SPARSE MODEL BUILDING FROM GENOME-WIDE VARIATION WITH GRAPHICAL MODELS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Benjamin Alexander Logsdon

January 2011

SPARSE MODEL BUILDING FROM GENOME-WIDE VARIATION WITH GRAPHICAL MODELS

Benjamin Alexander Logsdon, Ph.D.

Cornell University 2011

High throughput sequencing and expression characterization have lead to an explosion of phenotypic and genotypic molecular data underlying both experimental studies and outbred populations. We develop a novel class of algorithms to reconstruct sparse models among these molecular phenotypes (e.g. expression products) and genotypes (e.g. single nucleotide polymorphisms), via both a Bayesian hierarchical model, when the sample size is much smaller than the model dimension (i.e. $p \gg n$) and the well characterized adaptive lasso algorithm. Specifically, we propose novel approaches to the problems of increasing power to detect additional loci in genome-wide association studies using our variational algorithm, efficiently learning directed cyclic graphs from expression and genotype data using the adaptive lasso, and constructing genome-wide undirected graphs among genotype, expression and downstream phenotype data using an extension of the variational feature selection algorithm. The Bayesian hierarchical model is derived for a parametric multiple regression model with a mixture prior of a point mass and normal distribution for each regression coefficient, and appropriate priors for the set of hyperparameters. When combined with a probabilistic consistency bound on the model dimension, this approach leads to very sparse solutions without the need for cross validation. We use a variational Bayes approximate inference approach in our algorithm, where we impose a complete factorization across all parameters for the approximate posterior distribution, and then minimize the Kullback-Leibler divergence between the approximate and true posterior distributions. Since the prior distribution is non-convex, we restart the algorithm many times to find multiple posterior modes, and combine information across all discovered modes in an approximate Bayesian model averaging framework, to reduce the variance of the posterior probability estimates. We perform analysis of three major publicly available data-sets: the HapMap 2 genotype and expression data collected on immortalized lymphoblastoid cell lines, the genome-wide gene expression and genetic marker data collected for a yeast intercross, and genome-wide gene expression, genetic marker, and downstream phenotypes related to weight in a mouse F2 intercross. Based on both simulations and data analysis we show that our algorithms can outperform other state of the art model selection procedures when including thousands to hundreds of thousands of genotypes and expression traits, in terms of aggressively controlling false discovery rate, and generating rich simultaneous statistical models.

BIOGRAPHICAL SKETCH

Ben grew up in Palmer, Alaska where he graduated from high school in 2002. He moved to Pullman, Washington in 2002, to attend Washington State University, where he graduated Summa cum laude with a B.S. in biochemistry and minor in mathematics in 2006. At Washington State University he worked under the supervision of Richard Gomulkiewicz, Ph.D., studying the neutral evolution of multivariate heritable variation. After graduating from WSU, he moved to Ithaca, New York as a graduate student under Jason Mezey, Ph.D., developing novel statistical methods to learn the structure of models underlying complex phenotypic and genotypic variation. An avid runner, Ben has competed in many cross country and road races since high school, and was also actively involved in the student wind ensembles at both WSU and Cornell, participating in both places as lead French horn player.

This thesis is dedicated to my grandfather, Charles E Logsdon, Ph.D., who instilled a passionate interest in understanding complex and contingent causal processes.

ACKNOWLEDGEMENTS

I would first and foremost like to thank my parents. Without their constant support, I would not have overcome the many obstacles and challenges I have encountered thus far. I would also like to thank my brother, for his support and biting sense of humor, as well as intelligence and compassion. I owe a debt of gratitude to many people at Cornell, especially Gabe and Larsson, who both significantly helped me hone many of the ideas I had during my time at Cornell, as well as provide invaluable feedback on manuscripts, and other projects. In addition, Gabe significantly contributed to the results of Chapter 2, by implementing, and running the lasso algorithm. In addition I thank Larsson, Pavel, and Patrick, for the afternoon runs during the winters that helped keep me sane. I would also like to thank my funding sources from Cornell including the Presidential Life Science Fellowship, a Provost fellowship, a fellowship from the Center for Vertebrate Genomics, and a fellowship from the Cornell Center for Comparative and Population Genomics. I also wish to acknowledge my committee, Andrew Clark, Martin Wells, Carlos Bustamante, Adam Siepel, and my thesis advisor, Jason Mezey. I would also definitely not have made it this far without Brian, Attila, and Matt for their friendship and support. Finally I thank Heather, who has consistently been there for me; whose support and encouragement are invaluable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 The genotype-phenotype map

Identifying the mutations in the genetic code that give rise to variation in the phenotype of an organism has been a central goal of the field of genetics since its inception [42]. Specifically, the goal of mapping different regions of the genome as being related to a given phenotype is an important first step in understanding the biology of a phenotype through its genetic architecture [90]. There have been many study designs and associated statistical methodologies proposed to address this problem, including linkage mapping [1], family based association tests [41], and genome-wide association studies [63] among others. Phenotypes with Mendelian genetic architectures (i.e. few loci of strong effect/penetrance) are more amenable to such analyses, since the test statistic for a given region of the genome will be well powered, even for small sample sizes [36]. Alternatively, when the underlying genetic architecture of a phenotype is complex, i.e. includes many loci of small effect and/or complex epistatic interactions, then marginal test statistics (i.e. testing loci individually) can become significantly underpowered, especially for realistic sample sizes and multiple testing corrections [63, 36]. Many phenotypes of interest have an underlying complex genetic architectures including the human disease phenotypes of Crohn's disease, diabetes, hypertension, coronary artery disease, bipolar disorder [124, 132, 67, 126, 35, 152, 29], obesity [128], as well as height [139], among others.

If a phenotype has a complex, polygenic architecture, ideally one would use a multiple feature statistical model to test which loci are linked to the phenotype, because the ordering of the significance scores (e.g. P-values) can be different between a multiple feature model and a marginal test statistic model [45]. The reordering can happen, even in the case of a set of mutually orthogonal features, because the estimate of the variance of the error term for a marginal test statistic will be inflated from the other true, unmeasured features in the model. Given that the P-value of a given feature in a linear model is a direct function of this estimate of the error variance, the significance scores will be randomly resorted based on how much the variance of the error is inflated because of the variance of the predicted effect for all additional true features in the model. Unfortunately, standard multiple regression test statistics, such as a least squares estimator, are ill-conditioned for realistic genomic data sets, where the sample size is significantly smaller than the number of features being tested, (e.g. thousands of samples, and millions of features in Genome-wide Association Studies) [65, 142]. These observations are the main motivation for the development of the novel methods presented in this thesis, specifically development of novel variational Bayes sparse feature selection methodologies presented in Chapters 2 and 4.

With the advent of high-throughput genotyping and sequencing, there has also been an explosion in molecular phenotype characterization; specifically characterization of the levels of expression of genes within the cell [110, 27]. These intermediary phenotypes provide a window into possible modes of action for genetic variation through expression quantitative trait loci (eQTL), as suggested by previous authors [71, 108]. These different modes of action include both *cis* (e.g. physically local) and *trans* (e.g. physically distant) genetic effects [113, 16, 89, 161, 84]. The genomic distribution of these effects has been characterized in a variety of organisms, including mice, [58], humans [39, 23], as well as yeast [161] among others, where broad patterns such as numerous eQTL hotspots, as well as on average strong *cis* effects have been observed across organisms and cell types. It has also been proposed that these eQTL effects can be treated as perturbations of an underlying regulatory system, and can be used to learn the nature of regulatory relationships among gene expression products within the cell [108]. In chapter 3 we propose a novel methodology to leverage *cis*-eQTL effects to learn the structure of a broad class of graphical models, including directed cyclic networks.

The primary goal of this thesis is to motivate the results of the final chapter, where all the levels of variation (e.g. genotypic, expression, and downstream phenotype) for a biological system are integrated, and a highly sparse multi-feature model is generated where all the interactions identified have strong statistical evidence. This type of approach has been proposed previously by many authors [23, 39, 110, 112, 66, 161, 27, 91], but we provide the first computationally scalable and rigorously statistically significant multiple feature methodology that can be easily applied to these data. To arrive at that point we need to motivate certain aspects of the modeling approach we take; specifically we wish to define the rich statistical substrate of graphical models. In the following we give a very brief introduction to some of the definitions and ideas used in the field of graphical models. More detailed introductions are presented in Lauritzen [80], Jordan [136] and Bishop [9].

1.2 An introduction to graphical models

The field of graphical models has grown in popularity over the last two decades, with direct applications to genomics including modeling probabilistic regulatory networks [50], gene finders and comparative genomic analyses [121, 117], and even admixture analyses [18]. A generative graphical model can be defined for a particular joint distribution:

$$p\left(x_1, x_2, \ldots, x_n\right) \tag{1.1}$$

if the distribution satisfies a set of conditional dependence and independence relationships, that can be mapped onto the graph structure and vice versa []. For example, a conditional distribution that factorizes as:

$$p\left(x_1, x_2 | x_3\right) = p\left(x_1 | x_3\right) p\left(x_2 | x_3\right), \tag{1.2}$$

where the random variables x_1 and x_2 are independent when conditioning on the state of the variable x_3. This type of relationships could be mapped onto any of the graphs as shown in Figure 1.1, but not the graph shown in Figure 1.2, because conditioning on x_3 induces a dependence between the marginally independent variables x_1 and x_2. These mappings are generally represented as different types of separation criterion [120, 100]. They can be used to defined various Markov properties (i.e. conditional independence relationships) of a network, including the pairwise, local, or global Markov properties [80]. One of the first graphical models proposed was the covariance selection model [31], now known as the Gaussian graphical model [80]. The Gaussian graphical model is an alternative parameterization of a multivariate Gaussian random variable, where the zero/non-zero structure of the inverse covariance matrix acquires an additional interpretation in terms of conditional independence relationships among the random variables in the system. This type of model has

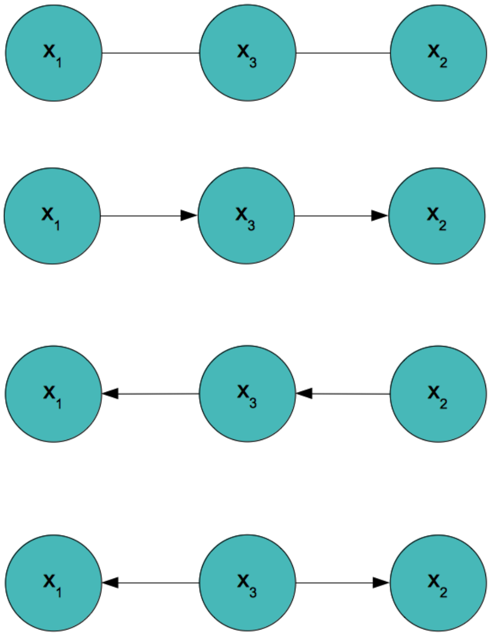

Figure 1.1: Different undirected and directed graphical models consistent with the conditional independence relationship specified by Equation 1.2

been proposed to be used for analysis of expression data [114, 79], as well as combined genotype and expression data by Chu et al. [25].

1.2.1 Markov random fields defined over undirected graphs

Consider an undirected graph $\mathcal{G}_{UG} = (\mathcal{V}_{UG}, \mathcal{E}_{UG})$ defined by a set of vertices, $\mathcal{V}_{UG}$ and a set of undirected edges $\mathcal{E}_{UG}$, consisting of unordered pairs of $\mathcal{V}_{UG}$. A Markov random field defined with respect to this graph, $\mathcal{G}_{UG}$, satisfies three equivalent levels of separation. The first level is pairwise separation, where for

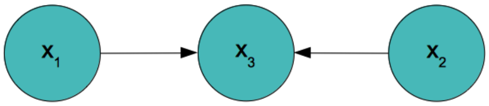

Figure 1.2: A graphical model that is inconsistent with the conditional independence relationship specified by Equation 1.2. This is also known as a v-structure

all pairs of random variables x_i and x_j, $i \neq j$ and $(i, j) \nexists \mathcal{E}_{UG}$:

$$p\left(x_i, x_j | x_{-(i,j)}\right) = p\left(x_i | x_{-(i,j)}\right) p\left(x_j | x_{-(i,j)}\right), \tag{1.3}$$

with $x_{-(i,j)}$ indicating all random variables except x_i and x_j. For a Gaussian graphical model this would correspond to element (i, j) of the inverse covariance matrix being zero [80]. The second level of separation is defined as local separation [80]:

$$p\left(x_i, x_{V\setminus cl(i)} | x_{ne(i)}\right) = p\left(x_i | x_{ne(i)}\right) p\left(x_{V\setminus cl(i)} | x_{ne(i)}\right), \tag{1.4}$$

with $ne(i)$ denoting all vertices in $\mathcal{V}_{UG}$ directly connected to i, and $cl(i) = i \cup ne(i)$ denoting the closure of the neighborhood of i. We illustrate an example of a local separating set for a node in an undirected graph, which is equivalent to the Markov blanket for that node, in Figure 1.3. Finally, the global Markov condition is satisfied if:

$$p(x_A, x_B | x_C) = p(x_A | x_C)\, p(x_B | x_C), \tag{1.5}$$

for disjoint subsets of $\mathcal{V}_{UG} : A, B, C,$ if all paths from A to B are blocked by elements of C [80]. Given the necessary and sufficient conditions from the Hammersley-Clifford theorem [80], a Markov random field can also be factorized based on the cliques (or complete sub-graphs) of the graph.

1.2.2 Directed acyclic graphs

Define a directed acyclic graph as $\mathcal{G}_{DAG} = (\mathcal{V}_{DAG}, \mathcal{E}_{DAG})$, where $\mathcal{V}_{DAG}$ is the set of vertices, and $\mathcal{E}_{DAG}$ is a set of directed edges among vertices, where there are no directed cycles defined by the edges. A probability distribution can be defined with respect to this type of graph if it satisfies the d-separation criterion for all disjoint subsets of variables. The d-separation criterion is stated as follows: for disjoint subsets of $\mathcal{V}_{DAG} : A, B, C,$ A is independent of B conditioned on C if all paths from A to B in $\mathcal{G}_{DAG}$ are blocked by C. In a directed acyclic graph a path from any node in A to any node in B is blocked if there exists a node on the path such that the arrows meet either head to tail or tail to tail at that node and it is in C. Or, a path is blocked if the arrows meet head to head at the node, and neither the node, nor any of its descendants exist in C [100, 9]. This criterion can be used to show that unlike in the undirected graph, where the simple neighborhood of a node can be used to separate it from the rest of the graph (i.e. become conditionally independent), in addition the co-parents of the node must be used, as

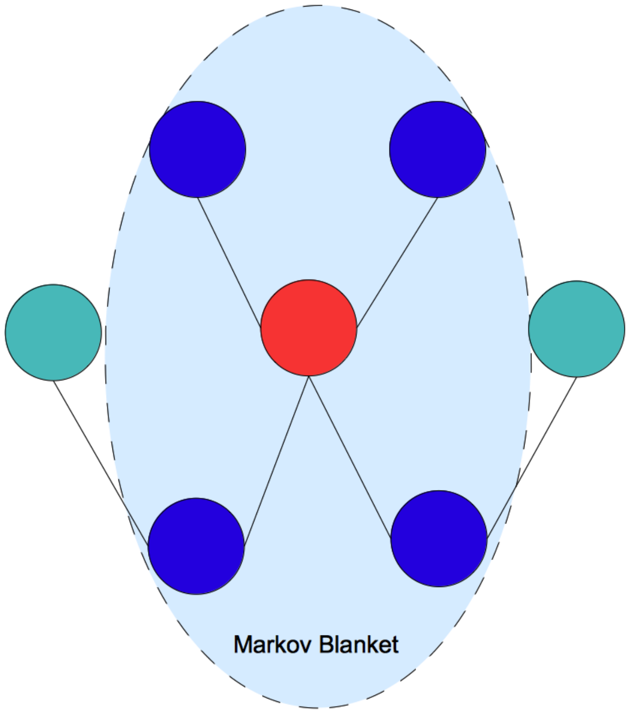

Figure 1.3: Markov blanket of an arbitrary node (depicted in red) in terms of its surrounding neighborhood (depicted in blue) in an undirected graph, where a Markov blanket is a minimal set of nodes that blocks all paths from the red node, to other nodes in the graph (e.g. the green nodes).

shown in Figure 1.4 [9]. This is also an example of a Markov blanket of a node for a directed acyclic graph.

In addition, for directed acyclic graphs, different orientations of edges can produce the same sampling distribution, i.e. there are equivalence classes of graphs in terms of the conditional independence and dependence statements they imply. There exists a graphical criterion to generate all equivalent graphs for a DAG because all equivalent DAGs have the same set of v-structures (i.e. the motif in Figure 1.2). Therefore any edge orientation can be reversed and pro-

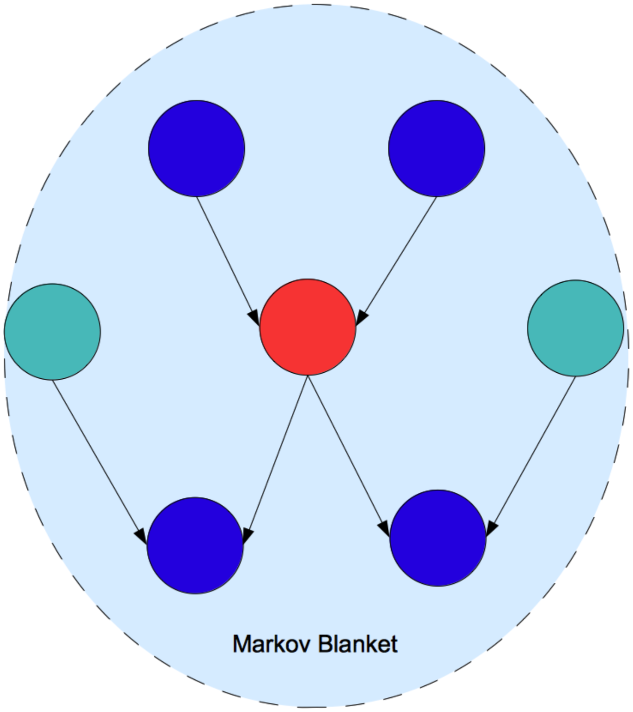

Figure 1.4: Markov blanket of an arbitrary node (depicted in red) in a directed acyclic graph. In this case the co-parents are also members of the Markov blanket, because conditioning on the children of the red node induces dependence between the red node and the green nodes.

duce an equivalent DAG, as long as it does not create or destroy a v-structure [100].

1.2.3 Directed cyclic graphs

For directed cyclic graphs, Spirtes showed that the same d-separation criterion defined for directed acyclic graphs is satisfied by directed cyclic graphs defined in the context of linear structural equations models [119]. In addition to this criterion, we show that the induced dependencies in the Markov blanket of a node for one directed cyclic graph can actually correspond to directed edges in an equivalent cyclic graph, restating and extending a result of Richardson [107] by characterizing equivalence relationships in directed cyclic graphs using signed permutation matrices. This phenomena never occurs for directed acyclic graphs, where induced edges are always absent from the underlying generating graph. We prove this property of equivalence for directed cyclic graphs in Chapter 3 in the build up to the 'Recovery' theorem. In addition, we show an example of this equivalence property, where two different graph structures satisfy the same set of conditional independence and dependence relations in Figure 1.5.

1.2.4 The importance of induced dependence

While in chapter four we focus on undirected graphs, which are fundamentally a less rich class of graphical model than directed graphs, in chapter three we propose a novel approach to learning the structure of a directed cyclic graph. Specifically we prove that the structure learning problem can be mapped from a curved exponential space (i.e. an exponential family model where the log likelihood has a polynomial parameterization) onto a linear exponential space (i.e. an exponential family model where the log likelihood has a linear parameterization) [19], with a sufficient set of additional perturbations in a conditional Gaussian model. We solve the structure inference problem explicitly using a regression approximation to the full likelihood problem with the adaptive lasso feature selection algorithm. While the general problem of learning DAGs is provably NP-hard [24], we propose a computationally efficient solution, when

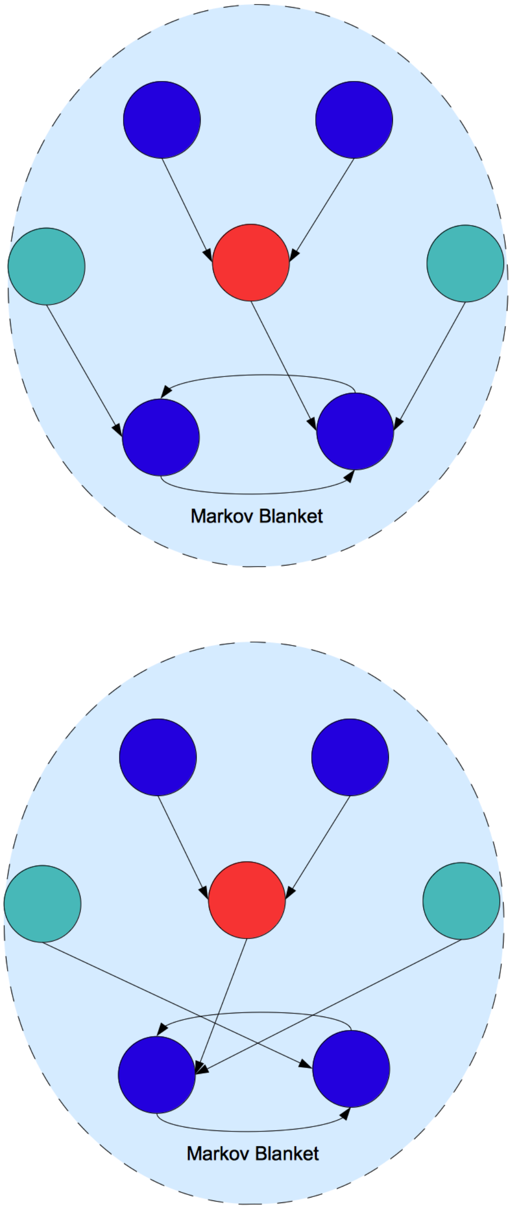

Figure 1.5: Two equivalent directed cyclic graphs, with the Markov blanket of the red node illustrated in both. Note that the skeleton (i.e. topology of the graph without directionality of edges) changes between equivalent models, which can not happen for equivalent directed acyclic graphs.

one has additional perturbation data. In addition, this mapping from a rich parameter space to a larger less rich parameter space is motivated by the phenomena of induced dependence, where variables are marginally independent, but conditionally dependent. We use this asymmetric relationship to restrict the class of possible directed cyclic graphs describing the data, given we know the direction of a subset of edges.

1.3 Approximations

We wish to produce complex multiple feature models, where all the identified relationships between features are highly statistically significant for any given neighborhood of a variable, be it an expression phenotype, a genotype, or downstream phenotype. In the case of genome-wide expression and genotypic variation, the number of variables is very high, i.e. in the tens of thousands to millions of variables [63]. This makes the full likelihood form of the problem for any of the above graphical models ill-conditioned, since the sample size is usually on the order of hundreds to thousands. To address this problem, we decouple the joint neighborhood selection problem into a set of individual, independent neighborhood selection problems by using a regression approximation to the full likelihood. One can imagine this as learning a type of series expansion of the first order effects driving the local behavior of variables in the model. In the context of neighborhood selection this approximation has been proposed previously by Meinshausen and Bühlmann for the penalization problem with the lasso [97], and has the advantage of being well-conditioned, and highly scalable [151].

1.3.1 Variational approximations

In addition, in the context of each of the neighborhood selection problems defined by the individual regressions we propose the other novel feature of this thesis, a variational approximation to the fully Bayesian inference problem for a spike a slab prior in a multiple feature model (i.e. a linear multiple regression). This prior has many known theoretical advantages over other feature selection priors including bounded shrinkage [72] and is likely model selection consistent [147]. We use a mean field, or variational Bayes [136, 9] approximation, where we can define an Expectation Maximization (EM) type algorithm on an approximating distribution for the fully Bayesian inference problem in Chapter 2 (for a slightly richer parameterization (with both positive and negative effect classes). In chapter 4, we simplify the underlying statistical model, by reducing the number of non-zero effect classes, and identify closed form solutions to some of the approximate updates from the original formulation presented in Chapter 2. We also enrich this model by proposing an approximate Bayesian model averaging step, as well as a rule for undirected network inference. Our algorithm is highly scalable, and will be a practical tool for practitioners interested in identifying richer sets of simultaneous strongly supported gene interactions from genome-wide gene expression, genotype, and downstream phenotype variation.

CHAPTER 2

A VARIATIONAL BAYES ALGORITHM FOR FAST AND ACCURATE MULTIPLE LOCUS GENOME-WIDE ASSOCIATION ANALYSIS

2.1 Abstract

Background: The success achieved by genome-wide association (GWA) studies in the identification of candidate loci for complex diseases has been accompanied by an inability to explain the bulk of heritability. Here, we describe the algorithm V-Bay, a variational Bayes algorithm for multiple locus GWA analysis, which is designed to identify weaker associations that may contribute to this missing heritability.[1]

Results: V-Bay provides a novel solution to the computational scaling constraints of most multiple locus methods and can complete a simultaneous analysis of a million genetic markers in a few hours, when using a desktop. Using a range of simulated genetic and GWA experimental scenarios, we demonstrate that V-Bay is highly accurate, and reliably identifies associations that are too weak to be discovered by single-marker testing approaches. V-Bay can also outperform a multiple locus analysis method based on the lasso, which has similar scaling properties for large numbers of genetic markers. For demonstration purposes, we also use V-Bay to confirm associations with gene expression in cell lines derived from the Phase II individuals of HapMap.

[1] This chapter was published as a methodology article in BMC bioinformatics on January 27, 2010 [87].

Conclusions: V-Bay is a versatile, fast, and accurate multiple locus GWA analysis tool for the practitioner interested in identifying weaker associations without high false positive rates.

2.2 Background

Genome-wide association (GWA) studies have identified genetic loci associated with complex diseases and other aspects of human physiology [37, 62]. All replicable associations identified to date have been discovered using GWA analysis techniques that analyze one genetic marker at a time [96]. While successful, it is well appreciated that single-marker analysis strategies may not be the most powerful approaches for GWA analysis [65]. Multiple locus inference is an alternative to single-marker GWA analysis that can have greater power to identify weaker associations, which can arise due to small allelic effects, low minor allele frequencies (MAF), and weak correlations with genotyped markers [65]. By correctly accounting for the effects of multiple loci, such approaches can reduce the estimate of the error variance, which in turn increases the power to detect weaker associations for a fixed sample size. Since loci with weaker associations may contribute to a portion of the so-called 'missing' or 'dark' heritability [69, 27, 92], multiple locus analyses have the potential to provide a more complete picture of heritable variation.

Methods for multiple locus GWA analysis must address a number of problems, including 'over-fitting' where too many associations are included in the genetic model, as well as difficulties associated with model inference when the number of genetic markers is far larger than the sample size [156]. Two general approaches have been suggested to address these challenges: hierarchical models and partitioning/classification. Hierarchical modeling approaches [148, 149, 150, 86, 155, 142] employ an underlying regression framework to model multiple marker-phenotype associations and use the hierarchical model structure to implement penalized likelihood [149], shrinkage estimation [144], or related approaches to control over-fitting. These methods have appealing statistical properties for GWA analysis when both the sample size and the number of true associations expected are far less than the number of markers analyzed, which is generally considered a reasonable assumption in GWA studies [156]. Alternatively, partitioning methods do not (necessarily) assume a specific form of the marker-phenotype relationships but rather assume that markers fall into non-overlapping classes, which specify phenotype association or no phenotype association [157, 155]. Control of model over-fitting in high dimensional GWA marker space can then be achieved by appropriate priors on marker representation in these classes [155].

Despite the appealing theoretical properties of multiple locus methods that make use of hierarchical models or partitioning, these methods have not seen wide acceptance for GWA analysis. There are at least two reasons for this. First, an ideal multiple locus analysis involves simultaneous assessment of all markers in a study and, given the scale of typical GWA experiments, most techniques are not computationally practical options [28, 40, 148, 149, 157]. Second, there are concerns about the accuracy and performance of multiple locus GWA analysis. This is largely an empirical question that needs to be addressed with simulations and analysis of real data.

Here we introduce the algorithm V-Bay, a (V)ariational method for (Bay)esian hierarchical regression, that can address some of the computational limitations shared by many multiple locus methods [28, 40, 148, 149, 157]. The variational Bayes algorithm of V-Bay is part of a broad class of approximate inference methods, which have been successfully applied to develop scalable algorithms for complex statistical problems, in the fields of machine learning and computational statistics [59, 61, 70, 10]. The specific type of variational method implemented in V-Bay is a mean-field approximation, where a high dimensional joint distribution of many variables (in this case genetic marker effects) is approximated by a product of many lower dimensional distributions [8]. This method is extremely versatile and can be easily extended to a range of models proposed for multiple locus analysis [150, 142, 93, 65].

The specific model implemented in V-Bay is a hierarchical linear model, which includes marker class partitioning control of model over-fitting. This is particularly well suited for maintaining a low false-positive rate when identifying weaker associations [155]. V-Bay implements a simultaneous analysis of all markers in a GWA study and, since the computational time complexity per iteration of V-Bay is linear with respect to sample size and marker number, the algorithm has fast convergence. For example, simultaneous analysis of a million markers, genotyped in more than a thousand individuals, can be completed using a standard desktop (with large memory capacity) in a matter of hours.

We take advantage of the computational speed of V-Bay to perform a simulation study of performance, for GWA data ranging from a hundred thousand to more than a million markers. In the Results we focus on the simulation results for single population simulations, but we also implement a version of the algorithm to accommodate known population structure and missing genotype data. We demonstrate that in practice, V-Bay consistently and reliably identifies both strong marker associations, as well as those too weak to be identified by single-marker analysis. We also demonstrate that V-Bay can outperform a recently proposed multiple locus methods that uses the least absolute shrinkage and selection operator (lasso) penalty [142], a theoretically well founded and widely accepted method for high dimensional model selection. V-Bay therefore provides a powerful complement to single-marker analysis for discovering weaker associations that may be responsible for a portion of missing heritability.

2.3 Results and Discussion

2.3.1 The V-Bay Algorithm

The V-Bay algorithm consists of two components: a hierarchical regression model with marker class partitioning and a variational algorithm for approximate Bayesian inference. The underlying hierarchical model of V-Bay is a Bayesian mixture prior regression [55] that has been previously applied to association and mapping problems [155]. The regression portion of this hierarchical model is a standard regression used to model genetic marker-phenotype associations, and allows for natural incorporation of population structure and other covariates. The model partitioning incorporates global features of genetic marker associations, which are assumed to be distributed among positive, negative, and zero effect classes. The zero effect class is used to provide a parametric representation of the assumption that most markers in GWA studies will not be linked to causative alleles and therefore do not have true associations with phenotype [155].

Approximate Bayesian inference with V-Bay is accomplished by an algorithm adapted from variational Bayes methods [135]. As with other variational Bayes methods, the goal of V-Bay is to approximate the joint posterior density of the hierarchical regression model with a factorized form and then to minimize the Kullback-Liebler (KL) divergence between the factorized form and the full posterior distribution [6]. This is accomplished by taking the expectation of the log joint posterior density, with respect to each parameter's density from the factorized form, and iterating until convergence [8]. The overall performance of V-Bay will depend on how well the factorized form approximates an informative mode of the posterior distribution of the hierarchical model. We have chosen a factorization with respect to each regression and hierarchical parameter, which appears to perform extremely well for identifying weak associations when analyzing simulated GWA data that include large numbers of genetic markers.

2.3.2 Computational speed

The computational efficiency of V-Bay derives from two properties: it is a deterministic algorithm and the objective function has a factorized form. Since V-Bay is deterministic it does not need the long runs of Markov chains required by exact Bayesian MCMC algorithms [54]. For GWA analysis, these latter stochastic algorithms can be very slow to converge, particularly when marker numbers are large and when there are complex marker correlations produced by linkage disequilibrium [156]. The factorized form of V-Bay means that the minimization is performed with respect to each parameter independently, where each iterative update satisfies consistency conditions for maximizing the lower bound, given the state of the other parameters. Unlike univariate update algorithms, which may not necessarily have efficient updates with respect to the likelihood gradient function [65], the consistency conditions produced by the factorized form ensure that the univariate updates produce a computationally efficient approach to a KL-divergence minimum.

More precisely, V-Bay has linear time complexity scaling with respect to both marker number and sample size per iteration (Appendix A.2, Methods). V-Bay therefore has better computational scaling properties than most currently proposed multiple locus algorithms for full likelihood or exact MCMC Bayesian analysis, when simultaneously considering all markers in a GWA study [28, 40, 148, 149, 157]. While the total time to convergence will depend on the true underlying genetic model, total computational times appear to be very tractable. As an example, using a dual-quad core Xeon 2.8Ghz, with 16 Gb of memory, V-Bay converges in less than four hours for data sets in the range of 1 million markers, for a sample size of 200, and has average convergence around ten hours for sample sizes of 1000.

2.3.3 Significance thresholds

We assessed significance of marker associations using $-\log_{10}$ p-vbay, the negative log posterior probability of a marker being in either the positive or negative effect class. This is a natural statistic for deciding significance, since p-vbay is the (approximate posterior) probability that the marker has an association with the phenotype. While different significance thresholds based on $-\log_{10}$ p-vbay can be assigned to control false positive rate, as illustrated in Figure 2.1, the distribution of this statistic has an appealing property. The statistic has a value of zero for most of the true hits and there is a large gap (about 1-2 orders of magnitude) between significant markers and those with less significant scores. This is true even when the individual heritabilities of the true hits are low. This property of V-Bay is remarkably robust. A GWA practitioner using V-Bay can therefore easily identify a significant association (a 'hit') in practice when applying a conservative significance threshold.

2.3.4 Performance of V-Bay compared to single-marker analysis

We empirically analyzed V-Bay performance on 150 simulated GWA data sets. Marker numbers for these data were one-hundred thousand, six-hundred thousand, or one million markers and were simulated using the approximate coalescent simulator MaCS [22]. We simulated a continuous phenotype with normally distributed error under the conditions listed in Table 2.1, where each GWA data set analyzed was produced by choosing a combination of these conditions. For these simulated data sets, we analyzed the performance of V-Bay compared to a single-marker analysis that was implemented by applying a linear regression model individually to each marker.

As illustrated in Table 2.2, V-Bay can perform better than single-marker anal-

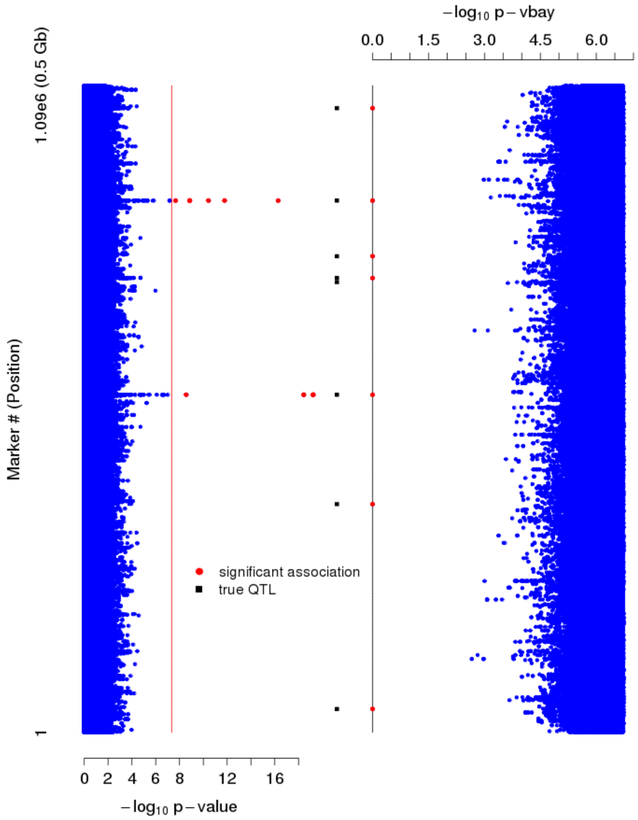

Figure 2.1: Manhattan plots of the results of a single-marker (left) and V-Bay analysis (right) of a simulated GWA data set. Data were simulated with a sample size of 200, one million markers, 8 loci with phenotype associations, and a total phenotype heritability of 0.9. The locations of the loci with phenotype associations are represented by the black squares. Each dot reflects the $-\log_{10}$ p-value resulting from single-marker analysis (left) and the $-\log_{10}$ p-vbay output of V-Bay (right), where non-significant associations are represented as blue dots. The markers above the red line for the single-marker analysis are significant when using a Bonferroni correction. The markers in red for the V-Bay analysis (connected by a black line) are significant using a conservative control of the false positive rate equal to a Bonferroni correction. In this case, the single-marker analysis correctly identifies two of the true associations, while V-Bay identifies 7 of the 8 true associations. This result was typical for our simulation analyses.

Table 2.1: Components and range of values used to simulate GWA data.

Component	Values
sample	200 or 1000
markers	0.1 to 1.0 million
missing	0% or 2%
loci	4, 8, or 32
effects	gamma(2,1) or fixed
heritability	0.5 or 0.9
populations	one or four

ysis given a sufficient sample size or a sufficient number of loci with high individual heritabilities. Both the number of true associations identified and the amount of heritable variation explained can be greater when employing highly conservative false positive tolerances. For example, when using a false positive rate approaching a Bonferroni correction, V-Bay can on average double the number of associations found by single-marker analysis and can explain 20% more of the variance in phenotype under the most favorable conditions simulated. The reason for this increase in performance is that V-Bay has greater power to detect weaker (true) associations by accounting for the effects of multiple loci.

Whether small associations are identified by V-Bay depends on the interplay between the sample size of the GWA study and the percentage of variation explained by the individual marker associations. For example, Figure 2.2a and 2.2b present the Receiver Operator Characteristic (ROC) curves comparing the

Table 2.2: Comparison of V-Bay and single-marker GWA analysis of simulated data for 1 million markers. Phenotypes were simulated with a fixed total heritability of 0.9. The false positive rate was controlled to be $< 10^{-7}$ for both the V-Bay analysis and the single-marker analysis. ($\overline{TP}$: average true positive rate). [a]Average, maximum, and minimum individual heritabilities of the individual loci. [b]The smallest individual heritability identified among the true positives. [c]The average total heritability accounted for by the true positives identified.

				V-Bay			single-marker	
sample	loci	$\overline{h_m^2}$ (min/max)[a]	$\overline{TP}$	min(h_m^2)[b]	%h^2[c]	$\overline{TP}$	min(h_m^2)[b]	%h^2[c]
200	4	0.24 (0.0032/0.75)	0.83	0.026	98.9	0.55	0.16	87.4
200	32	0.028 (6.7e-5/0.28)	0.053	0.033	26.9	0.072	0.050	35.3
1000	4	0.23 (0.0050/0.65)	1.00	0.0050	100	0.78	0.045	98.7
1000	32	0.028 (8.3e-5/0.30)	0.61	0.0037	95.6	0.32	0.0099	78.2

performance of V-Bay and single-marker analyses for 10 replicate simulations, with 4 or 32 loci affecting a phenotype, total heritability of 0.9, and sample sizes of 200 or 1000, respectively (note that we use these high heritability cases for exploratory purposes; we also consider a total heritability of 0.5 in other simulations). With a sample size of 200 (Figure 2.2a), V-Bay outperforms single-marker analysis for the 4 loci simulations, and is about the same for the 32 loci simulations. The reason for the relative decrease in performance of V-Bay in this latter case is the average individual heritability associated with each associated marker is lower. Most of the true associations are therefore too small to detect even when controlling for the largest effects with a multiple locus method like V-Bay (Figure 2.2c). With a larger sample size however, V-Bay is able to detect a much larger proportion of the weaker associations in the case of 32 contributing loci (Figure 2.2d). Also, since there are more loci to detect with 32 loci, V-Bay has far better performance than single-marker analysis overall at a highly conservative false positive rate ($< 10^{-7}$). Further simulations indicated that even for a uniform distribution of individual heritabilities (i.e. constant minor allele frequency and effect size), V-Bay performs better for similar sample sizes and individual heritabilities. For example, for 32 loci with a sample size of 1000, and false-discovery rate of 5.0% the average power of V-Bay was 93%. This is greater than the corresponding power of 72% for single-marker analysis with the same false-discovery rate. In general, regardless of sample size, if there are enough loci with associations that are not too weak, then V-Bay outperforms single-marker analysis.

V-Bay performance is a direct function of the individual heritabilities, and not the total heritability of the phenotype. The individual heritability is defined by both the minor allele frequency and the effect size (see Methods). Therefore loci with large effects may still have low individual heritabilities if the minor allele frequencies of the true loci are low (or vice versa). For example, for our simulations where the total heritability was controlled to be 0.5, and the individual heritabilities were shifted to be smaller overall, V-Bay performance was far closer to single-marker analysis. When we increased the individual heritabilities associated with associations in these simulations, while holding the total heritability at 0.5, V-Bay can outperform single-marker analysis. For all simulations, when an individual heritability falls below a certain threshold, neither approach could detect the association. There exists a limit to how weak an asso-

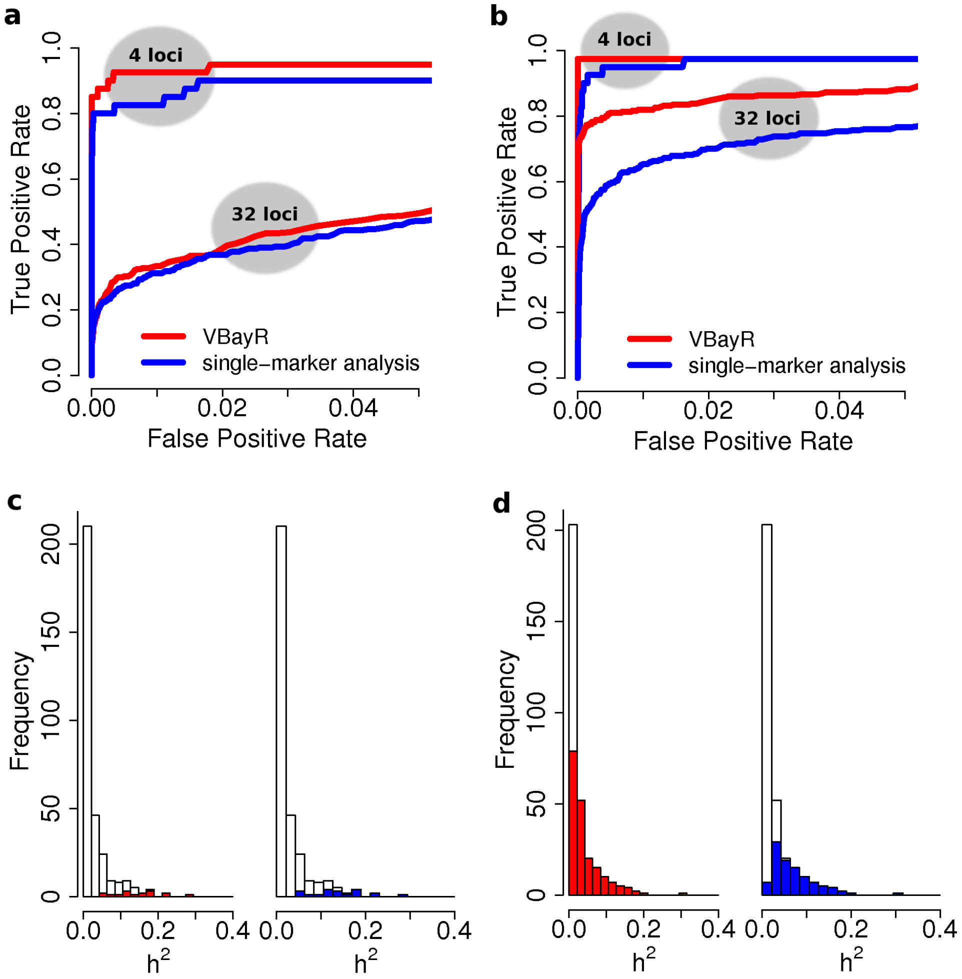

Figure 2.2: Comparison of V-Bay and single-marker analysis for simulated GWA data. The total heritability for the phenotype in each data set was controlled to be 0.9. The Receiver Operator Characteristic (ROC) curves in the upper graphs reflect the average across 10 replicate data sets that included (**a**) 200 samples and (**b**) 1000 samples. The lower graphs plot the distribution of individual heritabilities for the 32 loci simulations for the data sets that included (**c**) 200 samples and (**d**) 1000 samples, where the proportion of correctly identified loci for V-Bay are plotted in red and for single-marker analysis in blue when controlling the false positive rate at $< 10^{-7}$.

ciation can be and still be detected by V-Bay, given the sample size of the GWA study. Even in the worst case scenarios simulated, with many loci with small individual heritabilities and a small sample size, the performance of V-Bay was not significantly different from single-marker analysis across simulations. This result suggests that even if the number of loci were increased (i.e. the average individual heritability was decreased), the performance of V-Bay would at worst be the same as single-marker analysis.

The inset in Figure 2.3 illustrates another appealing property of V-Bay. In contrast to a single-marker analysis, where each marker in a linkage disequilibrium block containing a true association will have an inflated $-\log_{10}$ p-value, V-Bay identifies only a single marker as significant, which is in high linkage disequilibrium with the true association. We found in our single population simulations that, while the specific marker assigned depends on the update order of the algorithm, the correlation between the marker and the causative allele averages $r^2 = 0.75$, with 28% of hits on markers in perfect linkage disequilibrium, and 52% of markers with $r^2 \geq 0.9$. V-Bay can therefore provide high mapping resolution within a linkage disequilibrium block.

2.3.5 Comparison to the Lasso

The V-Bay algorithm was compared to the lasso, one of the only other currently proposed multiple locus methods that make use of a hierarchical regression model and have similar scaling properties to V-Bay [142]. For comparison to V-Bay, we use a form that implements a lasso type penalty [129], based on the algorithm presented in Wu et al. [142], modified to allow continuous pheno-

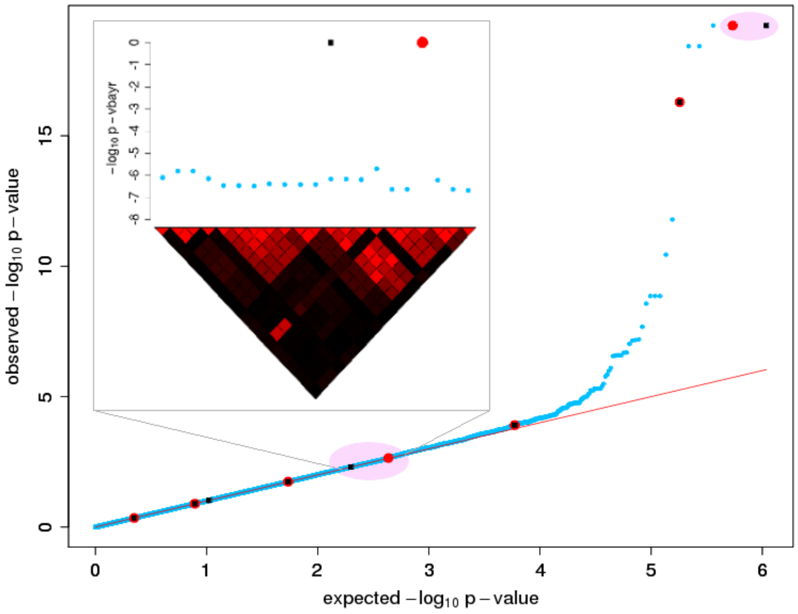

Figure 2.3: Quantile-Quantile plot of the genome-wide p-values obtained in the single marker analysis of the data presented in Figure 2.1. The seven associations correctly identified by V-Bay are circled in red. The locations of the loci with phenotype associations (black squares) and the results of the V-Bay analysis (red circles) are depicted with respect to their observed and expected quantiles from the single-marker analysis (blue circles). In this analysis, V-Bay is able to detect true associations that are undetectable with the single-marker analysis. The inset plot shows one of the hits from V-Bay that does not lie exactly on the marker in tightest linkage disequilibrium with the associated locus but is six SNPs away.

types.

Figure 2.4 presents the power of V-Bay, the lasso, and single-marker analysis for simulations with one-hundred thousand markers, 32 loci, and 1000 samples, when the false-discovery rate is controlled to 0%. V-Bay, the lasso, and single-marker analysis can all correctly detect a high proportion of loci in the upper tail of the distribution, where the individual heritabilities of associations are high. However, there is variability in the number of smaller heritability loci detected, with multiple locus methods performing better. The reason for this result is when multiple locus methods correctly identify loci with larger individual heritabilities, they directly account for the effect of these loci in the statistical model. This shrinks the estimate of the error term, which increases the power to detect loci with even weaker associations. For these simulations, V-Bay outperforms not only single-marker analysis, but also the lasso. We found V-Bay performed better than the lasso (and single-marker analysis) for additional architectures and sample sizes, when controlling the false discovery rate to 5.0% (Table 2.3).

2.3.6 Genome-wide association analysis of HapMap gene expression

To investigate the empirical properties of V-Bay, we performed a GWA analysis on gene expression levels measured in eternal lymphoblastoid cell lines, generated from the 210 unrelated individuals of Phase II of the International HapMap project [123]. Individuals in this sample were genotyped for upwards of 3.1 million SNPs and were derived from four populations: Caucasian with European

Table 2.3: Power comparison for V-Bay, the lasso, and single-marker GWA analysis from simulated data with 100,000 markers. Phenotypes were simulated with a fixed total heritability of 0.9. The false discovery rate was controlled to 5% for all three analyses.

sample	loci	V-Bay	the lasso	single-marker
200	4	90.0%	87.5%	47.5%
200	32	14.1%	4.69%	7.19%
1000	4	97.5%	77.5%	60.0%
1000	32	80.6%	65.0%	33.1%

origin (CEU), Chinese from Beijing (CHB), unrelated Japanese from Tokyo (JPT), and Yoruba individuals from Ibadan, Nigeria (YRI) [68]. In the original GWA analysis of these data, Stranger et al. used a single-marker testing approach, considering each population independently, and limiting the analysis to SNPs in the *cis*-regions of each gene to control the level of multiple test correction [123].

Using a version of V-Bay that accounts for population structure and missing genotype data, we analyzed the pooled data from these populations. We did not limit the analysis to *cis*-regions, although we did limit our analyses to SNPs with MAF > .10, leaving 1.03 million markers genome-wide. To minimize computational cost, we also limited our analysis to the 100 expression probes Stranger et al. found to have the most significant associations, and an additional 20 probes with the largest residual variance, after correcting for population structure. For

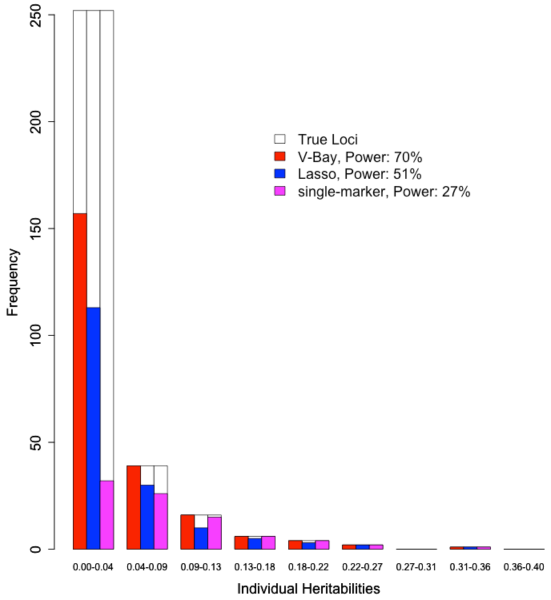

Figure 2.4: Histograms of loci identified by V-Bay, the lasso, and single-marker analysis as a function of individual heritability. The false-discovery rate is controlled to 0.0%. These graphs summarize the results of ten replicate simulated data-sets with 100,000 markers, 32 loci with associations, a sample size of 1000, and a total phenotype heritability of 0.9. The power for each method at 0.0% false-discovery rate is shown in the legend.

comparison, we also applied a single-marker analysis to these pooled data, for the 120 expression probes, incorporating a covariate to account for population structure.

On average, V-Bay was able to complete the GWA of each of these expression phenotypes in 1.5 hours using a dual-quad core Xeon 2.8Ghz (16 Gb of memory). In 90% of cases, where our single-marker analysis reproduced the most significant *cis*-associations reported by Stranger et al., V-Bay also identified the association. In addition, a total of 72 out of the 100 previously reported *cis*-

associations [123] were identified with V-Bay (Appendix A, Table A.1,A.2). A typical result from these analyses is presented in Figure 2.5. These Manhattan plots are for the HLA-DRB1 expression probe, which was not reported by Stranger et al. as having a strong *cis*-association. For this probe, V-Bay, the lasso, and our multiple population single-marker analysis indicated a strong *cis*-association. Since this association was also found with single-marker analysis, identification was not due to V-Bay but to the analysis of the pooled data from different populations (as opposed to testing within populations as in Stranger et al. [123]). Still, the increased sensitivity of V-Bay was suggested in this case by *trans*-associations identified by individual runs of V-Bay, which were not identified by the single-marker analysis or the lasso. However, we imposed the restrictive criteria that an association identified by V-Bay would only be considered significant if it was robust to missing data resampling and marker reordering runs. Using this conservative strategy, none of the putative *trans*-associations were robust enough to report. With an increased sample size, we believe that these *trans*-associations could be confidently assigned as true hits.

2.4 Conclusions

V-Bay addresses computational efficiency and performance concerns associated with many multiple locus GWA algorithms. While V-Bay currently utilizes a hierarchical partitioning model, the same approach could be used to implement scalable algorithms for a wide range of models. For example, different shrinkage or penalization models such as the lasso [150, 142], ridge regression [93], or a normal exponential gamma distribution penalty [65] are easily implemented by removing the partitioning and substituting the appropriate prior distribution.

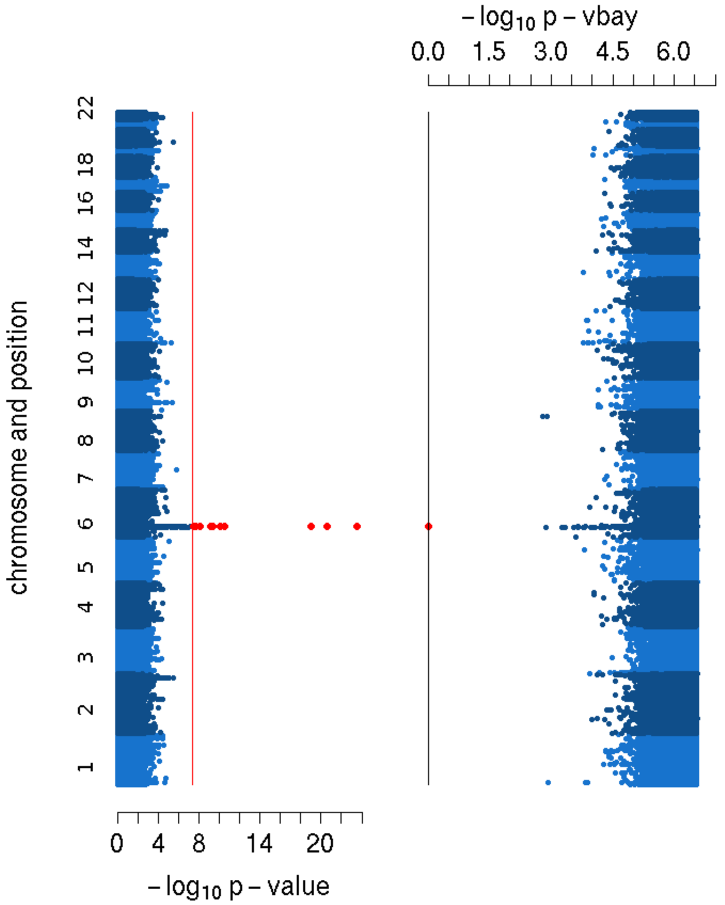

Figure 2.5: Manhattan plots of the results of a single-marker (left) and V-Bay GWA analysis (right) of the gene expression product HLA-DRB1 for individuals in HapMap. Each dot reflects the $-\log_{10}$ p-value resulting from the single-marker analysis (left) and the $-\log_{10}$ p-vbay output of V-Bay (right), where non-significant associations are represented as blue dots (alternating shades are used to distinguish chromosomes). The markers above the red line for the single-marker analysis are significant when using a Bonferroni correction. The marker in red for the V-Bay analysis (in the black line) is significant at an equivalently conservative false positive control. Note that the lasso was also able to identify this association. We did not incorporate the SNPs on the X and Y chromosomes in our analyses.

Further, the variational Bayes method used for computation does not require specific closed form integrals arising from hyperparameter distributions, which characterize many of the proposed algorithms for full penalized-likelihood or Bayesian GWA analysis [150, 93, 65]. There is therefore the potential for developing an entire class of scalable multiple locus algorithms for GWA analysis that could be tuned for different genetic and experimental conditions within the V-Bay framework.

2.5 Methods

2.5.1 V-Bay Algorithm

The V-Bay algorithm consists of two components, a hierarchical regression model with marker class partitioning and a variational Bayes computational algorithm. The hierarchical regression is adapted directly from Zhang et al. [155] with minor alterations. The first level of the hierarchical regression model for a sample of n individuals with m markers is a standard multiple regression model:

$$y_i = \mu + \sum_{j=1}^{m} x_{ij}\beta_j + e_i, \tag{2.1}$$

where y_i is the phenotype of the i^{th} individual, μ is the sample mean, x_{ij} is the genotype of the j^{th} marker of the i^{th} individual, β_j is the effect of the j^{th} marker, and $e_i \sim \mathrm{N}\left(0, \sigma_e^2\right)$. While we limit the current presentation of the model to continuous traits with normal error, more complex error structures and extensions to discrete traits is straightforward. Because equation (1) is a linear model, it can be easily expanded to test for dominance or epistasis using a standard mapping approach. In addition, confounding factors such as population structure can be accounted for by the addition of covariates. The effects of these additional covariates can be modeled within the hierarchical regression framework or can be treated simply as nuisance parameters and given uninformative priors. We used an uninformative prior $\left(\frac{1}{\sigma_e^2}\right)$ for the error parameter, σ_e^2, and a constant (improper) prior for the mean parameter μ.

The second level of the hierarchical model consists of a partitioning of markers into positive, negative, and zero effect classes and the prior control over the distributions of these classes. The partitioning is accomplished by modeling each of the regression coefficients using mixture prior distributions:

$$\beta_j \sim \begin{array}{c} (1 - p_{\beta_+} - p_{\beta_-})\mathrm{I}_{\{\beta_j=0\}} + p_{\beta_+}\mathrm{N}_+(0, \sigma^2_{\beta_+}) \\ +p_{\beta_-}\mathrm{N}_-(0, \sigma^2_{\beta_-}) \end{array}, \tag{2.2}$$

where $\mathrm{I}_{\{\beta_j=0\}}$ is an indicator function for β_j with a value of zero, and N_+ and N_- are positive and negative truncated distributions [155]. The priors on the population distribution of positive and negative effect probability hyperparameters (p_{β_+} and p_{β_-}) are:

$$\left(p_{\beta_+}, p_{\beta_-}, 1 - p_{\beta_+} - p_{\beta_-}\right) \sim \text{Dirichlet}\left(\theta_\beta, \phi_\beta, \psi_\beta\right). \tag{2.3}$$

In our analyses we chose an uninformative Dirichlet prior by setting the parameters $\theta_\beta, \phi_\beta, \psi_\beta$ all to one. The hyperparameters p_{β_+} and p_{β_-} reflect the partitioning aspect of the model. Within the positive and negative partitions, the population variance parameters ($\sigma^2_{\beta_+}$ and $\sigma^2_{\beta_-}$) have χ_1^{-2} priors. This choice of prior for the regression coefficients in the positive and negative effect classes increases the robustness to outliers. Assuming the number of markers in the GWA data set, m, is greater than the sample size, n, we truncate the Dirichlet distribution such that $p_{\beta_-} + p_{\beta_+} \leq \sqrt{n}/m$, where the truncation puts a lower bound on the harshness of shrinkage [156]. We found this truncation very important when considering data sets with large numbers of markers. Without truncation, the evidence in the data is too weak to enforce harsh enough shrinkage for desirable model selection.

The variational Bayes component of V-Bay is constructed by approximating the joint posterior density of the hierarchical model:

$$p(\beta_1,\beta_2,\ldots,\beta_m,p_{\beta_+},p_{\beta_-},\sigma^2_{\beta_+},\sigma^2_{\beta_-},\sigma^2_e,\mu|\mathbf{y},\mathbf{x}) \tag{2.4}$$

in terms of a factorized form:

$$q(\beta_1)\cdots q(\beta_m)q(p_{\beta_+},p_{\beta_-})q(\sigma^2_{\beta_+})q(\sigma^2_{\beta_-})q(\sigma^2_e)q(\mu) \tag{2.5}$$

and then minimizing the KL-divergence between the factorized and full form. Equation 2.5 is a natural factorization for the V-Bay hierarchical model since most of the priors are conjugate. The posterior factorized distributions all have closed form expressions and each parameter is completely characterized by an expected sufficient statistic [6] (Appendix A.2, Methods). The algorithm is therefore equivalent to updating these expected sufficient statistics.

Minimizing the KL-divergence between each marginal distribution (e.g. $q(\beta_j)$) and the full joint distribution is performed by considering the expectation of the full log joint distribution with respect to each parameter. For a generic parameter θ, the expectation step is equivalent to setting:

$$\log\{q(\theta)\} \propto \begin{array}{l} \mathrm{E}_{-\theta}\Big[\log\Big\{p(\beta_1,\beta_2,\ldots,\beta_m,p_{\beta_+}, \\ \quad p_{\beta_-},\sigma^2_{\beta_+},\sigma^2_{\beta_-},\sigma^2_e,\mu|\mathbf{y},\mathbf{x})\Big\}\Big] + \mathrm{C} \end{array} \tag{2.6}$$

with C some normalizing constant, and $\mathrm{E}_{-\theta}$ indicating expectation of the log of equation 2.4 with respect to every other parameter's factorized distribution, except $q(\theta)$. This defines a system of equations which can be iterated through until convergence [8, 6]. With the factorized form, it is a simple matter to demonstrate the time complexity of V-Bay is $O(nm)$ per iteration (Appendix A.2, Methods).

2.5.2 V-Bay Convergence

The factorization of equation 2.4 is used to define a function $\mathcal{L}(\theta)$ which lower bounds the log posterior probability of the data (i.e. the probability of the observed data after integrating out all parameters in the model). The lower bound $\mathcal{L}(\theta)$ is defined as the expectation of the log of equation 2.4 with respect to every factorized distribution plus the entropy of each factorized distribution. In the full form, the convergence of V-Bay to a local maximum of the lower bound $\mathcal{L}(\theta)$ is guaranteed because of the convexity of $\mathcal{L}(\theta)$ with respect to each parameter's approximate posterior distribution [13]. In the described implementation we used an approximation for some higher order expectation terms that we found increased computational efficiency (Appendix A.2, Methods).

Given that global convergence to a single stationary point is not guaranteed [135], the standard practice is to use multiple parameter initializations. We found that with random initializations of expectations of β_j, V-Bay finds local modes that correspond to over-fit (under-determined) models, while with initializations of only a few non-zero expectations of β_j's, V-Bay tends to update these values close to zero before converging. We therefore use the approach of setting all expectations of β_j parameters equal to zero as a starting point for all runs of V-Bay, an approach that has precedent in simultaneous marker analysis [65]. This also corresponds to appropriate starting estimates given our prior assumption that not too many markers are associated with a phenotype.

We have found that the order in which the parameters are updated can affect local convergence, particularly when there is missing genetic data. In general, the different association models we found using different orderings were not widely different from one another, often differing in whether they included one or two specific associations. For cases where we found ordering did make a difference, we ran V-Bay with multiple random orderings and used the conservative criteria of considering only associations found to be significant in at least 80% of the cases to be true positives for all simulations and data analyses compared to single-marker analysis. The cutoff of 80% corresponds directly to a false discovery rate of 0%. We also considered a less stringent cutoff and an observed false discovery rate of 5% in the comparison to the lasso.

2.5.3 V-Bay Software

An implementation of V-Bay is available at http://mezeylab.cb.bscb.cornell.edu/Software.aspx. The software has basic control parameters available to the user and only requires tab delimited genotype and phenotype files as input. The algorithm itself consists of the following steps: 1) randomize marker ordering, 2) initialize the expected sufficient statistics and expectations of parameters, 3) update the expected sufficient statistics for a particular parameter, given the expectations of all the other parameters, 4) update the expectations of a particular parameter given the expectations of all the other parameters, 5) repeat steps 3 and 4 for all the parameters in the model, 6) check convergence based on the current estimate of the lower bound, $\mathcal{L}(\theta)$. Further functional details are presented in Appendix A, Tables A.3-A.9. The main output from the algorithm is the $-\log_{10}$ of p-vbay= $p_{j+} + p_{j-}$ statistic for each marker, which can be used to assess significance of a marker association.

2.5.4 The Lasso

Originally proposed by Tibshirani [130], recently applied to GWA data by Wu et al. [142] and modified by Hoggart et al. [65], the lasso is a form of hierarchical regression that imposes a double exponential prior on the coefficients of each marker. Although expressed in a Bayesian context, maximum *a posteriori* (MAP) estimates are obtained by maximizing the following penalized log-likelihood:

$$\begin{aligned} \ell(\beta|Y,\lambda) &= \ell(\beta|Y) + \log p(\beta|\lambda) \\ &= \ell(\beta|Y) - \lambda \sum_{j=1}^{m} |\beta_j| \end{aligned} \tag{2.7}$$

where $\ell(\beta|Y)$ is the log-likelihood for the relevant generalized linear model. By penalizing the magnitude of each β_j coefficient, MAP estimates shrink the coefficient values compared to the estimates under the unpenalized model. This shrinkage causes most coefficients to be exactly zero, so that only very few markers are selected to be nonzero for a single value of λ. This penalty produces a convex log-likelihood surface with a single maximum even for underdetermined systems (i.e. when there are more markers than samples). Therefore, the lasso can jointly consider all markers in a single model and simultaneously account for variance in the response caused by multiple markers. The lasso model is fit for multiple values of λ and a single subset of coefficients is selected to be nonzero by 10-fold cross-validation. Confidence scores are obtained for each selected marker by comparing an unpenalized model with all selected markers to a model that omits each marker in turn. An F-test is performed for each marker, but note that these confidence scores cannot be interpreted as typical p-values since they are obtained from a two step procedure. Algorithmic details for fitting the LASSO model for the linear-Gaussian case are provided by [143, 47].

2.5.5 Simulation Study

GWA data were simulated under the set of conditions listed in Table 2.1. The genomic marker data were generated using MaCS [22], a scalable approximate coalescent simulator, using the default approximation tree width. For the comparison to single-marker analysis, three basic types of genotype data sets were simulated. For the first and second type, 0.5 Gb of DNA was simulated from a single diploid population with $N_e = 10000$, the population scaled mutation rate $4N_e\mu = \theta = 0.001$, and the genome-wide population scaled recombination rate $4N_e\kappa = \rho = .00045$, values taken from Voight et al. [133]. Samples of 200 and 1000 were sampled screening the minor allele frequency (MAF) to be 0.10, leaving more than one-million markers for analysis. For the third type, 200 diploid samples of 0.5 Gb were simulated from a simple four population migration model. The approximation $F_{st} = \frac{1}{(4N_eM+1)} = 0.12$, as observed in the overall Phase I HapMap analysis [2], was used to determine the population per generation migration rate for a simple symmetric island migration model, with populations of equal size. After screening MAF to be > 0.10, this left over 660 thousand markers for analysis. The final data included the addition of 2% missing data.

Given the simulated genotypic data, phenotypic data were produced with a simple additive linear model as shown in equation 2.1. The genotypes were represented in the linear model with a consistent dummy variable encoding of $\{0, 1, 2\}$ across loci. The additive effects were drawn independently from a $\Gamma(2, 1)$ distribution or from a model with fixed effects. The locations for loci were randomly sampled throughout the genome. For each genomic data set, 4, 8, or 32 loci with phenotype associations were simulated. The total heritability of the phenotype was fixed at either 0.5 or 0.9. The MAF is computed for each sampled locus in the genetic model since each locus is chosen from the SNPs generated by MaCS. By combining the MAF with the effects sampled for each locus in the genetic model, it is possible to determine the proportion of observed variation contributed by each locus. This individual heritability for each locus is defined as follows:

$$h_j^2 = \frac{2f_j(1 - f_j)\beta_j^2}{\sigma_p^2} \tag{2.8}$$

where f_j is the MAF of locus j, β_j is the additive effect of the locus j, and σ_p^2 is the total phenotypic variance of the trait.

GWA analysis of the simulated data were performed using both V-Bay and a linear regression single-marker analysis. When population structure was incorporated, the linear model (1) becomes a fixed effect ANOVA model, for both V-Bay and the single-marker analysis. The population means in V-Bay were treated as having normal priors centered on zero with a very large variance ($\tau = 1000$), and were updated in a similar fashion as the other parameters in the V-Bay algorithm. The V-Bay algorithm was run until the tolerance for the likelihood portion of the lower bound $\mathcal{L}(\theta)$ was $< 10^{-9}$. For the simulations with missing data, the minor allele frequency across loci ($f_j \ \forall j$) was estimated given the observed genotype data, and then the missing data points were sampled from a $Bin(n = 2, f_j)$, i.e. assuming Hardy-Weinberg equilibrium, for both V-Bay and single-marker analysis. We did random re-sampling of missing data to test the robustness of the output of V-Bay and the single-marker analysis (Appendix A.2, Methods).

The false positive and true positive rates were calculated for each set of replicate simulations. Care was taken to account for the effect of linkage disequilibrium on the test statistics, for both V-Bay and single-marker analysis. A simple window was computed around each marker to determine when the r^2 decayed to 0.4. The cutoff of 0.4 was used to be as generous to single-marker analysis as possible. Any marker in this window was considered a true positive. In the case where multiple recombination events occurred recently between different ancestral lineages, multiple blocks of markers in linkage disequilibrium were generated, that were separated by markers in low linkage disequilibrium. In these cases, a conservative rule for evaluating a true positive was implemented. If a marker had a p-vbay>0.99, or $-\log_{10}$ p-value for the single-marker analysis in excess of the Bonferroni correction, and the r^2 between the significant genetic marker and the true location was greater than 0.4, then the marker was considered a true positive.

For the comparison between V-Bay, the lasso, and single marker analyses, one-hundred thousand markers and samples sizes of 200 or 1000 for a single population were simulated (the reduced number of markers for these simulations was used to conserve CPU cycles). The genetic architectures were simulated as with the larger scale simulations, but with only 4 or 32 loci being sampled randomly from the one-hundred thousand markers, and effects sampled from a $\Gamma(2, 1)$ distributions for 10 replicated data sets. Eight random reorderings of the markers were used with the V-Bay analysis, and the false discovery rate for V-Bay was controlled based on the consensus of associations found across reorderings with p-vbay>0.99 (e.g. a false discovery rate of 5% corresponded to an association being found in at least 3 out of the 8 reorderings). The false discovery rate for the lasso (using F-statistics) and single-marker analysis were controlled based on the p-values computed for each method respectively.

2.5.6 Data Analysis

We performed a GWA analysis for gene expression levels measured in the eternal lymphoblastoid cell lines that were generated for the 210 unrelated individuals of Phase II of the International HapMap project [123]. This sample included 60 individuals sampled from Utah of European descent (CEU), 45 individuals sampled from Han Chinese population (CHB), 45 individuals sampled from Japanese population (JPT), and 60 individuals sampled from the Yoruban population in Africa (YRI). Expression data for these lines were available for 47,000 probes for (~17,000 genes) assayed with the Illumina bead array. For our analyses, we screened for MAF > 0.10 in all populations which left $1.03 * 10^6$ SNPs on chromosomes 1 to 22. The X and Y chromosomes were not analyzed by Stranger et al. and we ignored these chromosomes in our analyses as well. Stranger et al. [123] reported 879 gene expression probes with highly significant *cis*-eQTL associations, found by testing within populations, where every SNP in a 2Mb window around each gene was analyzed. We performed a GWA analysis, with both V-Bay and a single-marker regression, for their top 100 most significant expression probes. We combined genotypic data across populations, where we accounted for the effect of population structure in each case by including appropriate covariates. We also tested the top 20 probes, not in their association list that had the largest residual variance after correcting for population structure. Only 120 expression probes were analyzed to conserve CPU cycles; all 879 could easily be analyzed in a future study. The total missing data for this SNP set was 1.78%. We accounted for missing data using the same approach as with our simulated data analysis.

CHAPTER 3

GENE EXPRESSION NETWORK RECONSTRUCTION BY CONVEX FEATURE SELECTION WHEN INCORPORATING GENETIC PERTURBATIONS

3.1 Abstract

Cellular gene expression measurements contain regulatory information that can be used to discover novel network relationships. Here, we present a new algorithm for network reconstruction powered by the adaptive lasso, a theoretically and empirically well-behaved method for selecting the regulatory features of a network. Any algorithms designed for network discovery that make use of directed probabilistic graphs require perturbations, produced by either experiments or naturally occurring genetic variation, to successfully infer unique regulatory relationships from gene expression data. Our approach makes use of appropriately selected *cis*-expression Quantitative Trait Loci (*cis*-eQTL), which provide a sufficient set of independent perturbations for maximum network resolution. We compare the performance of our network reconstruction algorithm to four other approaches: the PC-algorithm, QTLnet, the QDG algorithm, and the NEO algorithm, all of which have been used to reconstruct directed networks among phenotypes leveraging QTL. We show that the adaptive lasso can outperform these algorithms for networks of ten genes and ten *cis*-eQTL and is competitive with the QDG algorithm for networks with thirty genes and thirty *cis*-eQTL, with rich topologies and hundreds of samples. Using this novel approach, we identify unique sets of directed relationships in *Saccharomyces cerevisiae* when analyzing genome-wide gene expression data for an intercross be-

tween a wild strain and a lab strain. We recover novel putative network relationships between a tyrosine biosynthesis gene (TYR1), and genes involved in endocytosis (RCY1), the spindle checkpoint (BUB2), sulfonate catabolism (JLP1), and cell-cell communication (PRM7). Our algorithm provides a synthesis of feature selection methods and graphical model theory that has the potential to reveal new directed regulatory relationships from the analysis of population level genetic and gene expression data.[2]

3.2 Introduction

Network analyses are increasingly applied to genome-wide gene expression data to infer regulatory relationships among genes and to understand the basis of complex disease [23, 39]. Probabilistic graphical techniques, which model genes as nodes and the conditional dependencies among genes as edges, are among the most frequently applied methods for this purpose. A diversity of such approaches have been proposed including Bayesian networks [50, 101, 160], undirected networks [95, 114, 79], and directed cyclic networks [82, 84, 20]. The popularity of these methods derives, in part, from the structure of these models that is well suited to algorithm development and because the network representation of these models can be used to construct specific biological hypotheses about the processes governing the activity of genes in a system [50]. As an example of this latter property, genes connected by an edge may indicate (at least) one of the genes is regulated by the other.

[2] This chapter was published in PLoS Computational Biology on December 2, 2010 as a research article [88].

In graphical network inference, a theoretical principle that is now well appreciated [134, 71, 111, 108, 20, 84, 21, 160, 161] is that 'perturbations' of the network can be leveraged to reduce the set of possible networks that can equivalently explain gene expression. In fact, since equivalent models can indicate conflicting regulatory relationships, perturbations are often necessary to extract regulatory relationships with any confidence. If the perturbations are controlled (e.g. knockouts of single genes), then a network among n genes can be recovered very efficiently with n knockouts [134]. Alternatively, perturbations that arise from naturally segregating variants, or combinations of genetic variants produced from carefully designed crosses, can also be leveraged [71, 111, 108, 20, 84, 4, 21, 98, 161, 160]. Perturbations of this type, caused by genetic polymorphisms in a population that alter the expression of genes across a population sample, are expression quantitative trait loci (eQTL) [108].

Despite the acknowledged importance of perturbations in network analysis, there has been little theoretical work concerning sets of perturbations that maximally limit the set of equivalent models for arbitrary directed networks. Limiting the set of equivalent models is of particular concern in cases where the true network has cyclic structure, where the set of statistically indistinguishable models may include drastically different topologies [107]. In this paper, we present theory concerning a minimally sufficient set of (genetic) perturbations to infer a maximally limited equivalent set of network architectures, which can subsequently be reconstructed using a single, convex optimization procedure. We demonstrate that for a specific type of network among both gene expression products and genotypes (an interaction or conditional independence network [80]), when including an appropriate set of genetic perturbations for the genotypes, specifically locally occurring *cis*-eQTL [111], the interaction network contains all the information necessary for directed network reconstruction. We can therefore estimate the regulatory relationships or features of a network directly from the interaction network with many different approaches [97, 48, 79, 3, 114]. Here, we use the adaptive lasso [162], a convex optimization procedure, to efficiently solve this model selection problem. This approach allows us to avoid the reliance on computationally inefficient heuristics [50, 101, 20, 84, 4, 21, 98, 160] with non-unique solutions, which can generate many possibly poor-fitting networks when considering sample sizes that are typical of experiments collecting genome-wide gene expression data.

Our algorithm includes three steps. First, an association analysis is carried out to identify strong local (*cis*-eQTL) perturbations of gene expression. Second, we combine the gene expression data and genotypes for the *cis*-eQTL, and use an adaptive lasso regression procedure [162, 79] to identify an interaction network [80] among gene expression products and *cis*-eQTL genotypes. The novel component of our algorithm is incorporated into this step, where we can immediately extract a unique, directed acyclic or cyclic network, given each gene in the network analysis has a unique *cis*-eQTL. Third, to ensure the edges in the interaction network correspond to the correct dependencies in the directed graph, we do a permutation test to ensure marginal independence between the *cis*-eQTL and the upstream gene based on the undirected edges recovered. We only use genetic perturbations that are *cis*-eQTL because of empirical evidence that local genetic polymorphism tends to have larger effects than *trans*-eQTL [113, 16, 122], and are therefore statistically more likely to be linked to locally causal variants. If the true network is a directed cyclic graph and if one uses *trans*-eQTL to attempt to find the true model, there can still be a larger equivalence class of models, since there is no way to know which gene a *trans*-eQTL actually feeds into in a cyclic graph because of equivalence (this is shown in the "Recovery" Theorem in the Methods). Our approach mirrors directed network inference approaches that seek to identify conditional independence and dependence relationships but avoids a computationally demanding step of iteratively testing for these relationships [73, 107, 20, 25].

To test this algorithm, we explore performance for simulated data. Specifically, the simulations are designed to capture scenarios where the underlying network is relatively sparse, and the strength of both the *cis*-eQTL and regulatory relationships is strong enough to detect given a relatively small numbers of samples, on the order of the number of genes being tested. We investigated networks of modest size (either 10 or 30 genes), since we wished to focus on cases where the set of genes being tested have strong *cis*-eQTL in linkage equilibrium, which in a typical eQTL genome-wide association study will be much smaller than the total number of genes being tested, [122, 16]. As a benchmark, we compare the performance of our algorithm to the PC-algorithm [120, 73], the QDG algorithm [20], the QTLnet algorithm [21], and the NEO algorithm [4]. We find that our algorithm can outperform all of these approaches in terms of controlling the false-discovery rate, and having greater power (given a large enough sample size) for the recovery of directed acyclic graphs and directed cyclic graphs. To empirically assess our algorithm, we also analyze data from a well powered intercross study in yeast [16]. From this analysis, we identify 35 genes with strong, independent *cis*-eQTL, and leveraged these perturbations to identify novel interactions. While we analyze the data from an intercross, both the theoretical results as well as the algorithm itself can be applied to natural populations as well.

3.3 Results

3.3.1 The gene expression network model

Biologically, our goal is to identify relationships between the expression of multiple genes, such as the case depicted in Figure 3.1. In this figure we see that the expression level of Gene A has an effect on the expression level of Gene B, mediated through some biological process (i.e. unobserved factors). Even though we do not directly observe all the factors involved in the regulatory interaction, we still want to be able to detect that there is a regulatory effect, including the relative magnitude, the presence, and direction of the effect. To resolve these relationships uniquely, we need perturbations of expression, which in this case arise from genetic polymorphisms affecting expression. Therefore, both gene expression and genotype data needs to be collected on the same set of individuals, for all genes of interest, as well as all genotypes that will possibly act as perturbations of expression. Overall, one can consider our model selection process as acting on the joint covariance between and within the gene expression products and genotypes identified as being strong QTL. In our algorithm we further focus on *cis*-eQTL, because of recent studies indicating that there are widespread genetic polymorphisms local (i.e. *cis*) to genes that cause significant changes in expression [113, 16, 122].

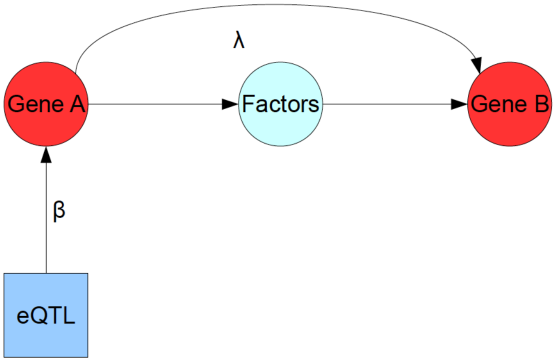

Figure 3.1: Example of biological relationships that can be reconstructed by the algorithm. An expression Quantitative Trait Locus (eQTL) directly alters the expression level of Gene A, a relationship that we represent in our network model with the parameter β. This gene in turn has an effect on Gene B through an unobserved pathway represented by the 'Factors' node. While these factors are unobserved we can still infer that there is a regulatory effect of Gene A on the downstream Gene B, which is represented in our network model by the parameter λ.

We want to identify the genes with strong *cis*-eQTL (x) with linear effects on gene expression (y) parametrized by genetic effect parameters (β), and then identify unique regulatory relationships among gene expression products parametrized by λ. For p measured gene expression phenotypes and m loci for which we have genotypes, the directed graphical model of the network has $p+m$ nodes and $(p(p-1)+pm)$ possible edges, representing $p(p-1)$ possible regulatory relationships among the genes, and pm possible perturbation effects of loci (eQTL) on each of the expression phenotypes. Written in matrix notation, the network model for a sample of n individuals can be represented as:

$$\mathbf{Y}_{nxp}\mathbf{\Lambda}_{pxp} = \mathbf{X}_{nxm}\mathbf{B}_{mxp} + \mathbf{E}_{nxp}, \tag{3.1}$$

where $\mathbf{Y}$ is a matrix of gene expression measurements, $\mathbf{\Lambda}$ is a matrix of regulatory effects, $\mathbf{X}$ is a matrix of observed perturbations, $\mathbf{B}$ is a matrix of genetic effect parameters, and $\mathbf{E} \sim N(\mathbf{0}, \mathbf{R})$, where $\mathbf{R}$ is a diagonal matrix. Non-zero elements of $\mathbf{\Lambda}$ and $\mathbf{B}$ are edges representing regulatory relationships and eQTL effects, respectively, where the size of the parameter indicates the strength of the resulting relationship, as shown in Figure 2.1. Versions of this model are used regularly in analysis of networks [50, 84, 79] when assuming that gene expression measurements are taken from independent and identically distributed (*iid*) samples, where the regulatory relationships can be approximated by a system of linear equations, and the distribution of expression traits across samples is well modeled with a multivariate normal distribution. Another common assumption we make use of in our algorithm is that most detectable eQTL effects will have a significant linear component, especially for *cis*-eQTL [16, 122], where the polymorphism has simple switch-like behavior, such as determining whether transcription of the gene is up or down regulated.

A potential pitfall of modeling expression traits using directed networks of the type in Equation 3.1 is the problem of likelihood equivalence between models. Figure 3.2 presents a simple example that illustrates the problems raised by equivalence for network inference. In this example, the true model, which is a linear pathway between four genes $x \rightarrow y \rightarrow z \rightarrow t$, is indistinguishable from three other equivalent models. Each of these equivalent models has a very distinct implication for regulatory relationships among these genes but they are indistinguishable, regardless of the sample size. To be able to distinguish between these models, one needs to either collect time-course data to determine the temporal sequence in which regulation occurs [163], or alternatively, perturb

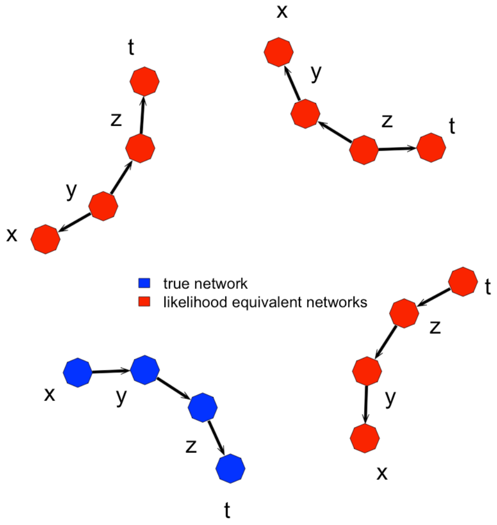

Figure 3.2: Example of a graphical model equivalence class when determining regulatory relationships among four genes $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{t})$. Edges represent the direction of regulation. In this case, the true regulatory network connecting the four genes (blue) has the same sampling distribution as the other three incorrect models (red). Without perturbations (i.e. eQTL), each of these models will equivalently describe the pattern of expression observed among these genes for any data-set.

the expression level of these genes in some fashion.

3.3.2 The Algorithm

Our goal is to identify a unique network underlying the observed expression and genotype data, especially when the sample size is at most 1,000 (a large, biologically realistic sample size). To accomplish this, in the Methods we prove a set of theorems to show that if each gene being considered has its own, unique eQTL, then one can go from the sample covariance among gene expression phenotypes and genotypes (defined as $\mathbf{S}$ in the Methods, see Figure 3.3a), to the inverse covariance (i.e. precision matrix or undirected network defined as $\mathbf{\Sigma}$ in the Methods, see Figure 3.3b), then subsequently to a directed cyclic network underlying the expression data (defined as $\mathbf{\Lambda}$, see Figure 3.3c), where the last step makes use of our "Recovery" Theorem. In the algorithm, we begin with a screening process to identify a set of expression traits with putative strong *cis*-eQTL (Step 1). We then make use of the adaptive lasso function for reconstruction of conditional independence networks (i.e. the structure of the inverse covariance matrix, Figure 3.3b) (Step 2) to identify genes with strong induced dependencies among *cis*-eQTL genotypes and gene expression phenotypes and reconstruct the unique directed acyclic or cyclic network that is a result of these induced edges. Finally, for each putative strong induced dependency, we further filter the induced edges based on a permutation test (Step 3), to ensure marginal independence between the upstream gene and the downstream *cis*-eQTL:

Step 1: Selection of expression phenotypes:

A standard genome-wide association analysis is performed on each expression trait, focusing on genetic polymorphisms in a *cis*-window around a gene (e.g. a 1Mb window) [122]. Each marker is tested individually using either a linear statistical model or non-parametric test statistic (e.g. Spearman rank-correlation), with a correction for multiple tests using either a control of false discovery rate [7], a conservative Bonferroni correction (i.e. α/n, where α is the significance

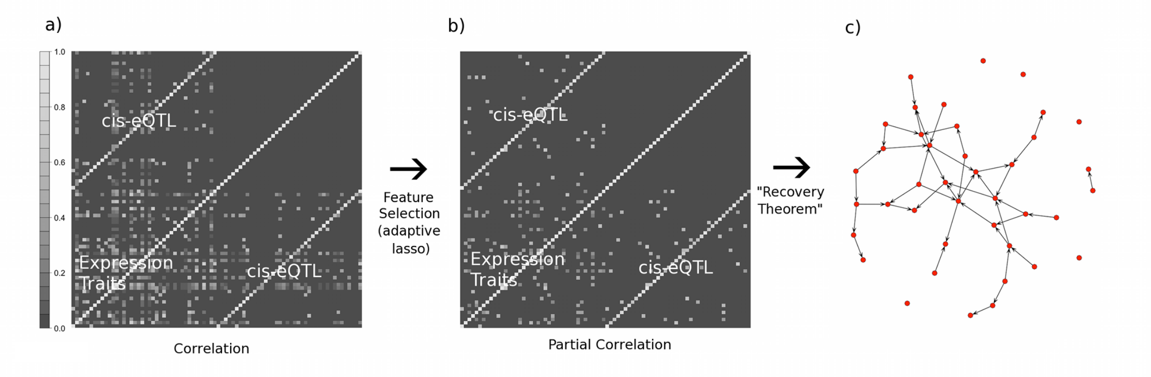

Figure 3.3: Outline of the structure of Step 2 of the algorithm. (a) After selection of phenotypes in Step 1, we produce a covariance matrix between observed gene expression products, and their associated unique *cis*-eQTL. (b) A convex feature selection method (the adaptive lasso) is used to learn the structure of the inverse covariance matrix, which is also the conditional independence or interaction network among gene expression products and *cis*-eQTL genotypes. (c) The directed cyclic network among expression products can then be recovered directly from the conditional independence network, using the "Recovery" Theorem. For Step 3, each of the induced edges between expression phenotypes and *cis*-eQTL, shown in (b), are tested to ensure marginal independence using a permutation test.

level and n is the number of tests), or through a permutation approach to compute significance based on the empirical distribution of test statistics after shuffling the data, as in Stranger et al. [122]. After this initial association analysis is performed, the remaining *cis*-eQTL and their associated genes are further filtered such that the *cis*-eQTL genotypes are strongly independent of one another. In our analyses we use the very conservative cutoff $r^2 <= 0.03$ between any pair. This ensures that each *cis*-eQTL represents a unique perturbation, which is especially important for small sample sizes, when the sampling variability of the entire data-set is high.

Step 2: Regulatory network reconstruction:

Once the set of expression phenotypes are identified, we combine the genotype and gene expression data, so as to infer a joint gene expression, *cis*-eQTL interaction network, (i.e. identifying which elements of the matrix $\mathbf{\Sigma}$ are non-zero). This model selection method is similar to the network recovery method proposed by [97], except using the adaptive lasso instead of the regular lasso [79]. The adaptive lasso procedure is performed by first solving the lasso problem:

$$\text{argmax}_{\alpha}\left\{-\sum_{i=1}^{n}(y_i - z_i\alpha)^2 - \eta\sum_{j=1}^{p+m-1}|\alpha_j|\right\} \tag{3.2}$$

then using the coefficients from this problem to solve the following adaptive lasso problem [162]:

$$\text{argmax}_{\zeta}\left\{-\sum_{i=1}^{n}(y_i - z_i\zeta)^2 - \eta\sum_{j=1}^{p+m-1}\hat{w}_j|\zeta_j|\right\} \tag{3.3}$$

for every phenotype, y_i in the reduced data-set, where $\hat{w} = |\hat{\alpha}|^{-1/2}$, z is the combined gene expression products and associated *cis*-eQTL genotypes, and α and ζ are the corresponding regression coefficients, whose non-zero structure should asymptotically be the same as $\mathbf{\Sigma}$, given an appropriate choice of the penalty parameter η. The penalty parameter η is chosen by five fold cross validation based on the mean-squared prediction error across both steps of the procedure. In addition, all variables are centered to have mean zero and rescaled to have variance one, so that the gene expression products and genotypes with small or large variances will not be penalized differently. After the interaction network is determined, we infer the directed regulatory network immediately from the interaction network structure, based on the results shown in the "Recovery" Theorem.

While we could make use of any undirected inference approach that infers the conditional independence network [73, 107, 20, 25] for Step 2, we use the adaptive lasso because of its theoretical advantages [162] and empirical performance, as far as finding sparse solutions with the lowest mean-squared error (by cross-validation) [79]. A lasso type procedure can be used for model selection [97] by shrinking parameters to exactly zero and is convex [129], providing computationally efficiency. However, there has been theoretical work showing that since the lasso shrinks non-zero parameters too harshly, it will not always return the true model asymptotically (i.e. as sample size goes to infinity). In fact the conditions under which it will return the correct model may be very unlikely for high dimensional problems [158]. The adaptive lasso was proposed to remedy this problem, and in general appears to have better properties as far as model selection both theoretically and in practice, without sacrificing the convexity of the lasso [162, 79].

Step 3: Edge interpretation and filtering:

The primary goal of the "Recovery" Theorem is to map the problem of learning a directed cyclic graph among a set of phenotypes onto the problem of learning an undirected graph among a set of phenotypes and appropriately selected genotypes (i.e. unique *cis*-eQTL), then determining the corresponding directed cyclic graphs from the original problem. Each edge in this idealized larger undirected graph between the genotypes and the phenotypes represents an induced dependency between a given *cis*-eQTL and the immediate upstream phenotype of that *cis*-eQTL's *cis*-gene. Yet in practice, some of these edges identified in the undirected graph may arise from *trans*-effects, i.e. a given *cis*-eQTL may also have a large marginal correlation with another gene expression product in the data-set, that is not explained away entirely by the relationships inferred among phenotypes. In this case a further test can be performed, to ensure that for any putative induced dependencies identified from the undirected graph, the *cis*-eQTL and upstream gene are marginally uncorrelated. For this we perform a resampling method of the marginal correlation between *cis*-eQTL and upstream phenotype, and only use the edges which are very likely induced dependencies, in this case where the probability of observing a larger marginal correlation, given that they are uncorrelated, is 0.90. This threshold of 0.90 was used as a highly conservative threshold for marginal independence.

3.3.3 Simulation analyses and comparison to other network recovery algorithms

To benchmark the performance of our algorithm, we compared it to the PC-algorithm [120, 73], the QDG algorithm [20], the QTLnet algorithm [21], and the NEO algorithm [4]. The other previously proposed cyclic algorithms either do not scale well (e.g. the approach of Li et al. [82]) or have prohibitively complex implementations (Richardson's cyclic recovery algorithm [107] or the algorithm of Liu et al. [84]). The PC-algorithm is designed to recover directed acyclic graphs using iterative tests of conditional dependence and independence, is a computationally efficient algorithm (scales to thousands of genes for sparse networks), and has competitive performance with other directed acyclic graph reconstruction algorithms [73, 131]. Additionally, the PC-algorithm forms the backbone of the QDG algorithm where it is used to construct an undirected graph (the skeleton of the directed acyclic graph) among expression phenotypes before orienting these edges using known QTL [20]. The QTLnet algorithm proposes a full Markov chain Monte Carlo approach to network inference, but does not scale above twenty phenotypes because of convergence rates of the Markov chain, and does not explicitly model directed cyclic graphs [21]. We also compared our algorithm to the NEO algorithm [4], and found that our approach controlled the false-discovery rate much better and had higher power for small networks ($p = 5$, results not shown), but the implementation of the NEO algorithm available from the author was not stable for our simulations of larger networks ($p >= 10$), and so we did not include it in a larger comparison.

To compare the performance we simulated data from the model presented in Equation 3.1 with strong *cis*-eQTL, low sample variances, and different topologies, representing a scenario where there are strong eQTL, and few direct interactions between genes, with sample networks illustrated in Figure 3.4. The four different classes of simulations included directed acyclic graphs for 10 phenotypes, with sparse and dense topologies (Figure 3.4a, 3.4b), and directed cyclic graphs for dense (Figure 3.4c) and intermediate topologies (Figure 3.4d), with 10 and 30 phenotypes respectively, for a total of 160 distinct network topologies generated across all the simulations. This simulation is biologically motivated by the need for strong, statistically independent *cis*-eQTL and interactions among genes, as observed in previous studies [113, 16, 122].

We simulated a set of either 10 or 30 expression phenotypes and genotypes for sample sizes of $n = 50, 100, 200, 300, 400, 600, 800$, and 1000 for both directed acyclic graphs and directed cyclic graphs. We simulated an F2 cross with the R package QTL [17], with either 10 or 30 independent known unique *cis*-eQTL of constant effect ($\text{diag}(\mathbf{B}) = 1$), and error variances of $1\text{x}10^{-2}$. The regulatory effects ($\mathbf{\Lambda}$) were sampled from a uniform distribution with parameters $(1/2, 1)$ or $(-1, -1/2)$ with equal probability. The network topologies were generated by randomly connected variables with equal probability, where the expected number of edges for each variable was either one, two, or three.

Five replicate simulations were performed, sampling a new network topology and parameterization each time, and the power and false-discovery rate were computed for the adaptive lasso, PC-algorithm, QDG algorithm, and QTLnet algorithm for 10 expression traits, and all except QTLnet for 30 expression traits (because of the scaling of QTLnet). In addition, because we simulate the QTL independently, with no *trans* effects, we do not perform the third step of our adaptive lasso algorithm. We compared the performance for both directed acyclic graphs as well as directed cyclic graphs. In Figure 3.5 and Figure 3.6 we show the power and false discovery rate for recovering the correct set of directed edges using these methods. While some of the power and false-discovery rate curves show large fluctuations with increasing sample size in Figure 3.5 and Figure 3.6, this is due to elevated sampling variability due to each replicate sim-

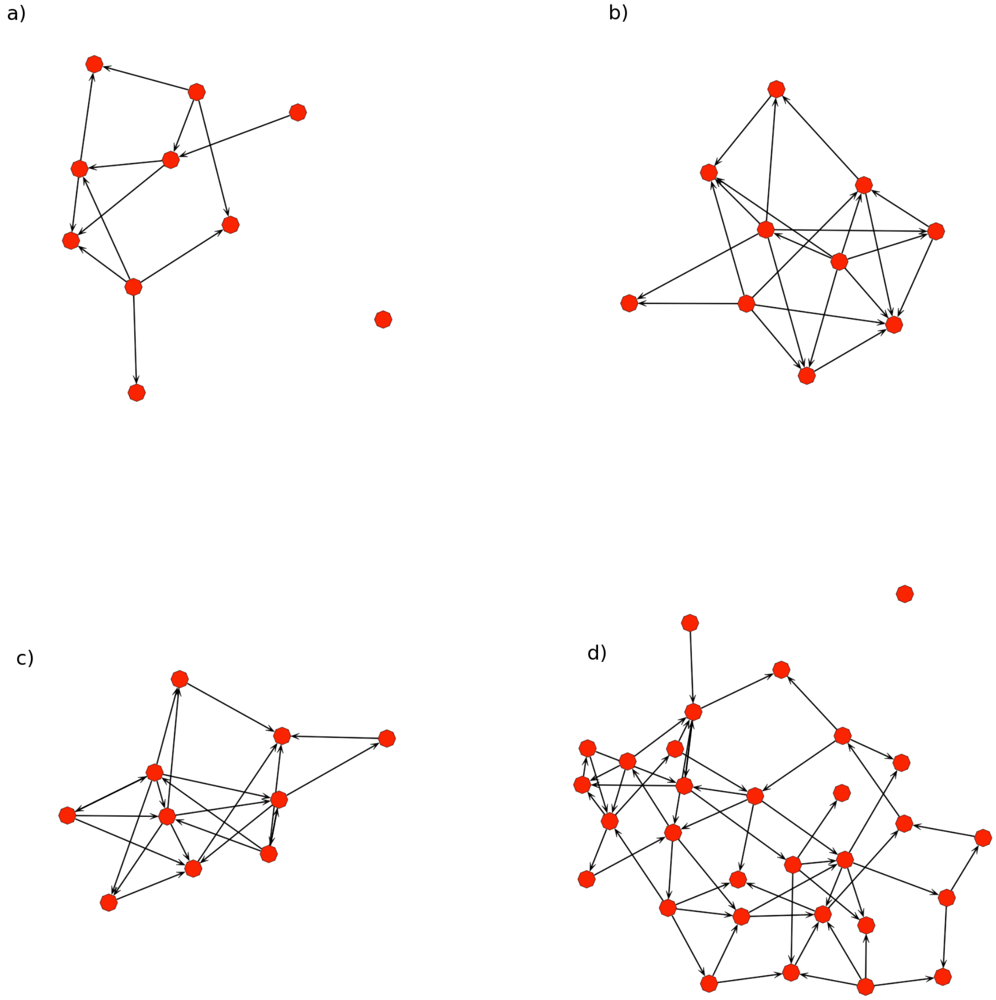

Figure 3.4: Examples of four network topologies used to simulate gene expression data from 160 total topologies. Sparse acyclic (a), dense acyclic (b), and dense cyclic (c) graphs were simulated for networks with 10 genes. Intermediately dense cyclic networks were simulated networks with 30 genes (d). Nodes represent expression levels of genes and the directed edges represent regulatory (conditional) relationships among genes, where the strength of the relationships were determined by sampling from a uniform distribution. Each phenotype (node) has a unique, independent cis-eQTL feeding into into it (not shown), with constant effect.

ulation having a unique topology and parameterization.

For two of these scenarios, we show that our algorithm using the adaptive lasso can outperform the PC-algorithm, the QDG algorithm, and QTLnet in terms of statistical performance (see Figure 3.5c, 3.5d and Figure 3.6a, 3.6b) with similar computational scaling. In general, only the QDG algorithm has competitive performance with the adaptive lasso (see Figure 3.6c, 3.6d). This indicates that the necessary sample size to have a significant performance gain over the QDG algorithm may be much larger than is biologically realistic for larger more complex networks. These are significant results in two ways, the first being that we show that a feature selection method using linear regression can 1) identify directed regulatory architecture (given sufficient perturbations) and 2) it can also outperform state of the art network reconstruction algorithms, given a sufficient samples size and appropriate model dimension.

The adaptive lasso approach appears to work the best for smaller problems (i.e. 10 phenotypes) with denser topologies (i.e. Figure 3.4b, 3.4c) and performs better than other approaches in such cases (see Figure 3.5c, 3.5d and Figure 3.6a, 3.6b). This may be because smaller dimensional problems behave asymptotically at a faster rate. Unfortunately, this suggests that for larger problems (e.g. hundreds to thousands of phenotypes), unless the true topology is relatively sparse, the adaptive lasso, and perhaps all of these approaches will have poor performance without unrealistically large sample sizes (e.g. thousands) for both directed acyclic and cyclic graphs. We also performed a simulation for a small network (e.g. 10 phenotypes and 10 *cis*-eQTL), with dense directed acyclic

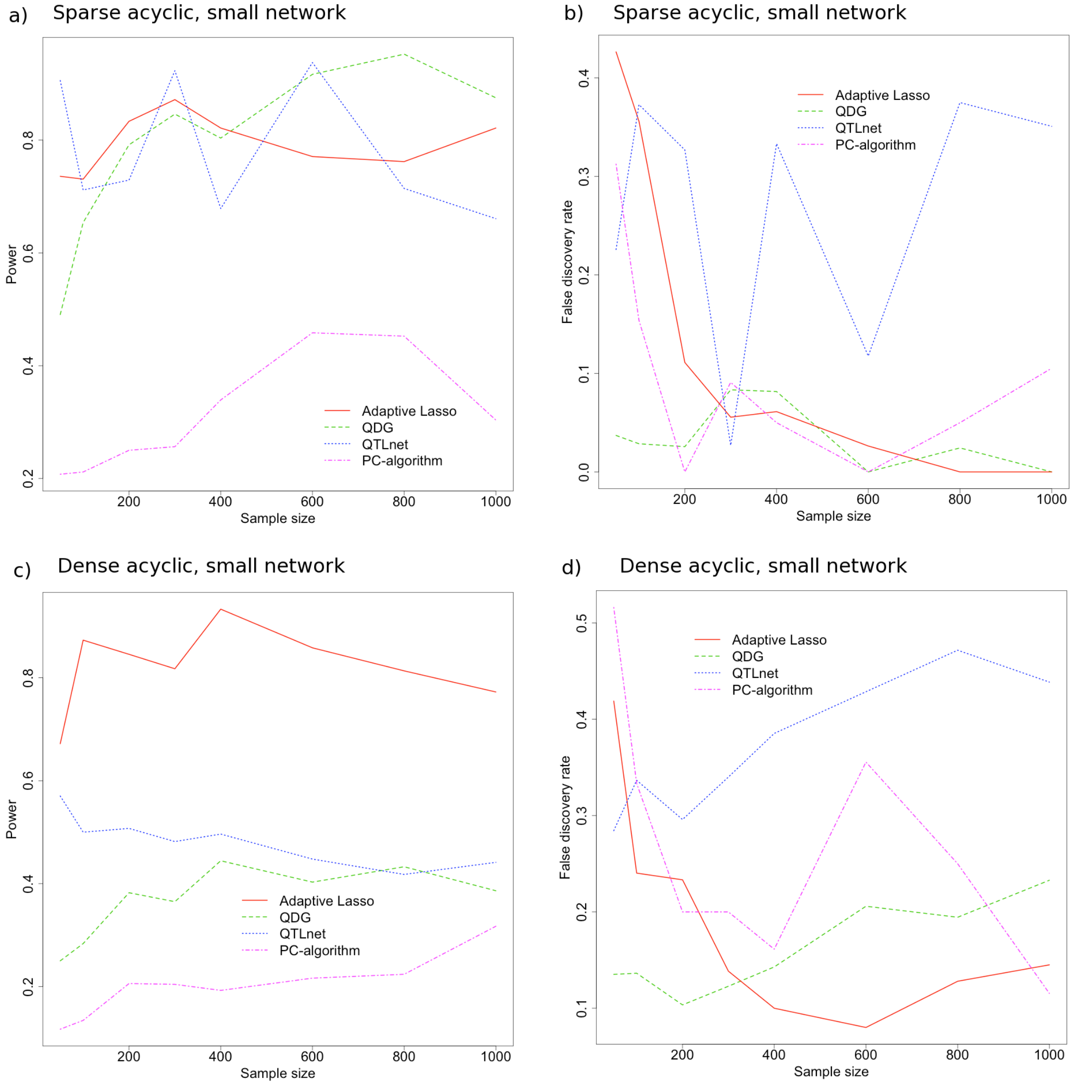

Figure 3.5: Performance of our algorithm using the adaptive lasso for directed acyclic graphs compared to other algorithms. Theses other algorithms include the PC-algorithm, the QDG algorithm, and the QTLnet algorithm for reconstructing different acyclic topologies of 10 genes. For a sparse directed acyclic topology (as in Figure 3.4a), the power (a) and false discovery rate (b) are plotted as a function of the sample size for five replicate simulations. Similarly, for a dense directed acyclic topology (as in Figure 3.4b), the power (c) and false discovery rate (d) are plotted.

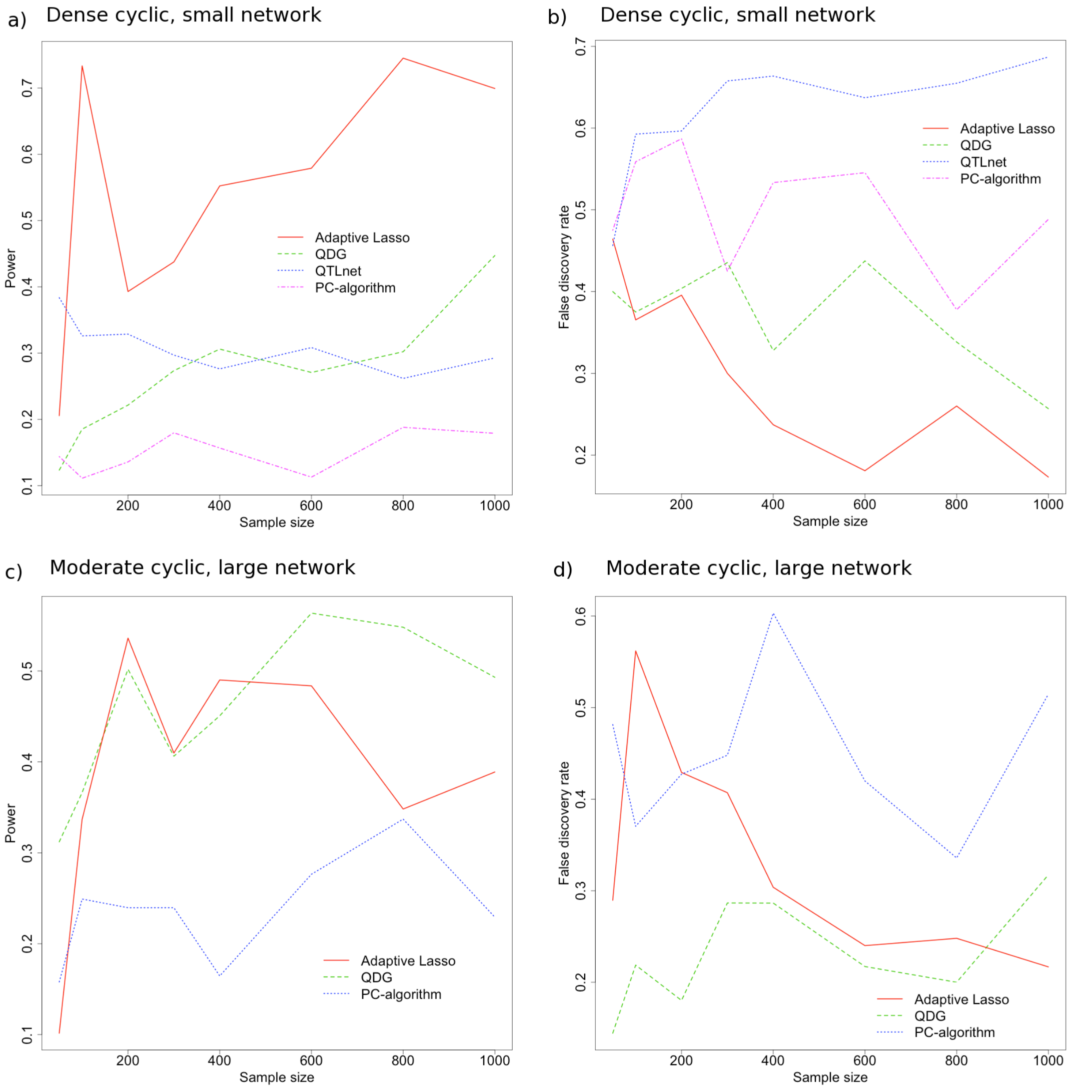

Figure 3.6: Performance of our algorithm using the adaptive lasso for directed cyclic graphs compared to other algorithms. These other algorithms include the PC-algorithm, the QDG algorithm, and the QTLnet algorithm for reconstructing different cyclic topologies of 10 genes (a) and (b) or 30 genes (c) and (d). For a dense directed cyclic topology (as in Figure 3.4c), the power (a) and false discovery rate (b) are plotted as a function of the sample size for five replicate simulations. Similarly, for an intermediately dense directed cyclic topology of 30 genes (as in Figure 3.4d), the power (c) and false discovery rate (d) are plotted.

topology and 200 or 1000 individuals with random variances and eQTL effects simulated from a $\Gamma(2,1)$ distribution. We found a uniform reduction in power (10-20%) across all methods, as well as a modest increase in false discovery rate (5-10%). Increased sample size appeared to correct for this additional randomness in the parameterization (results not shown).

3.3.4 Yeast Network Analysis

We used our algorithm to reconstruct network structure for genome-wide gene expression data and genetic markers assayed in 112 segregants of a cross between two strains of *Saccharomyces cerevisiae*, reported by Brem and Kruglyak [16]. This cross was between a lab strain (BY4716) and a wild strain (RM11-1a), with 2,957 genetic markers genotyped and expression levels for 5,727 genes measured. While the sample size is relatively small, the study was well powered, with many strong *cis*-eQTL and interactions among genes [16]. An individual marker analysis was run around the *cis* region of each gene (25 kb around the start site of the gene) to identify a set of gene expression products with strong *cis*-eQTL ($-\log_{10}$(p-value)$<$ 1x10^{-5}), which identified 262 genes. We further filtered this set to remove *cis*-eQTL genotypes with high linkage, by filtering for a set with pairwise $r^2 < 0.03$ between any two *cis*-eQTL genotypes. Additionally, we tested the robustness of the inferred edges by randomly sampling the flanking genetic markers 20 times for all *cis*-eQTL and refitting the model. The percentage recovery for the top six recovered directed edges for the 20 resamplings are shown in Table 3.1. All missing data for a given genotype or phenotype was set to the sample mean of the respective variable.

After the additional filtering described above, we were left with 35 genes with unique, independent *cis*-eQTL, with an undirected network shown in Figure 3.7a, and possibly directed network shown in Figure 3.7b. Performing the adaptive lasso procedure on these 35 gene expression phenotypes and 35 genotypes identified 91 possibly directed edges among these genes, and 145 undirected edges among the genes. These hits were further filtered to ensure they represented induced dependencies, leaving six edges with relatively strong evidence of directionality (see Table 3.1 and Figure 3.7b). These include four edges feeding out of the TYR1 gene, a gene involved in tyrosine biosynthesis [94]. Since TYR1 is also a hub in the undirected network (see Figure 3.7a), this suggests that amino acid biosynthesis, and perhaps anabolism in general is driving the expression of many of this particular subset of genes. The genes in which TYR1 appears to have direct effects on have diverse molecular and biological functions including endocytosis (RCY1), sulfonate catabolism (JLP1), cell-cycle checkpoint (BUB2), and cell-cell communication (PRM7) [141, 64, 46, 60].

Additionally PRM7 feeds into POC4, a proteasome chaperone protein [81], representing possible cross-talk between cell-cell communication response and protein processing. Finally, SEN1, a helicase indicated in RNA polymerase 2 termination [105], appears to robustly directly affect MST27 an integral membrane protein implicated in vesicle formation [109]. In the implied undirected graph, there were striking topological features, including an average degree of 8.28 (relatively dense), and four genes appeared to be major hubs of a sort, TYR1, NUP60, RDL1, and POC4. These hub genes may represent major axes of variation driving the expression of this subset of genes including processes such

Table 3.1: Directed regulatory edges identified by the adaptive lasso for *S. cerevisiae* cross.

Regulator gene	Target gene	Scaled effect	% Recovery from adjacent marker resamplings
TYR1	RCY1	0.035	0.05
TYR1	JLP1	0.123	0.35
TYR1	BUB2	-0.0056	0.55
TYR1	PRM7	0.0576	0.55
SEN1	MST27	-0.135	0.85
PRM7	POC4	0.154	0.15

as amino acid biosynthesis, information transfer across the nuclear envelope [32], and protein degradation. While most of the edges in the network were not orientable, there still appeared to be many dependencies (even with a possibly high false-discovery rate), indicating a potentially complex set of regulatory interactions, projected on this subset of genes, driving variation in expression. Additionally, there were many edges from eQTL that would appear to be *trans* associations (i.e. with large marginal correlations), demonstrating that many of the pathways that mediate these *trans* genetic effects are not captured in the observed sets of genes. Based on the simulation study, and the complexity of the recovered network (which most likely indicates a high false discovery rate), a much higher sample size would need to be collected to definitively resolve this possible set of regulatory interactions, and have increased confidence in the directional interpretation of the induced edges.

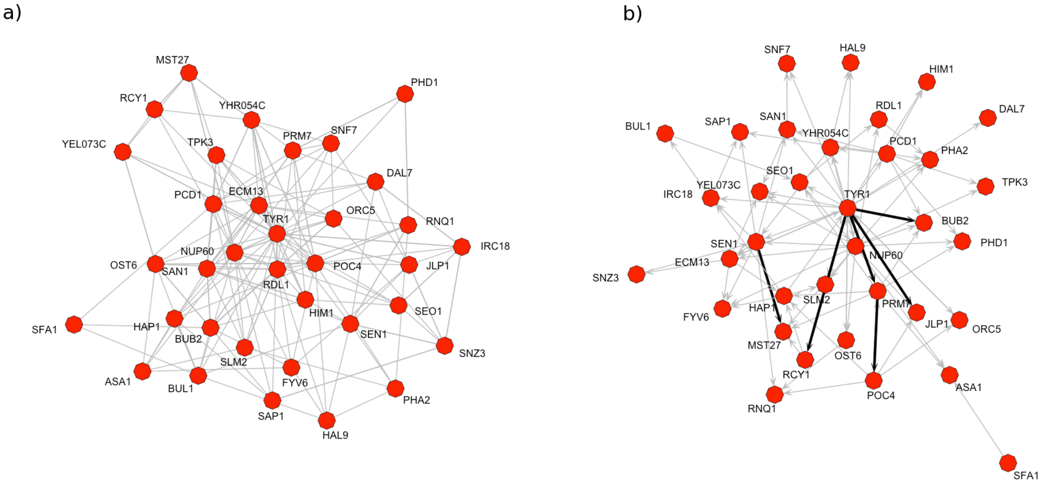

Figure 3.7: Sparse network reconstruction among 35 gene expression products. These genes were filtered for having strong, independent *cis*-eQTL (pairwise $r^2 \leq 0.03$) using the adaptive lasso algorithm for a *Saccharomyces cerevisiae* cross between a wild strain and lab strain [16], with 112 segregants (see text for details). (a) Recovered undirected network among these 35 gene expression products and (b) putative directed network reconstructed for the same genes, based on the edges between *cis*-eQTL (not shown) and the 35 genes. Bold edges represent directed edges with strong confidence based on a resampling procedure (see text for details).

3.4 Discussion

Our algorithm represents a novel approach to directed network recovery by making use of a convex optimization approach for regulatory feature selection when analyzing gene expression products and *cis*-eQTL. This is the first algorithm that makes use of sufficient sets of *cis*-eQTL to infer unique directed cyclic networks from gene expression data with a feature selection methodology. Our use of the adaptive lasso procedure for feature selection has significant computational and theoretical advantages, since the underlying optimization program is convex (ensuring a computationally efficient, unique solution), is model selection consistent, and has the oracle property (asymptotically, the estimates of the non-zero regression coefficients behave as if the model was known *a priori*) [162]. There have not been many algorithms proposed for genome-wide cyclic regulatory network recovery, [107, 82, 84, 20] and they all have either computational or theoretical challenges associated with them, including heuristic searches through regulatory network space with no guarantee to reach networks with the strongest evidence given the data [84, 20, 4], or lack sufficient perturbations to allow unambiguous regulatory inference [107, 82]. With respect to directed acyclic network recovery, we see in the simulations that our feature selection approach with sufficient perturbations outperforms the PC-algorithm, the QDG algorithm, and the QTLnet algorithm for dense, small scale problems as shown in Figure 3.5c, 3.5d and Figure 3.6a, 3.6b. This increase in performance is a direct function of the adaptive lasso procedure correctly identifying the children of a given node, which will then force an edge to appear between the additional co-parents of that node, and its unique cis-eQTL. Once all these induced edges are identified, the structure of the directed network can be elucidated, since all the expression parents of each gene will be known. Our algorithm also does this all in a single optimization procedure, avoiding sets of iterative tests, where type-I and type-II errors can build up at each stage, such as in the PC-algorithm. Alternatively for larger more complex graphs the performance appears to be similar to that of the QDG algorithm Figure 3.6c, 3.6d, perhaps because the asymptotic properties take much larger sample sizes to be practically realized.

For the analysis of the yeast data the topology of the identified network included many undirected cycles, with the few orientable edges being acyclic, as shown in Figure 3.7. In addition there were a set of genes which appeared to be hubs (the most connected being TYR1, NUP60, RDL1, POC4, and SEN1, PCD1, and SAN1 to a lesser extent). This phenomena is probably in part due to an inflation in false-positives because of the small sample size, and a complex underlying model with many unobserved variables. Yet a subset of these edges may represent hub genes capturing different broad patterns of variation across this entire sub-network. Even though most of the edges in this network are not orientable, an experiment could be devised where each of these hubs was perturbed, and given the topology it would produce a prediction about how a relatively large set of other genes in the hub's neighborhood would behave. More strongly, in the case of the TYR1 gene which had the most orientable edges, it suggests that if the process driving that gene's expression was stopped, many other genes would also be affected, but not vice-versa.

A number of assumptions concerning biological networks are implicit to our algorithm. These include assumptions that are common to most graphical modeling techniques, such as sparsity, faithfulness, linearity of regulatory relationships, and normally distributed error, as well as an assumption that is specific to our algorithm: the presence of known, independent perturbations from *cis*-eQTL. The common assumptions are reasonable when constructing a first approximation to regulatory network structure. Sparsity and faithfulness (i.e. the true network does not contain pathological parametrizations where there is parameter cancellation) are essential assumptions that are implicit in algorithms for both directed and undirected network inference algorithms [95, 73, 107, 20, 25, 21, 160]. Regulatory relationships are not linear, but linearity is the simplest approximation that provides biologically relevant information, i.e. there is a detectable relationship between two genes, or no relationship. An assumption of normality is conservative in terms of being the most ‘random’ distribution that could have generated the data, since given an observed covariance structure, normal distributions have maximum entropy [136]. Given the absence of knowledge about the specific biological process generating the distribution of expression measurement error, and barring any clear evidence of non-normality in data, such a conservative approximation is appropriate.

The assumption of independent, detectable *cis*-eQTL effects is the most restrictive assumption. Other methods have proposed to use *trans*-eQTL directly to increase the power to detect causal relationships and reduce the space of equivalent models [160, 84, 98, 21, 82, 4, 20]. We require the assumption of only *cis*-eQTL, because without it, there is no longer the exact isomorphism between the undirected graph among genotypes and phenotypes and the directed cyclic graph among phenotypes. This occurs because in the case of directed cyclic graphs, it is statistically impossible to know which phenotype in a network a *trans*-eQTL directly feeds into, unless their is prior knowledge about the true causal structure of the system, as with the assumption we make about *cis*-eQTL. This statistical degeneracy arises as a result of the “Recovery” Theorem, where when there is a set of equivalent models with independent, unique perturbations, that contains reversals of cycles, each equivalent directed cyclic graph will have an alternative perturbation topology (i.e. the mapping between unique eQTL and gene expression phenotypes, determining which eQTL causally affects which gene expression product).

Alternatively, as we show in real data, even if there do appear to be many *trans*-eQTL we can still detect a subset of edges from the *cis*-eQTL that behave how we would like (by using Step 3 of the algorithm). While this may reduce our power to detect directed cycles in practice, it ensures that for real data-analysis we are more confident in the edges we reconstruct. Another possible solution to the incorporation of *trans*-eQTL would be to use the adaptive lasso to generate the initial undirected graph among genotypes and phenotypes, then to orient the edges in the graph using an iterated testing approach, as in the NEO algorithm [4], the algorithm of Millstein et al. [98], or the QDG algorithm [20]. We do not expect the requirement of unique *cis*-eQTLs to be a good approximation for all regulatory modeling situations. However, this assumption also seems reasonable, given recent biological observations of strong local polymorphism associations with gene expression (eQTL) which are often not in linkage disequilibrium [16, 122, 38, 113, 89]. What is more, due to the structure of linkage disequilibrium in outbred populations (the correlation structure among genotypes) it is often possible to identify a large set of *cis*-eQTL that are uncorrelated and each have unique expression phenotypes, e.g. a set of eQTL that are present on different chromosomes or are far away from one another in terms of genetic map distance [122].

As a final comment, the theory of sufficient perturbations that maximize regulatory resolution, which is used as the foundation of our algorithm, is quite general, and could be used to integrate multiple data types to make predictions about putative causal regulators underlying complex phenotypes, such as disease [23, 39]. The "Recovery" Theorem defines a class of perturbation architectures where there is a direct isomorphism between two very different types of networks: the inverse covariance structure (an undirected network) with perturbations and a directed cyclic graph representing a regulatory network. The theory does not require perturbations to be *cis*, just that there be an appropriate set of perturbations that provide resolution. More complex perturbation sets, which include sufficient perturbations as a subset, can also provide maximum resolution. One could therefore construct algorithms similar to the algorithm presented in this paper, without the local *cis* perturbation restriction. Moreover, the specific topology of eQTL effects need not be known, if one is willing to accept the cost of larger network equivalence classes and therefore less total regulatory resolution. With this restriction lifted, it would be possible to jointly infer the genetic perturbation architecture simultaneously with regulatory architecture, although such a joint reconstruction would require much larger sample sizes.

3.5 Methods

3.5.1 The Network Model

The network model is presented in equation 3.1. For this model, we make the assumption that in the true network model, $\mathbf{\Lambda}$ is sparse. In addition, we assume that $\mathbf{R}$, the error covariance matrix of expression products, is diagonal, and $\operatorname{diag}(\mathbf{\Lambda}) = \mathbf{1}$, where the constraint on the diagonal of $\mathbf{\Lambda}$ ensures model identifiability. This constraint corresponds to a lack of self-loops, since the parameters representing self-loops are confounded with the error variance parameters specified by $\mathbf{R}$. These latter assumptions on $\mathbf{R}$ and $\mathbf{\Lambda}$ (i.e. no error covariance or self-loops) are standard, and used by all popular graphical network inference algorithms, directed and undirected, proposed to date [50, 111, 82, 84, 120, 73, 107, 20, 95]. The model depicted by Equation 3.1 is a completely observed structural equation model (SEM) [11].

3.5.2 Likelihood and Equivalence

The conditional log-likelihood of the model defined by Equation 3.1 can be written as:

$$\ell\left(\mathbf{Y}|\mathbf{X};\mathbf{\Lambda},\mathbf{B},\mathbf{R}\right) \propto \log\left\{\det\left(\mathbf{\Sigma_{yy}}\right)\right\} - \mathrm{Tr}\left(\mathbf{\Sigma S}\right), \tag{3.4}$$

where the full precision matrix $\mathbf{\Sigma}$ and empirical covariance matrix $\mathbf{S}$ are:

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma_{yy}} & \mathbf{\Sigma_{yx}} \\ \mathbf{\Sigma^{T}_{yx}} & \mathbf{\Sigma_{xx}} \end{bmatrix} = \begin{bmatrix} \mathbf{\Lambda R^{-1}\Lambda^{T}} & \mathbf{\Lambda R^{-1}B^{T}} \\ \mathbf{BR^{-1}\Lambda^{T}} & \mathbf{BR^{-1}B^{T}} \end{bmatrix} \tag{3.5}$$

$$\mathbf{S} = \frac{1}{n}\begin{bmatrix} \mathbf{Y^{T}Y} & \mathbf{Y^{T}X} \\ \mathbf{X^{T}Y} & \mathbf{X^{T}X} \end{bmatrix}, \tag{3.6}$$

with the data matrices $\mathbf{Y}$ and $\mathbf{X}$ re-centered.

We can define a fully parametrized model matrix $\mathbf{\Gamma}$:

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Lambda R^{-\frac{1}{2}}} \\ \mathbf{BR^{-\frac{1}{2}}} \end{bmatrix}, \tag{3.7}$$

since by definition $\mathbf{R} > \mathbf{0}$, and $\mathrm{diag}\left(\mathbf{\Lambda}\right) = \mathbf{1}$, both $\mathbf{\Lambda}$ and $\mathbf{B}$ can be rescaled by the positive square root of the error precision matrix $\mathbf{R^{-1}}$.

From Equation 3.5, Equation 3.6, and Equation 3.7 the relationship between the fully parametrized model matrix $\mathbf{\Gamma}$, and the full precision matrix $\mathbf{\Sigma}$ is

$$\mathbf{\Gamma}\mathbf{\Gamma}^{\mathbf{T}} = \mathbf{\Sigma}. \tag{3.8}$$

This defines a system of homogeneous polynomials of degree two which exactly specifies the relationship between the directed graph $\mathbf{\Gamma}$, which may contain no cycles (a directed acyclic graph or DAG) or may contain cycles (a directed cyclic graph or DCG), and the moralized undirected graph $\mathbf{\Sigma}$.

Definition of equivalence [100]:

Two sparse directed cyclic graphs specified by the model in Equation 3.1, with parametrization $\mathbf{\Gamma_1}$ and $\mathbf{\Gamma_2}$, are equivalent in distribution iff for all parametrizations $\mathbf{\Gamma_1}, \exists \mathbf{\Gamma_2} : \mathbf{\Gamma_2}\mathbf{\Gamma_2^T} = \mathbf{\Gamma_1}\mathbf{\Gamma_1^T}$ and for all parametrizations $\mathbf{\Gamma_2}, \exists \mathbf{\Gamma_1} : \mathbf{\Gamma_1}\mathbf{\Gamma_1^T} = \mathbf{\Gamma_2}\mathbf{\Gamma_2^T}$.

Intuitively, the parametrization defined by $\mathbf{\Gamma_1}$ and $\mathbf{\Gamma_2}$ provide a unified representation of the directed cyclic graph among gene expression products along with the set of perturbations of expression (i.e. genotypes). This definition of equivalence allows us to characterize our theory of sufficient perturbations.

3.5.3 "Recovery" Theorems

Given the importance of having as small a set of equivalent models as possible for making meaningful inference, and the necessity of perturbations for minimizing equivalence classes, it is of interest to know what will constitute a sufficient set of perturbations, i.e. to shrink the size of arbitrary equivalence classes as much as possible. In the following section we provide proofs of three theorems that describe such a set. We note that it should also be possible to use the work of Richardson on cyclic causal discovery [107] to arrive at the same theoretical condition concerning a set of sufficient perturbations, though it is beyond the scope of this work to show this connection. Here, we use an independent and simpler proof based on normal theory and matrix algebra. Our theory also provides a generalization of the work of Chaibub Neto et al. [20], which shows that sets of unique (or "driving") QTL for each phenotype can be used to uniquely orient edges in a directed cyclic network. Our approach allows us to represent the problem of directed network inference as a model selection problem within a regression equation for each phenotype. This allows us to avoid the reliance on computationally inefficient heuristics [50, 84, 20], which can generate many possibly poor-fitting networks depending on how the algorithm is run, when considering sample sizes that are typical of experiments collecting genome-wide gene expression data.

The "Recovery" Theorem demonstrates how the set of equivalent DCGs can be recovered from the precision matrix between expression phenotypes and loci (the matrix $\mathbf{\Sigma_{yx}}$). This last result is incorporated into our algorithm for inferring sparse network structure with a sufficient perturbation (eQTL) set. Note that while the algorithm depends on sparsity for efficient network recovery, the results of these theorems are general and do not require such a constraint. In addition, we note in a further Lemma that even in the case of directed cycles, if we know which phenotype a perturbation feeds into, we can further reduce the size of the equivalence class to a unique directed cyclic graph.

Theorem 1:

Given two distribution equivalent directed cyclic graphs, with equivalent parametrizations $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$, any matrix $\mathbf{A}$ which satisfies $\boldsymbol{\Gamma}_1\mathbf{A} = \boldsymbol{\Gamma}_2$, must be orthonormal (i.e. $\mathbf{A}\mathbf{A}^{\mathbf{T}} = \mathbf{I}$).

Proof of Theorem 1:

Since $\boldsymbol{\Gamma}_1\mathbf{A}\mathbf{A}^{\mathbf{T}}\boldsymbol{\Gamma}_1^{\mathbf{T}} = \boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2^{\mathbf{T}}$, and from the definition of equivalence, if $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ are equivalent, then $\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1^{\mathbf{T}} = \boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2^{\mathbf{T}}$. Therefore, $\boldsymbol{\Gamma}_1\mathbf{A}\mathbf{A}^{\mathbf{T}}\boldsymbol{\Gamma}_1^{\mathbf{T}} = \boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1^{\mathbf{T}}$. Left multiply by $\boldsymbol{\Gamma}_1^{\mathbf{T}}$ and right multiply by $\boldsymbol{\Gamma}_1$, then $\mathbf{C}\mathbf{A}\mathbf{A}^{\mathbf{T}}\mathbf{C} = \mathbf{C}\mathbf{C}$, where $\mathbf{C} = \boldsymbol{\Gamma}_1^{\mathbf{T}}\boldsymbol{\Gamma}_1$ is a positive definite invertible matrix of rank p. Left and right multiply by $\mathbf{C}^{-1}$, and $\mathbf{A}\mathbf{A}^{\mathbf{T}} = \mathbf{I}$.

The matrix $\mathbf{A}$ can be thought of as a linear operator that allows transformations between models which produce the same covariance (and inverse covariance) structure (even between models which are not faithful). We use this operator to prove the following theorem after rescaling the network and perturbation parameters as in Equation 3.7: $\boldsymbol{\Lambda}_{\mathbf{i}} = \boldsymbol{\Lambda}_{\mathbf{i}}\mathbf{R}_{\mathbf{i}}^{-\frac{1}{2}}$, $\mathbf{B}_{\mathbf{i}} = \mathbf{B}_{\mathbf{i}}\mathbf{R}_{\mathbf{i}}^{-\frac{1}{2}}$:

Theorem 2:

If there exists an ordered set $S = \left\{s_1, s_2, \ldots, s_p\right\}$ of rows of the perturbation graph parametrized by $\mathbf{B_1}$ such that $\mathbf{L_1} = \mathbf{B_1^{(S)}P_1}$, where $\mathbf{L_1}$ is a diagonal matrix of rank p and $\mathbf{P_1}$ is a signed permutation matrix, then 1) if $\mathbf{\Lambda_1}$ parametrizes a DAG, then for any parametrization $\mathbf{\Lambda_1}$ of any DAG, there does not exist an alternative equivalent DAG or DCG, and 2) if $\mathbf{\Lambda_1}$ parametrizes a DCG, then for any parametrization of any DCG, there exists a finite set of equivalent DCGs, where each equivalent DCG contains a reversed directed cycle with reference to the original DCG.

Proof of Theorem 2:

Given $\mathbf{L_1}$ exists, assume there exists an alternative equivalent model parametrized by $\mathbf{B_2}$ and $\mathbf{\Lambda_2}$. Then, by Theorem 1, there exists an orthonormal matrix $\mathbf{A}$ where $\mathbf{\Lambda_1 A} = \mathbf{\Lambda_2}$, $\mathbf{B_1 A} = \mathbf{B_2}$, and $\mathbf{L_1 A} = \mathbf{L_2}$. Because $\mathbf{L_1}$ and $\mathbf{L_2}$ are invertible, we have: $\mathbf{A} = \mathbf{L_1^{-1} L_2}$. This implies that $\mathbf{L_1 L_1^T} = \mathbf{L_2 L_2^T}$. Since $\mathbf{L_1}$ is diagonal for any parametrization $\mathbf{B_1}$, $\mathbf{L_1 L_1^T}$ and $\mathbf{L_2 L_2^T}$ must also be diagonal for all equivalent parametrizations $\mathbf{L_1}, \mathbf{L_2}$. If there does not exist a signed permutation matrix $\mathbf{P_2}$ such that $\mathbf{F} = \mathbf{L_2 P_2}$, with $\mathbf{F}$ diagonal, then there always exists a parametrization of $\mathbf{L_2}$ where $\mathbf{L_2 L_2^T}$ is not diagonal, and therefore not equivalent (since all non-zero elements of $\mathbf{L_2}$ are free to vary). Therefore $\mathbf{A} = \mathbf{P_2^T}$ is either an identity matrix or a signed permutation matrix. Now consider $\mathbf{\Lambda_1 A} = \mathbf{\Lambda_2}$. Because in this parametrization, $\mathrm{diag}\left(\mathbf{\Lambda}\right) = \mathrm{diag}\left(\mathbf{R}^{\frac{1}{2}}\right)$, the only allowable equivalent model transformations must have positive non-zero elements along the entire diagonal. Therefore, if $\mathbf{\Lambda}$ parametrizes a DAG, then $\mathbf{A} = \mathbf{I}$, and if $\mathbf{\Lambda}$ parametrizes a DCG, then $\mathbf{A} = \mathbf{P}$ where $\mathbf{P}$ is any signed permutation matrix which ensures non-zero positive elements along the diagonal of $\mathbf{\Lambda}$. This corresponds directly to reversing the order of any set of directed cycles in the graph.

This theorem allows us to understand constraints on possible equivalent models in the specific case when each node has at least one unique perturbation. In the next theorem, we focus on the structure of the moralized graph (i.e. the precision matrix $\mathbf{\Sigma}$) for these models, and see how it maps back to the set of possible unmoralized directed graphs that generated the moralized graph. We define the set of parents of a particular node, y_i, from the directed graph as $pa(y_i)$, and the set of all nodes in an undirected graph $\mathbf{\Sigma}$ that have edges to node z as $adj(\mathbf{\Sigma}, z)$.

"Recovery" Theorem:

If in $\mathbf{\Sigma}$ there exists an independent perturbation vertex set $x = (x_1, \ldots, x_q)$ and a response vertex set $y = (y_1, \ldots, y_q)$ where $\forall i, |adj(\mathbf{\Sigma_{yx}}, y_i)| \geq 1$ and $\exists x_j \in pa(y_i)$, then the only equivalent directed cyclic graphs among y that could have generated $\mathbf{\Sigma}$ contain permutations of cycles, and can be recovered from $\mathbf{\Sigma_{yx}}$.

Proof of the "Recovery" Theorem:

The existence of an independent perturbation vertex set and response vertex set that satisfies these conditions corresponds directly to a perturbation topology and parametrization specified by $\mathbf{L_1}$ from Theorem 2. Given this observation, Theorem 2 ensures the constraint on possible equivalent models. Finally, the reason the structure can be recovered from $\mathbf{\Sigma_{yx}}$ is apparent from Equation 3.5 and 3.7, where $\mathbf{\Sigma_{yx}} = \mathbf{\Lambda B^T}$, and therefore $\mathbf{\Sigma_{yx}^{L_1}} = \mathbf{\Lambda L_1^T}$ Since $\mathbf{L_1^T}$ is diagonal it won't change which elements of $\mathbf{\Sigma_{yx}^{L_1}}$ are zero or non-zero.

In the case of DAGs, a generalization of this theorem is trivial to prove for graphs defined over arbitrary probability measures, since the process of moralization of a graph connects all the parents of a given node. Since in this specific perturbation case, each node has at least one unique parent (from the perturbations), then a connection will be induced between the unique perturbation parent and each of its co-parents, indicating exactly what the unique set of parents are for that given node.

Alternatively, as we saw in Theorem 2, the assumptions of normality and linearity are key to showing that even for directed cyclic graphs that have unique perturbations, there still exists multiple equivalent models. In the "Recovery" Theorem we see that we can still determine these 'minimal' equivalence classes from the moralized graph. It is interesting to observe that the perturbation topology can completely change among equivalent directed cyclic graphs, whereas it cannot for directed acyclic graphs. If one knows which node each perturbation feeds into, then the following is true:

Lemma:

If the underlying perturbation topology, $\mathbf{B_1}$, is known, then the cardinality of all directed cyclic equivalence classes is reduced to one.

This further reduction of the equivalence relationships is apparent when one considers that each equivalent perturbation topology specifies exactly one member of the equivalence class (from the "Recovery" Theorem). Therefore, if one knows the true perturbation topology, then one knows the true regulatory model. This allows us to infer a unique directed cyclic graph in the case where we know which phenotype each genetic perturbation feeds into. Hence, the reason behind making our major biological assumption: to only consider the genetic effects of *cis*-eQTL and assume that the *cis*-eQTL feeds directly and uniquely (i.e. non-pleiotropically) into the local gene. With *trans*-eQTL, unless there is prior knowledge about exactly which gene each *trans*-eQTL affects (i.e. about the pathways in question), there is no way to reduce this equivalence class to a unique directed cyclic graph.

3.6 Algorithms

3.6.1 Adaptive lasso

For Step 1 of the algorithm, we perform an individual marker analysis of each genetic polymorphism in a window around the start site of the gene, and only include the markers that are significant given a Bonferroni correction for multiple testing. We then filter these sets of *cis*-eQTL such that they are effectively independent given the linkage disequilibrium structure of the data. For the analysis of the yeast data, we found that a maximum pairwise $r^2 <= 0.03$ between *cis*-eQTL genotypes was a very conservative threshold given a resampling test of random markers across the genome (results not shown).

For Step 2 of the algorithm, the lasso problems from Equation 3.2 and 3.3 are solved using the cyclic coordinate descent method of Friedman et al. [49], as implemented in the 'glmnet' package, called by the 'parcor' package [79]. While this method is an approximation to solving the adaptive lasso for the log-likelihood defined in Equation 3.4, there are theoretical connections between an exact solution to the problem, and this approximate solution which suggest that in some cases the approximation will not perform much worse than the exact solution (i.e. highly penalized cases) [48].

For Step 3 of the algorithm, we performed a permutation test to very conservatively ensure that the induced edge found between an upstream gene, and the *cis*-eQTL, did not arise from a *trans*-effect of the *cis*-eQTL. To do this we randomly resampled the genotype data 10,000 times for each induced edge, and determined the proportion of the time the absolute value of the marginal correlation between upstream gene and *cis*-eQTL under the empirical null model was greater than the absolute value of the observed marginal correlation. We only treated induced edges as representing a directed relationship between a pair of phenotypes if the probability of observing a greater value under the empirical null model was greater than 0.90.

3.6.2 PC-algorithm

While this is only designed to reconstruct directed acyclic graphs, it has been used in a combined gene expression and genotype context to reconstruct directed cyclic graphs [20]. The PC-algorithm reconstructs the skeleton (i.e. set of edges regardless of edge orientation) of a partially directed acyclic graph (PDAG) by performing forward tests of conditional independence. It first starts by constructing a correlation graph (i.e. a conditional independence graph where one conditions on the empty set), then in a forward step-wise manner, removing edges in the neighborhood of each node by increasing the size of the conditioning set based on the neighborhood of each node. Once the cardinality of the conditioning set is equal to or larger than the neighborhood for all nodes, the algorithm terminates. While this is being done, all identified v-structures (co-parents of a common child) are being tabulated, so that afterwards these edges can be oriented. Then, there is a set of rules, based on the seed v-structures which orient a small initial set of edges, which orient many additional edges in the network, by propagating the implications of the few initial oriented edges, with respect to the d-separation criterion defined for directed acyclic graphs [120, 73].

We applied the PC-algorithm by giving it the entire set of gene expression products with *cis*-eQTL as well as all of the *cis*-eQTL genotypes as well. There is one tuning parameter, α, for the implementation 'PCalg', which represents the level of significance each test of conditional independence has to pass to correspond to removing an edge from the skeleton of the network. We used a conservative value of $\alpha = 0.001$, based on simulation results presented in Kalisch et al. [73]. For directed acyclic graphs, the PC-algorithm will also use the *cis*-eQTL to orient each of the edges in the network correctly and uniquely. For directed cyclic graphs, the PC-algorithm will try to orient the edges to form a directed acyclic graph, but often will fail, and draw a random DAG instead. We also apply the PC-algorithm to directed cyclic network recovery by having it identify both the skeleton with perturbations, and then have it attempt to orient as many edges as possible, given that every regulatory relationship should be orientable with the PC-algorithm when there are sufficient, unique perturbations. While in some cases this will fail, especially as the sample size grows and it becomes more sensitive to variations away from the assumption of no cycles, in practice it is able to orient many edges correctly in a directed cyclic graph.

3.6.3 QDG algorithm

The default settings were used for the QDG algorithm, as provided by the authors [20]: $\alpha = 0.005$ for the PC-algorithm skeleton reconstruction step, the skeleton reconstruction method based on the PC-algorithm, and the number of random restarts of iterative testing of different global edge orientations was set to ten. The QDG algorithm uses either the PC-algorithm or UDG algorithm

[116] to generate a skeleton among phenotypes [20]. Then, the QDG algorithm orients edges between phenotypes based on a LOD score computed by leveraging each phenotype's known QTL. To find a globally optimal orientation of edges, an iterative search over orientations is performed to find all possibly cyclic networks which fit the data well [20]. We tried both methods in the QDG algorithm to generate the skeleton, and did not see a significant difference in performance for our simulations (results not shown).

3.6.4 QTLnet algorithm

The default settings were used for the QTLnet algorithm, as provided by the authors [21]: we ran it for 20,000 iterations, sampling every 20^{th} iteration after a burn-in of 2,000 iterations. The QTLnet algorithm uses a fully Bayesian Markov chain Monte Carlo approach to solve the problem of joint phenotype genotype network inference, constraining the proposed graph transitions to directed acyclic graphs [21]. In our analyses, we use the Bayesian model averaged output of the QTLnet algorithm, and include an edge only if its posterior probability of inclusion is greater than 0.50.

3.6.5 NEO algorithm

We used the default settings for the NEO algorithm, based on the code available from the author's website: http://www.genetics.ucla.edu/labs/horvath/aten /NEO/ [4]. The NEO algorithm uses multiple QTL to orient edges between an arbitrary pair of phenotypes based on different structural equation model based statistics [4], but has no mechanism to remove edges among phenotypes by conditioning on other phenotypes, and will therefore often have high false-discovery rate for recovery of the network generating the data among phenotypes. This was another justification, aside from the scaling of the algorithm, for why we did not include it in our broader comparison of alternative methods.

CHAPTER 4

AGGRESSIVE FALSE POSITIVE CONTROL FOR GENOME-WIDE FEATURE SELECTION IN A MOUSE F2 CROSS

4.1 Abstract

Complex disease risks can arise from a variety of genetic, molecular, and environmental factors. We propose a novel machine learning algorithm to aggressively control false positive rates for sparse models of the simultaneous effect of genetic variants and molecular phenotypes on downstream phenotypes, as well as sparse models within molecular phenotypes (i.e. network models). While other sparse feature selection procedures have been proposed in the fields of machine learning and statistics, such as the lasso, our algorithm is the first that is explicitly designed to be both highly scalable, as well as contain a natural metric for stringent control of false positives, all within a single statistical model. We use our algorithm to characterize the effect of genetic markers and liver expression traits on weight, cholesterol, glucose, and free fatty acid levels in an F2 mouse intercross. A sparse simultaneous model among >24,000 phenotypes and genotypes was generated in less than 72 hours on a single workstation where we identify a set of genes, Zfp69, Crhr1, Qpctl, Vcam-1, Cnr2, Gabarabl1, Gch1, and a quantitative trait loci (QTL) on chromosome 7 near the Atp10a gene, all directly linked to weight, and all previously associated with obesity. Strikingly, these specific interactions with previous known associations to obesity were not identified as significant in the initial and subsequent analyses of this specific data set. In addition we identify a set of nine genes previously associated with obesity-related phenotypes, including insulin resistance,

hypertension, and diabetes, as well as 78 novel genes and QTL affecting weight, cholesterol, glucose, and free fatty acid levels.

4.2 Introduction

Recent studies have shown that complex disease risk can be mediated by many interacting molecular pathways, and these pathways can be identified through statistical methodologies and algorithms [23, 39, 110, 112, 66, 161, 27, 91]. With the advent of high throughput sequencing and molecular phenotype characterization technologies it is now possible to collect genome-wide profiles of both sequence variation, and gene activity variation in a sample of individuals [108]. Given that all sources of systematic error are controlled, these data present the unique possibility to identify novel regulatory mechanisms and pathways underlying disease [110, 23, 39, 112]. For example, Yang et al. [146] validated three novel genes involved in obesity and obesity related traits in an F_2 mouse cross, based on predictions made from models generated on combined genome-wide molecular phenotype and genotype variation. Yet, many of the models used to generate these predictions are inaccessible to the practical user, or may suffer from poor performance for realistic data (i.e. the strongest scoring interactions may still be heavily enriched for false positives), which can make the process of validation of novel interactions prohibitively expensive [110].

Previous statistical model generation approaches have focused on different levels in the hierarchy of variation when considering combined gene expression and genotype data, from broad patterns (i.e. ensemble behavior of groups of genes), to specific conditional relationships (i.e. network models of interactions among genes). An example of an ensemble method that has been proposed is gene expression module identification using clustering techniques [154]. This type of method has been used to identify specific modules associated with quantitative trait loci (mQTL), as well as the effect of those modules on disease risk [161, 23, 58, 51]. Additionally, authors have proposed methods based on the inference of the topology of a network, which fall into two broad classes, undirected networks [95, 97, 159, 114] and directed networks [50, 101, 73, 160, 111]. These models are able to capture complex and conditional relationships among genes but will be more sensitive to sampling variation and the effects of systematic error [88]. In the context of directed networks, one goal has been to leverage QTL, to reduce the space of models with equivalent sampling distributions, to gain possibly causal interpretation for the direction of the edges in the inferred graph [98, 4, 84, 111, 71, 108]. One constraint of all of the previously proposed approaches is that they can suffer from high false positive rates for the identification of any specific edge in a statistical network underlying the observed data [88], when considering genome-wide network reconstruction. Our algorithm augments the previous approaches for generating novel statistical models of genomic variation in two ways. First, we focus on the specific and practical problem of generating simultaneous statistical models (i.e. undirected network models), where the false positive rate can be stringently controlled, all within a single statistical feature selection framework. Second, our approach is highly scalable in that it can be run on a single workstation, with any level of relevant genomic variation included in the model. We propose a novel variational Bayes algorithm for undirected network inference, generalizing the algorithm proposed in Logsdon et al. [87].

Statistically, these data-sets exist in the realm of “large p, small n”, in terms of having large number of features (e.g. genetic polymorphisms, gene expression phenotypes, protein quantification phenotypes), and relatively few samples [140, 78]. The false discovery rate (FDR) can be controlled directly, assuming a uniform distribution of the null p-values [7] in a parametric or semi-parametric model, yet when the null distribution of p-values deviates from uniformity, more sophisticated methods must be employed [125]. In addition, statistical problems in the “large p, small n” realm have generally focused on controlling the FDR of marginal, or individual tests whereas if the underlying model being tested includes multiple features simultaneously, the problem of controlling FDR becomes even more challenging [78, 156, 87]. We seek to propose a method which can aggressively control the FDR in complex, multiple feature statistical models.

To address the “large p, small n” problem, a handful of machine learning algorithms have been proposed with respect to learning the structure of a statistical undirected graph underlying observed molecular phenotype and genotype variation, using a penalized or regularized objective function. With respect to learning purely generative models, the lasso [97], and adaptive lasso [159] penalized regression has been proposed for neighborhood selection. In addition Schäfer and Strimmer have proposed a shrinkage estimator [114], based on a shrunken estimate of inverse covariance matrix (in a Gaussian graphical model this defines the structure of the graph; see Equations 4.1-3), which was expanded by Chu et al. [25], to analyze both expression and genotypic data. Friedman et al. [48], and Fan et al. [43] proposed algorithms to solve the full penalized likelihood version of the lasso and adaptive lasso Gaussian graphical model (GGM) problems. It has has been argued that in the $p \gg n$ realm, solving the full penalized likelihood version of the problem is ill-conditioned because of the positive semidefinite constraint on the structure of the inferred inverse covariance matrix. Additionally, full semidefinite programming solutions to this type of constraint are not computationally practical on the scale of tens of thousands of features[151].

We therefore propose a highly scalable algorithm based on a penalized regression model for feature selection that has significant computational and performance advantages over these other approaches. Our algorithm is related to a mixture penalty, or spike and slab model [56], in a fully Bayesian hierarchical model, similar to Zhang et al. [155]. The form of the penalty function is illustrated in the first panel of Figure 4.1, as compared to the popular lasso penalty, (or a Laplace prior in a Bayesian context) shown in the second panel of Figure 4.1. The spike and slab prior we use is a Bayesian representation of a mixture L_0, L_2^2 norm, with an additional probabilistic interpretation. This approach has certain theoretical advantages over the lasso [155, 156, 147], and is equally as scalable. To demonstrate this property we analyze genome wide genetic variation and expression variation, and their effects on downstream phenotypes related to weight in a mouse intercross [58].

In our data analysis we demonstrate that our approach can generate better performance than the lasso and the adaptive lasso in terms of identifying strictly statistically relevant features for a simultaneous feature model. (We also validate this property on simulated data, as shown in Appendix B). We can outperform the lasso because it can suffer from poor model selection performance

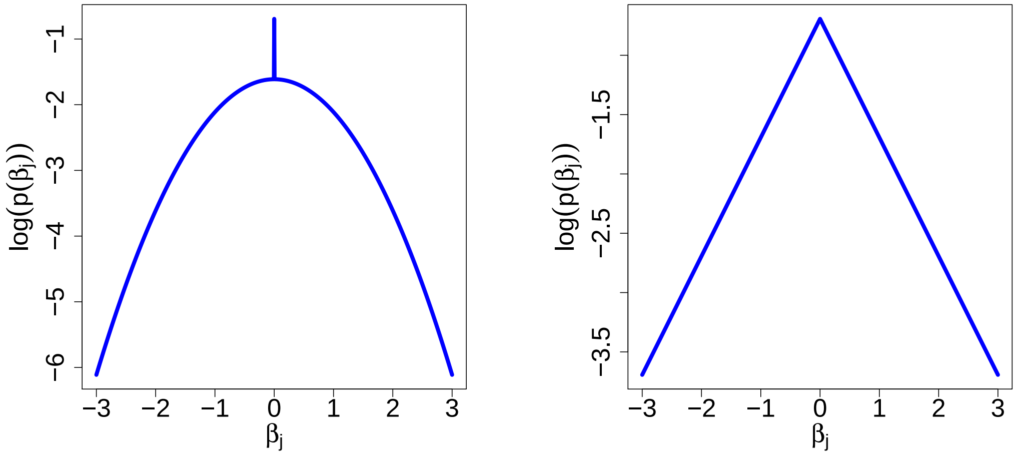

Figure 4.1: The spike and slab prior ($L_0 + L_2^2$ penalty) and the log Laplace prior (L_1/lasso penalty) used in the regularized regression procedures for model selection.

with low false-discovery rate, especially for models with many correlated features. These drawbacks include model selection inconsistency (when the irrepresentability condition is met [158]), where even asymptotically the correct model will not be selected, because of pathological correlations between features in the true model, and features that are not in the true model. While other penalties have been proposed to remedy this problem including the adaptive lasso [158], the smooth clipped absolute deviation penalty (SCAD) [44], most of those penalties still have significant drawbacks in terms of requiring asymptotics to take advantage of properties like model selection consistency, as well as the problem of the choice of model complexity penalty parameter, which must be done through a heuristic process like minimizing some metric of cross validation error, or through minimizing a model complexity measure such as AIC or BIC [43], for data when $p \gg n$.

4.3 Results

4.3.1 Mouse Network Analysis

We analyzed the F2 progeny of a cross between the C57BL/6J (B6) and C3H/HeJ (C3H) strains on an apolipoprotein E null (ApoE – / –) background (BXH.ApoE$^{-/-}$), as presented in Ghazalpour et al. and Wang et al. [58, 138]. This cross was generated to investigate metabolic syndrome associated phenotypes [58, 138]. We focused on the gene expression data that was collected in the liver of the mice where expression was assayed on 23,574 custom probes [58]. In addition there were 22 downstream phenotypes that were assayed, including weight, cholesterol, glucose, free fatty acid, among other metabolic phenotypes, as well as 1,347 genetic markers [58]. After filtering down to a common set of individuals with both expression and markers collected we were left with 298 individuals. Previous authors have shown with this data that there is antagonistic sex effects [138], i.e. the effect of a risk locus is opposite between males and females. To address the sex specific effects, as well as other possibly confounding factors, we included both the sex, as well as the 20 first eigenvectors computed across samples for expression phenotypes as fixed effects in our linear model, as depicted in Equation 4.6.

We ran our analysis in two steps. First, we ran the variational algorithm on each of the 22 obesity related downstream phenotypes individually, where we performed sparse feature selection on all genes products, genetic markers, given the 20 top principal components, and sex, with 1,000 random restarts of the algorithm that let us identify many possible models, and their associated evidence (see Appendix B). The variational algorithm produced a sparse set of expression and genetic markers for each downstream phenotype, with the phenotypes with more than seven expression or genotype features identified shown in Table 4.1 (with a cutoff of $\hat{p}_j > 0.99$). In addition, we ran the lasso and the adaptive lasso with five-fold cross validation for the same set of downstream phenotypes, as shown in Table 4.1. While the size of the identified neighborhoods of each downstream phenotype was on average much larger for the lasso and the adaptive lasso, the variational algorithm identifies additional features, with only 55% overlap with the lasso, and 30% overlap with the adaptive lasso for these seven phenotypes shown in Table 4.1. In addition, when the features were extracted for each method, and analyzed in an independent, non-penalized linear multiple regression model, (for all phenotype except weight), both the lasso and the adaptive lasso contained many features that were not statistically significant at the $P < 0.05$ significance level. Alternatively, for the variational method, the features identified were all statistically significant. This strengthens the results of the simulation study, where at the $\hat{p}_j > .99$ cutoff, all the returned features make significant statistical contributions to the model, and are therefore more likely to be biologically relevant.

In the second step, we generated an expression-eQTL undirected network, by solving the neighborhood selection problem defined in Equation 4.6 for each gene expression product individually, against all other genes and genetic markers, given the top 20 principal components, and the sex, with 50 restarts. We resolved the neighborhoods of the expression-eQTL network very conservatively, by averaging the $\hat{p}_j$ scores in both directions of regression for the expression phenotypes, and only declaring an interaction between genes present

Table 4.1: Model size, intersections, and proportion of significant associations based on an independently fit linear model, between the variational method, the lasso, and the adaptive lasso, where 'Vari' indicates the variational method, Adalasso indicates the adaptive lasso, Vari ∩ Lasso indicates the intersection of the features returned by the variational method, and the lasso, and % P < 0.05 indicates the percentage of features with P-values less than 0.05 in an independently fit linear simultaneous statistical model.

Pheno	Vari	Lasso	Adalasso	Vari ∩ Lasso	Vari ∩ Adalasso	%P < 0.05 Vari	%P < 0.05 Lasso	%P < 0.05 Adalasso
Weight	13	107	14	6	3	100%	29%	100%
Total Chol	12	176	55	7	4	100%	17%	69%
HDL	9	186	24	7	5	100%	16%	88%
UC	13	172	39	8	4	100%	20%	77%
FFA	9	194	23	5	2	100%	10%	96%
Glucose	8	196	100	3	3	100%	22%	65%
LDL+VLDL	12	187	127	6	2	100%	13%	33%

in the model if the averaged $\hat{p}_j$ scores were greater than 0.99. To determine the most relevant aspects of this sparse network with respect to weight and other related phenotypes, we combined the neighborhoods produced for each of the downstream phenotypes, and the expression-eQTL undirected network, to depict the local sub-networks associated with each downstream phenotype, as shown in Figure 4.2 for weight, and Figure 4.3 for total cholesterol, high density lipoprotein (HDL) cholesterol, unesterified cholesterol (UC), free fatty acids (FFA), glucose levels, and low density lipoprotein + very low density lipoprotein (LDL+VLDL) levels. Table 4.2 summarizes identified genes which have been previously implicated in obesity, or related diseases and pathologies.

Most strikingly, in Figure 4.2 we see that the expression of the zinc-fingered protein 69 (Zfp69) is directly linked to weight, in this conditional regression

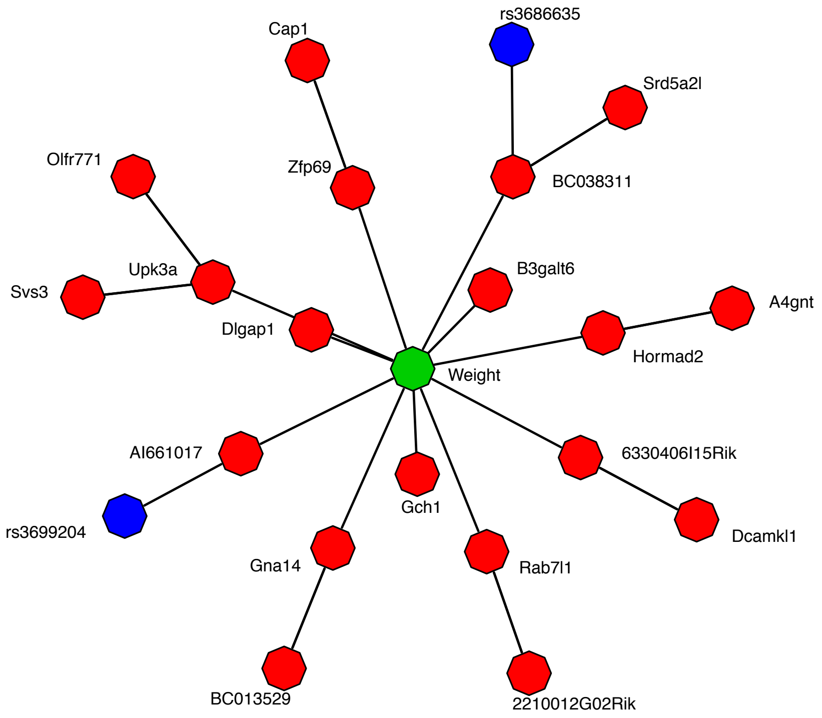

Figure 4.2: A sparse, reconstructed sub-network with the variational algorithm for all expression products and genetic markers associated with weight in a BXH.ApoE$^{-/-}$ F2 mouse intercross. The blue nodes represent genetic markers, the red nodes expression traits, and the green node is the weight phenotype.

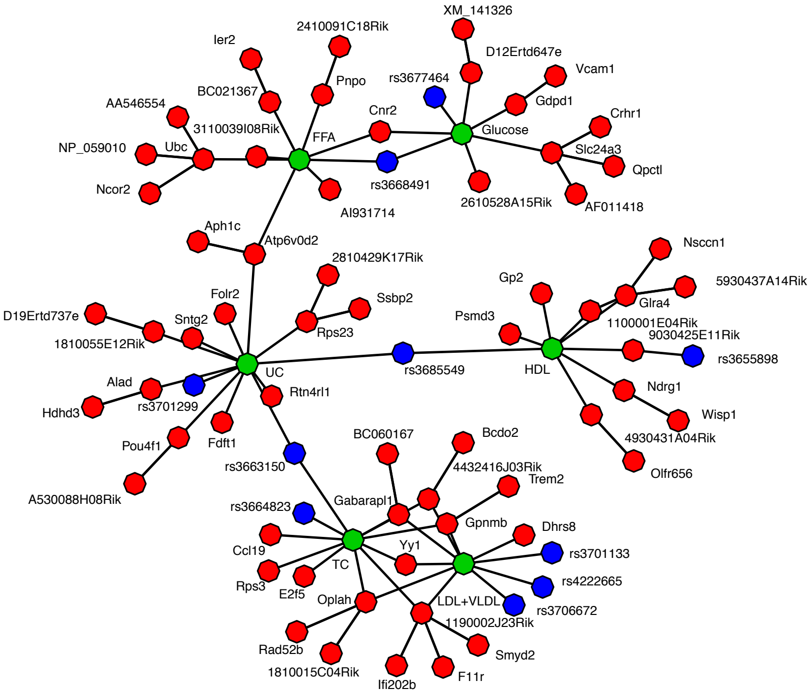

Figure 4.3: A sparse, reconstructed sub-network with the variational algorithm for total cholesterol (TC), high density lipoprotein cholesterol (HDL), unesterified cholesterol (UC), low-density lipoprotein cholesterol (LDL+VLDL), free fatty acids (FFA), and glucose levels. As with Figure 4.2, the blue nodes represent genetic markers, the red nodes expression products, and the green nodes are downstream metabolic phenotypes.

Table 4.2: Interactions identified by the Variational method with previous evidence as being associated with obesity, or obesity related traits.

Gene/SNP	Disease	Organism(s)	Reference
Zfp69	Candidate gene for diabetes associated with obesity	Mouse and Human	[115]
Gna14	Association study of hypertension	Human	[77]
F11r	Induces hypertension in the brain stem	Rat	[137]
Gabarabl1	Regulator of insulin dependent hepatic autophagy	Mouse	[85]
Wisp1	Association study of hypertension	Human	[145]
Fdft1	Squalene (cholesterol) biosynthesis gene	Mouse and Human	[75, 102]
Ier2	Induced gene in insulin signaling pathways	Rat	[74]
Slc24a3	Down regulated in diet sensitive obesity	Human	[57]
Crhr1	Candidate obesity gene possibly affecting feeding behavior	Mouse and Human	[26, 103]
Qpctl	Association study identified candidate obesity gene	Human	[118]
Vcam-1	Atherosclerotic plaque associated gene	Human	[30, 99]
Gch1	Identified in linkage studies of maximal sedentary oxygen uptake	Human	[12]
Dlgap1	Type-2 diabetes associated gene	Human	[5]
Yy1	Type-1 diabetes associated gene	Rat	[76]
Ccl9	Adipocyte inflammation	Human	[153]
Cnr2	Obesity associated adipocyte inflammation	Mouse	[33]
Atp10a/rs3664823	Obesity associated gene	Mouse	[34]

neighborhood selection model. This gene has previously been identified as a candidate gene, for the diabetogenic effect of the Nidd/SJL loci in obese mice [115]. In addition, the expression of the genes GTP cyclohydrolase 1 (Gch1), discs, large (Drosophila) homolog-associated protein 1 (Dlgap1), and guanine nucleotide binding protein, alpha 14 (Gna14), are all also directly linked to weight, where Gch1 was previously identified in a linkage scan for maximal sedentary oxygen uptake [12], Dlgap1 has been identified as possibly associated to Type-2 diabetes in humans [5], and Gna14 has been identified as associated with hypertension [77], which are all diseases with related etiologies with obesity. Furthermore, we identify two SNPs in the sub-network directly connected to the expression products associated with mouse weight. First, rs3699204 is a *cis*-eQTL linked to Al661017 (Tmem71), a transmembrane protein on chromosome 15, and second, rs3686635 is a *cis*-eQTL for BC038311 (Cox18), a cytochrome c oxidase assembly homolog.

In Figure 4.3, we see that the gene F11r, also known as junctional adhesion molecule-1 (JAM-1) is related to both the total cholesterol levels, as well as the combined LDL and VLDL cholesterol levels, through the 1190002J23Ri probe i.e. kelch domain containing 9 (Klhdc9) gene. The F11r gene has been previously identified in rats as having a role in hypertension, where over-expression of the gene significantly increased blood pressure [137]. The gene gamma-aminobutyric acid (GABA) A receptor-associated protein-like 1 (Gabarabl1) was also identified as being directly connected to total cholesterol and LDL+VLDL cholesterol levels, and is a known regulator of autophagy [85]. This gene was previously identified as being down-regulated in the liver of mice that were induced to be insulin resistant [85]. The gene Yin yang 1 (Yy1) has been previously associated with type-1 diabetes in rats [76] and the gene WNT1 induced signaling pathway protein 1 (Wisp1) was connected with HDL levels through the N-myc downstream regulated gene 1 (Ndrg1) gene. This gene was recently identified as being associated with hypertension in a Japanese population from a longitudinal analysis [145].

The gene farnesyl diphosphate farnesyl transferase 1 (Fdft1) is a known squalene (cholesterol) synthesis gene, and high levels of this gene are known to be associated with visceral obesity [102], and is known to be up-regulated in mice on a high fat diet [75] and is directly linked to the unesterified cholesterol levels. The gene immediate early response 2 (Ier2) also known as Pip92, is known to be induced by insulin signaling [74], and is linked through BC021367 (a transmembrane protein also known as Tmem161a) to the levels of free fatty acids. Solute carrier family 24, member 3 (Slc24a3), has been previously identified as having significantly decreased expression in a panel of individuals with diet-sensitive obese women, and is directly linked to glucose levels in our network [57]. Furthermore, both the genes corticotropin releasing hormone receptor 1 (Crhr1) and glutaminyl-peptide cyclotransferase-like (Qpctl) are directly linked to Slc24a3, and have both been previously implicated as candidate obesity genes [26, 103]. Additionally, the gene vascular cell adhesion molecule 1 (Vcam1), has been previously shown to be at expressed in human atherosclerosis lesions [30, 99], and is linked to the levels of glucose through the expression of the gene glycerophosphodiester phosphodiesterase domain containing 1 (Gdpd1) in our network.

We also see that the gene chemokine ligand 19 (Ccl19) is directly linked to total cholesterol levels, where this gene has been shown to be up-regulated under induced endoplasmic reticulum (ER) stress in adipocyte tissue [153], where obesity is known to put ER stress on adipocyte tissue [153]. We also identify cannabinoid receptor 2 (Cnr2) as being directly connected to both the levels of free fatty acids and glucose, where Cnr2 has been shown to be directly mediate an innate immune response leading to inflammation in obese mice adipocytes [33]. Finally, we also identified a possible genetic variant that has been previously linked to increased obesity, with the SNP rs3664823 on chromosome 7, interacting with the total levels of cholesterol [34].

4.4 Discussion

Identifying multiple, sparse, strongly supported statistical interactions among different levels of a biological system is a very challenging and relevant problem, especially when the number of features greatly exceeds the sample size, as is common in genome-wide assays of joint molecular genetic, phenotypic, and downstream phenotype variation [23, 39, 110, 112, 66, 161, 27, 91]. This goal is especially relevant when attempting to characterize the specific drivers of a disease, where the underlying disease aetiology can be complex and conditional [110]. Having a statistical model which can account for a rich set of interactions, as with a conditional Gaussian graphical model, is necessary to both leverage all the available information in the data, and generate a nuanced understanding of how previously uncharacterized features fit into the context of a specific biological system. For example, we were able to generate the prediction that a *cis*-eQTL near rs3686646 effects Cytochrome c assembly, which in turn may have an impact on weight. To this end we have identified a rich set of genes whose expression is strongly statistically linked to both weight (Figure 4.2), as well as weight associated phenotypes (Figure 4.3). We have demonstrated that our novel methodology can outperform other commonly used penalized regression approaches in the F2 mouse intercross (Table 4.1) in terms of only returning interactions that have very strong statistical support, given the available information in the data. We also demonstrate this for simulated data, as shown in Appendix B, with figures B.1, and B.2.

Some authors have suggested that even though not all the features identified by the lasso or the adaptive lasso are statistically relevant (Table 4.1), one could implement a two stage procedure where after running either method, one performs statistical tests on the identified features (where the penalty parameter has been chosen based on cross-validation or an information criterion) [142]. Yet, if the initial number of features returned by the model is large, this could lead to ill-behaved test-statistics (e.g. if the effective degrees of freedom in an unpenalized model is low); additionally in the case of the analysis we present in Table 4.1, the variational method consistently identifies a sparse subset where all the features are significant, and not a strict subset of any of the features identified by the lasso or adaptive lasso. For example, neither the lasso nor adaptive lasso (which by definition returns a strict subset of the features identified by the lasso) identify the gene Zpf69, a known obesity candidate gene, as being directly connected to weight, whereas the variational method does. Furthermore, the regression model is fit in a single procedure with the variational method, where the model dimension is adaptively determined by an inference procedure in the hierarchical Bayesian model based on the evidence in the data, and the probabilistic consistency bound on model size.

We focus on the reconstruction of sparse undirected graphs, because they are more amenable to highly scalable, sparse feature selection methods based on penalized probabilistic models [79]. Directed regulatory networks represent a richer class of statistical models (e.g. most parametric representations of Bayesian networks are in the curved exponential family [53]), and are therefore not only more challenging to learn (NP-hard, [24]), but the sampling variability of directed inferences will also be much higher, because the possible set of models is much richer. In the realm of $p \gg n$ this is a significant challenge, since the sampling error may dominate most of the observed variation and covariation. In addition, when the goal is specifically novel interaction discovery, with low false-discovery rates, sacrificing inference of the direction of the interaction, while gaining the knowledge of a novel contributing factor is a worthwhile compromise.

We propose a mean-field, or variational Bayes, type approximation to the full posterior inference problem, where data analyses and simulations indicate that this approximation is appropriate for the highly sparse model selection problem. The deterministic algorithm defined through the variational Bayes approximation allows us to scale our approach to problems beyond the scope of current exact inference approaches (e.g. Markov chain Monte Carlo (MCMC) algorithms [56, 155, 156]). Also, because the best-subset selection problem can be unstable in certain circumstances [15], there will be many local solutions. This translates into a highly multi-modal posterior surface. Applying the ridge like penalty to the non-zero coefficients with the 'slab' component of the mixture can regularize or stabilize the problem. Unfortunately, there is still a fundamental computationally intense problem of finding many possible models or modes, and determining the relative evidence of each model or mode. Because the algorithm is very fast, we can run it many times (up to thousands) and identify many models, along with the relative evidence of each model identified, based on the lower bound (Equation B.14), and integrate the evidence across the models through approximate Bayesian model averaging, as in equation B.16. Frequentist approaches do not inherently have the option of Bayesian model averaging, and therefore in the case of model selection when there are many competing models with similar evidence, we believe that the fully Bayesian approach is the most effective at integrating this information, and generating the best estimates of which interactions are most strongly supported by the data. Hence even the recently proposed Minorization-Maximization (MM) algorithm of [147] which approximates a similar spike and slab prior as depicted in this paper for a likelihood version of the regression problem, could still suffer from identifying many unstable solutions (modes of the likelihood surface) with no natural way to regularize across modes or models.

This mixture penalty in a Bayesian framework has attractive theoretical properties, including bounded shrinkage and indications that it may approach optimal efficiency for sparse underlying parameter spaces [72]. Plus, recent theoretical work suggests that the spike and slab penalty is still model selection consistent when the irrepresentability condition is met [147]. In addition, as previously proposed by Zhang et al. [155, 156], we incorporate a very stringent model complexity control through a probabilistic consistency bound of the total number of features allowed in the model of $O\left(\sqrt{n}/m\right)$. This provides a consistency bound to ensure asymptotically optimal mean-squared error [156]. One of the main advantages of this approach is that the hierarchical model can adaptively shrink the penalty to match the sparsity of the underlying parameter space, without having to resort to prediction based metrics like cross-validation or possibly heuristic model complexity measures based on information criterion such as AIC or BIC.

One of the most important novel contributions of our algorithm over the previously proposed version [87] is an approximate Bayesian model averaging step [104]. Because this penalty is non-convex, the posterior surface can be highly multi-modal, where each mode in the posterior density can represent a different set of identified features (i.e. neighborhood). One well characterized weakness of the L_0 type penalty, also known as best subset selection, is that the solution can be highly unstable to perturbations of the data [14]. By performing Bayesian model averaging across the identified modes, we can additionally regularize the solution, by reweighting the assignments to different posterior probabilities, proportional the the volume underneath the mode (also a measure of the relative evidence of the given model). In the machine learning literature this is also known as bagging [15], and it allows us to additionally regularize our solution based on the set of identified solutions.

In addition we use a novel maximization step to correct for potentially globally confounding factors in the model by including the k first principal components. The possibility of globally confounding variables is well known in the genome-wide association study literature when population structure can broadly bias association test statistics [106]. While in this model we take a naive approach through the incorporation of principal components to correct for global patterns of variation across expression phenotypes, a more nuanced model could be defined that includes explicit hidden factors, or a full mixed model of regression such as been suggested by Listgarten et al. [83]. Correcting for these large scale patterns is essential, since they may arise from confounding effects like sex, batch effect, or other sources of systematic error, which prevents the identification of specific local interactions within and among genetic variants and downstream phenotypes, with low false discovery rates.

4.5 Methods

4.5.1 The network model

For a set of p gene expression traits m genetic markers and fixed effects we define the log-likelihood of the conditional Gaussian Graphical model as follows [48]:

$$\log\left(\mathbf{Y}|\mathbf{X},\mathbf{\Theta}\right) \propto \log\left\{\det\left(\mathbf{\Theta_{yy}}\right)\right\} - \mathrm{Tr}\left(\mathbf{S\Theta}\right), \tag{4.1}$$

where:

$$\mathbf{\Theta} = \begin{bmatrix} \mathbf{\Theta_{yy}} & \mathbf{\Theta_{yx}} \\ \mathbf{\Theta_{yx}^T} & \mathbf{\Theta_{xx}} \end{bmatrix}, \tag{4.2}$$

and

$$\mathbf{S} = \frac{1}{n}\begin{bmatrix} \mathbf{Y^TY} & \mathbf{Y^TX} \\ \mathbf{X^TY} & \mathbf{X^TX} \end{bmatrix}, \tag{4.3}$$

being the sample covariance matrix, $\mathbf{X}$ and $\mathbf{Y}$ mean-centered, $\mathbf{X}$ being an n x m matrix of genotypes and fixed effects, and $\mathbf{Y}$ being an n x p matrix of expression or downstream phenotypes. The matrix $\mathbf{\Theta}$ represents the pairwise Markov dependencies of the random variables $\mathbf{Y}$ [80]. Intuitively, the set of non-zero θ_{ij} parameters for a given random variable y_i, defines the set of other phenotypes once conditioned on, make y_i probabilistically independent from the rest of the variables in the model (also known as the neighborhood of y_i). In this model everything is conditional on the state of the entire set of genotypes and fixed effects. The non-zero structure of the $\mathbf{\Theta_{yy}}$ sub-matrix specifies a conditional Markov random field among the expression phenotypes. Accordingly, the element θ_{yy}^{ij} for $i \neq j$ of the $\mathbf{\Theta_{yy}}$ matrix is zero iff

$$p\left(y_i, y_j|\mathbf{Y}_{-(i,j)}, \mathbf{X}\right) = p\left(y_i|\mathbf{Y}_{-(i,j)}, \mathbf{X}\right) p\left(y_j|\mathbf{Y}_{-(i,j)}, \mathbf{X}\right), \tag{4.4}$$

i.e. the probability distribution satisfies the local Markov property with respect to an undirected graph $\mathcal{G} = (\mathrm{V}, \mathrm{E})$, with $\mathbf{Y}_{-(i,j)}$ indicating the set of other phenotypes, excluding the variables y_i and y_j. Since this is a Markov random field conditioned on $\mathbf{X}$, the non-zero structure of the $\mathbf{\Theta_{yx}}$ sub-matrix does not imply a factorization over an underlying probability density, but the element θ_{yx}^{ij} is zero iff

$$cov\left(x_i, y_j|\mathbf{X_{-i}}, \mathbf{Y_{-j}}\right) = 0, \tag{4.5}$$

i.e. the covariance between x_i and y_j is zero, when conditioning on all other variables. Finally, since this is a conditional Markov random field, the rank of the matrix $\mathbf{\Theta}$ is p and $\mathbf{\Theta_{xx}} = \mathbf{\Theta_{xy}}\mathbf{\Theta_{yy}^{-1}}\mathbf{\Theta_{yx}}$.

To infer the structure of the underlying undirected graph, many authors have proposed putting different forms of element-wise penalties on the $\mathbf{\Theta}$ matrix, such as the lasso (L_1 norm). Additionally, as other authors have noted [151], the positive-semi definite constraint on $\mathbf{\Theta}$ imposed by the $\log\{\det\}$ function in the log likelihood makes optimization of the full likelihood problem challenging for large scale problems, especially when the number of phenotypes and genotypes $p + m$ greatly exceeds the sample size, n. Therefore, instead of solving the full likelihood optimization problem, we follow the general strategy of Meinshausen and Bühllman, Zhou et al., and Kraemer et al. [97, 159, 79], and treat the structure learning problem as a neighborhood identification problem; i.e. we perform model selection on a set of uncoupled regression equations, where each expression phenotype is regressed on every other phenotype, and genotype. At the end of this process we resolve the neighborhoods of each gene expression product by averaging the posterior probabilities of edge inclusion in both directions of regression.

We define a given multiple regression equation as:

$$y_i = \mu + \sum_{j}^{p+m-1} z_{ij}\beta_j + \sum_{l}^{k} t_{ik}\alpha_k + e_i, \tag{4.6}$$

where y_i is i^{th} sample of a given phenotype, z_{ij} is the i^{th} sample of the j^{th} feature, out of the combined phenotype and genotypes, excluding the phenotype y, β_j is the effect of the j^{th} feature, t_{ik} is the i^{th} sample of the k^{th} non-penalized effect, α_k is the effect of this k^{th} feature, and e_i is the residual error term, assumed to be normally distributed with mean zero, and variance σ_e^2. In addition, the population mean is modeled as a fixed effect, μ.

4.5.2 Bayesian hierarchical model for sparse feature selection

Given the regression equation defined in equation 4.6, we define the following hierarchical model, similar in vein to Zhang et al. and Logsdon et al. [155, 156, 87]:

$$\beta_j \sim p_{\beta=0}\mathrm{I}\left[\beta = 0\right] + p_{\beta\neq 0}\mathrm{N}\left(0, \sigma_\beta^2\right), \tag{4.7}$$

$$p_{\beta=0}, p_{\beta\neq 0} \sim \mathrm{Beta}\left(1, 1\right), \tag{4.8}$$

$$\sigma_\beta^{-2} \sim \Gamma\left(2, 1/2\right), \tag{4.9}$$

$$\sigma_e^{-2} \sim \Gamma\left(2, 1/2\right), \tag{4.10}$$

with the additional truncation restriction on the prior distribution over p_β of $p_\beta \leq \sqrt{n}/m$.

4.6 Network recovery algorithms

4.6.1 Variational spike and slab algorithm

The variational Bayes approximation for an arbitrary parameter θ is given as follows:

$$q_{\theta_j}^{t+1}\left(\theta_j\right) = \frac{1}{Z_{\theta_j}} p\left(\theta\right) \exp\left\{\int q_{\theta}^{t}\left(\theta\right) d\theta \log\left\{p\left(\mathbf{y}|\theta, \mathbf{X}\right)\right\}\right\}, \tag{4.11}$$

where a factorization is defined over the joint approximate posterior distribution of parameters:

$$q_{\theta}\left(\theta\right) = \prod_{i} q_{\theta_i}\left(\theta_i\right), \tag{4.12}$$

and the integral in equation 4.11 at iteration t is taken with respect to every approximate distribution except $q_{\theta_j}^{t}\left(\theta_j\right)$. The details of this approximation for each density are presented in Appendix B. A probability of inclusion statistic, $\hat{p}_j$ is computed after the algorithm converges, and this statistic is averaged across all models identified (i.e. modes in the posterior surface), based on the total evidence for each model (i.e. equation B.17). This model averaged probability of inclusion statistic, $\hat{p}_j$ is used to determine whether the j^{th} feature is included in the model, at a given threshold.

4.6.2 Lasso and Adaptive lasso

First proposed by Tibshirani [129], the lasso, and the adaptive lasso [162], were implemented using the R package ‘parcor’ [79]. There are very efficient cyclic coordinate descent algorithms for the penalized regression problem to solve the lasso [49]:

$$\text{argmax}_{\beta}\left\{-\sum_{i=1}^{n}(y_i - z_i\beta)^2 - \eta\sum_{j=1}^{p+m-1}|\beta_j|\right\}, \tag{4.13}$$

and then using the coefficients from this problem to solve the following adaptive lasso problem [162]:

$$\text{argmax}_{\zeta}\left\{-\sum_{i=1}^{n}(y_i - z_i\zeta)^2 - \eta\sum_{j=1}^{p+m-1}\hat{w}_j|\zeta_j|\right\}, \tag{4.14}$$

where $\hat{w} = |\hat{\alpha}|^{-1/2}$, z is the combined gene expression products genetic marker genotypes, and α and ζ are the corresponding regression coefficients. The model tuning parameter, η is determined as in Kraemer et al., with ten-fold cross validation [79]. Additional algorithms used for comparison with analysis of simulated data are shown in Appendix B.

APPENDIX A

A VARIATIONAL BAYES ALGORITHM FOR FAST AND ACCURATE MULTIPLE LOCUS GENOME-WIDE ASSOCIATION ANALYSIS

A.1 Supplementary Results

A.1.1 Data analysis.

We chose a subset of eleven gene expression phenotypes from the Stranger et al. study [123] that contained putative *trans*-associations from a single run of the algorithm (for a single sampling of missing data). We did both completely random reordering of the markers as well as resampling of the missing data. In all cases the putative *trans*-associations were not robust under the random re-orderings and resampling, hence we only reported the *cis*-associations (identified by both V-Bay and our single-marker reanalysis). We report an additional 61 *cis*-associations (along with these 11) that were identified with both V-Bay and single-marker analysis under a single run of V-Bay in Tables A.1, A.2.

A.2 Supplementary Methods

A.2.1 V-Bay algorithm steps.

The algorithm proceeds as follows: 1) Initialize all expected sufficient statistics and expectations for β_j parameters and the expectation of μ to 0. Initialize expectations for $p_{\beta+}$, $p_{\beta-}$ parameters to $\frac{1}{3}$. Initialize expectations of variance param-

Table A.1: HapMap Phase II gene expression reanalysis results.

GENE	SNP ID	Position	Chromosome
FLJ10781	rs12978825	51666146	19
UGT2B17	rs3100645	69525783	4
UGT2B11	rs2708697	69031629	4
GSTM1	rs366631	110052995	1
C14orf52	rs10132742	64456675	14
KIAA1463	rs3742062	49415099	12
UBA2	rs2314664	18552942	19
GSTT1	rs407257	22676550	22
Hs.396207	rs3014241	45860466	1
UGT2B7	rs2708697	69031629	4
FLJ46603	rs1014390	72224481	17
LOC284293	rs6567407	5978905	18
LOC51240	rs1233276	190393027	2
USMG5	rs11191688	105182560	10
MRPL43	rs10786612	102643755	10
MGC2752	rs7249714	63749895	19
PKHD1L1	rs1026437	110562125	8
PHACS	rs2074040	44049899	11
LOC283970	rs6499292	68661559	16
IRF5	rs10229001	128386633	7
hmm1412	rs747172	70242799	11
NUDT2	rs4310287	34364979	9
FLJ21616	rs1487969	28941580	8
PTER	rs4748302	16595868	10
Hs.400876	rs752775	36488195	20
AXIN1	rs214249	288688	16
TINP1	rs6883061	74128556	5
LOC284184	rs11150780	76878755	17
FLJ21347	rs6504675	45989356	17
RPL37A	rs284565	217067051	2
LOC375097	rs752775	36488195	20
C21orf107	rs2836934	39486755	21
LCMT1	rs7188975	25044950	16
MRPL43	rs10786612	102643755	10
hmm8232	rs3863641	46615803	1
CCNDBP1	rs2412752	41127265	15

Table A.2: HapMap Phase II gene expression reanalysis results.

GENE	SNP ID	Position	Chromosome
PLOR2E	rs3787016	1041803	19
LOC378075	rs2419490	64907258	7
LOC400642	rs9948693	5237432	18
KIAA1913	rs4897398	130649264	6
XRRA1	rs2298746	74231482	11
EIF2S1	rs1078194	66777549	14
SYNGR1	rs909685	38077617	22
PEX6	rs2274514	43042478	6
FLJ90036	rs6814287	112323	4
QRSL1	rs6568448	107200445	6
LOC339804	rs1177303	61241859	2
Hs.453941	rs880034	119702442	8
CDK5RAP2	rs2297454	122211576	9
NUDT2	rs7039222	34322740	9
VPS13A	rs1054368	78981613	9
LOC339229	rs3830068	77233294	17
PPA2	rs13108489	106512574	4
KIAA1712	rs4695916	175437965	4
Hs.519979	rs3862293	2992845	6
MGC12458	rs1979568	243259095	1
MGC22773	rs792310	74438081	1
STK25	rs2240482	242053684	2
HSRTSBETA	rs2305995	683977	18
UGT2B10	rs3100651	69495658	4
LOC400933	rs10854876	48396574	22
dJ383J4.3	rs1951626	172158724	1
LOC197322	rs12931350	87735816	16
Hs.6637	rs2293577	47393768	11
FLJ32112	rs12046885	54329878	1
Hs.379903	rs9891938	15855797	17
Hs.26039	rs2279327	10709785	5
WBSCR27	rs4304218	72890916	7
ST7L	rs7415820	112970972	1
HLA-DQA2	rs9275312	32773706	6
hmm26268	rs11118858	220019581	1
OAS1	rs7134391	11851074	12

eters σ_e^2 and hyperparameters $\sigma_{\beta+}^2$,$\sigma_{\beta-}^2$ to 1. 2) Compute the likelihood portion of the lower bound, $\mathcal{L}(\theta)$. The likelihood component of $\mathcal{L}(\theta)$ was a very practical convergence diagnostic in terms of computational efficiency. 3) Update the expected sufficient statistics and expectation of the μ parameter. 4) Update the expected sufficient statistics and expectations for each β_j parameter. 5) Update the expected sufficient statistics and expectations for the error term σ_e^2. 6) Update the expected sufficient statistics and expectations for the variance hyperparameters $\sigma_{\beta+}^2$, $\sigma_{\beta-}^2$. 7) Update the expected sufficient statistics and expectations for the probability of effect hyperparameters $p_{\beta+}$, $p_{\beta-}$. 8) Repeat steps 2-7) until the difference in lower bound between updates is less than 10^{-9}. 9) Return the sufficient statistics, specifically the p_{j+} and p_{j-} parameters (see Table A.4).

A.2.2 Expected sufficient statistics and expectations of parameters.

The population mean μ has a normal approximate factorized posterior, and is therefore characterized by a mean μ_μ and variance σ_μ^2 illustrated in Table A.3. The expectation $\mathrm{E}[\mu]$ is just the mean statistic, μ_μ.

The factorized approximate posterior density for each β_j parameter is a mixture distribution characterized by six sufficient statistics, a (posterior) positive effect mean μ_{j+}, a (posterior) negative effect mean μ_{j-}, a (posterior) positive effect variance σ_{j+}^2, a (posterior) negative effect variance σ_{j+}^2, a (posterior) probability of positive effect p_{j+}, and a (posterior) probability of negative effect p_{j-}. Table A.4 shows the expectations of these sufficient statistics in terms of the expectations of other parameters in the model. The functions $\phi(x)$ and $\Phi(x)$ are the standard Normal probability density and cumulative density functions respectively. Once the expected sufficient statistics for β_j are computed, the necessary expectations of β_j can be computed, $\mathrm{E}\left[\beta_j\right]$ and $\mathrm{E}\left[\beta_j^2\right]$ as shown in Table A.5. Note that $\mathrm{E}\left[p_{\beta+}\right]$ and $\mathrm{E}\left[p_{\beta-}\right]$ are used instead of $\exp\left(\mathrm{E}\left[\log\left\{p_{\beta+}\right\}\right]\right)$ and $\exp\left(\mathrm{E}\left[\log\left\{p_{\beta+}\right\}\right]\right)$ respectively for computational convenience, which we found did not affect the performance of the algorithm significantly (results not shown).

The approximate factorized posterior for the inverse of the error variance, σ_e^{-2} is characterized by a Gamma distribution, hence has shape and scale sufficient statistics, $\nu_e = \frac{n}{2}$ and ρ_e shown in Table A.6. Some of the higher order terms are dropped here from ρ_e (specifically the $\mathrm{E}\left[\beta_j^2\right]$ terms) for ease of computation. Again, we found that this did not significantly affect the performance of the algorithm (results not shown). The expectation for σ_e^{-2} is therefore $\mathrm{E}\left[\sigma_e^{-2}\right] = \nu_e\rho_e$.

Next we turn to the expected sufficient statistics for the positive and negative effect class variance hyperparameters, $\left(\sigma_{\beta+}^{-2}, \sigma_{\beta-}^{-2}\right)$. Since the priors for these are χ_1^2, the approximate posterior distribution for each parameter is a Gamma distribution, characterized by two sufficient statistics, a shape statistic ν_+ or ν_-, and a scale statistic ρ_+ or ρ_- as shown in Table A.7. Then the expectations, $\left(\mathrm{E}\left[\sigma_{\beta+}^{-2}\right], \mathrm{E}\left[\sigma_{\beta-}^{-2}\right]\right)$, can be computed as shown in Table A.8.

In addition, the expected sufficient statistics for the probability of membership in the positive, negative, and zero effect classes $\left(\Theta_\beta, \phi_\beta, \Psi_\beta\right)$ are illustrated in

Supplementary Table A.9. A uniform prior was assumed for this Dirichlet distribution. To compute the expectations with a truncated Dirichlet prior we used the property that any pairwise marginal distribution of the Dirichlet distribution is a Beta distribution. We used the marginal distribution with the positive and negative effect classes pooled since this is how the truncation was defined: $p_{\beta+} + p_{\beta-} \leq \frac{\sqrt{n}}{m}$. With the Beta distribution we used the GNU Scientific Library [52] to access the incomplete Beta function to compute the necessary expectations, $\left(\mathrm{E}\left[p_{\beta+} + p_{\beta-}\right], 1 - \mathrm{E}\left[p_{\beta+} + p_{\beta-}\right]\right)$ for a truncated Beta distribution. Then, we used the relative proportion of evidence in the positive and negative effect class to evaluate the expectation for the classes (e.g. $\mathrm{E}\left[p_{\beta+}\right] = \mathrm{E}\left[p_{\beta+} + p_{\beta-}\right] \frac{\Theta_{\beta}}{\Theta_{\beta}+\Phi_{\beta}}$).

A.2.3 Population structure.

The population structure version of the algorithm has additional population mean parameters α_l for k populations incorporated into the linear model in Equation 2.1 in the main text. The same factorization as in equation Equation 2.5 in the main text is assumed over the posterior distribution of the α_l parameters. A normal prior with large variance is applied to each α_l parameter, leading to update equations similar to those in Table A.4, except the approximate posterior density is no longer a mixture density but just a Normal distribution characterized by mean and variance sufficient statistics (not shown, available on request).

A.2.4 Additional algorithmic details.

For data sets with marker number $\geq 100,000$, the numerical library we used to evaluate the truncated expected sufficient statistics for the $\mathrm{E}\left[p_{\beta}\right]$ terms did not have high enough numerical precision to prevent underflow for large values of the sufficient statistics Θ_{β} and Φ_{β}. We therefore used a harsh update $\frac{\sqrt{n}}{100m}$ for each expectation $\mathrm{E}\left[p_{\beta+}\right]$ and $\mathrm{E}\left[p_{\beta-}\right]$ for the initial iterations where the sufficient statistics were large. In general, once the sufficient statistics shrink sufficiently (so as to be on the order of $\frac{\sqrt{n}}{m}$), then the truncated expectations $\mathrm{E}\left[p_{\beta}\right]$ can be computed exactly. We empirically found this to be the best trade-off between converging to a suboptimal over-fit model with hundreds of significant markers for too weak of an approximate update and too harsh of an update where we lose significant power (results not shown). For smaller marker numbers, $m \leq 100,000$, this was not a problem.

A.2.5 $O(nm)$ complexity for a single update.

As demonstrated by the form of μ_{j+} or μ_{j-} in Table A.4, the most computationally intensive step in the algorithm occurs during the update of the expected sufficient statistic for the β_j parameters. This is because the residual term $\sum_{l \neq j} \mathrm{E}\left[\beta_l\right] x_{li}$ must be recomputed for each step. Most of the terms in summation in this expression stay the same for updates of different β_j parameters. Hence, we store this residual term as a vector of length n and for any particular update of the expected sufficient statistics of a new β_j parameter we add or subtract the nec-

Table A.3: Expected Sufficient Statistics for μ.

	Expected Statistic
μ_μ	$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m}\mathrm{E}\left[\beta_j\right]x_{ji}\right)$
σ_μ^2	$\frac{\mathrm{E}\left[\sigma_e^2\right]}{n}$

essary terms (e.g. subtract the term $\mathrm{E}\left[\beta_j\right]x_{ji}\forall\ i$ from this residual when updating μ_{j+} and μ_{j-} for the β_j parameter). The update of a single β_j expected sufficient statistic therefore has $O(n)$ complexity. Because there are m β_j parameters the total time complexity of a single update of all β_j parameters is $O(nm)$. The complexity of updating the other parameters is either linear in terms of the sample size n or marker number m (e.g. updating σ_e^2 has complexity $O(n)$, and updating $\sigma_{\beta+}^2$ has complexity $O(m)$.) Therefore the total time complexity of the algorithm is $O(nm)$.

A.3 Supplementary Tables

Table A.4: Expected Sufficient Statistics for β_j.

	Expected Statistic
μ_{j+}	$\frac{\mathrm{E}\left[\sigma^2_{\beta+}\right]\sum_{i=1}^n x_{ji}\left(y_i - \mathrm{E}\left[\mu\right] - \sum_{l\neq j}\mathrm{E}\left[\beta_l\right]x_{li}\right)}{\mathrm{E}\left[\sigma^2_e\right] + \mathrm{E}\left[\sigma^2_{\beta+}\right]\sum_{i=1}^n x^2_{ji}}$
μ_{j-}	$\frac{\mathrm{E}\left[\sigma^2_{\beta-}\right]\sum_{i=1}^n x_{ji}\left(y_i - \mathrm{E}\left[\mu\right] - \sum_{l\neq j}\mathrm{E}\left[\beta_l\right]x_{li}\right)}{\mathrm{E}\left[\sigma^2_e\right] + \mathrm{E}\left[\sigma^2_{\beta-}\right]\sum_{i=1}^n x^2_{ji}}$
σ^2_{j+}	$\frac{\mathrm{E}\left[\sigma^2_{\beta+}\right]\mathrm{E}\left[\sigma^2_e\right]}{\mathrm{E}\left[\sigma^2_e\right] + \mathrm{E}\left[\sigma^2_{\beta+}\right]\sum_{i=1}^n x^2_{ji}}$
σ^2_{j-}	$\frac{\mathrm{E}\left[\sigma^2_{\beta-}\right]\mathrm{E}\left[\sigma^2_e\right]}{\mathrm{E}\left[\sigma^2_e\right] + \mathrm{E}\left[\sigma^2_{\beta-}\right]\sum_{i=1}^n x^2_{ji}}$
p_{j+}	$\frac{2\mathrm{E}\left[p_{\beta+}\right]\frac{\sigma_{j+}}{\sqrt{\mathrm{E}\left[\sigma_{\beta+}\right]}}\Phi\left(\frac{\mu_{j+}}{\sigma_{j+}}\right)\exp\left\{\frac{\mu^2_{j+}}{2\sigma^2_{j+}}\right\}}{1 - \mathrm{E}\left[p_{\beta+}\right] - \mathrm{E}\left[p_{\beta-}\right] + 2\mathrm{E}\left[p_{\beta+}\right]\frac{\sigma_{j+}}{\sqrt{\mathrm{E}\left[\sigma_{\beta+}\right]}}\Phi\left(\frac{\mu_{j+}}{\sigma_{j+}}\right)\exp\left\{\frac{\mu^2_{j+}}{2\sigma^2_{j+}}\right\} + 2\mathrm{E}\left[p_{\beta-}\right]\frac{\sigma_{j-}}{\sqrt{\mathrm{E}\left[\sigma_{\beta-}\right]}}\Phi\left(-\frac{\mu_{j-}}{\sigma_{j-}}\right)\exp\left\{\frac{\mu^2_{j-}}{2\sigma^2_{j-}}\right\}}$
p_{j-}	$\frac{2\mathrm{E}\left[p_{\beta-}\right]\frac{\sigma_{j-}}{\sqrt{\mathrm{E}\left[\sigma_{\beta-}\right]}}\Phi\left(-\frac{\mu_{j-}}{\sigma_{j-}}\right)\exp\left\{\frac{\mu^2_{j-}}{2\sigma^2_{j-}}\right\}}{1 - \mathrm{E}\left[p_{\beta+}\right] - \mathrm{E}\left[p_{\beta-}\right] + 2\mathrm{E}\left[p_{\beta+}\right]\frac{\sigma_{j+}}{\sqrt{\mathrm{E}\left[\sigma_{\beta+}\right]}}\Phi\left(\frac{\mu_{j+}}{\sigma_{j+}}\right)\exp\left\{\frac{\mu^2_{j+}}{2\sigma^2_{j+}}\right\} + 2\mathrm{E}\left[p_{\beta-}\right]\frac{\sigma_{j-}}{\sqrt{\mathrm{E}\left[\sigma_{\beta-}\right]}}\Phi\left(-\frac{\mu_{j-}}{\sigma_{j-}}\right)\exp\left\{\frac{\mu^2_{j-}}{2\sigma^2_{j-}}\right\}}$

Table A.5: Expectations of β_j.

	Expectations
$\mathrm{E}\left[\beta_{j+}\right]$	$\mu_{j+} + \frac{\sigma_{j+}\phi\left(-\frac{\mu_{j+}}{\sigma_{j+}}\right)}{1 - \Phi\left(-\frac{\mu_{j+}}{\sigma_{j+}}\right)}$
$\mathrm{E}\left[\beta_{j-}\right]$	$\mu_{j-} - \frac{\sigma_{j-}\phi\left(-\frac{\mu_{j-}}{\sigma_{j-}}\right)}{\Phi\left(-\frac{\mu_{j-}}{\sigma_{j-}}\right)}$
$\mathrm{E}\left[\beta_j\right]$	$p_{j+}\mathrm{E}\left[\beta_{j+}\right] + p_{j-}\mathrm{E}\left[\beta_{j-}\right]$
$\mathrm{V}\left[\beta_{j+}\right]$	$\sigma^2_{j+}\left(\frac{1 - \frac{\mu_{j+}}{\sigma_{j+}}\phi\left(-\frac{\mu_{j+}}{\sigma_{j+}}\right)}{1 - \Phi\left(-\frac{\mu_{j+}}{\sigma_{j+}}\right)} - \left(\frac{\phi\left(-\frac{\mu_{j+}}{\sigma_{j+}}\right)}{1 - \Phi\left(-\frac{\mu_{j+}}{\sigma_{j+}}\right)}\right)^2\right)$
$\mathrm{V}\left[\beta_{j-}\right]$	$\sigma^2_{j-}\left(\frac{1 + \frac{\mu_{j-}}{\sigma_{j-}}\phi\left(-\frac{\mu_{j-}}{\sigma_{j-}}\right)}{\Phi\left(\frac{\mu_{j+}}{\sigma_{j-}}\right)} - \left(\frac{\phi\left(-\frac{\mu_{j-}}{\sigma_{j-}}\right)}{\Phi\left(-\frac{\mu_{j-}}{\sigma_{j-}}\right)}\right)^2\right)$
$\mathrm{E}\left[\beta^2_j\right]$	$p_{j+}\left(\mathrm{V}\left[\beta_{j+}\right] + \mathrm{E}\left[\beta_{j+}\right]^2\right) + p_{j-}\left(\mathrm{V}\left[\beta_{j-}\right] + \mathrm{E}\left[\beta_{j-}\right]^2\right)$

Table A.6: Expected Sufficient Statistics for σ_e^{-2}.

	Expected Statistic
ρ_e	$2\left(\sum_{i=1}^{n}\left(y_i - \mathrm{E}\left[\mu\right] - \sum_{j=1}^{m}\mathrm{E}\left[\beta_j\right]x_{ji}\right)^2\right)^{-1}$

Table A.7: Expected Sufficient Statistics for $\left(\sigma_{\beta+}^{-2}, \sigma_{\beta-}^{-2}\right)$.

	Expected Statistic
ν_+	$\frac{1}{2} + \sum_{j=1}^{m} p_{j+}$
ρ_+	$\left(\frac{1}{2} + \frac{1}{2}\sum_{j=1}^{m}\mathrm{E}\left[\beta_j^2\right]p_{j+}\right)^{-1}$
ν_-	$\frac{1}{2} + \sum_{j=1}^{m} p_{j-}$
ρ_-	$\left(\frac{1}{2} + \frac{1}{2}\sum_{j=1}^{m}\mathrm{E}\left[\beta_j^2\right]p_{j-}\right)^{-1}$

Table A.8: Expectations of $\left(\sigma_{\beta+}^{-2}, \sigma_{\beta-}^{-2}\right)$.

	Expectations
$\mathrm{E}\left[\sigma_{\beta+}^{-2}\right]$	$\nu_+\rho_+$
$\mathrm{E}\left[\sigma_{\beta-}^{-2}\right]$	$\nu_-\rho_-$

Table A.9: Expected Sufficient Statistics for $\left(p_{\beta+}, p_{\beta-}\right)$.

	Expected Statistic
Θ_β	$1 + \sum_{j=1}^{m} p_{j+}$
Φ_β	$1 + \sum_{j=1}^{m} p_{j-}$
Ψ_β	$1 + m - \sum_{j=1}^{m} p_{j+} - \sum_{j=1}^{m} p_{j-}$

APPENDIX B

AGGRESSIVE FALSE POSITIVE CONTROL FOR GENOME-WIDE FEATURE SELECTION IN A MOUSE F2 CROSS

B.1 Variational spike and slab updates

The variational Bayes expectation (VBE) steps are defined as:

$$q_{\beta_j}^{t+1}\left(\beta_j\right) = (1 - p_j^t)\mathrm{I}\left[\beta_j = 0\right] + p_j^t\mathrm{N}\left(\mu_j^t, \sigma_j^{2^t}\right) \tag{B.1}$$

with sufficient statistics:

$$\begin{aligned} \mu_j^t &= \frac{\sum_{i=1}^n z_{ij}\langle r_{-j}\rangle}{\sum_{i=1}^n x_{ij}^2 + \langle\sigma_\beta^{-2}\rangle/\langle\sigma_e^{-2}\rangle}, \\ \sigma_j^{-2^t} &= \langle\sigma_e^{-2}\rangle\sum_{i=1}^n z_{ij}^2 + \langle\sigma_\beta^{-2}\rangle, \\ p_j^t &= \frac{1}{1 + C_j^t}, \end{aligned} \tag{B.2}$$

and expectations:

$$\begin{aligned} \langle\beta_j\rangle &= p_j^t\mu_j^t, \\ \langle\beta_j^2\rangle &= p_j^t\left(\mu_j^{2^t} + \sigma_j^{2^t}\right), \end{aligned} \tag{B.3}$$

where $\langle r_{-j}\rangle$ is the expectation of the linear residual term with respect to each approximate distribution, and

$$\begin{aligned} C_j^t = \left(1 - \exp\left\{\langle\log\left(p_\beta\right)\rangle\right\}\right) / \\ \left(2\sigma_j^t\exp\left\{\frac{1}{2}\left(2\langle\log\left(p_\beta\right)\rangle + \mu_j^{2^t} / \left(\sigma_j^{2^t}\right) + \langle\log\left(\sigma_\beta^{-2}\right)\rangle\right)\right\}\right). \end{aligned} \tag{B.4}$$

The update for the distribution of the approximate error variance is

$$q_{\sigma_e^{-2}}^{t+1}\left(\sigma_e^{-2}\right) = \Gamma\left(\eta_1, \eta_2\right), \tag{B.5}$$

with sufficient statistics:

$$\begin{aligned}\eta_1 &= \frac{n+1}{2}\\ \eta_2 &= \frac{\langle U\rangle + 1}{2}\end{aligned} \tag{B.6}$$

with expectations:

$$\begin{aligned}\langle\sigma_e^{-2}\rangle &= \frac{\eta_1}{\eta_2}\\ \langle\log\left(\sigma_e^{-2}\right)\rangle &= \psi\left(\eta_1\right) + \log\left(1/\eta_2\right)\end{aligned} \tag{B.7}$$

with $\langle U\rangle$ the expectation of the residual sum of square errors with respect to each current approximating distribution, and $\psi\left(x\right)$ the digamma function.

$$q^{t+1}_{\sigma_\beta^{-2}}\left(\sigma_\beta^{-2}\right) = \Gamma\left(\zeta_1, \zeta_2\right), \tag{B.8}$$

with sufficient statistics:

$$\begin{aligned}\zeta_1 &= \frac{\sum p_j^t + 1}{2},\\ \zeta_2 &= \frac{\sum\langle\beta_j^2\rangle + 1}{2},\end{aligned} \tag{B.9}$$

with expectations:

$$\begin{aligned}\langle\sigma_\beta^{-2}\rangle &= \frac{\zeta_1}{\zeta_2},\\ \langle\log\left(\sigma_\beta^{-2}\right)\rangle &= \psi\left(\zeta_1\right) + \log\left(1/\zeta_2\right),\end{aligned} \tag{B.10}$$

$$q^{t+1}_{p_\beta}\left(p_\beta, 1 - p_\beta\right) = \text{TBeta}\left(\rho_1, \rho_2, \frac{\sqrt{n}}{m}\right) \tag{B.11}$$

with sufficient statistics:

$$\rho_1 = \sum p_j^t + 1$$

,

$$\rho_2 = m - \sum p_j^t + 1, \tag{B.12}$$

and expectations $\langle p_\beta\rangle$ and $\langle\log\left(p_\beta\right)\rangle$ computed by numerical integration (with $\text{TBeta}\left(x, y, z\right)$ being a truncated Beta distribution with parameter x, y and support on $[0, z]$). A maximization step of the effect α of the fixed covariates, $\mathbf{T}$, is defined as:

$$\hat{\alpha}^{t+1} = \left(\mathbf{T}'\mathbf{W}\mathbf{T}\right)^{-1}\mathbf{T}'\mathbf{W}\left(\mathbf{y} - \mathbf{X}\langle\beta\rangle\right) \tag{B.13}$$

with $\mathbf{W} = \text{diag}\left(\langle\sigma_e^{-2}\rangle\right)$. The lower bound:

$$\mathcal{L}^{t+1} = \langle U\rangle + \langle\theta\rangle, \tag{B.14}$$

is computed every iteration, and stored after the algorithm converges (in practice when the per iteration change $\Delta\mathcal{L} \leq 0.0001$), where

$$\begin{aligned} \langle\theta\rangle = \rho_2 \log\left(1 - \exp\left\{\langle\log\left(p_\beta\right)\rangle\right\}\right) + \\ \frac{1}{2}\rho_1\left(\langle\log\left(\sigma_\beta^{-2}\right)\rangle + 1 + 2\langle\log\left(p_\beta\right)\rangle\right) - \\ \zeta_2\left(\langle\sigma_\beta^{-2}\rangle + 1\right) + \frac{1}{2}\left(n+1\right)\langle\log\left(\sigma_e^{-2}\right)\rangle. \end{aligned} \tag{B.15}$$

Multiple re-orderings of markers are initialized, and an approximate posterior distributions over possible models is generated by assigning the following probabilities to every unique model discovered (i.e. unique mode in the approximate posterior surface):

$$p\left(M_i\right) = \frac{\exp\left(\mathcal{L}_i\right)}{\sum \exp\left(\mathcal{L}_i\right)} \tag{B.16}$$

And approximate Bayesian model averaged results are generated for all the sufficient statistics/expectations:

$$\hat{p}_j = \sum_i p_{ji} p_{M_i} \tag{B.17}$$

Therefore, if there is a high degree of model uncertainty in the estimate, then this will reduce the variance of estimates. Before we run the algorithm, we rescale and recenter all the features in equation 6 to have mean zero, and variance one, so as not to penalize each possible feature (i.e. expression trait of genetic marker) differently based on the scale of the feature. We also set sporadic missing data to the empirical mean of the observed data for all variables. We include a weak pre-filtering step, where we only include the genetic markers and gene expression phenotypes which have $P < 0.1$ from a marginal test in a linear model (conditioned on the fixed effects, $\mathbf{T}$). This is motivated by recent theoretical work by Fan et al. [45], that suggests that a simple marginal test statistic can be used to effectively screen features that are not relevant, i.e. not in the full model. The size of the reduced filtered set of features is used when computing the truncation of the distribution over p_β.

B.1.1 Shrinkage estimator

The shrinkage estimator of Schäfer and Strimmer [114] involves taking a convex combination of a unregularized estimate of the sample covariance matrix, $\mathbf{\Sigma}$, and combines it with a low-rank regularized estimate, $\mathbf{T}$, with a weighting parameter λ. The inverse covariance matrix is inferred based on this regularized estimate of the sample covariance. The number of significant non-zero parameter is determined based on an empirical estimation of the false discovery rate [114].

$$\hat{\mathbf{\Sigma}}_\lambda = \lambda\hat{\mathbf{T}} + (1-\lambda)\,\hat{\mathbf{\Sigma}} \tag{B.18}$$

B.1.2 Partial least squares estimator

The partial least squares estimator in regularized regression proposed by Tenenahaus et al. [127], and is defined as a constrained optimization problem to identifying a small set of orthogonal predictors, with maximal covariance with the response variable in question. The regularization from this approach arises by choosing a small subset of orthogonal predictors, where the number of predictors is chosen by cross-validation [79].

B.1.3 Ridge estimator

As opposed to an L_1 lasso penalty, Kraemer et al. [79] propose a ridge penalty (i.e. an L_2^2 penalty), and use an empirical control of false discovery rate (FDR) based on semi-parametric estimation of either the tail area-based FDR or local FDR [125]:

$$\text{argmax}_{\beta} \left\{ -\sum_{i=1}^{n} (y_i - z_i\beta)^2 - \eta \sum_{j=1}^{p+m-1} \beta_j^2 \right\} \tag{B.19}$$

B.1.4 Simulation analyses and comparison to other network recovery algorithms

To test our method on simulated data, we compared it to a set of other methods that were recently proposed and combined [79] for sparse, regularized undirected network inference. To allow a consistent comparison with Kraemer et al., we use the same set of simulation parameters, network connectivity, and implementation of the five methods aforementioned methods. In this case we simulated data from a network of 100 gene expression phenotypes (with no genotypes), with a density of 0.05 (i.e. 248 undirected edges were randomly assigned between pairs of variables, across 10 replicate simulations). A constant variance of one is assumed across all phenotypes, and the edges weights are simulated as uniformly distributed between -1 and 1; and the full $\mathbf{\Theta_{yy}}$ matrix is rescaled such that $\mathrm{diag}\left(\mathbf{\Theta_{yy}}\right) = 1$. The performance of all the methods for varying sample sizes is illustrated in Figure B.1 in terms of a Precision-Recall curve in terms of the combined estimated regression coefficients for each method (i.e. the β coefficients). We see that the three best methods are the adaptive lasso, the lasso, and the variational spike and slab algorithm, which all appear to be close in terms of performance for the range of high true discovery rate (TDR or Recall), for similar power (Precision).

While this performance is similar among these three methods as a function of the estimated regression coefficient, there is still a fundamental problem associated with how to decide which nonzero elements are statistically significantly different from zero, where the null distribution of any test statistic would be non-trivial [142]. In Figure B.2 we see the true discovery rate for different sample sizes in terms of the features that are very confidently returned by the variational spike and slab, the lasso and the adaptive lasso algorithms. For the variational spike and slab algorithm, we see that if we choose only the edges which have posterior probability, $\hat{p}_j > 0.99$, we can aggressively control the FDR. Whereas, if we include every feature that is proposed by either the lasso or the adaptive lasso, we have a significant number of false positives that get carried along. This suggests that we can use the variational spike-and-slab to identify edges between genes only when there is very strong statistical support for the interaction in the data.

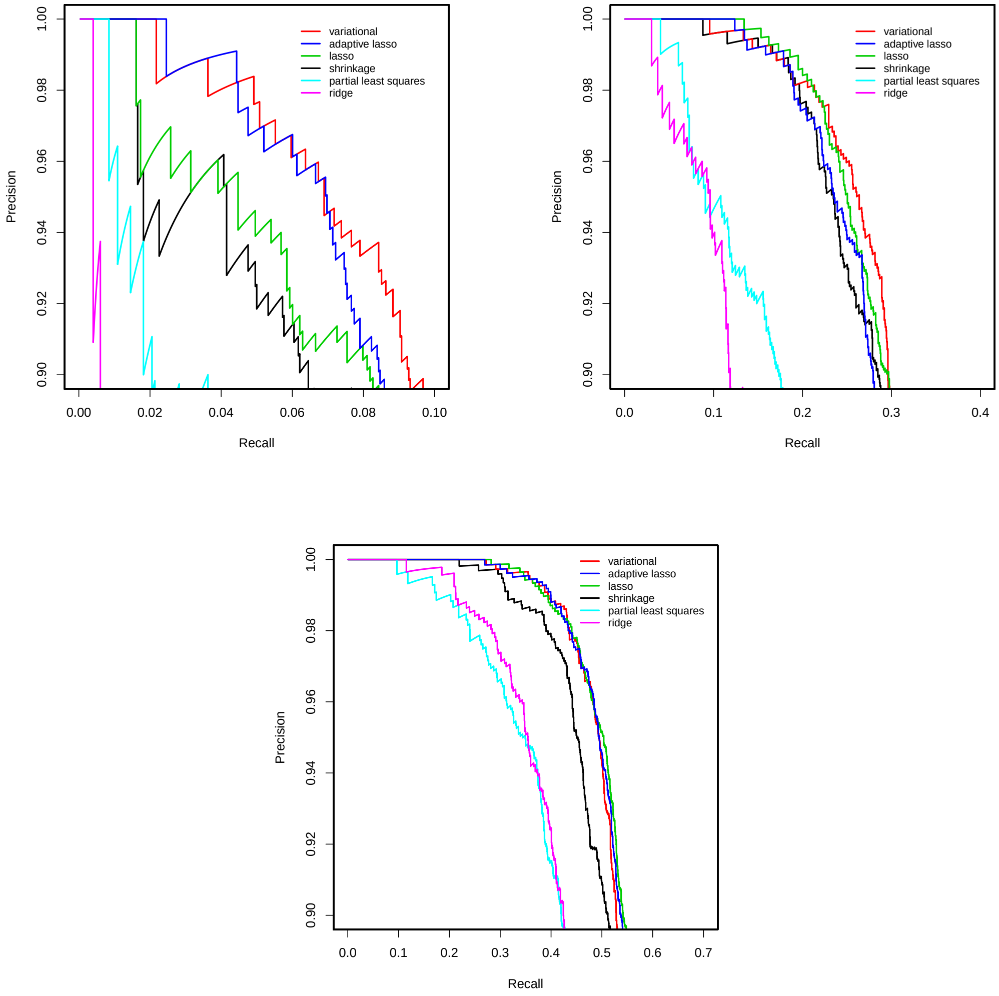

Figure B.1: A comparison across network reconstruction methods for a sparse (density of 0.05), undirected graph with 100 gene expression products. The precision (power) v.s. the recall (true discovery rate) is plotted for sample size 50 (A), 100 (B), and 200 (C), as determined by the estimated regression coefficients for each method.

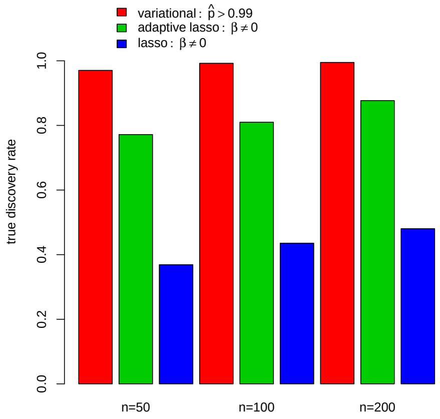

Figure B.2: True discovery rate as a function of sample size for the variational spike and slab method, the adaptive lasso, and the lasso, under the same simulations as shown in Figure 2.

BIBLIOGRAPHY

[1] L. Almasy and J. Blangero. Multipoint quantitative-trait linkage analysis in general pedigrees. *The American Journal of Human Genetics*, 62(5):1198–1211, 1998.

[2] D. Altshuler, L.D. Brooks, A. Chakravarti, F.S. Collins, M.J. Daly, and P. Donnelly. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[3] S. Anjum, A. Doucet, and C.C. Holmes. A boosting approach to structure learning of graphs with and without prior knowledge. *Bioinformatics*, 25(22):2929–2936, 2009.

[4] 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(1):34, 2008.

[5] Y.S. Aulchenko, J. Pullen, W.P. Kloosterman, M. Yazdanpanah, A. Hofman, N. Vaessen, P.J.L.M. Snijders, D. Zubakov, I. Mackay, M. Olavesen, et al. LPIN2 is associated with type 2 diabetes, glucose metabolism, and body composition. *Diabetes*, 56(12):3020, 2007.

[6] M.J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, 2003.

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

[8] C. M. Bishop. *Pattern recognition and machine learning*. Springer Science, New York, 2006.

[9] C.M. Bishop and SpringerLink (Online service). *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

[10] D.M. Blei and M.I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

[11] K.A. Bollen. *Structural equations with latent variables*. Wiley. NY, NY, 1989.

[12] C. Bouchard, T. Rankinen, Y.C. Chagnon, T. Rice, L. Perusse, J. Gagnon, I. Borecki, P. An, A.S. Leon, J.S. Skinner, et al. Genomic scan for maximal oxygen uptake and its response to training in the HERITAGE Family Study*. *Journal of Applied Physiology*, 88(2):551, 2000.

[13] S. Boyd and L. Vandenberghe. *Convex opimization*. Cambridge University Press New York, New York, 2004.

[14] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[15] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[16] 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.

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

[18] K. Bryc, A. Auton, M.R. Nelson, J.R. Oksenberg, S.L. Hauser, S. Williams, A. Froment, J.M. Bodo, C. Wambebe, S.A. Tishkoff, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences*, 107(2):786, 2010.

[19] G. Casella and R.L. Berger. *Statistical inference*. Thomson Brooks/Cole, 1990.

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

[21] 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. *The Annals of Applied Statistics*, 4(1):320–339, 2010.

[22] G.K. Chen, P. Marjoram, and J.D. Wall. Fast and flexible simulation of DNA sequence data. *Genome Res*, 19(1):136–142, Jan 2009.

[23] Y. Chen, J. Zhu, P.Y. Lum, X. Yang, S. Pinto, D.J. MacNeil, C. Zhang,

J. Lamb, S. Edwards, S.K. Sieberts, et al. Variations in DNA elucidate molecular networks that cause disease. *Nature*, 452(7186):429–435, 2008.

[24] D.M. Chickering, D. Geiger, and D. Heckerman. Learning Bayesian networks is NP-hard. *Microsoft Research*, pages 94–17, 1994.

[25] J. Chu, S.T. Weiss, V.J. Carey, and B.A. Raby. A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. *BMC Systems Biology*, 3(1):55, 2009.

[26] R.D. Cone. Editorial: The Corticotropin-Releasing Hormone System and Feeding Behavior–A Complex Web Begins to Unravel. *Endocrinology*, 141(8):2713, 2000.

[27] W. Cookson, L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184–194, 2009.

[28] H.J. Cordell and D.G. Clayton. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *The American Journal of Human Genetics*, 70(1):124–141, 2002.

[29] N. Craddock, MC Odonovan, and MJ Owen. The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *Journal of medical genetics*, 42(3):193, 2005.

[30] M.J. Davies, JL Gordon, AJH Gearing, R. Pigott, N. Woolf, D. Katz, and A. Kyriakopoulos. The expression of the adhesion molecules ICAM-1, VCAM-1, PECAM, and E-selectin in human atherosclerosis. *The Journal of Pathology*, 171(3):223–229, 1993.

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

[32] D. Denning, B. Mykytka, N.P.C. Allen, L. Huang, et al. The nucleoporin Nup60p functions as a Gsp1p–GTP-sensitive tether for Nup2p at the nuclear pore complex. *The Journal of Cell Biology*, 154(5):937–950, 2001.

[33] V. Deveaux, T. Cadoudal, Y. Ichigotani, F. Teixeira-Clerc, A. Louvet, S. Manin, J.T.V. Nhieu, M.P. Belot, A. Zimmer, P. Even, et al. Cannabinoid CB2 receptor potentiates obesity-associated inflammation, insulin resistance and hepatic steatosis. *PLoS One*, 4(6):e5844, 2009.

[34] M. Dhar, L.S. Webb, L. Smith, L. Hauser, D. Johnson, and D.B. West. A novel ATPase on mouse chromosome 7 is a candidate gene for increased body fat. *Physiological Genomics*, 4(1):93, 2000.

[35] A.F. Dominiczak, N. Brain, F. Charchar, M. McBride, N. Hanlon, and W.K. Lee. Genetics of hypertension: lessons learnt from mendelian and polygenic syndromes. *Clinical And Experimental Hypertension*, 26(7-8):611–620, 2004.

[36] P. Donnelly. Progress and challenges in genome-wide association studies in humans. *Nature*, 456(7223):728–731, 2008.

[37] P. Donnelly. Progress and challenges in genome-wide association studies in humans. *Nature*, 465(7223):728–731, Dec 2008.

[38] S. Doss, E.E. Schadt, T.A. Drake, and A.J. Lusis. Cis-acting expression quantitative trait loci in mice. *Genome Research*, 15(5):681–691, 2005.

[39] V. Emilsson, G. Thorleifsson, B. Zhang, A.S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G.B. Walters, S. Gunnarsdottir, et al. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428, 2008.

[40] D.M. Evans, J. Marchini, A.P. Morris, and L.R. Cardon. Two-stage two-locus models in genome-wide association. *PLoS Genet*, 2(9):e157, 2006.

[41] W.J. Ewens and R.S. Spielman. The transmission/disequilibrium test: history, subdivision, and admixture. *American Journal of Human Genetics*, 57(2):455, 1995.

[42] D.S. Falconer and T.F.C. Mackay. Introduction to quantitative genetics. 1996.

[43] J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive LASSO and SCAD penalties. *Annals*, 3(2):521–541, 2009.

[44] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[45] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[46] R. Fraschini, E. Formenti, G. Lucchini, and S. Piatti. Budding yeast Bub2 is localized at spindle pole bodies and activates the mitotic checkpoint via a different pathway from Mad2. *The Journal of cell biology*, 145(5):979–991, 1999.

[47] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

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

[49] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[50] 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.

[51] T.F. Fuller, A. Ghazalpour, J.E. Aten, T.A. Drake, A.J. Lusis, and S. Horvath. Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome*, 18(6):463–472, 2007.

[52] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi. GNU Scientific Library Reference Manual , ISBN 0954161734. *URL: http://www.gnu.org/software/gsl/*, 2003.

[53] D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: graphical models and model selection. *Annals of Statistics*, 29(2):505–529, 2001.

[54] A.J. Gelman, J.B Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. Chapman and Hall, Boca Raton, Florida, 2004.

[55] E.I. George and R.E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, Sept 1993.

[56] E.I. George and R.E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[57] M.F. Gerrits, S. Ghosh, N. Kavaslar, B. Hill, A. Tour, E.L. Seifert, B. Beauchamp, S. Gorman, J. Stuart, R. Dent, et al. Distinct skeletal mus-

cle fiber characteristics and gene expression in diet-sensitive versus diet-resistant obesity. *Journal of lipid research*, 51(8):2394, 2010.

[58] A. Ghazalpour, S. Doss, B. Zhang, S. Wang, C. Plaisier, R. Castellanos, A. Brozell, E.E. Schadt, T.A. Drake, A.J. Lusis, et al. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet*, 2(8):e130, 2006.

[59] M. Girolami. A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13(11):2517–2532, 2001.

[60] M.G. Heiman and P. Walter. Prm1p, a pheromone-regulated multispanning membrane protein, facilitates plasma membrane fusion during yeast mating. *The Journal of Cell Biology*, 151(3):719–730, 2000.

[61] G. Hermosillo, C. Chefd'Hotel, and O. Faugeras. Variational methods for multimodal image matching. *International Journal of Computer Vision*, 50(3):329–343, 2002.

[62] L.A. Hindorff, H.A. Junkins, J.P. Mehta, and T.A. Manolio. A Catalog of Published Genome-Wide Association Studies. *http://www.genome.gov/gwastudies*, April Accessed 2009.

[63] J.N. Hirschhorn and M.J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

[64] D.A. Hogan, T.A. Auchtung, and R.P. Hausinger. Cloning and characterization of a sulfonate/alpha-ketoglutarate dioxygenase from Saccharomyces cerevisiae. *Journal of bacteriology*, 181(18):5876–5879, 1999.

[65] C.J. Hoggart, J.C. Whittaker, M. De lorio, and D.J. Balding. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, 4(7):e1000130, Jul 2008.

[66] A.L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–690, 2008.

[67] V. Hyttinen, J. Kaprio, L. Kinnunen, M. Koskenvuo, and J. Tuomilehto. Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs. *Diabetes*, 52(4):1052, 2003.

[68] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, Oct 2007.

[69] SK Iyengar and RC Elston. The genetic basis of complex traits: rare variants or "common gene, common disease"? *Methods in molecular biology (Clifton, NJ)*, 376:71, 2007.

[70] T.S. Jaakkola and M.I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

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

[72] I.M. Johnstone and B.W. Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, pages 1594–1649, 2004.

[73] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, 8:613–636, 2007.

[74] A.B. Keeton, M.O. Amsler, D.Y. Venable, and J.L. Messina. Insulin signal transduction pathways and insulin-induced gene expression. *Journal of Biological Chemistry*, 277(50):48565, 2002.

[75] S. Kim, I. Sohn, J.I. Ahn, K.H. Lee, Y.S. Lee, and Y.S. Lee. Hepatic gene expression profiles in a long-term high-fat diet-induced obesity mouse model. *Gene*, 340(1):99–109, 2004.

[76] N. Klöting and I. Klöting. Genetic variation in the multifunctional transcription factor Yy1 and type 1 diabetes mellitus in the BB rat. *Molecular genetics and metabolism*, 82(3):255–259, 2004.

[77] K. Kohara, Y. Tabara, J. Nakura, Y. Imai, T. Ohkubo, A. Hata, M. Soma, T. Nakayama, S. Umemura, N. Hirawa, et al. Identification of hypertension-susceptibility genes and pathways by a systemic multiple candidate gene approach: the millennium genome project for hypertension. *Hypertension Research*, 31(2):203–212, 2008.

[78] M.R. Kosorok and S. Ma. Marginal asymptotics for the" large p, small n" paradigm: with applications to microarray data. *Annals of Statistics*, 35(4):1456, 2007.

[79] N. Kraemer, J. Schafer, and A.L. Boulesteix. Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC bioinformatics*, 10(1):384, 2009.

[80] S.L. Lauritzen. *Graphical models*. Oxford University Press. NY, NY, 1996.

[81] B. Le Tallec, M.B. Barrault, R. Courbeyrette, R. Guérois, M.C. Marsolier-Kergoat, and A. Peyroche. 20S proteasome assembly is orchestrated by two distinct pairs of chaperones in yeast and in mammals. *Molecular Cell*, 27(4):660–674, 2007.

[82] 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 Genet*, 2(7):e114, 2006.

[83] J. Listgarten, C. Kadie, E.E. Schadt, and D. Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 107(38):16465, 2010.

[84] 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.

[85] H.Y. Liu, J. Han, S.Y. Cao, T. Hong, D. Zhuo, J. Shi, Z. Liu, and W. Cao. Hepatic autophagy is suppressed in the presence of insulin resistance and hyperinsulinemia. *Journal of Biological Chemistry*, 284(45):31484, 2009.

[86] J. Liu, Y. Liu, X. Liu, and H.W. Deng. Bayesian mapping of quantitative trait loci for multiple complex traits with the use of variance components. *Am J Hum Genet*, 81(2):304–320, July 2007.

[87] B.A. Logsdon, G.E. Hoffman, and J.G. Mezey. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC bioinformatics*, 11(1):58, 2010.

[88] B.A. Logsdon and J. Mezey. Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Computational Biology*, 6(12):429–435, 2010.

[89] P.Y. Lum, Y. Chen, J. Zhu, J. Lamb, S. Melmed, S. Wang, T.A. Drake, A.J. Lusis, and E.E. Schadt. Elucidating the murine brain transcriptional net-

work in a segregating mouse population to identify core functional modules for obesity and diabetes. *Journal of Neurochemistry*, 97:50–62, 2006.

[90] M. Lynch and B. Walsh. *Genetics and analysis of quantitative traits*, volume 24. Sinauer Sunderland, MA, 1998.

[91] T.F.C. Mackay, E.A. Stone, and J.F. Ayroles. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10(8):565–577, 2009.

[92] B. Maher. Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18, 2008.

[93] N. Malo, O. Libiger, and N.J. Schork. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *The American Journal of Human Genetics*, 82(2):375–385, 2008.

[94] G. Mannhaupt, R. Stucka, U. Pilz, C. Schwarzlose, and H. Feldmann. Characterization of the prephenate dehydrogenase-encoding gene, TYR1, from Saccharomyces cerevisiae. *Gene*, 85(2):303–311, 1989.

[95] A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

[96] M.I. McCarthy, G.R. Abecasis, L.R. Cardon, D.B. Goldstein, J. Little, J.P.A. Ioannidis, and J.N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

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

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

[99] M. Otsuki, K. Hashimoto, Y. Morimoto, T. Kishimoto, and S. Kasayama. Circulating vascular cell adhesion molecule-1 (VCAM-1) in atherosclerotic NIDDM patients. *Diabetes*, 46(12):2096, 1997.

[100] J. Pearl. *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press, 2000.

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

[102] P. Peltola, J. Pihlajamäki, H. Koutnikova, E. Ruotsalainen, U. Salmenniemi, I. Vauhkonen, S. Kainulainen, H. Gylling, T.A. Miettinen, J. Auwerx, et al. Visceral Obesity is Associated with High Levels of Serum Squalene&ast. *Obesity*, 14(7):1155–1163, 2006.

[103] L. Perusse, T. Rankinen, A. Zuberi, Y.C. Chagnon, S.J. Weisnagel, G. Argyropoulos, B. Walts, E.E. Snyder, and C. Bouchard. The human obesity gene map: the 2004 update. *Obesity*, 13(3):381–490, 2005.

[104] A.E. Raftery, D. Madigan, and J.A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

[105] T.P. Rasmussen and M.R. Culbertson. The putative nucleic acid helicase Sen1p is required for formation and stability of termini and for maximal rates of synthesis and levels of accumulation of small nucleolar RNAs in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 18(12):6885–6896, 1998.

[106] D. Reich, A.L. Price, and N. Patterson. Principal component analysis of genetic data. *Nature genetics*, 40(5):491–492, 2008.

[107] T. Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pages 454–61, 1996.

[108] M.V. Rockman. Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature*, 456(7223):738–744, 2008.

[109] T. Sandmann, J.M. Herrmann, J. Dengjel, H. Schwarz, and A. Spang. Suppression of coatomer mutants by a new protein family with COPI and COPII binding motifs in Saccharomyces cerevisiae. *Molecular Biology of the Cell*, 14(8):3097–3113, 2003.

[110] E.E. Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–223, 2009.

[111] E.E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S.K. Sieberts, S. Monks, M. Reitman, C. Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7):710–717, 2005.

[112] E.E. Schadt, C. Molony, E. Chudin, K. Hao, X. Yang, P.Y. Lum, A. Kasarskis, B. Zhang, S. Wang, C. Suver, et al. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*, 6(5):e107, 2008.

[113] E.E. Schadt, S.A. Monks, T.A. Drake, A.J. Lusis, N. Che, V. Colinayo, T.G. Ruff, S.B. Milligan, J.R. Lamb, G. Cavet, et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, 2003.

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

[115] S. Scherneck, M. Nestler, H. Vogel, M. Blüher, M.D. Block, M.B. Diaz, S. Herzig, N. Schulz, M. Teichert, S. Tischer, et al. Positional cloning of zinc finger domain transcription factor Zfp69, a candidate gene for obesity-associated diabetes contributed by mouse locus Nidd/SJL. *PLoS genetics*, 5(7):593–596, 2009.

[116] B. Shipley. *Cause and correlation in biology*. Cambridge, UK: Cambridge University Press, 2002.

[117] A. Siepel, G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.D.W. Hillier, S. Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034, 2005.

[118] E.K. Speliotes, C.J. Willer, S.I. Berndt, K.L. Monda, G. Thorleifsson, A.U. Jackson, H.L. Allen, C.M. Lindgren, J. Luan, R. Mägi, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*, 2010.

[119] P. Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 491–498. Citeseer, 1995.

[120] P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, prediction, and search*. Boston, MA: The MIT Press, 2001.

[121] M. Stanke and S. Waack. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(Suppl 2), 2003.

[122] B.E. Stranger, M.S. Forrest, M. Dunning, C.E. Ingle, C. Beazley, N. Thorne, R. Redon, C.P. Bird, A. de Grassi, C. Lee, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, 2007.

[123] B.E. Stranger, M.S. Forrest, M. Dunning, C.E. Ingle, C. Beazley, N. Thorne, R. Redon, C.P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S.W. Scherer, S. Tavare, P. Deloukas, M.E. Hurles, and E.T. Dermitzakis. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, Feb 2007.

[124] M. et al. Stratton. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[125] K. Strimmer. A unified approach to false discovery rate estimation. *BMC bioinformatics*, 9(1):303, 2008.

[126] M. Stumvoll, B.J. Goldstein, and T.W. van Haeften. Type 2 diabetes: principles of pathogenesis and therapy. *The Lancet*, 365(9467):1333–1346, 2005.

[127] A. Tenenhaus, V. Guillemot, X. Gidrol, and V. Frouin. Gene association networks from microarray data using a regularized estimation of partial correlation based on PLS regression. *IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2008.

[128] G. Thorleifsson, G.B. Walters, D.F. Gudbjartsson, V. Steinthorsdottir, P. Sulem, A. Helgadottir, U. Styrkarsdottir, S. Gretarsdottir, S. Thorlacius, I. Jonsdottir, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature genetics*, 41(1):18–24, 2008.

[129] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[130] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[131] I. Tsamardinos, L.E. Brown, and C.F. Aliferis. The max-min hill-

climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

[132] C. Tysk, E. Lindberg, G. Järnerot, and B. Floderus-Myrhed. Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut*, 29(7):990, 1988.

[133] B.F. Voight, A.M. Adams, L.A. Frisse, Y. Qian, R.R. Hudson, and A. Di Rienzo. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences*, 102(51):18508–18513, 2005.

[134] A. Wagner. How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. *Bioinformatics*, 17(12):1183–1197, 2001.

[135] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational methods. *New Directions in Statistical Signal Processing. MIT Press*, 2005:138, 2003.

[136] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[137] H. Waki, B. Liu, M. Miyake, K. Katahira, D. Murphy, S. Kasparov, and J.F.R. Paton. Junctional adhesion molecule-1 is upregulated in spontaneously hypertensive rats: evidence for a prohypertensive role within the brain stem. *Hypertension*, 49(6):1321, 2007.

[138] S. Wang, N. Yehya, E.E. Schadt, H. Wang, T.A. Drake, and A.J. Lusis. Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet*, 2(2):e15, 2006.

[139] M.N. Weedon, H. Lango, C.M. Lindgren, C. Wallace, D.M. Evans, M. Mangino, R.M. Freathy, J.R.B. Perry, S. Stevens, A.S. Hall, et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nature genetics*, 40(5):575–583, 2008.

[140] M. West. Bayesian factor regression models in the large p, small n paradigm. *Bayesian statistics*, 7(2003):723–732, 2003.

[141] A. Wiederkehr, S. Avaro, C. Prescianotto-Baschong, R. Haguenauer-Tsapis, and H. Riezman. The F-box protein Rcy1p is involved in endocytic membrane traffic and recycling out of an early endosome in Saccharomyces cerevisiae. *The Journal of Cell Biology*, 149(2):397–410, 2000.

[142] T.T. Wu, Y.F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714, 2009.

[143] T.T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat*, 2(1):224–244, 2008.

[144] S. Xu. Estimating polygenic effects using markers of the entire genome. *Genetics*, 163(2):789–801, Feb 2003.

[145] Y. Yamada, F. Ando, and H. Shimokata. Association of polymorphisms of SORBS1, GCK and WISP1 with hypertension in community-dwelling Japanese individuals. *Hypertension Research*, 32(5):325–331, 2009.

[146] X. Yang, J.L. Deignan, H. Qi, J. Zhu, S. Qian, J. Zhong, G. Torosyan, S. Majid, B. Falkard, R.R. Kleinhanz, et al. Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nature genetics*, 41(4):415–423, 2009.

[147] T.J. Yen. A Majorization-Minimization Approach to Variable Selection Using Spike and Slab Priors. *Arxiv preprint arXiv:1005.0891*, 2010.

[148] N. Yi and S. Banerjee. Hierarchical generalized linear models for multiple quantitative trait locus mapping. *Genetics*, 181(3):1101–1113, Jan 2009.

[149] N. Yi and D. Shriner. Advances in Bayesian multiple quantitative trait loci mapping in experimental crosses. *Heredity*, 100(3):240–252, March 2008.

[150] N. Yi and S. Xu. Bayesian Lasso for quantitative trait loci mapping. *Genetics*, 179(2):1045–1055, May 2008.

[151] M. Yuan. Efficient computation of the l1 regularized solution path in Gaussian graphical models. *J. Comput. Graph. Stat*, 17:809–826, 2006.

[152] S. Yusuf, S. Hawken, S. Ôunpuu, T. Dans, A. Avezum, F. Lanas, M. McQueen, A. Budaj, P. Pais, J. Varigos, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the

INTERHEART study): case-control study. *The Lancet*, 364(9438):937–952, 2004.

[153] M. Zeyda, K. Gollinger, E. Kriehuber, FW Kiefer, A. Neuhofer, and TM Stulnig. Newly identified adipose tissue macrophage populations in obesity with distinct chemokine and chemokine receptor expression. *International Journal of Obesity*, 2010.

[154] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128, 2005.

[155] M. Zhang, K.L. Montooth, M.T. Wells, A.G. Clark, and D. Zhang. Mapping multiple quantitative trait loci by Bayesian classification. *Genetics*, 169(4):2305, 2005.

[156] M. Zhang, D. Zhang, and M.T. Wells. Variable selection for large p small n regression models with incomplete data: mapping QTL with epistases. *BMC bioinformatics*, 9(1):251, 2008.

[157] Y. Zhang and J.S. Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39(9):1167–1173, Aug 2007.

[158] P. Zhao and B. Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[159] S. Zhou, S. Van De Geer, and P. Bühlmann. Adaptive Lasso for high dimensional regression and Gaussian graphical modeling. *Arxiv preprint arXiv:0903.2515*, 2009.

[160] 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.

[161] 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.

[162] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[163] M. Zou and S.D. Conzen. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71–79, 2005.