor

i binds to a putative target gene j, into a posterior probability for the presence and directionally of an edge in a Bayesian network. Following Bernard and Hartemink [2005] we assume that the p-value P_{ij} follows a truncated exponential distribution with mean λ when the transcription factor i binds to a target gene j ($G_Y(i, j) = 1$) and a uniform distribution when the transcription factor does not bind to a target gene ($G_Y(i, j) = 0$).

$$P_\lambda(P_{ij} = p \mid G_Y(i, j) = 1) = \frac{\lambda e^{-\lambda p}}{1 - e^{-\lambda}},$$
$$P_\lambda(P_{ij} = p \mid G_Y(i, j) = 0) = 1.$$

The probability of the directed edge before observing any biological data is assumed to be $P(G_Y(i, j) = 1) = 1/2$ so that without any biological data, the probability of the presence of the edge only depends on the expression data. By the Bayes' rule, the probability of presence of an edge after observing a p-value is

$$P_\lambda(G_Y(i, j) = 1 \mid P_{ij} = p) = \frac{\lambda e^{-\lambda p}}{\lambda e^{-\lambda p} + (1 - e^{-\lambda})}.$$

Here λ is assumed to be uniformly distributed over the interval $[\lambda_L, \lambda_H]$ and the integration over λ is performed to obtain the probability of the presence of an edge,

$$P(G_Y(i, j) = 1 \mid P_{ij} = p) = \frac{1}{\lambda_H - \lambda_L} \int_{\lambda_L}^{\lambda_H} \frac{\lambda e^{-\lambda p}}{\lambda e^{-\lambda p} + (1 - e^{-\lambda})} d\lambda.$$

This can be solved numerically, for instance, by choosing λ in the range $[0, 10000]$. We should thus obtain the following estimate: $B(i, j) = P(G_Y(i, j) = 1 \mid P_{ij} = p)$.

2.3.3.2 Protein-protein interaction

Since protein-protein interaction is non directional, we put the same probability on both directions. If we do not consider the diverse reliabilities of protein-protein interaction from several experiments, we set $B(i, j)$ and $B(j, i)$ to be $\delta > 1/2$ when we find any interaction on any experiment. If there are gold standards for positive and negative protein-protein interactions, and experiments have diverse reliabilities, then we can use the Bayes classifier proposed by Jansen et al. [2003] to combine heterogeneous data. Positive gold standards are well-known true protein-protein interactions while negative gold standards are interactions which cannot happen such as a pair of

proteins in different subcellular compartments. An interaction experimental data set is a collection of observations over all pairs of proteins by binaries whether the interaction is present or absent for each pair. Suppose there are L interaction experimental data sets with different false positive rates. We can calculate the posterior odds of an interaction from binary observations $f_1, \ldots, f_L$,

$$\begin{aligned} O_{posterior} &= \frac{P(pos \mid f_1, \ldots, f_L)}{P(neg \mid f_1, \ldots, f_L)} = O_{prior} \times LR \\ &= \frac{P(pos)}{P(neg)} \times \frac{P(f_1, \ldots, f_L \mid pos)}{P(f_1, \ldots, f_L \mid neg)}. \end{aligned}$$

In the positive gold standard interactions, we can find a set of interactions which have the observed values $f_1, \ldots, f_L$. The likelihood under the positive gold standard can be defined to be the proportion of the set with the values $f_1, \ldots, f_L$ in the positive gold standard. Similarly we define the likelihood $P(f_1, \ldots, f_L \mid neg)$ under the negative gold standard. Then we can take the ratio of the two likelihoods to calculate the likelihood ratio (LR). The prior odds O_{prior} can be defined by an expert. The encoding of B can be obtained by transforming the posterior odds into a posterior positive rate:

$$B(i, j) = B(j, i) = \frac{O_{posterior}}{1 + O_{posterior}}.$$

When the posterior odds is equal to 1, $B(i, j)$ and $B(j, i)$ are equal to $1/2$. As the posterior odds increases, the values of $B(i, j)$ and $B(j, i)$ also increase.

2.3.3.3 Gene ontology

The Gene Ontology (GO) [Ashburner et al., 2000] is a well controlled vocabulary of terms describing the molecular functions, biological processes and cellular components of a gene. A GO ontology is structured as a directed acyclic graph where each node represents a GO term. The GO terms annotate a large fraction of genes. The distance between two genes can be defined in terms of their GO annotations. One well defined distance is Lord's similarity [Lord et al., 2003]. This measure takes into account the hierarchy of GO ontology and GO term occurrences in the myriad of genes. If two genes share a more specific GO term positioned in the lower part of the GO hierarchy, they are more likely to be similar. However, even if the shared GO terms lie in the same

level of the hierarchy, the frequencies of the GO terms in the whole genes are different and it affects the similarity. Consider that two GO terms c_1 and c_2 lie in the same level of the hierarchy. Suppose there are one hundred genes annotated with term c_1 and there are one thousand genes annotated with term c_2. Then the chance of two genes sharing the term c_2 is higher than the chance of sharing the term c_1. Therefore, it implies that the term c_1 is more informative. The information content $IC(c)$ for a GO term c is defined to be the negative logarithm of the number of times the term or any of its descendant terms occurs in the myriad of genes divided by the total occurrences of GO terms. The root of the hierarchy will have zero information content while the leaf of the hierarchy will have high information content. Once the information content $IC(c)$ for each node in the GO ontology is set up, we can define GO term similarity and gene similarity. The similarity between two GO terms is defined to be the maximum information content among the shared parents of the two terms, which is

$$sim(c_1, c_2) = \max_{c \in (pa(c_1) \cap pa(c_2))} IC(c).$$

Then, since a gene is annotated with a set of GO terms, the similarity between two genes g_1 and g_2 can be defined as the average similarities of all pairs of GO terms between two genes. That is,

$$sim(g_1, g_2) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} sim(c_{1,i}, c_{2,j})}{nm}.$$

This Lord's measure can be used as an encoding of B if it is rescaled to be in the interval $[0, 1]$.

2.4 Simulations

We performed a simulation study for comparing the proposed method (QTLnet-prior) with three other methods - QTLnet [Chaibub Neto et al., 2010a], WH-prior [Werhli and Husmeier, 2007], and Expression. Table 2.3 provides a summary of these four methods in terms of using the genetic variation information and biological knowledge. The QTLnet was implemented using R/QTLnet, the QTLnet-prior was implemented with prior setting on R/QTLnet, the WH-prior was programmed as in Werhli and Husmeier [2007] with a modification of approximating the marginal

likelihood with the BIC score instead of using the BGe score[4] [Geiger and Heckerman, 1994]. Expression was programmed by modifying R/QTLnet excluding QTL mapping.

Table 2.3: Four methods which differ in the use of genetic variation information and biological knowledge.

Method	Use of Genetic Variation Information	Use of Biological Knowledge
QTLnet-prior	YES	YES
QTLnet	YES	NO
WH-prior	NO	YES
Expression	NO	NO

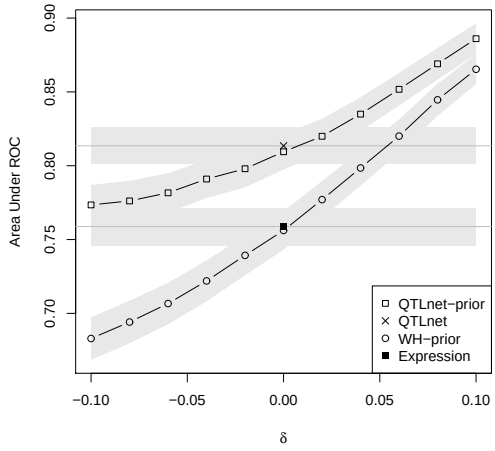
We simulated expression data and *a priori* knowledge matrix according to the network structure in Fig. 2.1 and produced 100 simulated data sets. To generate expression data based on the network in Fig. 2.1, the genetic information was simulated first. The genetic map described 5 chromosomes of 100 cM with 10 equally spaced markers in each chromosome and the markers were simulated for 500 mice in an F2 population using R/qtl [Broman et al., 2003]. We assumed QTL Q_t is located in the middle of chromosome t. Then, each expression data set of F2 population was generated with different genetic effects and partial regression coefficients between phenotypes. Genetic additive effects were sampled from a uniform distribution $U[0, 0.5]$ and dominance effects were sampled from $U[0, 0.25]$. The partial regression coefficients β_{ut} were sampled from $U[-0.5, 0.5]$. The residual phenotypic variance was 1. Biological knowledge matrix B was

[4]BGe stands for *B*ayesian metric for *G*aussian networks having score *e*quivalence. The BGe score is developed as a scoring metric for a Bayesian network of continuous variables under the assumption that the data is sampled from a multivariate Gaussian distribution. The BGe score is first derived for a complete Bayesian network where every pair of distinct nodes is connected by a direct edge. It assumes a prior on parameter to be a normal-Wishart distribution so that one can obtain a closed-form marginal distribution. Under the assumption of a parameter independence and modularity, the BGe score for an arbitrary Bayesian network is derived to be $p(Y \mid G) = \prod_{t=1}^{T} \frac{p(Y_t, Y_{pa(t)} \mid G_c)}{p(Y_{pa(t)} \mid G_c)}$ where G_c is any complete DAG model such that each node has the same parents as in G. It is known that Markov equivalent DAGs have the same BGe score. See Geiger and Heckerman [1994] for details.

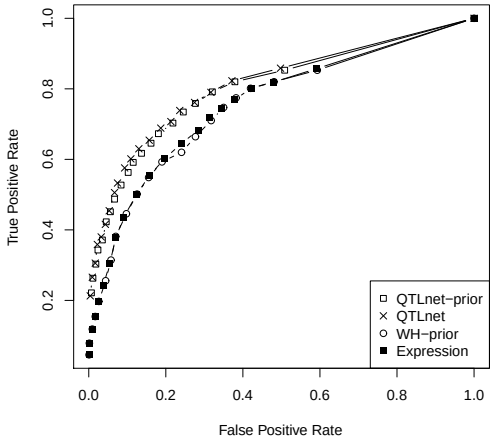
generated for several cases. The value $B(t,u)$ was generated from one of two $[0,1]$-truncated normal distributions $\mathcal{N}_{\pm}(0.5 \pm \delta, 0.1)$ [Geier et al., 2007]. The distribution is truncated at 0 and 1 to guarantee the value $B(t,u)$ ranging from 0 to 1. When no biological knowledge is available, the natural choice of $B(t,u)$ is 1/2. Consequently, the evidence for the presence (resp. absence) of edge $t \to u$ is necessarily specified through a value of $B(t,u)$ greater (resp. lower) than 1/2. In the simulation, the $B(t,u)$ value of true edge was generated from $\mathcal{N}_{+}$ and the $B(t,u)$ value of false edge was generated from $\mathcal{N}_{-}$. The parameter δ controls the accuracy of prior knowledge. We denote the generated biological knowledge to be positive knowledge, non informative knowledge or negative knowledge based on the sign of δ: $+, 0, -$, respectively. We examined eleven cases of different accuracies of prior knowledge: $\delta \in \{\pm 0.1, \pm 0.08, \pm 0.06, \pm 0.04, \pm 0.02, 0\}$. In the extreme case when δ is equal to 0.5, the prior knowledge almost correctly reflects the network structure while when δ is equal to -0.5, the prior knowledge is incorrectly reflecting the network structure almost in the opposite way. When δ is equal to 0, the information is generated with no distinction between true and false edges. For each simulated data set, we ran a Markov chain Monte Carlo for 30300 iterations, discarded the first 300 iterations, sampled every 10 iterations, and generated 3000 samples.

We assessed these four methods by using receiver operator characteristic (ROC) curves of the proportion of recovered and spurious edges. Bigger areas under the ROC curve generally indicate better performance, as the area represents the probability that the classifier ranks true edges higher than false edges [Fawcett, 2006]. The ROC curves are obtained from the set of proportions of recovered edges and spurious edges for various posterior probability thresholds ranging from 0 to 1.

If we are more interested in getting the true edges, precision-recall curves are useful. Since the area under ROC curve could be high if the false edges are not included frequently in the network but the true edges are not so much included. Precision-recall curves compare the proportion that the inferred edge is true (precision) and the proportion that true edge is inferred (recall). Figure 2.5 shows that as positive knowledge is incorporated, the performance to get the true edges is improved. In addition, QTLnet-prior works better than WH-prior in estimating the true edges.



(a) The areas under ROC curves of QTLnet-prior, QTLnet, WH-prior, and Expression. The areas under ROC curves of QTLnet-prior and WH-prior are plotted against the accuracy of biological knowledge, δ. Since QTLnet and Expression do not incorporate biological knowledge, they are plotted in a single point each (×, ■). The shaded area indicates the standard error of the area under ROC curve.



(b) The ROC curves of QTLnet-prior and WH-prior are drawn when non informative biological knowledge ($\delta = 0$) is incorporated. They are compared with the ROC curves of QTLnet and Expression which do not incorporate biological information.

Figure 2.4: The comparison of four methods by the area under the ROC curves with respect to the accuracy of biological knowledge.

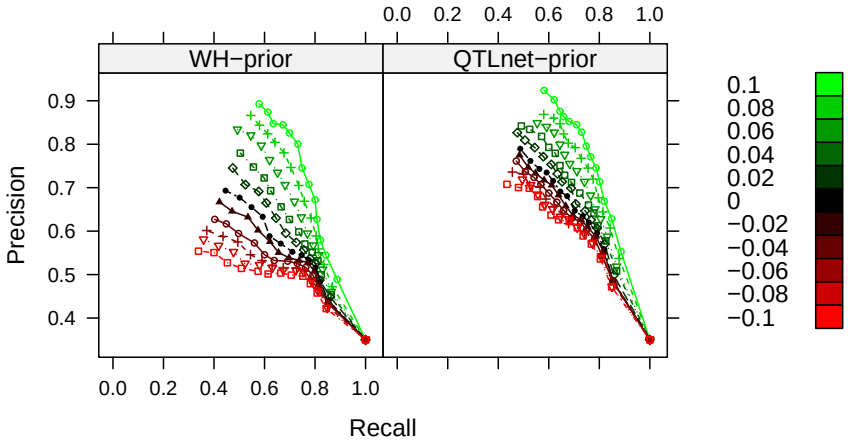


Figure 2.5: Precision-recall curves show the trade-off between the probability that an inferred edge is the true edge and the probability that a true edge is included. As the biological knowledge gets positively informative, the precision-recall curve moves to the upper right which means good performance. The performance is better when the network is estimated by QTLnet-prior compared to the network by WH-prior.

First, we evaluate the effect of incorporating genetic variation information. The effect of QTL mapping can be tested by comparing QTLnet-prior and WH-prior. QTLnet-prior is more effective in recovering the network structure than WH-prior in Fig. 2.4a and we can conclude that QTL mapping increases the effectiveness. Better causal QTLs can be inferred by conditioning on the phenotype network. Similarly, a better phenotype network can be inferred by conditioning on causal QTLs. The gain is more emphasized when the biological knowledge is negative.

Second, we evaluate the effect of incorporating biological knowledge. In Fig. 2.4a, when δ is positive, QTLnet-prior performs better than QTLnet and WH-prior performs better than Expression, whereas when δ is negative, QTLnet-prior performs worse than QTLnet and WH-prior performs worse than Expression. With a positive δ, as the accuracy of knowledge increases, QTLnet-prior and WH-prior benefit by the prior knowledge incorporation. However, a negative δ, indicating that the knowledge disagrees with the true network structure, makes QTLnet-prior and WH-prior be harmed by prior knowledge incorporation. The decreased performances in QTLnet-prior and WH-prior bring in the attention whether W can effectively control the influence of negative knowledge. Figure 2.6 shows that the median of W in the posterior sample is close to 0 with negative knowledge. It implies that the weight W can effectively control the use of negative knowledge to some extent but not completely, based on the decreased recovery observed in comparison to the case of non informative knowledge evidenced in Fig. 2.4a. In comparison with QTLnet and Expression, the reduced performance of QTLnet-prior and WH-prior can be explained by the remaining uncontrolled effect of prior probability incorporating negative knowledge. When non informative knowledge is incorporated, there is no significant difference in area under ROC curve between QTLnet and QTLnet-prior (p-value=0.82) and between Expression and WH-prior (p-value=0.89) as shown in Fig. 2.4a and Fig. 2.4b.

2.5 Analysis of yeast cell cycle genes

We applied QTLnet-prior to reconstruct a network of 26 genes involved in the cell cycle in yeast (*Saccharomyces cerevisiae*), previously chosen by Bernard and Hartemink [2005] for cell

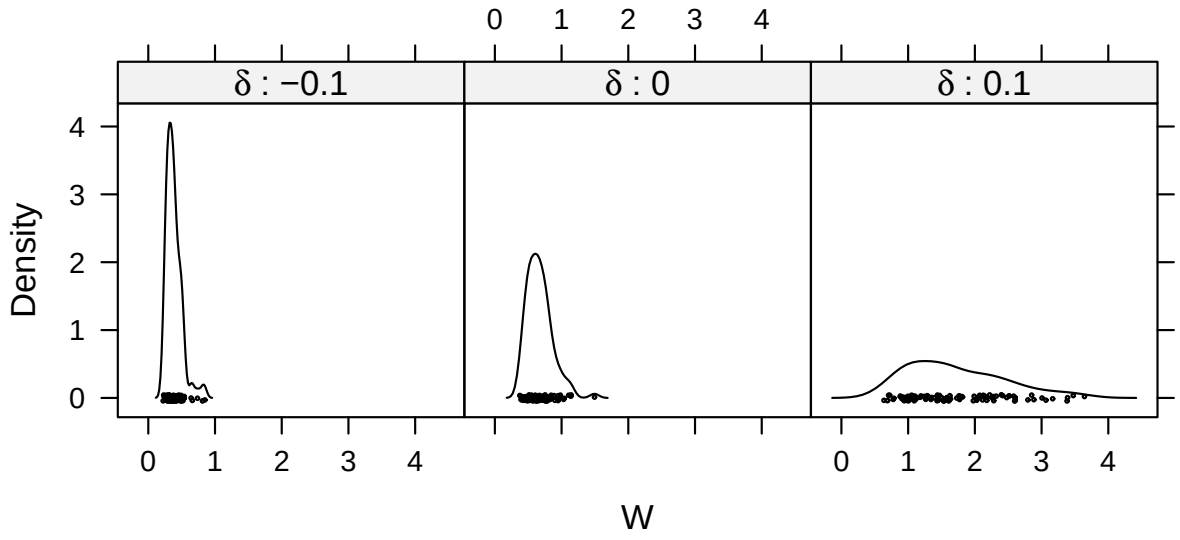


Figure 2.6: The distribution of median weight W of posterior sample by QTLnet-prior inference. Each panel shows the median W distribution when biological knowledge is defective ($\delta = -0.1$), non informative ($\delta = 0$), and informative ($\delta = 0.1$).

cycle network analysis with time-dependent expression data and transcription factor binding information. Genes express periodically by cell cycle phases (genome duplication phase, gap phase 2, cell division phase and gap phase 1) and there are transcription factors to regulate some periodical genes [Bähler, 2005]. The gene expression data and genetic variation information were obtained from a backcross population of 112 segregates between BY4716 and RM11-1a [Brem and Kruglyak, 2005]. Brem and Kruglyak [2005] extracted the gene expression data by constructing a backcross, isolating RNA, and hybridizing cDNA to microarrays. They also genotyped the population at 2957 genetic markers for genetic variations. In addition to gene expression data and genetic variations, we incorporated transcription factor binding information as biological knowledge for QTLnet-prior analysis. The p-value for evidence of transcription factor binding from chromatin immunoprecipitation with microarray experiments is available for 106 transcription factors from Lee et al. [2002]. For the 26 genes in our analysis, 11 of them are transcription factors (TF) and the rest are known targets of one or more transcription factors. We transformed the p-values into biological knowledge matrix B as described in Section 2.3.3.

The construction of the causal network focused on the 26 phenotypes of 112 yeast segregates, incorporating genetic variation information at 2957 markers and biological knowledge of transcription factor binding. We ran an MCMC for 760,000 iterations, discarded the first 200,000 iterations, sampled every 100 iterations, and finally got 5,600 samples used for estimation. The computation took around 14 days of CPU time on a 2.66GHz Intel(R) Core(TM)2 Quad running Red Hat 4.1.2-50. To examine the mixing and convergence of the MCMC chain, we first computed the autocorrelation of BIC scores and autocorrelation of W, respectively. As shown in Fig. A.2 in the appendix, both autocorrelation values get close to 0. It indicates that the MCMC chain may not suffer from a slow mixing rate. Furthermore, we calculated Geweke's convergence statistics [Geweke, 1992] to check the convergence of the Markov chain. The Geweke's statistics is asymptotically $\mathcal{N}(0,1)$ when it is equal for the two means of the first 10% and the last 50% parts of the Markov chain. The Geweke's statistics on BIC score is 0.34 and is -0.25 on W, suggesting the convergence of the chain. Figure 2.7 shows the causal phenotype network reconstructed by

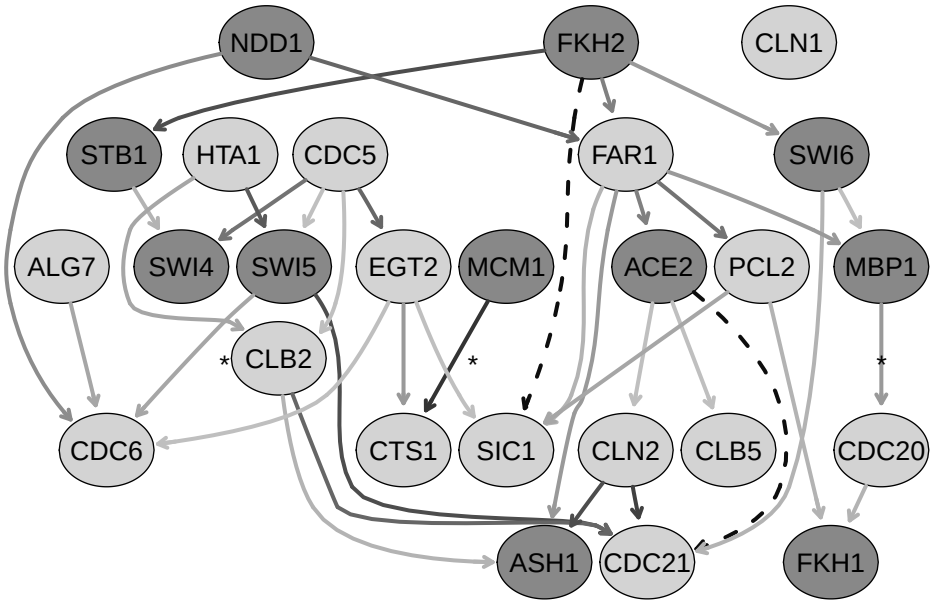


Figure 2.7: Yeast cell cycle phenotype network by QTLnet-prior, integrating transcription factor binding information. A solid edge is the inferred edge with its posterior probability over 0.5 and the darkness of the edge is in proportion to the posterior probability. Dark nodes are transcription factors. The edge consistent with transcription factor binding information is marked with a star, $*$. The TF binding relation recovered by an indirect path in the inferred network is represented by a dashed edge.

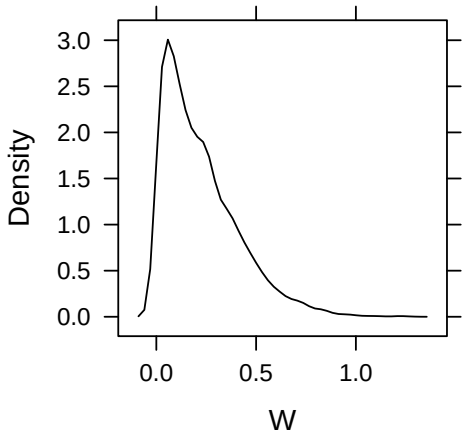


Figure 2.8: The posterior distribution of weight *W* of transcription factor information in reconstructing a yeast cell cycle network by QTLnet-prior.

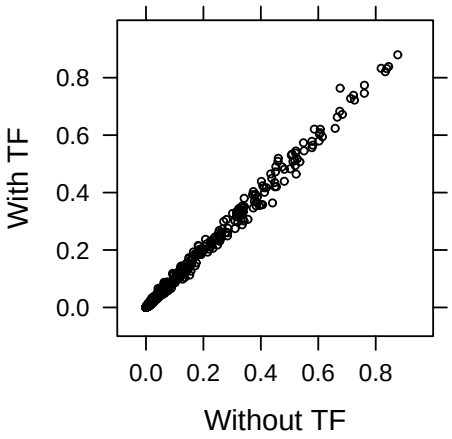


Figure 2.9: Comparison of the posterior probability of every possible directed edge between the network inferred by QTLnet-prior and the network inferred by QTLnet.

QTLnet-prior. The full network of phenotypes and causal QTLs can be found in Fig. A.1 in the appendix.

In the TF biological knowledge matrix B, we defined a pair (i.e. an edge from node i to node j) to be *significant* if its $B(i, j)$ value is over 0.5. There are 44 significant TF pairs in the TF biological knowledge B. For the constructed network with 36 inferred edges as in Fig. 2.7, we found 3 significant direct TF pairs (MBP1 → CDC20, SWI5 → CDC6 and MCM1 → CTS1) and 2 significant indirect TF pairs (ACE2 → CDC21 and FKH2 → SIC1). Interestingly, we did not find any reverse relations – causal relation from target gene to TF – in the inferred network. The remaining 39 TF causal relations in B were not inferred in the phenotype network.

Figure 2.8 shows the posterior density distribution of the weight W of TF information, which has a mode of approximately 0 with right skewness. To further examine the contribution of TF information on phenotype network reconstruction, we applied QTLnet to construct the phenotype network without using TF information. For the two networks inferred by QTLnet-prior and QTLnet, the posterior probability of every possible directed edge is very similar to each other as shown in Fig. 2.9. Although the TF knowledge did not improve the reconstruction of the cell cycle network, it did not have a negative impact on reconstruction, either. The weight parameter was actually effective in protecting the network reconstruction against the inconsistent TF information. The inconsistency between TF information and expression data may be due to any of the following reasons: (1) inconsistency between physical regulation of transcription binding and transcriptional regulation level of expression changes; (2) necessary post-translational modification of TF or construction of TF complex with other proteins for regulation of target genes; (3) cell cycle phase or tissue dependent TF binding or false transcription binding information; (4) incapability of capturing cyclicity of the cell cycle network from static expression data of a single time point.

2.6 Conclusion

We have developed a phenotype network inference method (QTLnet-prior) to incorporate genetic variation information and biological knowledge. Genotypes are known to control phenotypes but not the other way and thereby can help to distinguish phenotype network structures. Biological

knowledge can improve the clustering and directional inference between phenotypes. The simulation study shows that the proposed method can improve the reconstruction of the gene network by integrating genetic variation information and biological knowledge as long as knowledge agrees with data. When biological knowledge does not agree with data, the weight of knowledge controls the contribution of prior probability of biological knowledge on the likelihood of data, reducing (to some extent) the negative impact of the defective knowledge. We applied QTLnet-prior to estimate a yeast cell cycle network of 26 genes with causal QTLs by integrating transcription factor binding information, and compared its performance to QTLnet. The distribution of weight suggests that TF binding information was inconsistent with expression data. Nonetheless, comparison with QTLnet's output showed fairly similar result, suggesting the weight parameter of knowledge was effective in controlling the negative impact of inconsistent knowledge in this case.

When we interpret the inferred networks, we need to be cautious. Even though, in theory, the incorporation of causal QTLs allows us to distinguish network structures that would otherwise be likelihood equivalent, in practice some of the detected expression-to-expression causal relationships might be invalid. The problem is that the inferred expression network represents a projection of real causal relationships that might take place outside the transcriptional regulation level. For instance, the true causal regulations could be due to transcription factor binding, direct protein-protein interaction, phosphorylation, methylation, etc. and might not be well reflected at the gene expression level. The incorporation of diffused biological knowledge, mined from different levels of biological regulation, could potentially improve the reconstruction of gene-expression regulatory networks. In any case, the inference of these networks can still play an important role in generating hypothetically possible causal relations.

There are several factors that could change the inference by QTLnet-prior. One is the prior distribution specification. We have used the Gibbs distribution as a prior distribution for network structures in eqn (2.6) in terms of an absolute distance measure in eqn (2.5) to incorporate biological knowledge. The exponential distribution is used for the weight of biological knowledge in eqn (2.7) with the rate parameter (see Section 2.3.1). However, we could consider different

choices of network structure distributions, measures to incorporate information, weight distributions, and hyperparameters. Another factor is the sample size of expression data. As the sample size increases, the contribution of biological knowledge will be generally reduced. This shows the limited contribution of biological knowledge on the reconstruction of networks, even though biological knowledge B can also be obtained from a number of experiments as discussed in Werhli and Husmeier [2007]. The third factor is the global control of biological knowledge on network reconstruction. Illustrated by the yeast cell cycle network, every TF/target regulation was controlled by the same weight parameter. It may have resulted in no contribution of any biological knowledge even though 5 TF/target regulations were inferred to be consistent with expression data. This suggests incorporation of biological knowledge by local control parameters when reconstructing a network. Finally, the encoding of biological knowledge plays an important role. We have proposed to use the encoding for transcription factor and its targets by Bernard and Hartemink [2005], protein-protein interaction by Jansen et al. [2003], and gene ontology annotations by Lord et al. [2003]. These encodings are mainly about direct relationships in separate biological regulation levels. As discussed in the previous paragraph, this diffused biological knowledge can improve the Bayesian network reconstruction.

There are shortcomings of QTLnet-prior framework inherited from QTLnet. One of the assumptions of QTLnet is no presence of latent variables. Latent variables can make it impossible to find the marginalized model in the class of DAG as shown in Richardson and Spirtes [2002] and can induce erroneous relations. Suppose there are three nodes y_1, y_2, y_3, and y_1 and y_2 have a common parent c_1 while y_2 and y_3 have a common parent c_2. If the common parents c_1 and c_2 are not observed, we obtain the following independence relations: $y_1 \perp y_3$ and $y_1 \not\perp y_3 \mid y_2$. Then we mistakenly infer that y_1 and y_3 are parents of y_2. To address this problem, one can consider the more general class of ancestral graphs, which takes care of latent variables. Ancestral graphs open up the possibility of latent variables while they do not explicitly include the latent variables in the network structures [Richardson and Spirtes, 2002].

A persistent challenge in Bayesian network analysis is to cope with large networks since the DAG space size grows super-exponentially with the number of nodes. Approaches based on

Markov blankets with and without restrictions on the number of parent nodes have been proposed [Riggelsen, 2005, Schmidt et al., 2007, Perrier et al., 2008]. Jaakkola et al. [2010] approximated the Bayesian network problem to a linear programming problem. Tamada et al. [2011] developed a parallel algorithm that infers subnetworks restricted on a Markov blanket and merges the subnetworks. Likewise, in phylogeny estimation, the supertree reconstruction from small trees has been studied [Bininda-Emonds et al., 2002]. We think the rigorous development of super Bayesian network methodology to integrate small subnetworks is a promising direction to infer a large network since the inference of small subnetworks is computationally inexpensive and multiple subnetworks can be parallelized for computation. In the era of vast biological data and knowledge in various aspects, integrating them reasonably in a large scale can be an interesting topic for future research.

Chapter 3

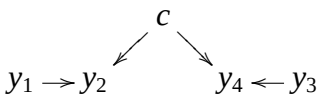
Causal network incorporating genetic variations with latent phenotypes

3.1 Motivation

In the previous chapter, it was discussed that Bayesian network inference assumes that all variables are included. In other words, there are no latent variables, no hidden variables, or no unmeasured variables. This assumption may not hold because data cannot be always obtained for all the constituents of the network. For example, a microarray measures genome-wide mRNA expression levels and we could reconstruct a gene network on mRNA gene expressions. However, it is possible that some mRNA levels are governed by other molecules' levels such as metabolites, proteins and signaling molecules. In this case, even if we have the genome-wide expression levels, the important variables are missing when constructing a network. Another reason is that we may focus on a subset of observed variables to be computationally practical in inferring a network. For instance, QTL mapping on expression levels of mRNAs provides us a list of genes that are controlled by the same genomic loci, which co-maps with interesting clinical traits such as insulin levels, glucose levels, and so on. To decipher how the insulin level is controlled by genetic variations and mRNA levels, we can reconstruct a gene network with genes co-mapped with the insulin level. However, it is not always computationally feasible to reconstruct a causal gene network on hundreds of co-mapped genes and in practice, we take a subset of genes to reconstruct a gene network.

Then, what happens to the inference of networks on observed variables when some variables in the true data-generating Bayesian network are not measured? The answer is that the class of

Bayesian networks is not sufficient to properly represent the conditional independence relations on observed variables and we need to consider an extension of Bayesian networks. In the following Bayesian network, suppose y_1, y_2, y_3 and y_4 are observed but c is unmeasured. The conditional



independence relations on observed variables are like these:

$$y_2 \not\perp y_4 | \{y_1, y_3\}$$

$$y_1 \not\perp y_2 \qquad\qquad y_3 \not\perp y_4$$

$$y_1 \perp y_4 \qquad\qquad y_3 \perp y_2$$

$$y_1 \not\perp y_4 | y_2 \qquad\qquad y_3 \not\perp y_2 | y_4.$$

The conditional independence relations on y_1, y_2, y_4 can orient edges to be $y_1 \rightarrow y_2 \leftarrow y_4$ while the conditional independencies on y_2, y_3, y_4 can orient edges to be $y_3 \rightarrow y_4 \leftarrow y_2$. These orientations have conflicting edges between y_2 and y_4, and hence there is no Bayesian network representing the conditional independencies on observed variables y_1, y_2, y_3 and y_4. Therefore, we consider a class of ancestral graph Markov models, an extension of the class of DAG Markov models (Bayesian networks), to properly represent conditional independencies on observed variables.

An ancestral graph Markov model is a graphical independence model whose graph consists of vertices and three kinds of edges between vertices – directed, undirected and bidirected edges [Richardson and Spirtes, 2002]. Each vertex corresponds to a variable and each edge shows the relationship between variables. Directed edges are associated with causal relations as in DAGs, bidirected edges are associated with marginalization over latent variables and undirected edges are associated with conditioning on selection variables. Conditioning on selection variables means that the data is collected by the selecting criteria of the selecting variables. This data is sampled on a selected sub-population, but not on the whole population. The details of ancestral graphs will be described in Section 3.2.

Here we focus on the cases when there could be latent variables between phenotypes but not selecting variables, and hence we restrict a phenotype network to directed ancestral graphs whose edges are either directed or bidirected. We extend the phenotype network with genotypes and show that the extended network is an ancestral graph. It is proved that the addition of QTLs helps to distinguish Markov equivalent phenotype networks. We model the extended network by recursive linear equations with correlated Gaussian errors. We develop a Markov chain Monte Carlo (MCMC) algorithm to search over the extended phenotype networks allowing latent variables

The details of ancestral graph inference for phenotypes and genotypes are organized as follows. In Section 3.2 we briefly introduce the definition, terminology and properties of ancestral graphs. Section 3.3 formulates a statistical model of a gene network with genotypes represented by an ancestral graph and presents its graphical and statistical properties. Section 3.4 proposes an MCMC method to infer a directed ancestral graph of phenotypes and QTLs. Section 3.5 conducts the simulation study of the proposed methods for gene network inference and Section 3.6 applies the developed method on F2 mice data. Finally, in Section 3.7, we discuss the pros and cons of our approach.

3.2 Ancestral graph

A graphical independence model is a model to represent the independence relations entailed by a graph. The independence relations are obtained by applying a separation criterion to a graph, which produces a global Markov property. For example, a Bayesian network is a directed acyclic graph (DAG) Markov model and its independencies are entailed by applying d-separation on the DAG as defined in Chapter 2. Likewise, the independence relations of an ancestral graph is entailed by m-separation criterion. We will first describe terminology, define an ancestral graph and explain the m-separation. Note that this section summarizes the paper of Richardson and Spirtes [2002] which introduced and defined ancestral graphs and proved several properties of ancestral graphs.

Terminology In an ancestral graph G, three types of edges describe the relationship between vertices.

$$\text{If } \begin{pmatrix} x \to y \\ x \leftarrow y \\ x - y \\ x \leftrightarrow y \end{pmatrix}, \text{ then } x \text{ is a } \begin{pmatrix} parent \\ child \\ neighbor \\ spouse \end{pmatrix} \text{ of } y.$$

The set of parents of y is denoted by $pa(y)$, the set of neighbors of y is denoted by $ne(y)$, and the set of spouses of y is denoted by $sp(y)$. A *path* between two vertices is a sequence of distinct adjacent vertices. A vertex x is an *ancestor* of a vertex y, $x \in an(y)$, if either there is a directed path $x \to \cdots \to y$ or $x = y$. In an ancestral graph, x is an *anterior* of y, $x \in ant(y)$, if there is an *anterior path* from x to y which is either $x - \cdots - y$, $x \to \cdots \to y$, or $x - \cdots - \to \cdots \to y$. Note that an anterior path cannot be in the form of $x \to \cdots \to - \cdots - y$ because it will be contradictory to the definition of an ancestral graph in the next paragraph. A *collider* at η on the path is defined as $\to \eta \leftarrow$, $\leftrightarrow \eta \leftrightarrow$, $\leftrightarrow \eta \leftarrow$, $\to \eta \leftrightarrow$, shortly, $\circ\!\!\to \eta \leftarrow\!\!\circ$, where $\circ\!\!\to$ can be either $\leftrightarrow$ or $\to$ and $\leftarrow\!\!\circ$ can be either $\leftrightarrow$ or $\leftarrow$. Vertices other than a collider are defined to be non-colliders.

Definition 3.1 (Definition 2.1 in Ali et al. [2009])**.** An *ancestral graph* is a graph whose vertices are connected by at most one of undirected (—), directed (→) or bidirected (↔) edges, holding the following conditions:

1. there are no directed cycles;
2. whenever there is an edge $x \leftrightarrow y$, then there is no directed path from x to y or from y to x;
3. if there is an undirected edge $x - y$ then x and y have no spouses or parents.

The third condition makes the configurations $x - y \leftrightarrow z$ and $x - y \leftarrow z$ impossible. We can split all vertices V in an ancestral graph G into a set of vertices with no arrowheads pointing to it (un_G) and the remaining vertices ($V \setminus un_G$). Explicitly, let $un_G = \{x \mid pa(x) \cup sp(x) = \emptyset\}$. The schematic view of an ancestral graph separated by un_G is like in Figure 3.1.

An edge takes a specific form whether the endpoints of the edge is in un_G or not. If both endpoints x and y of an edge are in un_G, then the edge should be $x - y$. If $x \in un_G$ and $y \in V \setminus un_G$,

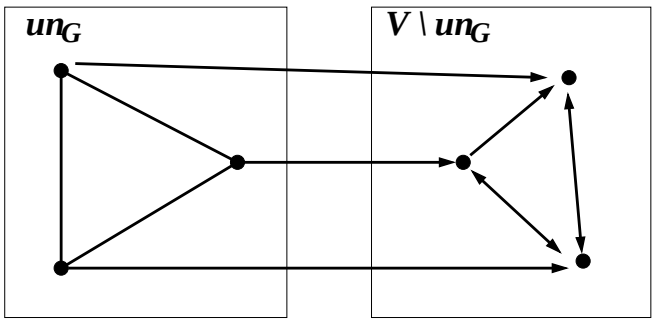


Figure 3.1: Schematic view of an ancestral graph. Reproduced from from Fig. 4 in Richardson and Spirtes [2002].

then the edge takes the form $x \rightarrow y$. This decomposition by un_G will be used in parameterization of an ancestral graph.

m-separation The separation criterion for an ancestral graph is *m-separation*. A path π between x and y along graph G is m-connected given Z, with $x, y \notin Z$, if

1. every non-collider on the path π is not in Z, and
2. every collider on the path π is in $ant(Z)$.

If there is no m-connecting path between x and y given Z, then x and y are said to be m-separated given Z. By applying m-separation on ancestral graph G, we can read off a set of independencies $\mathcal{J}(G)$ such that $< X, Y \mid Z >\in \mathcal{J}(G)$ interpreted as X and Y are independent conditional on Z. The second condition is equivalent to "every collider on the path is in $an(Z)$" because a vertex η in $ant(Z)$ but not in $an(Z)$ is connected to a vertex by an undirected edge and then the vertex η cannot have any parents or spouses by the third condition in the definition of an ancestral graph, which thereby cannot form a collider.

We have presented the definition of an ancestral graph, its special property and m-separation criterion. Now we will briefly show how marginalization of latent variables and conditioning on selection variables are associated with an ancestral graph.

Marginalization and Conditioning The ancestral graph is motivated by the case in which the data generating process is a directed acyclic graph (DAG) model but only a subset of variables are observed (corresponding to marginalization) and there are selection effects (corresponding to conditioning) [Richardson and Spirtes, 2002]. As an example of the effect of marginalization over latent variables, consider a network $X \leftarrow C \rightarrow Y$ where C is latent and X and Y are observed. The variables X and Y are independent conditional on C in the underlying network while we cannot recover the conditional independence of X and Y from the observed variables only. The marginalization over C will transform the graph $X \leftarrow C \rightarrow Y$ into $X \leftrightarrow Y$. As an example of the effect of conditioning on selection variables, suppose that we want to investigate whether the hours

of workout (W) and the amount of food eaten (F) are related. Let the truth to be that W and F affect the body weight B respectively and W and F are independent: $W \rightarrow B \leftarrow F$. If the data is collected randomly, we would observe the independence between W and F. However, if units in the data are sampled depending on some criteria such as sampling only overweight people, $B >$ 300lbs, in other words, the data is collected conditional on the selecting variable B, a misleading conclusion will be made that W and F are inversely dependent. Conditioning on the selecting variable B will transform the graph into W — F.

These changes of dependencies by marginalization and selection are handled by forming an ancestral graph with bidirected edges and undirected edges, respectively. We will show how a DAG is transformed after marginalization and conditioning.

Theorem 3.2 (Richardson and Spirtes [2002])**.** *Let G be an ancestral graph with vertex set V and L and S are arbitrary disjoint sets in V. After marginalizing out the vertices in L and conditioning on the vertices in S, G is transformed into the ancestral graph $G[^S_L$ with vertex set $V \setminus (S \cup L)$. The edge in $G[^S_L$ between x and y is present if*

$$x \not\perp\!\!\!\perp y \mid Z \cup S, \quad \text{for all } Z \text{ such that } Z \subset V \setminus (S \cup L \cup \{x, y\}).$$

Its edge type is specified by the following way:

$$\text{If} \quad \begin{pmatrix} x \in ant(\{y\} \cup S); y \in ant(\{x\} \cup S) \\ x \notin ant(\{y\} \cup S); y \in ant(\{x\} \cup S) \\ x \in ant(\{y\} \cup S); y \notin ant(\{x\} \cup S) \\ x \notin ant(\{y\} \cup S); y \notin ant(\{x\} \cup S) \end{pmatrix} \quad \text{then} \quad \begin{pmatrix} x - y \\ x \leftarrow y \\ x \rightarrow y \\ x \leftrightarrow y \end{pmatrix}.$$

Proposition 4.13 in Richardson and Spirtes [2002] shows that the transformed ancestral graph from a DAG will have undirected edges only when there are selection variables S. Likewise, Proposition 4.14 shows that bidirected edges will only be present if there are latent variables L. These two propositions justify the statement that bidirected edges are associated with marginalization over latent variables and undirected edges are associated with conditioning.

By Theorem 3.2, we can obtain the transformed ancestral graph $G[^S_L$ after marginalizing over L and conditioning on S, and then we can obtain its independence relations $\mathcal{J}(G[^S_L)$ by applying

m-separation on $G[^S_L$. Another track to obtain the marginalized and conditioned independence relations would be by marginalizing out L and conditioning on S from the independence relations of the original graph G. The set of independence relations $\mathcal{J}(G)[^S_L$ after marginalizing out L and conditioning on S is defined to be:

$$\mathcal{J}(G)[^S_L = \{< X, Y \mid Z > \mid < X, Y \mid Z \cup S > \in \mathcal{J}; < X \cup Y \cup Z > \cap (S \cup L) = \varnothing\}.$$

Theorem 4.18 in Richardson and Spirtes [2002] shows that the independence relations $\mathcal{J}(G[^S_L)$ entailed by the transformed ancestral graph $G[^S_L$ are the same as the independence relations $\mathcal{J}(G)[^S_L$ after marginalizing and conditioning the independencies $\mathcal{J}(G)$ of the original graph G.

We have described fundamentals of ancestral graphs so far. Unlike a DAG, a missing edge in an ancestral graph does not mean an existence of conditional independencies. We will introduce a subclass of ancestral graphs where a missing edge corresponds to at least one set of conditional independencies.

Maximal ancestral graph A graph is defined to be *maximal* if for every pair of vertices (x, y) that are not adjacent in G, there is a set Z such that $x \perp y \mid Z$ for $x, y \notin Z$. Hence, a missing edge in a maximal ancestral graph means that the corresponding two vertices are independent by conditioning on some set. A way to make a maximal ancestral graph from a non-maximal ancestral graph G is by constructing $G[^{\varnothing}_{\varnothing}$ because $G[^S_L$ is always maximal. This will add bidirected edges without changing independence relations of G. Having the same set of independencies means that two graphs are *Markov equivalent*. Therefore, G and $G[^{\varnothing}_{\varnothing}$ are Markov equivalent. Even two different maximal ancestral graphs can be Markov equivalent to each other. By augmenting missing dependent edges with bidirected edges, a maximal ancestral graph has the following pairwise Markov property: If there is no edge between x and y in G, then

$$x \perp y \mid ant(\{x, y\}) \setminus \{x, y\}.$$

In a non-maximal ancestral graph, two non-adjacent vertices, for which no m-separating set Z exists, will be joined by an inducing path. An *inducing path* π between vertices x and y in an

ancestral graph G is a path on which every vertex other than x and y is both a collider on π and an ancestor of at least one of x or y.

We will link an ancestral graph with probability distributions. Richardson and Spirtes [2002] suggested a natural Gaussian parameterization via recursive equations with correlated errors.

Gaussian Parameterization A probability density of an ancestral graph G can be factorized into the undirected component and the remaining component,

$$P(Y_V) = P(Y_{un_G})P(Y_{V \setminus un_G} | Y_{un_G}).$$

This factorization is enabled by the decomposition of an ancestral graph by the subgraph of un_G and the remaining subgraph, mentioned after the definition of an ancestral graph, diagramed in Figure 3.1. The undirected subgraph of un_G can be parameterized by an undirected graphical Gaussian model [Dempster, 1972]. Alternatively, un_G can be defined to be a set of vertices connected by undirected edges, $\{x \mid ne(x) \neq \emptyset\}$, and the factorization still works.

The remaining subgraph can be parameterized by a set of recursive equations in a Gaussian distribution family as follows. First, for directed edges,

$$Y_t = \mu_t + \sum_{v \in pa(t)} \beta_{tv} Y_v + \epsilon_t.$$

Second, for bidirected edges:

$$\text{If there is no edge in } t \leftrightarrow s \text{ in } G, Cov(\epsilon_t, \epsilon_s) = 0.$$

$$\text{Otherwise, } Cov(\epsilon_t, \epsilon_s) = 0.$$

Finally, we will briefly show the difference in the inference of DAGs and ancestral graphs. When there is no noise in the data, an algorithm to infer a DAG correctly is PC algorithm [Spirtes et al., 2000]. Similarly, an algorithm to infer an ancestral graph correctly is FCI algorithm [Spirtes et al., 2000]. Both algorithms build graphs from conditional independencies and begins with undirected graphs that are fully connected. They first infer the skeleton by sequentially removing edges if

there exists a set that makes the two nodes to be conditionally independent. The key element to infer directions is by the relations that *v*-structure specifically have. The PC algorithm orients $X — Y — Z$ with nonadjacent X and Z as $X \rightarrow Y \leftarrow Z$ if and only if Y does not make X and Z to be conditionally independent. The FCI algorithm orients the relaxed version ($X * — * Y * — * Z$) to be $X* \rightarrow Y \leftarrow *Z$ where $*$ can be either an arrowhead, tail, or no information. This way, ancestral graphs can have bidirected edges. Another difference between PC algorithm and FCI algorithm is that additionally FCI should consider the conditioning sets that are connected to the two nodes but not directly. Further explanation can be found in Spirtes et al. [2000].

when there is no noise in the data differ differences in inferring DAGs and ancestral graphs. algorithmic differences to get DAGs and ancestral graphs. we will show differences in algorithms which correctly infer DAGs and ancestral graphs respectively when there are no noises in the data. One algorithm to infer DAGs correctly is PC algorithm [Spirtes et al., 2000] and one algorithm to infer ancestral graphs correctly is FCI algorithm [Spirtes et al., 2000]. Both algorithms test conditional independencies and begin with complete graphs where all nodes connected by undirected edges. The PC algorithm works like this.

1.

3.3 Ancestral graph for phenotypes and genotypes

We will develop a statistical model for a network of phenotypes and genotypes based on ancestral graphs in Section 3.3.1. The parametric family of the proposed statistical model is determined in Section 3.3.3 and the graphical properties of the ancestral graph of phenotypes and genotypes are established in Section 3.3.2. Lastly, Section 3.3.4 conjectures that the Markov equivalence of two MAGs implies the same set of distributions corresponding to each graph in a Gaussian distribution and the parametric family of the proposed model.

3.3.1 Model

In a traditional experimental cross study, QTL mapping is done to identify genetic locations (QTL) associated with a phenotype trait Y. To refine QTL positions, pseudomarkers Q are augmented from flanking markers m with recombination rates [Broman, 2001]. A phenotype is often modeled with pseudomarker genotypes Q such that

$$P(Y, Q \mid m) = P(Y \mid Q, m)P(Q \mid m) = P(Y \mid Q)P(Q \mid m), \tag{3.1}$$

where the second equality holds because conditioned on pseudomarker genotypes, marker information would not give additional information about phenotypes. The recombination model $P(Q \mid m)$ estimates the probability of pseudomarker genotypes from flanking markers m.

Multiple phenotypes can be observed such as gene expression from microarray data. We focus on building a network of phenotypes Y and pseudomarker genotypes Q on how a genotype changes a phenotype level and how a phenotype level changes other phenotype levels. In building a network of phenotypes and genetic variations, a Bayesian network has often been used [Li et al., 2006, Zhu et al., 2008, Chaibub Neto et al., 2008, 2010a]. The inference of a Bayesian network assumes that all variables constituting the Bayesian network are observed. The assumption may not hold because 1) there could be unmeasured variables in a network such as metabolite levels controlling gene expressions, or 2) we may take a subset of genes in building a network instead of working on thousands of genes, and in this process we may omit important genes comprising a Bayesian network. This is why we consider an ancestral graph, which can properly represent conditional independencies after marginalizing out latent variables.

Equation (3.1) still holds when Y are various observed phenotypes. The linkage model $P(Q \mid m)$ is rather fixed and we will show how to model $P(Y \mid Q)$ when there could be latent variables between phenotypes. For $i = 1, \ldots, n$ and $t = 1, \ldots, T$, let Y_{ti} be the value of the phenotype for individual i and trait Y_t. Each phenotype Y_{ti} can be modeled as follows:

$$Y_{ti} = \mu^*_{ti} + \sum_{v \in pa(t)} \beta_{tv} Y_{vi} + \epsilon_{ti}, \tag{3.2}$$

where μ_{ti}^* is a genetic architecture and $\sum_{v \in pa(t)} \beta_{tv} Y_{vi}$ is the effect of its parental phenotypes. The genetic architecture μ_{ti}^* is modeled to be

$$\mu_{ti}^* = \mu_t + \sum_{k=1}^{K} \gamma_{tk} \theta_{tk} X_{ki}, \tag{3.3}$$

where μ_t is the overall mean for trait Y_t, γ_{tk} represents the inclusion ($\gamma_{tk} = 1$) or exclusion ($\gamma_{tk} = 0$) of the k-th pseudomarker (Q_k), X_{ki} is a column vector of coded variables of pseudomarker genotypes for individual i and the vector θ_{tk} is a row vector of several types of genetic effects of pseudomarker Q_k. When $\gamma_{tk} = 1$, Q_k is identified to be a QTL for Y_t. Different from a Bayesian network model, the errors $\epsilon = (\epsilon_1, \ldots, \epsilon_T)'$ can be correlated as follows:

$$\epsilon \sim N_T(0, \Omega), \tag{3.4}$$

where $\Omega(t, s) = 0$ if and only if there is no bidirected edge between Y_t and Y_s. The bidirected edges are due to latent variables between phenotypes.

This Gaussian parameterization by eqn (3.2) and eqn (3.4) along with the linkage model $P(Q \mid m)$ corresponds to a graph G of phenotypes Y and psuedomarkers Q. The schematic view of G is in Figure 3.2. The graph G can be split into three subgraphs — a linkage map G_Q, a QTL mapping $G_{Q \to Y}$ and a phenotype network G_Y. First, a linkage map G_Q is constructed to have undirected edges between neighboring pseudomarkers. It is modeled by $P(Q \mid m)$. Second, a QTL mapping $G_{Q \to Y}$ is constructed to have directed edges from Q to Y. The directed edge from Q_k to Y_t is present when $\gamma_{tk} = 1$ in eqn (3.3). Third, a phenotype network G_Y is constructed to have directed edges and bidirected edges among phenotypes Y. The directed edge from Y_v to Y_t is present when $v \in pa(t)$ in eqn (3.2) and the bidirected edge between Y_t and Y_s is present when $\Omega(t, s) \neq 0$ in eqn (3.4). Table 3.1 summarizes the relationship between graphical representation and the parameters in the statistical model.

Since we only allow latent variables but not selection variables between phenotypes, the class of phenotype networks, G_Y, we consider is a directed ancestral graph, which consists of directed edges and bidirected edges. Without losing any possible set of conditional independencies that directed ancestral graphs can entail, we restrict to the phenotype network to be a directed maximal

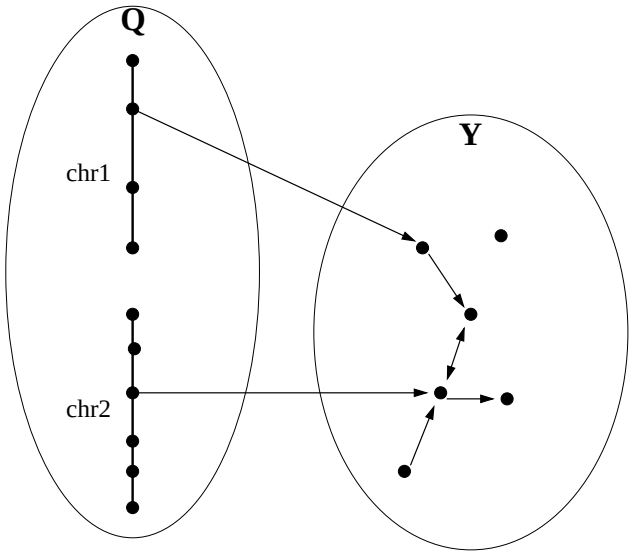


Figure 3.2: Schematic view of an ancestral graph of phenotypes Y and pseudomarkers Q.

Table 3.1: Decomposition of ancestral graph G of genotypes and phenotypes into $G_Q, G_{Q\to Y}$ and G_Y. Nodes and edges consisting of each subgraph are summarized and the corresponding values in the statistical model are summarized.

subgraph	meaning	nodes	values	edges	values
G_Q	Linkage map	Q_k	Q_k	Q_k — Q_{k+1}	modeled by $P(Q\|m)$
$G_{Q\to Y}$	QTL mapping	Q_k	X_{ki}	$Q_k \to Y_t$	$\gamma_{tk}\theta_{tk} \neq 0$
		Y_t	Y_{ti}		
G_Y	Phenotype network	Y_t	Y_{ti}	$Y_v \to Y_t$	$\beta_{tv} \neq 0$
				$Y_s \leftrightarrow Y_t$	$\Omega(t,s) \neq 0$

ancestral graph (DMAG). We will prove that the extended network G from the phenotype network G_Y with $G_{Q\to Y}$ and G_Q is a maximal ancestral graph (MAG) in the next section. Note that even though undirected edges in G_Q are not associated with conditioning on selection variables, they fit well in the definition of ancestral graphs. First, Q always precede Y because of the assumption that genotypes affect phenotypes but not the other way around. Second, the undirected Markov property holds for $P(Q \mid m)$ such that non-adjacent pseudomarkers are independent given other pseudomarkers. Third, $P(Y, Q \mid m) = P(Q \mid m)P(Y \mid Q, m) = P(Q \mid m)P(Y \mid Q)$, which is in analogous to the factorization of the probability function of an ancestral graph into the undirected graph and the remaining.

3.3.2 Graphical properties of extended network

We assume a phenotype network G_Y to be a directed maximal ancestral graph (DMAG). We prove that the extended network G by QTLs mapping and the linkage map is also a maximal ancestral graph (MAG) in Theorems 3.3 and 3.4. The converse also works that if G is a MAG, then G_Y is a DMAG in Theorem 3.5.

Theorem 3.3. *Suppose G_Y is a directed ancestral graph consisting of bidirected and directed edges. If we add QTLs $G_{Q\to Y}$ and a recombination graph G_Q to G_Y, then the extended network G is an ancestral graph.*

Proof. To prove G is an ancestral graph, we need to check that the extended network G satisfies the conditions in the definition of ancestral graphs. The condition 1 in Definition 3.1 is satisfied in G. If adding $Q \to Y$ makes a directed cycle, then, there exists a directed path from Y to Q. However, there is no directed edge from a phenotype to a genotype, therefore, there is no directed path from Y to Q. The condition 2 is satisfied as follows: Suppose $Y_1 \leftrightarrow Y_2$. We need to prove that there is no directed path from Y_1 to Y_2 (or from Y_2 to Y_1) in the extended network. We know that there is no directed path from Y_1 to Y_2 only through phenotypes from the assumption of G_Y. In the path via Q, the triple (Y_m, Q, Y_{m+1}) cannot make a directed path from Y_1 to Y_2 because the only possible configuration of the triple is $Y_m \leftarrow Q \to Y_{m+1}$. Hence, condition 2 is satisfied. The

condition 3 is satisfied naturally because Q only have neighbors in Q or children in Y. Therefore, the extended network is an ancestral graph. □

Note that if there is an undirected edge $Y_1 — Y_2$ in G_Y, then, adding a directed edge $Q \rightarrow Y_1$ does not satisfy the conditions of an ancestral graph.

Theorem 3.4. *If $G_{Q \rightarrow Y}$ and G_Q is added to DMAG G_Y, then the extended network is a maximal ancestral graph (MAG).*

Proof. We need to prove that there is no inducing path between non-adjacent vertices in the extended network.

- Between Y_1 and Y_2: If there is an inducing path $\pi =< Y_1, v_1, \ldots, v_k, Y_2 >$, there are two cases,
  - $v_i \in \mathbf{Y}, 1 \leq i \leq k$: If v_i is an ancestor of Y_1 or Y_2, solely through the directed path on phenotypes, it contradicts the assumption of maximality in G_Y. If the directed path goes through Q, it contradicts the assumption of the directionality between Q and Y.
  - there exists $v_i \in \mathbf{Q}$: Q cannot be a collider in the collider path π.
- Between Q and Y: If there is an inducing path $\pi =< Q, v_1, \ldots, v_k, Y >$, the configuration of the collider path is $Q \rightarrow v_1 \leftrightarrow \ldots v_k \leftarrow\!\circ\ Y$ where $\leftarrow\!\circ$ can be either $\leftrightarrow$ or $\leftarrow$. By the condition for an inducing path and the condition for an ancestral graph, v_k should be an ancestor of Q, which contradicts the directionality between pseudomarkers and phenotypes and undirectedness between pseudomarkers.

□

Theorem 3.5. *If G is a MAG, then $G_Y = G[^Q_\varnothing$ and hence, G_Y is a MAG. Since we do not allow selection variables, G_Y is a DMAG.*

Proof. First, when there is an edge between x and y in G_Y, then, for all $Z \supset Q$, $x \not\perp y|Z$, and hence, there is an edge between x and y in $G[^Q_\varnothing$. By lemma 3.9 in Richardson and Spirtes [2002], the edge type will be conserved in $G[^Q_\varnothing$.

Second, when there is no edge between x and y in G_Y, the maximality of G ensures that there exists Z such that $x \perp y|Z$. We want to show that the relation $x \perp y|Z \cup Q$ also holds. Suppose that the relation $(x \perp y|Z \cup Q)$ does not hold. Then, there is an m-connecting path π between x and y given $Z \cup Q$, which satisfies both conditions:

(i) every noncollier is not in $Z \cup Q$, and

(ii) every collider on the path is in $ant(Z \cup Q)$.

If π goes through any of Q, the corresponding variables in Q are noncolliers and it violates the condition (i). Therefore, there is no m-connecting path given $Z \cup Q$ and $x \perp y|Z \cup Q$ holds. Let's consider that π does not go through Q and conditions (i) and (ii) are satisfied. Since π does not contain Q, the condition (i) can be stated that every noncollider is not in Z. Also, every collider on the path is in $ant(Z)$. The satisfaction of two conditions makes that π is an m-connecting path given Z, which is contradictory. Hence, when $x \perp y|Z$ holds, $x \perp y|Z \cup Q$ holds. This shows that when there is no edge between x and y in G_Y, then there is no edge between x and y in $G[^Q_\varnothing$.

It proves that $G_Y = G[^Q_\varnothing$ and thereby, G_Y is maximal. □

We have shown that the extended network G is also a maximal ancestral graph (MAG) if G_Y is a directed maximal ancestral graph (DMAG). We will look at how adding QTLs to G_Y can break Markov equivalence of G_Y in Lemma 3.7 and Theorem 3.8. Ali et al. [2009] states that if G_1 and G_2 are maximal ancestral graphs (MAGs), then G_1 and G_2 are Markov equivalent ($G_1 \equiv G_2$) if and only if G_1 and G_2 have the same adjacencies and the same colliders with order. An order for a triple is defined recursively.

Definition 3.6 (Ali et al. [2009])**.** Let $D_i (i \geq 0)$ be the set of *triples of order i* in a MAG G defined recursively as follows:

Order 0. A triple $< a, b, c > \in D_0$ if a and c are not adjacent.

Order i+1. A triple $< a, b, c > \in D_{i+1}$ if

1. for all $j < i + 1$, $< a, b, c > \notin D_j$, and,

2. there is a discriminating path $< x, q_1, \ldots, q_p, b, y >$ for b with either $< a, b, c >=< q_p, b, y >$ or $< a, b, c >=< y, b, q_p >$ and the p colliders

$$< x, q_1, q_2 >, \ldots, < q_{p-1}, q_p, b > \in \cup_{j \leq i} D_j.$$

A *discriminating path* $< x, q_1, \ldots, q_p, b, y >$ between non-adjacent x and y for b is actually a collection of paths

$$x \circ\!\!\rightarrow q_1 \leftrightarrow \cdots \leftrightarrow q_j \rightarrow y \quad (1 \leq j \leq p),$$
$$x \circ\!\!\rightarrow q_1 \leftrightarrow \cdots \leftrightarrow q_p \leftarrow\!\!\circ\, b \circ\!\!\rightarrow y,$$

where $\circ\!\!\rightarrow$ can be either $\leftrightarrow$ or $\rightarrow$ and $\leftarrow\!\!\circ$ can be either $\leftrightarrow$ or $\leftarrow$.

Lemma 3.7. *Suppose two vertices Y_1 and Y_2 are adjacent by either a directed edge or a bidirected edge. There are three Markov equivalent directed ancestral graphs for a pair of Y_1 and Y_2: $Y_1 \rightarrow Y_2$, $Y_1 \leftarrow Y_2$, and $Y_1 \leftrightarrow Y_2$. (i) If a QTL Q_1 affects Y_1 but not Y_2, then it partially distinguishes the extended graphs: $Q_1 \rightarrow Y_1 \rightarrow Y_2$ versus $Q_1 \rightarrow Y_1 \leftarrow Y_2$ or $Q_1 \rightarrow Y_1 \leftrightarrow Y_2$. (ii) Further, if a QTL Q_2 affects Y_2 but not Y_1, then it distinguishes all three extended graphs: $Q_1 \rightarrow Y_1 \rightarrow Y_2 \leftarrow Q_2$ versus $Q_1 \rightarrow Y_1 \leftarrow Y_2 \leftarrow Q_2$ versus $Q_1 \rightarrow Y_1 \leftrightarrow Y_2 \leftarrow Q_2$.*

Proof. (i) The extended networks $Q_1 \rightarrow Y_1 \rightarrow Y_2$, $Q_1 \rightarrow Y_1 \leftarrow Y_2$ and $Q_1 \rightarrow Y_1 \leftrightarrow Y_2$ are MAGs. They all have the same adjacencies. $Q_1 \rightarrow Y_1 \rightarrow Y_2$ has no collider while $Q_1 \rightarrow Y_1 \leftarrow Y_2$ and $Q_1 \rightarrow Y_1 \leftrightarrow Y_2$ have one unshielded collider. Since an unshielded collider is order 0, $Q_1 \rightarrow Y_1 \leftarrow Y_2$ and $Q_1 \rightarrow Y_1 \leftrightarrow Y_2$ are Markov equivalent.
(ii) The further extended networks are MAGs and have the same adjacencies. The triple $< Y_1, Y_2, Q_2 >$ in $Q_1 \rightarrow Y_1 \leftarrow Y_2 \leftarrow Q_2$ is not a collider while the triple $< Y_1, Y_2, Q_2 >$ in $Q_1 \rightarrow Y_1 \leftrightarrow Y_2 \leftarrow Q_2$ is a collider with order 0. Therefore, these two networks are not Markov equivalent. □

Similar to Result 2 in Chaibub Neto et al. [2010a], adding QTLs can distinguish Markov equivalent MAGs.

Theorem 3.8. *Consider a class of Markov equivalent DMAGs $\mathcal{G}_{\mathcal{Y}}$. Let Y_1 and Y_2 be any two adjacent nodes in the graphs in $\mathcal{G}_{\mathcal{Y}}$. Assume that for each such pair there exists at least QTLs, Q_1 directly affecting Y_1 but not Y_2 and Q_2 directly affecting Y_2 but not Y_1. Let $\mathcal{G}$ represent the class of extended graphs. Then the graphs in $\mathcal{G}$ are not Markov equivalent.*

Proof. For each pair of two adjacent phenotypes, we can apply the above lemma so that the extended subgraphs have the same adjacencies but different colliders with order. Since this holds for each pair, the graphs in $\mathcal{G}$ have the same adjacencies but different colliders with order, therefore, they are not Markov equivalent. □

3.3.3 The parametric family of the extended network

By extending a phenotype network G_Y with pseudomarkers, the data is mixed with continuous variables Y and discrete variables Q. We will show in Proposition 3.10 that the parametric family of phenotypes and pseudomarkers defined by eqn (3.2), eqn (3.4) and $P(Q \mid m)$ is a homogeneous conditional Gaussian (HCG) distribution. The conditional Gaussian (CG) distribution is named by the fact that the joint distribution of continuous variables are Gaussian conditional on discrete variables. The CG model is defined as in Definition 3.9 [Lauritzen, 1996].

Definition 3.9 (Cowell et al. [2003])**.** The conditional Gaussian (CG) model is

$$\log f(z) = \log f(q, y) = g(q) + h(q)'y - y'K(q)y/2,$$

where $z = (q, y)$, $q \in \Delta$ (discrete variables), $y \in \Gamma$ (continuous variables), $g(q)$ is a real number, $h(q)$ is a vector in $\mathcal{R}^{\Delta}$ and $K(q)$ is a positive definite matrix.

It is equivalent to

$$p(q) = P(Z_\Delta = i) > 0, \quad L(Z_\Gamma \mid Z_\Delta = q) = N_{|\Gamma|}(\xi, \Sigma(q)),$$

where

$$\Sigma(q) = K(q)^{-1}, \quad \xi(q) = K(q)^{-1}h(q).$$

It is called homogeneous when the covariance is independent of q, $\Sigma(q) \equiv \Sigma$, or, equivalently, $K(q) \equiv K$. The next proposition shows that our model for phenotypes and genotypes defined in eqn (3.2) and eqn (3.4) with $p(Q \mid m)$ in eqn (3.1) corresponds to a homogeneous conditional Gaussian (HCG) model.

Proposition 3.10. *The proposed model for phenotypes and pseudomarkers factored into a pseudomarker reconstruction model $p(Q \mid m)$ and a set of linear equations in eqn (3.2) with correlated Gaussian errors in eqn (3.4) falls into a homogeneous conditional Gaussian (HCG) family.*

Proof. Equations (3.2) and (3.4) can be written to be

$$Y \mid Q \sim N_T(\mu^* + Y\boldsymbol{\beta}, \Omega),$$

where $Y = (Y_1, \ldots, Y_T)'$, $\mu^* = (\mu_1^*, \ldots, \mu_T^*)'$ and $\boldsymbol{\beta}(t, s) = \beta_{ts} 1(s \to t)$. The joint distribution of phenotypes Y conditional on pseudomarker genotypes Q is

$$p(Y \mid Q) = (2\pi)^{-T/2} |\Omega|^{-1/2} \exp(-\frac{1}{2}(Y - \mu^* - \boldsymbol{\beta}Y)'\Omega^{-1}(Y - \mu^* - \boldsymbol{\beta}Y)).$$

Then, the joint distribution of phenotypes and genotypes is

$$\begin{aligned} \log p(Y, Q \mid m) &= \log p(Y \mid Q) + \log p(Q \mid m) \\ &= \log p(Q \mid m) - \frac{1}{2} T \log(2\pi) - \frac{1}{2} \log |\Omega| + (I) + (II), \end{aligned}$$

where

$$\begin{aligned} (I) &= -\frac{1}{2}(\mu^{*'}\Omega^{-1}\mu^* - 2\mu^{*'}\Omega^{-1}Y + 2\mu^{*'}\Omega^{-1}\boldsymbol{\beta}Y) \\ &= -\frac{1}{2}\mu^{*'}\Omega^{-1}\mu^* + \mu^{*'}\Omega^{-1}(I - \boldsymbol{\beta})Y, \end{aligned}$$

and

$$\begin{aligned} (II) &= (Y - \boldsymbol{\beta}Y)'\Omega^{-1}(Y - \boldsymbol{\beta}Y) \\ &= ((I - \boldsymbol{\beta})Y)'\Omega^{-1}((I - \boldsymbol{\beta})Y) \\ &= Y'[(I - \boldsymbol{\beta})'\Omega^{-1}(I - \boldsymbol{\beta})]Y \\ &= Y'\Omega^{\diamond}Y, \end{aligned}$$

where $\Omega^{\diamond} = (I - \beta)'\Omega^{-1}(I - \beta)$.

It can be rewritten to be

$$\log p(Y, Q \mid m) = g(Q) + h'(Q)Y - Y'K(Q)Y/2,$$

where $g(Q) = \log p(Q \mid m) - \frac{1}{2}(\mu^{*'}\Omega^{-1}\mu^{*} + T\log(2\pi) + \log|\Omega|)$, $h(Q) = \mu^{*'}\Omega^{-1}(I - \beta)$ and $K(Q) = \Omega^{\diamond}$. Therefore, the joint distribution of phenotypes and genotypes is a CG family. Since $\Omega^{\diamond}$ is independent of Q, it is a homogeneous CG family with (g, h, K). □

3.3.4 Conjecture: Distribution equivalence is the same as the Markov equivalence for Gaussian family and HCG family

When comparing the graphs, two graphs can be equivalent in terms of conditional independence relations, distributions, or likelihoods. Two graphs are *Markov equivalent* if they represent the same set of conditional independencies. Two graphs G_1 and G_2 are *distribution equivalent* with respect to the family F if for every θ_{G_1}, there exists a θ_{G_2} such that $p(Y \mid \theta_{G_1}, G_1) = p(Y \mid \theta_{G_2}, G_2)$, and vice versa, representing the same set of joint probability distributions. When two graphs are distribution equivalent with respect to F, it is often reasonable to expect that data can not help to discriminate G_1 and G_2, that is, $p(D \mid G_1) = p(D \mid G_2)$ for any data set D, called *likelihood equivalence*. Distribution equivalence with respect to F implies Markov equivalence however the converse does not hold in general.

We conjecture that there is a reparameterization between Markov equivalent DMAGs in a Gaussian family in Conjecture 3.12. If this conjecture holds, Conjecture 3.14 for the reparameterization between Markov equivalent MAGs for our gene network in a HCG family follows. Before going to the conjectures, we introduce a known graphical transformation between two Markov equivalent DMAGs, which will be an important piece in proving the conjectures.

Theorem 3.11 (Zhang and Spirtes [2005])**.** *For two Markov equivalent DMAGs, there is a sequence of legitimate mark changes between them. A legitimate mark change is defined to be as*

follows.

$$G : \textit{DMAG with } t \to v$$
$$G' : \textit{identical to } G \textit{ except } t \leftrightarrow v.$$

Then, G' is DMAG and Markov equivalent to G if and only if

1. *there is no directed path from t to v other than $t \to v$ in G.*
2. *For any $C \to t$ in G, $C \to v$ is also in G, and for any $D \leftrightarrow t$ in G, either $D \to v$ or $D \leftrightarrow v$ is in G.*
3. *there is no discriminating path for t on which v is the endpoint adjacent to t in G.*

Conjecture 3.12. *For two Markov equivalent DMAGs G_1 and G_2, there exists a reparameterization θ_2 for G_2 to have the same likelihood $L(\theta_1|G_1) = L(\theta_2|G_2)$ in a Gaussian distribution family.*

Proof. Sketch of proof 1. The likelihood of a DMAG in terms of (μ, Σ) where $\Sigma = C^{-1}\Omega C^{-T}$: In a Gaussian distribution family, the model can be written by a set of recursive equations with correlated Gaussian errors,

$$Y = \mu + B(Y - \mu) + \epsilon \text{ where } \epsilon \sim N(0, \Omega).$$

The coefficient matrix B is lower-triangle matrix such that $B(\alpha, \beta)$ is the coefficient for $\beta \to \alpha$ and 0 otherwise. The covariance Ω of errors is that $\Omega(\alpha, \beta) = 0$ if there is no $\alpha \leftrightarrow \beta$ when $\alpha \neq \beta$. Its joint probability distribution is

$$Y \sim N_{|V|}(\mu, \Sigma),$$
$$\Sigma = (I - B)^{-1}\Omega(I - B)^{-T} = C^{-1}\Omega C^{-T} \text{ for } C = I - B,$$

where $|V|$ is the number of vertices and I is the $|V| \times |V|$ identity matrix. Hence, the likelihoood is a function of (μ, Σ) where $\Sigma = C^{-1}\Omega C^{-T}$.

2. Markov equivalence by a sequence of legitimate mark changes : By Theorem 3.11, there exists a sequence of legitimate mark changes from G_1 to G_2. Therefore, a reparameterization

between two Markov equivalent DMAGs by a legitimate mark change can be sequentially applied for a reparameterization from G_1 to G_2. Let G and G^* be Markov equivalent by a legitimate mark change such that G and G^* are DMAGs and identical to each other except for one edge that is $t \to v$ in G and $t \leftrightarrow v$ in G^*. Denote (μ, Σ) and (μ^*, Σ^*) to be the sets of mean and covariance of G and G^*, respectively. In addition, C is a matrix to represent the directed edges and Ω is a matrix to represent the bidirected edges of G, where $\Sigma = C^{-1}\Omega C^{-T}$. Similarly, (C^*, Ω^*) represents G^*.

3. Reparameterization from (μ, C, Ω) of G to (μ^*, C^*, Ω^*) of G^* to have the same likelihood : If $\mu = \mu^*$ and $\Sigma = \Sigma^*$, then G and G^* have the same likelihood. We will propose a reparameterization from (C, Ω) to (C^*, Ω^*) to satisfy $\Sigma = \Sigma^*$ in addition to $\mu = \mu^*$, which will guarantee the same likelihood of G and G^*. First, we present the reparameterization of C^* from C and set the constraint on the reparameterization to fulfill the graphical structures of G and G^*. Second, we induce the relation between Ω and Ω^* from $\Sigma = \Sigma^*$ where $\Sigma = C^{-1}\Omega C^{-T}$ and $\Sigma^* = C^{-1}\Omega C^{-T}$. Lastly, we set the constraint on Ω^* to fulfill the graphical graphical structure of G^*.

3-1. $C^* = C + 1_v h$: We construct C^* from C such that the element in C^* corresponding to $t \to v$ is 0, the removal of $t \to v$ from G can be transferred into $pa(v) \to v$ and other elements remain the same.

$$C^* = C + 1_v h \tag{3.5}$$

$$h[t] = b_{vt} \quad \text{and } h[x] = 0 \text{ for } x \notin pa(v), \tag{3.6}$$

where 1_v is a column vector with zero entries except at the v-th entry as 1, h is a row vector and $h[x]$ is the x-th element in h. Then, C^* complies with the graphical structure of G^* such that

$$C^*[v,t] = C[v,t] + 1_v[v] * h[t] = -b_{vt} + b_{vt} = 0$$
$$C^*[v,x] = C[v,x] + 1_v[v]h[x] = 0 + 1 \cdot 0 = 0, \quad x \notin pa(v),$$

where $C[v,t]$ is a submatrix of C formed by row(s) v and column(s) t.

3-2. $\Sigma = \Sigma^*$: To hold $\Sigma = \Sigma^*$, C^* and Ω^* should satisfy that

$$\Sigma = (C^*)^{-1}\Omega^*(C^*)^{-T},$$

and we induce Ω^* such that

$$\begin{aligned}\Omega^* &= (C^*)\Sigma(C^*)^T = (C + 1_v h)\Sigma(C + 1_v h)^T \\ &= C\Sigma C^T + 1_v h\Sigma C^T + (1_v h\Sigma C^T)^T + 1_v h\Sigma h^T (1_v)^T \\ &= \Omega + 1_v h\Sigma C^T + (1_v h\Sigma C^T)^T + 1_v h\Sigma h^T (1_v)^T. \end{aligned} \tag{3.7}$$

3-3. Ω^* : Ω^* needs to comply with the graphical structure of G^* such that the elements of zero in Ω and Ω^* differ only between $t \to v$ in G and $t \leftrightarrow v$ in G^*. In eqn (3.7),

- the last term $1_v h\Sigma h^T (1_v)^T$ is a $|V| \times |V|$-matrix with zero entries except at the (v, v)-th entry, and thereby, it satisfies the constraint on Ω^*;
- the term $1_v h\Sigma C^T$ is a $|V| \times |V|$-matrix with zero entries except at the v-th row. The v-th row vector, $h\Sigma C^T$, should satisfy the constraint on Ω^* such that

$$s[x] := (h\Sigma C^T)[x] = 0 \quad \text{for } x \notin sp(v) \cup \{t, v\}. \tag{3.8}$$

4. Solve for h: We want to solve for h, satisfying the constraints in eqns (3.8) and (3.6) simultaneously. We have a set of $|V|$-linear equations:

$$s = h\Sigma C^T, \tag{3.9}$$

for $|sp(v)| + 2$ unknowns in s and $|pa(v)| - 1$ unknowns in h. There is a bigger number of linear equations than the number of unknowns. Let a vertex set V is divided into disjoint sets such as

$$\begin{aligned} V &= pa(v) \cup \{v\} \cup sp(v) \cup \text{others} \\ &= \{t\} \cup (pa(v) \setminus \{t\}) \cup \{v\} \cup sp(v) \cup \text{others}, \end{aligned}$$

and Σ and C^T are rearranged accordingly. Equation (3.9) can be expressed with the constraints that

$$\begin{aligned} &(s[t], s[pa(v) \setminus \{t\}], s[v], s[sp(v)], s[\text{others}]) \\ &\quad = (h[t], h[pa(v) \setminus \{t\}], h[v], h[sp(v)], h[\text{others}])\ \Sigma C^T \\ &\Longrightarrow \\ &(s[t], 0, s[v], s[sp(v)], 0) = (b_{vt}, h[pa(v) \setminus \{t\}], 0, 0, 0)\ \Sigma C^T. \end{aligned}$$

Since both Σ and C are non-singular, we can multiply the inverse of ΣC^T on the both sides and the equation becomes that

$$(s[t], 0, s[v], s[sp(v)], 0)\,(\Sigma C^T)^{-1} = (b_{vt}, h[pa(v) \setminus \{t\}], 0, 0, 0)$$

$$\Longrightarrow$$

$$s[\{t, v\} \cup sp(v)]\,(\Sigma C^T)^{-1}[\{t, v\} \cup sp(v), \cdot] = (b_{vt}, h[pa(v) \setminus \{t\}], 0, 0, 0).$$

Let $\triangle := \{t, v\} \cup sp(v)$. We can divide the equations into three sets:

$$(b_{vt}, 0, 0) = s[\triangle]\,(\Sigma C^T)^{-1}[\triangle, \triangle] \tag{3.10}$$

$$h[pa(v) \setminus \{t\}] = s[\triangle]\,(\Sigma C^T)^{-1}[\triangle, pa(v) \setminus \{t\}] \tag{3.11}$$

$$0 = s[\triangle]\,(\Sigma C^T)^{-1}[\triangle, \text{others}]. \tag{3.12}$$

We proceed in this order.

(I) Solve for $s[\triangle]$ in eqn (3.10):

$$s[\triangle] = (b_{vt}, 0, 0)((\Sigma C^T)^{-1}[\triangle, \triangle])^{-1},$$

provided the inverse of $(\Sigma C^T)^{-1}[\triangle, \triangle]$ exists.

(II) Plug in $s[\triangle]$ to get $h[pa(v) \setminus \{t\}]$ in eqn (3.11):

$$h[pa(v) \setminus \{t\}] = s[\triangle]\,(\Sigma C^T)^{-1}[\triangle, pa(v) \setminus \{t\}].$$

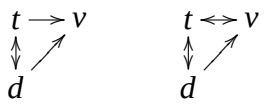
(III) Check eqn (3.12) is satisfied.

Things to be proved : If this sketch of proof works, we need to prove two properties to complete the proof: (1) The inverse of $(\Sigma C^T)^{-1}[\triangle, \triangle]$ exists when we solve for eqn (3.10); (2) Equation (3.12) is satisfied. It seems that eqn (3.12) is satisfied when the conditions in a legitimate mark change are satisfied.

5. Concluding remarks : The solution of h is added to C to get $C^* = C + 1_v h$ and Ω^* can be obtained accordingly. (C^*, Ω^*) will comply with the graphical structure of G^*. □

The following example applies the conjectured reparameterization from G to G^* and confirms that the reparameterization satisfies the constraints on the coefficient matrix C^* and the covariance matrix on errors Ω^*.

Example 3.13. *Consider the following two Markov equivalent graphs, G and G^*, where G^* is transformed by a legitimate mark change from G.*



(a) G (b) G^*

The graph G can be represented by recursive equations,

$$Y_t = \epsilon_t$$
$$Y_d = \epsilon_d$$
$$Y_v = b_{tv}Y_t + b_{dv}Y_d + \epsilon_v,$$

with correlated Gaussian error $cov(\epsilon_t, \epsilon_v) = \sigma_{td}$. *Let* $b_{tv} = b_{dv} = 1$, $var(\epsilon) = 1$ *and* $\sigma_{td} = 0.5$. *It is written to be*

$$C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix}, \qquad \Omega = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then, we get

$$\Sigma = C^{-1}\Omega C^{-T} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.5 & 1.5 \\ 0.5 & 1 & 1.5 \\ 1.5 & 1.5 & 4 \end{pmatrix}.$$

If we modify $t \to v$ to $t \leftrightarrow v$ as in G^*, to get the reparameterization $C^* = C + 1_v h$ and Ω^* for G^*, we first need to calculate ΣC^T,

$$\Sigma C^T = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 1.5 & 1.5 & 1 \end{pmatrix}, \qquad (\Sigma C^T)^{-1} = \begin{pmatrix} 4/3 & -2/3 & 0 \\ -2/3 & 4/3 & 0 \\ -1 & -1 & 1 \end{pmatrix}.$$

Following the procedure for reparameterization,

(I) Solve for $s[\triangle]$ where $\triangle := \{t, v\} \cup sp(v) = \{t, v\}$ in eqn (3.10):

$$\begin{aligned} s[\{t,v\}] &= h[\{t,v\}] * (\Sigma C^T)^{-1}[\{t,v\},\{t,v\}] \\ &= (1,0) * \begin{pmatrix} 4/3 & 0 \\ -1 & 1 \end{pmatrix}^{-1} \\ &= (1,0) * \begin{pmatrix} 0.75 & 0 \\ 0.75 & 0 \end{pmatrix} \\ &= (0.75, 0). \end{aligned}$$

(II) Plug in $s[\triangle]$ to get $h[pa(v) \setminus t]$ in eqn (3.11):

$$\begin{aligned} h[d] &= s[\{t,v\}](\Sigma C^T)^{-1}[\{t,v\}, pa(v) \setminus \{t\}] \\ (0.75, 0) * \begin{pmatrix} -2/3 \\ -1 \end{pmatrix} &= -1/2. \end{aligned}$$

Then, $h = (1, -1/2, 0)$ and we get the following reparameterized C^* and Ω^* for G^*:

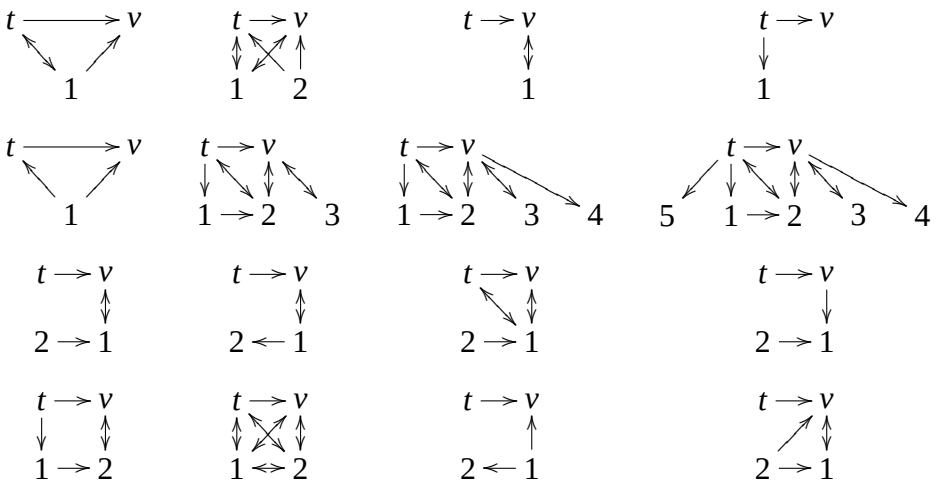
$$C^* = C + 1_v h = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1.5 & 1 \end{pmatrix}$$

$$\Omega^* = C^* \Sigma (C^*)^T = \begin{pmatrix} 1 & 0.5 & 0.75 \\ 0.5 & 1 & 0 \\ 0.75 & 0 & 1.75 \end{pmatrix}$$

where C^* and Ω^* satisfy the constraints on G^*.

We have applied the reparameterization on 16 DMAGs with the edge $t \to v$ in Table 3.2 and checked that the reparameterized parameters, (C^*, Ω^*), comply with their transformed DMAGs with the edge $t \leftrightarrow v$, respectively. In addition, we have applied the reparameterization on DMAGs in Table 3.3 that the assumptions on the legitimate mark change are violated — $pa(t) \in pa(v)$, $sp(t) \in pa(v) \cup sp(v)$, and DMAG is maximal — and we checked that the reparameterized parameters do not comply with DMAGs with $t \leftrightarrow v$.

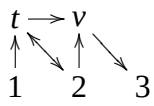
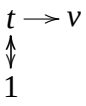
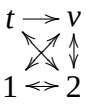
Table 3.2: A DMAG that the reparameterization complies with the transformed DMAG by a legitimate mark change from $t \to v$ to $t \leftrightarrow v$.



For Markov equivalent DMAGs, we can apply a sequence of legitimate mark changes to graphically transform from one DMAG to another DMAG and the intermediate DMAGs are also Markov equivalent. In general, there is no known transformation for MAGs. However, MAGs that we consider here have a nature that undirected edges for marker recombination will not be transformed to directed or bidirected edges and the directed edges from marker to phenotypes cannot be changed. The only configuration change in the edge should be between phenotypes and the subgraph for phenotypes is a DMAG. Hence, we can apply a sequence of legitimate mark changes between phenotypes to transform between Markov equivalent MAGs that we consider here.

Conjecture 3.14. *For two Markov equivalent MAGs G_1 and G_2 as a gene network that markers are connected by undirected edges, markers and phenotypes are connected by directed edges*

Table 3.3: DMAGs violating assumptions in a legitimate mark change

		
$pa(t) \not\subset pa(v)$	$sp(t) \not\subset pa(v) \cup sp(v)$	It is not maximal.

from markers to phenotypes and phenotypes can be connected by directed or bidirected edges, there exists a reparameterization θ_2 for G_2 to have the same likelihood $L(\theta_1|G_1) = L(\theta_2|G_2)$ in a homogeneous conditional Gaussian (HCG) distribution family.

Proof. Since we only allow legitimate mark changes between phenotypes and the QTLs enter the HCG model through the mean, the reparameterization follows by replacing μ_t by $\mu_t^* = \mu_t + \sum_{k=1}^{K} \gamma_{tk}\theta_{tk}X_{ki}$ in Conjecture 3.12. □

By proving that Markov equivalence implies distribution equivalence, we are ensured that Markov equivalent graphs would have the same set of distributions and their maximum likelihoods would be the same.

3.4 Algorithms for ancestral graph inference of phenotypes and QTLs

We have looked at the parametric family and the graphical class of the extended network. To infer a network, there are two approaches — one is constraint-based and the other is search-and-score based. The constraint-based method such as fast causal inference (FCI) algorithm Spirtes et al. [2000] infers a network by conditional independence tests on the pairs of vertices. The search-and-score method scores the graph by BIC or likelihood and searches over the space of graphs. We develop a score-based search algorithm for DMAGs.

3.4.1 Score-based search – MCMC

We can explore the ancestral graph space by Markov chain Monte Carlo (MCMC). The MCMC is constructed to jump over DMAGs and stays in DMAGs with higher scores (maximum likelihood of the DMAG).

First, we develop a Metropolis-Hastings algorithm to explore DMAG space in general.

1. Divide a DMAG G_{old} into bidirected graph G^B_{old} and directed graph G^D_{old}.
2. Propose a new directed graph G^D_{new} from G^D_{old} by a DAG proposal distribution $R(G^D_{new}|G^D_{old})$.
3. Propose new bidirected edges G^B_{new} as long as it does not violate the ancestral graph assumption. Construct G_{new} by stitching G^D_{new} and G^B_{new} together. Its proposal distribution is $R(G_{new}|G^D)$.
4. Accept the new DMAG G_{new} with a probability,
$$\min\{1, \frac{P(Y|G_{new})}{P(Y|G_{old})} \frac{R(G^D_{old}|G^D_{new})R(G_{old}|G^D_{old})}{R(G^D_{new}|G^D_{old})R(G_{new}|G^D_{new})}\}.$$
5. Iterate until the chain converges.

In step 2, the DAG proposal distribution is a mixture of single edge proposal of a DAG [Husmeier, 2003] and edge reversal proposal of a DAG [Grzegorczyk and Husmeier, 2008]. The single edge proposal is either addition of an edge, removal of an edge, or reversal of an edge as long as it does not make cycles in new DAG. In addition to the possible proposed DAGs by single edge proposal, we makes it possible for the old DAG to be equally sampled as other DAGs in order to explore the bidirected edge space better. The edge reversal proposal is done by choosing an edge ($t \to v$), deleting all incoming edges to t and j, and making new parents including v for t and new parents for v. Since edge reversal proposal is not sufficient to explore all DAGs, it is mixed with single edge proposal [Grzegorczyk and Husmeier, 2008].

In step 3, given the new directed graph G^D, new bidirected edges can be freely added between two nodes where two nodes are not in an ancestor relationship. Otherwise, it violates the ancestral graph condition. To propose new bidirected edges, it begins with excluding node pairs where they are in an ancestor relation in G^D. The remaining node pairs are possible new bidirected edge sites. For each bidirected edge site, the edge is included with a probability p_B, e.g., 0.5.

In step 4, we merge the directed graph G^D and the bidirected graph G^B into a directed ancestral graph G. Since G may not be maximal, we maximize G by adding bidirected edges if two nodes

are always unconditionally and conditionally dependent. During the maximization procedure, the same maximized graph $Max(G)$ can be formed from different bidirected graphs. Hence, the proposal probability $R(Max(G)|G^D)$ is the summation of proposal probabilities of bidirected edges who would be maximized to become the same $Max(G)$.

Steps from 1 to 4 make the move from G_0 to G_1 with a proposal probability $R(G_1^D|G_0^D)R(G_1|G_1^D)$. Hence, in step 5, the new DMAG ($Max(G)$) is accepted in proportion to the multiplication of likelihood ratio $\frac{P(Y|G_1)}{P(Y|G_0)}$ and the Hastings ratio $\frac{R(G_0^D|G_1^D)R(G_0|G_0^D)}{R(G_1^D|G_0^D)R(G_1|G_1^D)}$.

In step 4, the maximum likelihood of a DMAG does not have a closed form. Let Y be a set of vertices in V and $B = (\beta_{ij})$ be a $V \times V$ matrix such that $\beta_{ij} \neq 0$ only if $j \to i$. The corresponding set of recursive linear equations with correlated errors is

$$Y = BY + \epsilon, \qquad \text{where } \epsilon \sim N(0, \Omega).$$

We can define the covariance matrix of Y to be $\Sigma = (I - B)^{-1}\Omega(I - B)^{-T}$ where I is a $V \times V$ identity matrix. Then, $Y \sim N(0, \Sigma)$. One popular way to calculate the maximum likelihood is the following: First, the parameters in B associated with directed edges are calculated by the coefficient in a linear regression. Second, the parameters in $\Omega = (\omega_{ij})$ associated with bidirected edges are calculated by the covariance on the residuals after subtracting directed edge effects. However, the obtained Ω is not always positive definite because it puts zero for no bidirected edge. Drton [2004] proposed an iterative conditional fitting for Gaussian ancestral graph models to find an MLE of (B, Ω). For a fixed vertex i, it computes a submatrix $\Omega_{-i,-i}$ for the remaining vertices and $\beta_{pa(v),j}$ for $j \neq i$, and residuals of ϵ_{-i}. In the conditional Gaussian distribution of Y_i conditional on ϵ_{-i}, the conditional expectation and variance equal to respectively,

$$Var(Y_i|\epsilon_{-i}) = \omega_{ii.-i} - \Omega_{i,-i}(\Omega_{-i,-i})^{-1}\Omega_{-i,i} := w_{ii.-i} \tag{3.13}$$

$$E(Y_i|\epsilon_{-i}) = \sum_{j \in pa(i)} \beta_{ij} E(Y_j|\epsilon_{-i}) + E(\epsilon_i|\epsilon_{-i}) \tag{3.14}$$

$$= \sum_{j \in pa(i)} \beta_{ij} Y_j + \sum_{k \in sp(i)} \omega_{ik} Z_k, \tag{3.15}$$

where Z_k is the pseudo-variable in the k-th row in $Z_{sp(i)} = [(\Omega_{-i,-i})^{-1}]_{sp(i),-i}\epsilon_{-i}$. After computing Z_k from the fixed $\Omega_{-i,-i}$, it fits the linear regression in eqn (3.14) of Y_i with respect to $Y_{pa(i)}$ and

Z_k to get β_{ij} for $j \in pa(i)$ and w_{ik} for $k \in sp(i)$. It also calculates the conditional variance $\omega_{ii.-i}$ in eqn (3.13) from the residual of the fitted regression in eqn (3.14). Next, w_{ii} is calculated from the relation in eqn (3.13). It moves to next vertex iteratively until the MLE of (B, Ω) converges. There is a different method to compute the maximum likelihood for ancestral graphs using only least squares computations [Drton et al., 2009] and a Bayesian inference of parameters for a given graph by setting priors on parameters [Silva and Ghahramani, 2009].

Now we apply the developed MCMC algorithm for the extended network of phenotypes and genotypes. As the linkage map G_Q is fixed, we only need to explore the extended networks of G_Y and $G_{Q\to Y}$ and G_Y is a DMAG. Here we denote G to be the extended network of G_Y and $G_{Q\to Y}$. Steps 1 and 2 are modified to be

1. Propose a new DMAG G'_Y from G_Y.

2. Sample a new genetic architecture $G'_{Q\to Y}$ given G'_Y and get an extended network G' compiosed of $G'_{Q\to Y}$ and G'_Y.

3. Accept the extended network G' with a probability

$$\min\{1, \frac{P(Y|G', Q)}{P(Y|G, Q)} \frac{R(G_Y|G'_Y)}{R(G'_Y|G_Y)} \frac{R(G_{Q\to Y}|G_Y)}{R(G'_{Q\to Y}|G'_Y)}\}.$$

In step 2, sampling a new genetic architecture can be approximated by the interval mapping of each phenotype given the parental phenotypes in G'_Y for faster convergence of MCMC chains.

In step 3, The BIC score is used for the score of the model $P(Y|G)$. The BIC of an ancestral graph is defined to be [Richardson and Spirtes, 2002]

$$BIC = -2 \log L(\hat{B}, \hat{\Omega}) + \log(n)(2|V| + |E|),$$

where $L(\hat{B}, \hat{\Omega})$ is the maximum likelihood function, $|V|$ is the number of vertex, $|E|$ is the number of edges, and n is the sample size. In the DMAG model of phenotypes and QTLs, $|V|$ is defined to be the summation of number of phenotypes and number of identified QTLs and $|E|$ is defined to be the summation of number of directed edges and bidirected edges in the phenotype network and number of directed edges from QTL to phenotypes. The maximum likelihood of a DMAG is approximated

by plugging in the coefficients in the linear regression and the constrained covariance matrix of the residuals of the fitted regression. When the constrained covariance matrix in accordance with bidirected edges is not positive definite, the iterative conditional fitting [Drton, 2004] is used to get a positive definite matrix.

3.4.2 Summarizing MCMC samples

After running an MCMC, we get several sample DMAGs and need a way to summarize the MCMC chain. One way is to choose the most frequently sampled DMAG. This method is inefficient when the model space is too huge. Another way is Bayesian model averaging [Hoeting et al., 1999]. The averaged network is expressed by the posterior probabilities of edge types $(\rightarrow, \leftarrow, \leftrightarrow)$ for every edge. We construct the skeleton first by thresholding the posterior probability of being any edge type at 0.5. Then, we assign an edge type with the highest posterior probability. We use this criteria for model selection in the simulation.

3.5 Simulation

A simulation study is conducted to show that MCMC algorithm for DMAG works. Phenotype and genotype data from the network in Figure 3.3 are generated for 100 simulations. In Figure 3.3 phenotypes are $Y_1, Y_2, Y_3, Y_4, Y_5, c$ but c is hidden and QTLs are Q_1, Q_2, Q_3. We assume that the data is from F2 population and QTL_i is located in the middle of chromosome i. In each simulation, the effects of parental phenotypes are sampled from $0.5U[0.2, 0.5]+0.5U[-0.5, -0.2]$. The additive QTL effects are sampled from $0.5U[0.1, 0.5] + 0.5U[-0.5, -0.1]$ and dominance QTL effects are sampled from $U[-0.25, 0.25]$. The genetic marker data is generated for F2 population of 500 individuals at 10 unequally distributed locations on 3 chromosomes of length 100cM. Based on the sampled effect sizes, phenotype data is generated by a set of linear equations with random errors from $N(0, 1)$.

We consider c to be hidden and the corresponding graph after marginalizing c is in Figure 3.4.

Each simulated data omitting the phenotype c is run with MCMC for 66000 iterations, burned in for the first 6000 iterations, thinned at every 20 iteration, and finally resulted in 3000 samples.

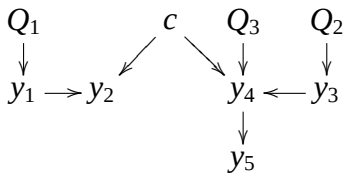


Figure 3.3: A network for simulation

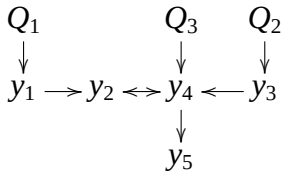


Figure 3.4: A network after marginalizing c out

We apply the model selection criteria by Bayesian model averaging. First, the edges in the true skeleton are recovered as in Figure 3.5. From 100 simulations, the edge Y_1 — Y_2 is recovered in all simulations, Y_2 — Y_4 is recovered in 54 simulations, Y_4 — Y_3 in all simulations, and Y_4 — Y_5 in all simulations. Other edges not in the true skeleton are detected in less than 8 simulations. The reason the edge Y_2 — Y_4 is detected in only half the simulations is that their relation is rather weak due to the hidden variable c. Figure 3.6 shows that the relation between Y_2 and Y_4 is weak as we observe that correlations between Y_2 and Y_4 are distributed closer to 0 than correlations between Y_1 and Y_2 are.

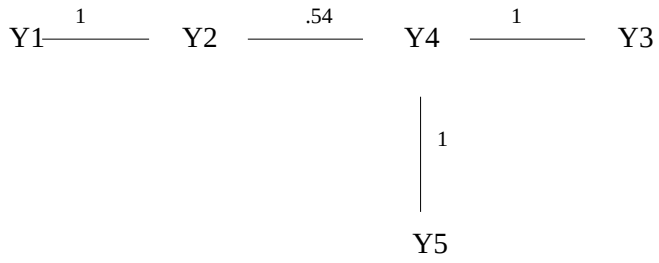


Figure 3.5: Frequency of detection of each edge in the true skeleton

Second, the true edge type is detected more often than other edge types as shown in Figure 3.7. Between Y_1 and Y_2, $Y_1 \rightarrow Y_2$ is assigned in 64 simulations, $Y_1 \leftarrow Y_2$ is assigned in 19 simulations, and $Y_1 \leftrightarrow Y_2$ is assigned in 17 simulations.

3.6 Real data analysis

We applied our method to reconstruct a causal gene network allowing latent variables in F2 mice population. The gene expression and genotype data were obtained from a F2 mice population between diabetes-resistant (B6) and diabetes-susceptible (BTBR) inbred lines, generated by Alan D. Attie's biochemistry lab in University of Wisconsin - Madison. The genome-wide QTL mapping of mRNA transcripts in islet of 491 mice showed several hotspots where several gene expressions co-map. Among them, we are interested in one hotspot around 112.102 cM

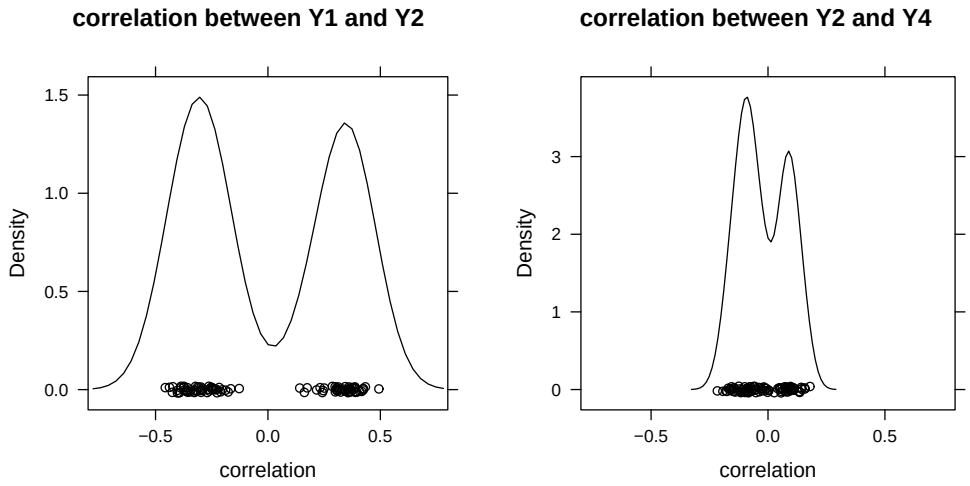


Figure 3.6: Correlation between Y_1 and Y_2 and correlation between Y_2 and Y_4.

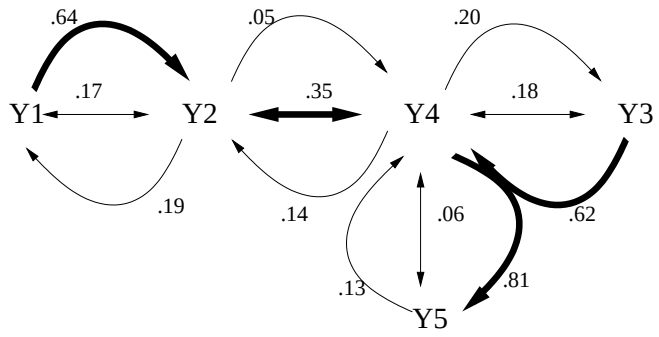


Figure 3.7: Frequency of edge types. Thick arrows correspond to true directions.

(165.74292Mb) in chromosome 2 where 593 genes are mapped to by the single QTL interval mapping. We remove sex effects and batch effects of date by subtracting these effects in advance.

Since we cannot use our MCMC method to reconstruct a gene network of 593 genes, we need to select a subset of genes. We first constructed an undirected graph of 593 genes by estimating a sparse covariance matrix using a graphical lasso method (glasso) [Friedman et al., 2008]. The undirected graph identifies connections between genes such that two genes are correlated conditional on all the other genes. The identified connections do not have the directionality inference but they imply that connected genes may have a regulatory relationship, a correlation that cannot be explained by other genes, or common downstream genes. In either way, connected genes are closely related in a network. We used an arbitrary penalization parameter $\rho = 0.5$, which gives us not too sparse and not too dense undirected networks and the result is shown in Figure 3.8. We observe that there two highly connected clusters of genes and some genes positioned in the outlier are not connected to other genes. The gene of interest is Nfatc2 because it is a transcription factor and it has many target genes in the hotspot. In Figure 3.8, Nfatc2 is numbered to be 428 in red. There are 6 genes that are directly connected to Nfatc2 in Figure 3.9 and 104 genes are connected by at most 2 steps from Nfatc2 in Figure 3.10.

We took Nfatc2 and two directly connected genes (Iqsec1, Pcnt) in Figure 3.9 and constructed a causal gene network allowing latent variables. In Figure 3.10, the top two highly connected genes are Iqsec1 and Pcnt. Iqsec1 is connected to 63 genes and eight of them overlap with the target gene of Nfatc2 and Pcnt is connected to 24 genes and two of them are the target genes of Nfatc2. The gene information of Nfatc2, Iqsec1 and Pcnt are like this: Nfatc2 is a nuclear factor of activated T-cells. Iqsec1 is a protein containing IQ motif and SEC7 domain and it accelerates GTP gamma S binding by ARFs and preferentially functions as a guanine nucleotide exchange protein for ARF6, mediating internalisation of beta-1 integrin. Pcnt is pericentrin which can be bound by calmoduline.

The estimated network by the MCMC method is shown in Figure 3.11. All pairs of Nfatc2, Iqsec1 and Pcnt are connected by bidirected edges, which imply that there could be latent common parents for them. We confirmed that the estimated network is actually the network of the highest

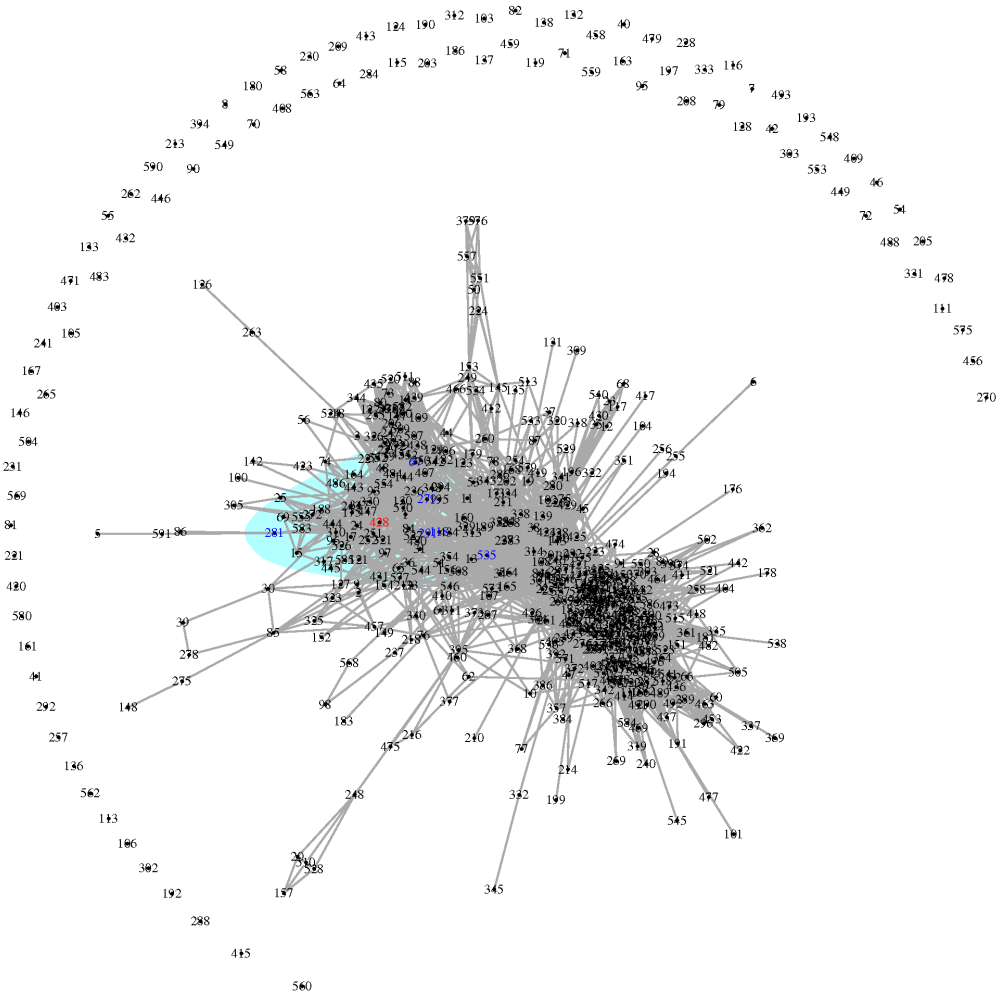


Figure 3.8: 497 genes out of 593 genes are connected to each other in the estimated undirected graph by glasso. The gene of interest (Nfatc2) is numbered to be 428 in red and directly connected genes are colored in blue.

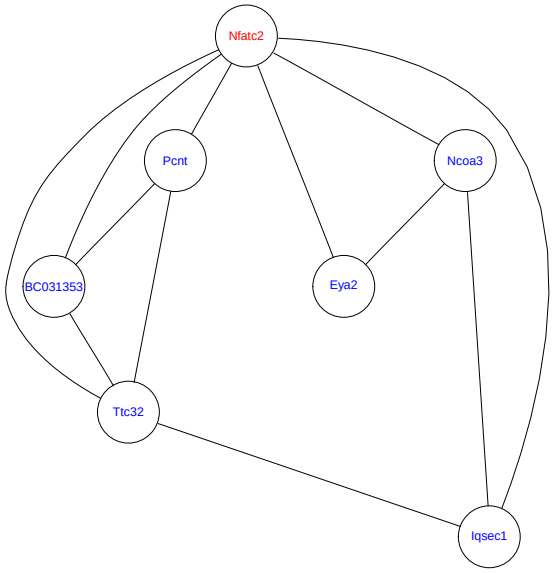


Figure 3.9: 6 genes are directly connected to Nfatc2

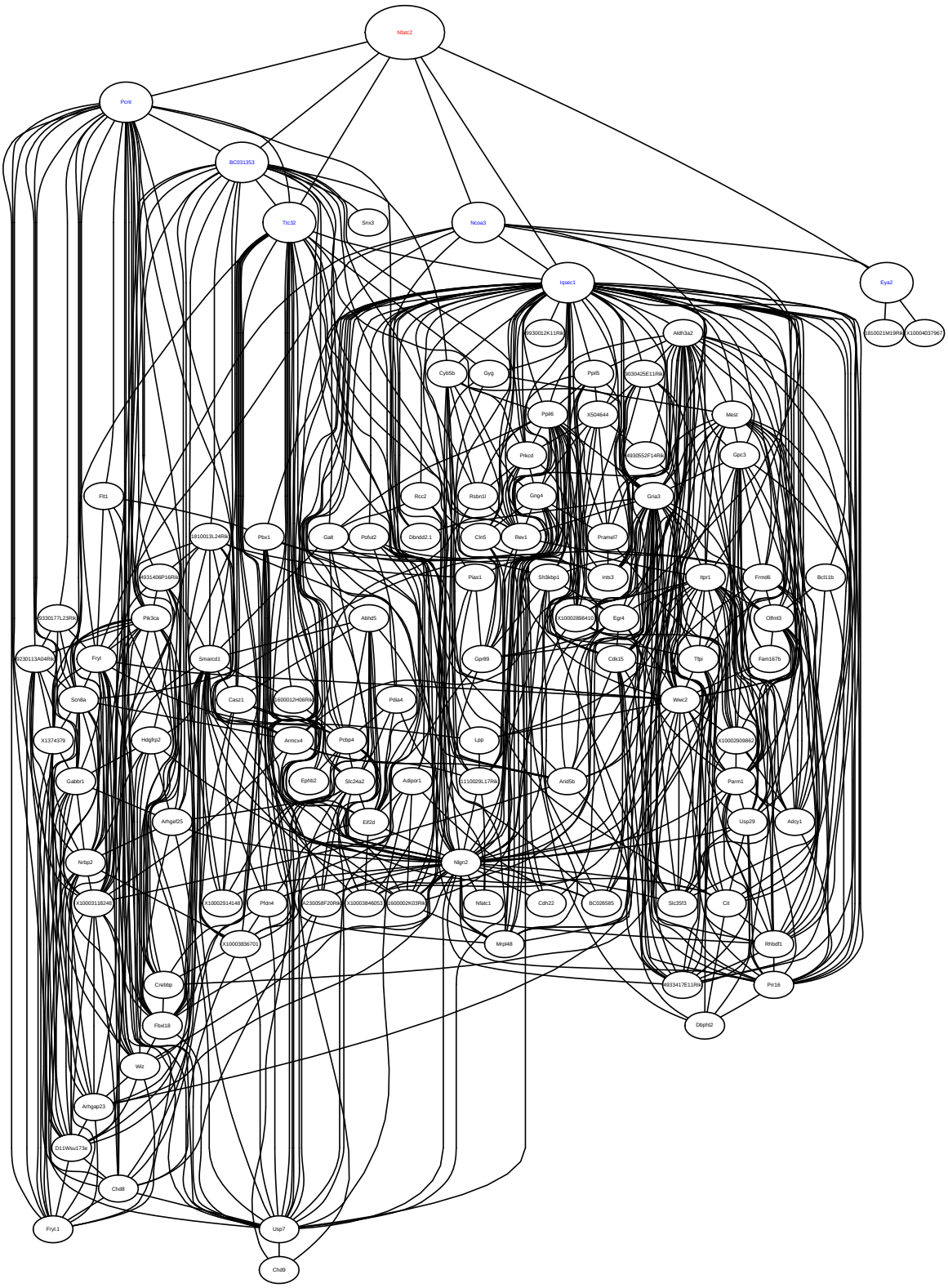


Figure 3.10: 104 genes are connected to Nfatc2 on the top in red by at most 2 steps. 6 genes that are directly connected to Nfatc2 is in blue.

likelihood, and hence MCMC converged well. Figure 3.12 shows that three genes are highly correlated even after adjusting for QTL effects. Three genes are also highly correlated conditional on the other gene respectively, rejecting the conditional dependence with p-value less than 0.0005, and thereby the only suitable network in terms of ancestral graphs would be Figure 3.11.

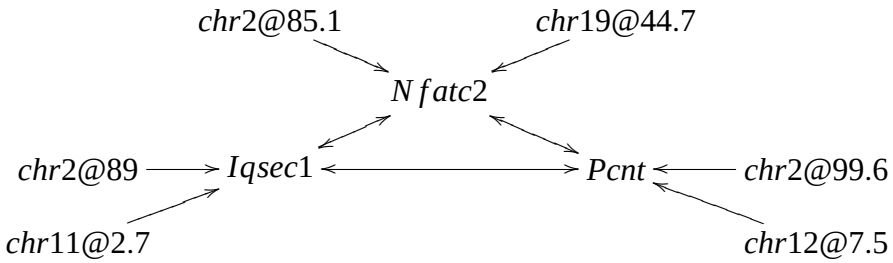


Figure 3.11: A causal network of Nfatc2, Iqsec1 and Pcnt allowing latent variables. chrK@x are identified QTLs at x cM in chromosome K conditional on the phenotype network.

3.7 Conclusion

We have developed a causal gene network with genetic variations allowing latent variables. Since Bayesian networks are not closed under marginalization, ancestral graph Markov models are the proper class of graphical models to consider latent variables. Ancestral graphs can be decomposed into an undirected graph and the remaining graph composed of directed and bidirected edges, where there is no directed or bidirected edges pointing to the undirected graph. This allows the probability of an ancestral graph to be factored into $P(Y_{un_G})P(Y_{V\setminus un_G}|Y_{un_G})$. We use this decomposition property in an analogous way to the decomposition into the recombination linkage map and the causal gene network with QTLs and latent variables in eqn (3.1). Conditional on the recombination linkage map, we model a gene network of QTLs and phenotypes allowing latent variables in eqn (3.2, 3.3, 3.4). The presence of nonzero off-diagonal elements in the covariance of residuals handles the marginalization over latent variables.

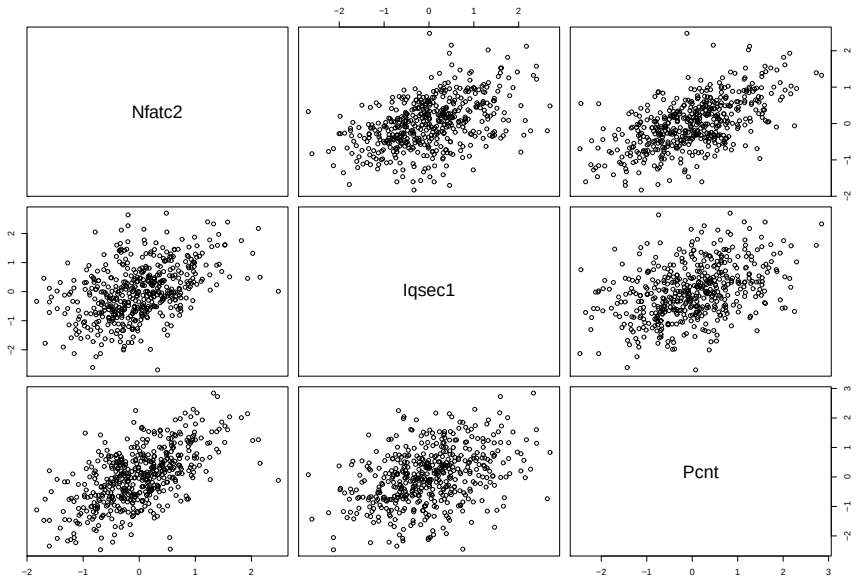


Figure 3.12: Scatter plots of gene expression after adjusting for sex, batch and QTL effects

Based on the ancestral graph modeling, we have proved that 1) QTLs help to distinguish Markov equivalent DMAGs and 2) our parameterization of the network is in a homogeneous conditional Gaussian (HCG) family. We conjecture that distribution equivalence implies the Markov equivalence and vice versa in a Gaussian family and HCG family and suggest the reparameterization between Markov equivalent graphs. We also developed an MCMC for DMAGs. The simulation study shows that the bidirected edge associated with a latent variable is correctly detected in 35% of the simulations in the given simulation setting. Directed edges are correctly detected between 60% to 80% of the simulations. The lower detection rate of the bidirected edge can be explained by lower correlation. The real data analysis on a subset of genes co-mapped to a hotspot in islet, on the contrary, identifies bidirected edges between genes. These genes show high correlations and partial correlations with each other, and hence any combination of directed edges cannot explain their relationships better than the bidirected graph. The bidirected graph can imply an hypothesis that there could be one or more gene products or signaling molecules that may govern these genes.

As ancestral graphs are more flexible than directed acyclic graphs, there is price to pay for allowing latent variables in ancestral graphs. The number of directed maximal ancestral graphs, for the number of nodes $p = 1, 2, \cdots, 5$, is 1, 4, 56, 2492, 328924 while the number of directed acyclic graphs is 1, 3, 25, 543, 29281. The space for ancestral graphs grows faster than the space for DAGs, which already grows super-exponentially. Hence, we need a very efficient algorithm to search over the ancestral graph space. In the proposed MCMC algorithm, the better proposal distribution of new DMAG is desirable which not only jumps from the current DAG subgraph but also jumps from the current bidirected graph.

Chapter 4

Summary

4.1 Summary of my work

In Chapter 2, we incorporated biological knowledge to reconstruct a causal gene network from genotypes and phenotypes. A Bayesian network is modeled for a causal network of genotypes and phenotypes. Biological knowledge sets the prior on phenotype network, which is a sub-graph of the whole causal gene network, that is, a causal network among phenotypes. The weight parameter is introduced as a hyper parameter in the prior of phenotype network to control the contribution of biological knowledge and the hyper prior for the weight is also set. The proposed model, QTLnet-prior, shows the improvement of recovery of network when correct biological knowledge is incorporated in a simulation study. When misleading biological knowledge is incorporated, there is a decrease in the recovery of networks, however, the decrease is controlled somewhat by the weight parameter so that the influence of biological knowledge could be negligible. By using the weight parameter, QTLnet-prior is robust to noisy biological knowledge. The application of QTLnet-prior on 26 yeast cell cycle genes with transcription factor binding information infers a network that does not make much use of transcription factor binding information. There are several explanations for this result. One is transcription factor binding information is noisy, especially since the data is generated by Chip-chip in 2002. Another explanation is that transcription factor binding information is not consistent with the underlying network to generate gene expression. It is also possible that there is a biological network with proteins, transcription factors, mRNAs and metabolites as elements and mRNA expression alone may not reflect the underlying biological network well.

In Chapter 3, we constructed a causal gene network of genotypes and phenotypes allowing latent variables. An ancestral graph is modeled to take into account the possibility of latent variables. The possibility of latent variables is considered because the measurements may omit significant variables in the network, and we often take a subset of variables to construct a network. We showed that 1) QTLs help to distinguish Markov equivalent ancestral graphs and 2) our model is a homogeneous conditional Gaussian (HCG) distribution. We proposed an MCMC algorithm to explore the space of DMAGs and presented the recovery of networks in the simulation study. We applied the algorithm to real data of 3 genes (Nfatc2, Iqsec1, Pcnt) in F2 mice population and the results suggested latent variables for all pairs of 3 genes.

The two chapters develop methods for the causal gene network inference from experimental cross study. They are intended to decipher how genotypes cause the change in gene expressions and how gene expressions cause the change in other gene expressions. The first method, QTLnet-prior, incorporates biological knowledge to get more comprehensive causal gene networks. The second method prevents us from making spurious causal relations by latent variables as much as we can. Using both methods together will generate hypothetical regulatory relationships and indicate hypothetical master regulators, which could be verified through biological experiments.

There are some limitations of our approach. One is that the currently developed algorithms cannot handle a large number of genes. Second is that both methods are based on linear models, but nonlinear relationships are important too.

4.2 Future work

The construction of large networks will be valuable to get a big picture of how gene products regulate each other. To accomplish this goal, we need to do the following: 1) development of a fast code using low-level programming language such as C and efficient algorithms such as Metropolis-coupled MCMC, and 2) development of a parallel computing algorithm through the divide-and-conquer strategy. The parallel computing algorithm will first divide the genes into sets of genes that would be closely connected in the undirected graph estimated by glasso [Friedman et al., 2008]. Second, it will construct causal subnetworks on each set in parallel. Third, the theory

in ancestral graphs about marginalization will be incorporated to develop a method to integrate subnetworks to infer one large network. Another aspect to improve in network inference to reflect biological relationship is the nonlinear relationship, since not all relationship can be approximated by linear parameterization. We could discretize variables and fit discretized values to capture nonlinear relationships. Nonparametric or semiparametric approach for network inference could be developed as well.

Causal network analysis on different conditions and developmental stages is of great interest. For example, hematopoietic stem cells (HSCs) give rise to various blood cells and these HSCs underwent development from embryo to adult [McKinney-Freeman et al., 2012]. By deciphering the causal networks on gene expressions at different developmental stages, we could identify key elements for each stage and the identified key elements could be used to induce the differentiation of blood cells from induced pluripotent stem cells (iPSCs). Since causal networks at each developmental stage may share some common features and turning off an upstream gene will affect the downstream genes, the joint prior probability on networks can be constructed to satisfy these two properties.

There are several statistically interesting problems arising from causal network inference. One is the violation of faithfulness assumption. The faithfulness assumption is that there are no conditional independencies other than the graph entails. The set of unfaithfulness distributions has Lebesgue measure zero where coefficients of the graph can cancel out making additional conditional independencies. To estimate a network, a strong faithfulness is assumed due to sampling error. The strong faithfulness assumption modifies the faithfulness assumption in that the conditional independence is defined to be in the small neighborhood of 0. Uhler et al. [2012] proved that there is a large volume of strong unfaithful distributions and the inference of causal networks based on conditional independence tests such as PC-algorithm will have fundamental limitations. We interpret this limitation as a multiple testing problem and it will be worthwhile to investigate the relationship between the likelihood-based method such as glasso [Friedman et al., 2008] and faithfulness assumption. Another interesting problem arises when we summarize the MCMC result of networks. Since networks can be Markov equivalent to each other, careful summarizing methods

will be needed. Since Markov equivalent models have the same set of conditional independencies, we can summarize the posterior probability of conditional independencies and reconstruct a network from significant conditional independence relations.

Since the fundamental questions in biology are causal relations, causal network analysis will be applied more and more. A statistical theory to support the consistent causal analysis will enhance the wide use of causal networks.

APPENDIX
Additional Figures

A.1 Inferred yeast cell cycle network with causal QTLs integrating TF information by QTLnet-prior

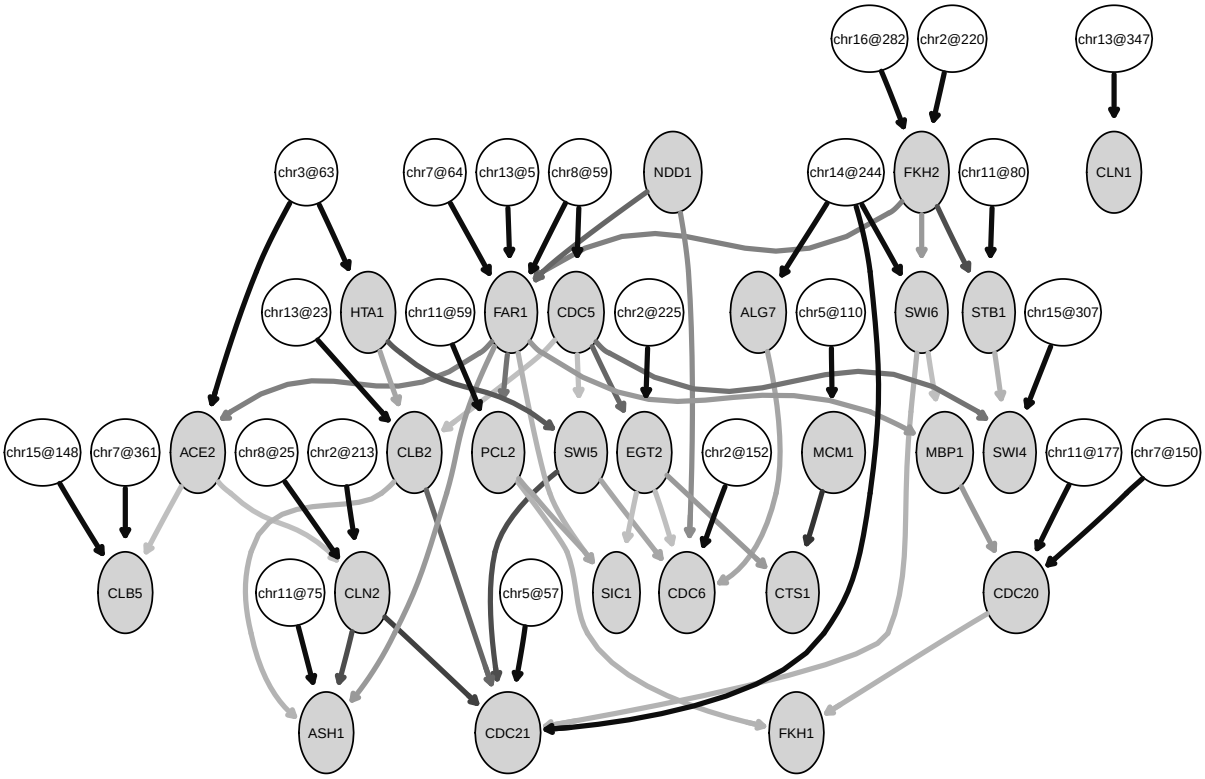


Figure A.1: Yeast cell cycle network integrating transcription factor binding information inferred by QTLnet-prior. The edge darkness is in proportion to the posterior probability.

A.2 Convergence diagnostics of yeast cell cycle network

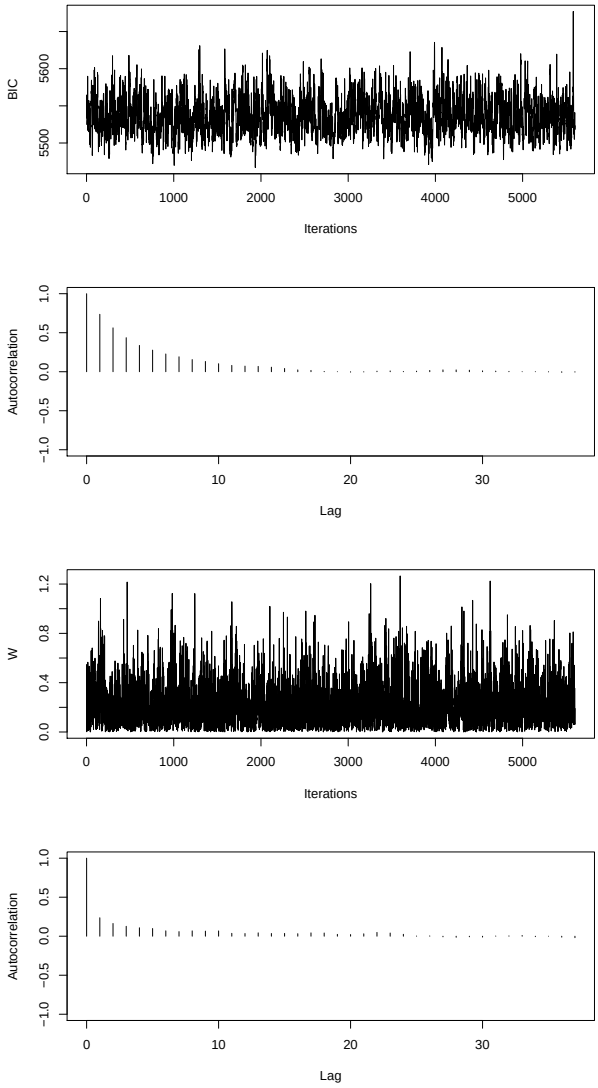


Figure A.2: The top two figures are the trace plot and the autocorrelation plot of BIC scores for sampled causal networks. The bottom two figures are the trace and the autocorrelation plots of the sampled weights (W) on transcription factor binding information.

Bibliography

R. A. Ali, T. S. Richardson, and P. Spirtes. Markov equivalence for ancestral graphs. *Annals of Statistics*, 37(5B):2808–2837, 2009.

J. D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao. Comparing statistical methods for constructing large scale gene networks. *PLoS ONE*, 7(1), 2012.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, and Consortium Gene Ontology. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

J. E. Aten, T. F. Fuller, A. J. Lusis, and S. Horvath. Using genetic markers to orient the edges in quantitative trait networks: The NEO software. *BMC Systems Biology*, 2:34, 2008.

G. D. Bader, D. Betel, and C. W. V. Hogue. BIND: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250, 2003.

J. Bähler. Cell-cycle control of gene expression in budding and fission yeast. *Annual Review of Genetics*, 39(1):69–94, 2005.

M. Bansal, G. D. Gatta, and D. di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–822, 2006.

A. Bauer and B. Kuster. Affinity purification-mass spectrometry - Powerful tools for the characterization of protein complexes. *European Journal of Biochemistry*, 270(4):570–578, 2003.

A. Bernard and A. J. Hartemink. Informative structure priors: Joint learning of dynamic regulatory networks from multiple types of data. *Pacific Symposium on Biocomputing*, pages 459–470, 2005.

O. R. P. Bininda-Emonds, J. L. Gittleman, and M. A. Steel. The (Super)tree of life: Procedures, problems, and prospects. *Annual Review of Ecology and Systematics*, 33:265–289, 2002.

R. B. Brem and L. Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences*, 102(5):1572–1577, 2005.

K. W. Broman. Review of statistical methods for QTL mapping in experimental crosses. *Laboratory Animals*, 30(7):44–52, 2001.

K. W. Broman and T. R. Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 64(4):641–656, 2002.

K. W. Broman, H. Wu, S. Sen, and Gary A. G.A. Churchill. R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19(7):889–890, 2003.

E. Chaibub Neto, C. T. Ferrara, A. D. Attie, and B. S. Yandell. Inferring causal phenotype networks from segregating populations. *Genetics*, 179(2):1089–1100, 2008.

E. Chaibub Neto, M. P. Keller, A. D. Attie, and B. S. Yandell. Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Annals of Applied Statistics*, 4(1):320–339, 2010a.

E. Chaibub Neto, M. P. Keller, A. T. Broman, A. D. Attie, and B. S. Yandell. Causal model selection tests in systems genetics. Technical Report 1157, Department of statistics, University of Wisconsin-Madison, 2010b.

L. S. Chen, F. Emmert-Streib, and J. D. Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology*, 8(10):R219, 2007.

S. Christley, Q. Nie, and X. Xie. Incorporating existing network information into gene network inference. *PLoS ONE*, 4(8):e6799, 2009.

G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

Robert G. Cowell, Philip A. Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems (Information Science and Statistics)*. Springer, New York, 2003.

H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9:67–103, 2002.

J. De Las Rivas and C. Fontanillo. Protein-Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Computational Biology*, 6(6), 2010.

A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

M. Drton. Iterative conditional fitting for gaussian ancestral graph models. In *In M. Chickering and J. Halpern (Eds.), Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 130–137. Morgan Kaufmann, 2004.

M. Drton, M. Eichler, and T. S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *J. Mach. Learn. Res.*, 10:2329–2348, 2009.

J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, and Mouse Genome Database Group. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Research*, 40(D1):D881–D886, 2012.

T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

C. Francke, R. J. Siezen, and B. Teusink. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, 13(11):550 – 558, 2005.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3–4):601–620, 2000.

N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. pages 139–147. Morgan Kaufmann, 1998.

T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629):102–105, 2003.

A.-C. Gavin, K. Maeda, and S. Kuehner. Recent advances in charting protein-protein interaction: mass spectrometry-based approaches. *Current Opinion in Biotechnology*, 22(1):42–49, 2011.

F. Geier, J. Timmer, and C. Fleck. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Systems Biology*, 1:11, 2007.

D. Geiger and D. Heckerman. Learning Gaussian networks. Technical Report MSR-TR-94-10, Microsoft Research, 1994.

J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, pages 169–193. University Press, 1992.

M. Grzegorczyk and D. Husmeier. Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, 71:265–305, 2008.

M. Grzegorczyk and D. Husmeier. Improvements in the reconstruction of time-varying gene regulatory networks: Dynamic programming and regularization by information sharing among genes. *Bioinformatics*, 27(5):693–699, 2011.

R. S. Hageman, M. S. Leduc, R. Korstanje, B. Paigen, and G. A. Churchill. A Bayesian framework for inference of the genotype-phenotype map for segregating populations. *Genetics*, 187:1163–1170, 2011.

C. S. Haley and S. A. Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69(4):315–324, 1992.

A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 422–33, 2001.

D. Heckerman and D. Geiger. Likelihoods and parameter priors for Bayesian networks. Technical Report MSR-TR-95-94, Microsoft Research, 1996.

D. Heckerman, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. In Dawn Holmes and Lakhmi Jain, editors, *Innovations in Machine Learning*, volume 194 of *Studies in Fuzziness and Soft Computing*, pages 1–28. Springer Berlin / Heidelberg, 2006.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–417, 1999.

D. Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19(17):2271–2282, 2003.

S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Journal of Bioinformatics and Computational Biology*, 2(1):77–98, 2004.

S. Imoto, S. Kim, T. Goto, S. Miyano, S. Aburatani, K. Tashiro, and S. Kuhara. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *Journal of Bioinformatics and Computational Biology*, 1(2):231–52, 2003.

R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4), 2003.

T. Jaakkola, D. Sontag, A. Globerson, and M. Meila. Learning Bayesian network structure using LP relaxations. *Journal of Machine Learning Research*, 9:358–365, 2010.

R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, 2003.

R. C. Jansen and J.-P. Nap. Genetical genomics: the added value from segregation. *Trends in Genetics*, 17(7):388 – 391, 2001.

R. C. Jansen and P. Stam. High-resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, 136(4):1447–1455, 1994.

B. H. Junker and F. Schreiber. *Analysis of Biological Networks*. Wiley-Interscience, 2008.

C.-H. Kao and Z.-B. Zeng. Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics*, 160:1243–1261, 2002.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90 (430):773–795, 1995.

D. Kulp and M. Jagalur. Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC Genomics*, 7(1):125, 2006.

E. S. Lander and D. Botstein. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–199, 1989.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804, 2002.

R. Li, S.-W. Tsaih, K. Shockley, I. M. Stylianou, J. Wergedal, B. Paigen, and G. A. Churchill. Structural model analysis of multiple quantitative traits. *PLoS Genetics*, 2(7):e114, 2006.

B. Liu, A. de la Fuente, and I. Hoeschele. Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, 178(3):1763–1776, 2008.

Benjamin A. Logsdon and Jason Mezey. Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Comput Biol*, 6(12):e1001014, 12 2010.

P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232, 1995.

A. Manichaikul, J.-Y. Moon, S. Sen, B. S. Yandell, and K. W. Broman. A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genetics*, 181(3):1077–1086, 2009.

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

S. McKinney-Freeman, P. Cahan, H. Li, S. A. Lacadie, H.-T. Huang, M. Curran, S. Loewer, O. Naveiras, K. L. Kathrein, M. Konantz, E. M. Langdon, C. Lengerke, L. I. Zon, J. J. Collins, and G. Q. Daley. The transcriptional landscape of hematopoietic stem cell ontogeny. *Cell Stem Cell*, 11(5), 2012.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

J. Millstein, B. Zhang, J. Zhu, and E. Schadt. Disentangling molecular relationships with a causal inference test. *BMC Genetics*, 10(1):23, 2009.

K. Murphy and S. Mian. Modelling gene expression data using dynamic Bayesian networks. Technical report, 1999.

J. H. Nadeau and A. M. Dudley. Systems genetics. *Science*, 331(6020):1015–1016, 2011.

N. Nariai, S. Kim, S. Imoto, and S. Miyano. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pacific Symposium on Biocomputing 2004*, pages 336–47, 2004.

S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, A. K. Johnson, M. T. Maurano, R. Humbert, E. Rynes, H. Wang, S. Vong, K. Lee, D. Bates, M. Diegel, V. Roach, D. Dunn, J. Neri, A. Schafer, R. S. Hansen, T. Kutyavin, E. Giste, M. Weaver, T. Canfield, P. Sabo, M. Zhang, G. Balasundaram, R. Byron, M. J. MacCoss, J. M. Akey, M. A. Bender, M. Groudine, R. Kaul, and J. A. Stamatoyannopoulos. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489 (7414):83–90, 2012.

M. Olson, L. Hood, C. Cantor, and D. Botstein. A common language for physical mapping of the human genome. *Science*, 245(4925):1434 – 1435, 1989.

O. Ourfali, T. Shlomi, T. Ideker, E. Ruppin, and R. Sharan. SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, 23(13):i359–i366, 2007.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference*. Morgan Kaufmann, 1988.

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

D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(suppl 1):S215–S224, 2001.

T. Peleg, N. Yosef, E. Ruppin, and R. Sharan. Network-free inference of knockout effects in yeast. *PLoS Computational Biology*, 6(1):e1000635, 2010.

C. A. Penfold and D. L. Wild. How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6):857–870, 2011.

J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

E. Perrier, S. Imoto, and S. Miyano. Finding optimal Bayesian network given a super-structure. *Journal of Machine Learning Research*, 9:2251–2286, 2008.

M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13, 2012.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4): 962–1030, 2002.

C. Riggelsen. MCMC learning of Bayesian network models by Markov blanket decomposition. In *Machine Learning: ECML 2005*, volume 3720 of *Lecture Notes in Computer Science*, pages 329–340. Springer Berlin / Heidelberg, 2005.

S. Rogers, M. Girolami, W. Kolch, K. M. Waters, T. Liu, B. Thrall, and H. S. Wiley. Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics*, 24(24):2894–2900, 2008.

J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. M. Li, J. S. Albala, J. H. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

J. Ruan, A. K. Dean, and W. Zhang. A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC Systems Biology*, 4, 2010.

E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. S. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. M. Yang, J. Castle, H. Y. Zhu, S. F. Kash, T. A. Drake, A. Sachs, and A. J. Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37 (7):710–717, 2005.

J. Schäfer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

C. Schlötterer. The evolution of molecular markers - just a matter of fashion? *Nature Reviews Genetics*, 5(1):63–69, 2004.

M. Schmidt, A. Niculescu-Mizil, and K. Murphy. Learning graphical model structure using L1-regularization paths. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, pages 1278–1283. AAAI Press, 2007.

E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19:i273–i282, 2003.

S. Sen and G. A. Churchill. A statistical framework for quantitative trait mapping. *Genetics*, 159 (1):371–387, 2001.

S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.

R. Silva and Z. Ghahramani. The hidden life of latent variables: Bayesian learning with mixed graph models. *J. Mach. Learn. Res.*, 10:1187–1238, 2009.

L. M. Silver. *Mouse Genetics: Concepts and Applications*. Oxford University Press, 1995.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, second edition, 2000.

U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

B. Suter, S. Kittanakom, and I. Stagljar. Two-hybrid technologies in proteomics research. *Current Opinion in Biotechnology*, 19(4):316–323, 2008.

A. C. Syvänen. Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, 2(12):930–942, 2001.

Y. Tamada, S. Imoto, and S. Miyano. Parallel algorithm for learning optimal Bayesian network structure. *Journal of Machine Learning Research*, 12:2437–2459, 2011.

Y. Tamada, S. Kim, H. Bannai, S. Imoto, K. Tashiro, S. Kuhara, and S. Miyano. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, 19(suppl 2):ii227–ii236, 2003.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 2006.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of faithfulness assumption in causal inference. *ArXiv e-prints*, July 2012.

B. D. Valente, G. J. M. Rosa, G. de los Campos, D. Gianola, and M. A. Silva. Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics*, 185: 633–644, 2010.

T. Verma and J. Pearl. *Equivalence and synthesis of causal models. In Readings in Uncertain Reasoning (G. Shafer and J. Pearl eds.)*. Morgan Kaufmann, 1990.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

A. V. Werhli and D. Husmeier. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology*, 6:15, 2007.

C. J. Winrow, D. L. Williams, A. Kasarskis, J. Millstein, A. D. Laposky, H. S. Yang, K. Mrazek, L. Zhou, J. R. Owens, D. Radzicki, F. Preuss, E. E. Schadt, K. Shimomura, M. H. Vitaterna, C. Zhang, K. S. Koblan, J. J. Renger, and F. W. Turek. Uncovering the genetic landscape for multiple sleep-wake traits. *PLoS ONE*, 4(4):e5161, 2009.

H. Yang, Y. Ding, L. N. Hutchins, J. Szatkiewicz, T. A. Bell, B. J. Paigen, J. H. Graber, F. P.-M. de Villena, and G. A. Churchill. A customized and versatile high-density genotyping array for the mouse. *Nature Methods*, 6(9):663–U55, 2009.

C. H. Yeang, T. Ideker, and T. Jaakkola. Physical network models. *Journal of Computational Biology*, 11(2–3):243–262, 2004.

N. J. Yi. A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics*, 167(2):967–975, 2004.

N. J. Yi, D. Shriner, S. Banerjee, T. Mehta, D. Pomp, and B. S. Yandell. An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. *Genetics*, 176:1865–1877, 2007.

N. J. Yi, B. S. Yandell, G. A. Churchill, D. B. Allison, E. J. Eisen, and D. Pomp. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, 170(3):1333–1344, 2005.

M. Zahurak, G. Parmigiani, W. Yu, R. B. Scharpf, D. Berman, E. Schaeffer, S. Shabbeer, and L. Cope. Pre-processing Agilent microarray data. *BMC Bioinformatics*, 8, 2007.

B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4:17, 2005.

J. Zhang and P. Spirtes. A transformational characterization of Markov equivalence for directed maximal ancestral graphs. Technical Report 168, Department of Philosophy, Carnegie Mellon University, 2005.

H. Zhao and T. P. Speed. On genetic map functions. *Genetics*, 142(4):1369–1377, 1996.

J. Zhu, P. Y. Lum, J. Lamb, D. GuhaThakurta, S. W. Edwards, R. Thieringer, J. P. Berger, M. S. Wu, J Thompson, A. B. Sachs, and E. E. Schadt. An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenetic Genome Research*, 105 (2–4):363–374, 2004.

J. Zhu, M. C Wiener, C. Zhang, A. Fridman, E. Minch, P. Y. Lum, J. R. Sachs, and E. E. Schadt. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Computational Biology*, 3(4):e69, 2007.

J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, and E. E. Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, 40(7):854–861, 2008.