BMC Genomics

METHODOLOGY ARTICLE

Open Access

A new statistical framework for genetic pleiotropic analysis of high dimensional phenotype data

Panpan Wang[1,2], Mohammad Rahman[1], Li Jin[2*] and Momiao Xiong[1,3*]

Abstract

Background: The widely used genetic pleiotropic analyses of multiple phenotypes are often designed for examining the relationship between common variants and a few phenotypes. They are not suited for both high dimensional phenotypes and high dimensional genotype (next-generation sequencing) data.
To overcome limitations of the traditional genetic pleiotropic analysis of multiple phenotypes, we develop sparse structural equation models (SEMs) as a general framework for a new paradigm of genetic analysis of multiple phenotypes. To incorporate both common and rare variants into the analysis, we extend the traditional multivariate SEMs to sparse functional SEMs. To deal with high dimensional phenotype and genotype data, we employ functional data analysis and the alternative direction methods of multiplier (ADMM) techniques to reduce data dimension and improve computational efficiency.

Results: Using large scale simulations we showed that the proposed methods have higher power to detect true causal genetic pleiotropic structure than other existing methods. Simulations also demonstrate that the gene-based pleiotropic analysis has higher power than the single variant-based pleiotropic analysis. The proposed method is applied to exome sequence data from the NHLBI's Exome Sequencing Project (ESP) with 11 phenotypes, which identifies a network with 137 genes connected to 11 phenotypes and 341 edges. Among them, 114 genes showed pleiotropic genetic effects and 45 genes were reported to be associated with phenotypes in the analysis or other cardiovascular disease (CVD) related phenotypes in the literature.

Conclusions: Our proposed sparse functional SEMs can incorporate both common and rare variants into the analysis and the ADMM algorithm can efficiently solve the penalized SEMs. Using this model we can jointly infer genetic architecture and casual phenotype network structure, and decompose the genetic effect into direct, indirect and total effect. Using large scale simulations we showed that the proposed methods have higher power to detect true causal genetic pleiotropic structure than other existing methods.

Keywords: Structural equations, Causal inference, Multiple phenotypes, Quantitative trait, Next-generation sequencing, Pleiotropic analysis

* Correspondence: lijin.fudan@gmail.com; Momiao.Xiong@uth.tmc.edu
[2]State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China
[1]Human Genetics Center, Department of Biostatistics, University of Texas School of Public Health, Houston, TX 77030, USA
Full list of author information is available at the end of the article

Background

In the past several years, a large number of statistical methods for association analysis of both qualitative and quantitative traits with next-generation sequencing data were developed [1–14]. Most genetic analyses of quantitative traits focus on association analysis of a single trait, analyzing each phenotype individually and independently [15]. However, multiple phenotypes are correlated. For example, metabolism of lipoproteins involves cholesterol, triglycerides, very low density lipoproteins (VLDL), low density lipoproteins and high density lipoproteins. These multiple traits are dependent. The integrative analysis of correlated phenotypes often increase statistical power to identify genetic associations [16, 17]. The association analysis of multiple phenotypes is expected to become popular in the near future [18].

Three major approaches are commonly used to explore association of genetic variants with multiple correlated phenotypes: multiple regression methods, integration of p values of univariate analysis, and dimension reduction methods [16]. Despite their differences in selection of specific methods for estimation, all these estimation methods share the following common features. First, many methods were designed for common variants and hence may not be appropriate for rare ones. Second, the results of all these analyses are difficult to interpret. They do not provide information to indicate which phenotypes the genetic variants are significantly associated [15]. Third, all these methods estimate the effect of the genetic variant on each phenotype individually and do not explore the dependency patterns of genetic effects among the phenotypes and do not provide a detailed characterization of the relationships among the genetic effects. Fourth, all these estimations only estimate the effects of the genetic variants on the phenotypes. However, the genetic effects can be classified into three types of effects: direct, indirect and total effects. These methods are unable to reveal mechanisms underlying the genetic structures of multiple phenotype association analysis [19]. The direct effect is the measurement of the influence of a genetic variant on a phenotype that is not mediated by any other phenotypes in a system. The indirect effect of a genetic variant measures the sensitivity of a phenotype to change of a genetic variant that is mediated by at least one intervening variable (phenotype). The total effect is the sum of the direct and indirect effects. The most popular multivariate association methods are lack of ability to decompose total effect into direct effect and indirect effect and ignore indirect effects through other mediating phenotypes and risk factors. Therefore, they cannot discover how the effect of the genetic variant on the phenotype is mediated by other phenotypes and the effect path from the initially affected phenotype by the genetic variant through a number of mediating phenotypes to the targeted phenotype. Pleiotropic effect is a context dependent genetic effect and plays an important role in multivariate trait association studies and evolution analysis [20]. The pleiotropic effect of a specific genetic variant on multiple phenotypes may be due to either direct contribution of the genetic variant to the multiple phenotypes or phenotype correlations (mediations). The multivariate trait association studies cannot distinguish the paths connecting multiple phenotypes and genetic effects [21].

In the past several years, there have been increasing interests in modeling the complex structures among phenotypes, risk factors and genotypes which are referred to as the genotype-phenotype networks and therefore overcome these limitations. Current methods for inference of genotype-phenotype networks can be classified into two categories: whole network scoring methods and local analysis methods [22–29]. Network scoring approaches assign a score to the network model for measuring how well the network fits the data and develop algorithms to search the network with the best score. Local analysis methods analyze small sets of variables that are pieced together into networks from multiple causality tests between variables.

One of network scoring methods is structure equations that can be used as a tool to model the complex network structures among phenotypes, risk factors and genotypes [19–21, 30–32]. A graphical model in which the variables are represented as nodes and the relationships between variables are represented by edges between the nodes can be used to model the genotype-phenotype networks. Structural equations can generate biological interpretations of relations among variables and uncover the mechanism structure underlying phenotypic and genotypic relationships. To date, in applications of the structural equation model (SEM) in quantitative genetics, the causal structure was assumed to be known as a priori, or partially specified, thereby allowing selection of the causal structure for a small set of variables from the data [21]. There are two major approaches to estimate the causal structure from the data. One approach is based on the conditional independence and the notion of Markov equivalence of directed acyclic graphs (DAGs) [33]. DAGs encode causal structure. However, a DAG is not, in general, identifiable from observational data. Conditional independence only determines the skeleton of the DAG which is the undirected graph of the DAG by removing its directions of all edges, and the v structure of the DAG where two nodes are directed to a common node (collider) [34]. A number of algorithms such as PC-algorithms have been used to estimate the equivalence class of DAGs [35]. A second approach is to use the notion of 'sparse' and develop sparse SEMs for estimating the causal structures [36]. By incorporating the penalized constraints of the parameters into the likelihood function to enforce the network sparsity, we could estimate the causal structure. Coordinate ascent algorithms are often used to maximize the penalized likelihood functions.

Despite their successful application to joint analysis of genetic architecture and causal phenotype networks, current approaches often demand intensive computations and are lack of efficient computational algorithms for implementing penalization of network structure parameters. Therefore, they cannot be used for large-scale causal inference. Most current approaches are designed for common variants and are difficult to be applied to next generation sequencing (NGS) data. The purpose of this paper is to overcome these limitations. We first develop novel functional SEMs where exogenous genotype profiles across a genomic region or a gene are represented as a function of the genomic position for genetic association analysis of multiple quantitative traits which is referred to as multivariate quantitative traits locus (QTL) analysis. The functional SEMs for multivariate QTL analysis consist of three components. The first component is a phenotype network that is modeled as a directed graph. The second component is a genotype network that is represented as an undirected graph. The third component is connections between the genotype network and phenotype network with direction from genotype nodes to phenotype nodes. To make the network sparse and reduce the burden of computations, we develop the novel sparse SEMs for genotype-phenotype networks and an efficient computational algorithms based on ADMM to search the causal structure and estimate the parameters [37, 38]. We will estimate the direct, indirect and total effects of the genetic variants on the phenotypes using estimated directed graph and intervention calculus [39] and explore the relationships between direct, indirect and total effects estimated from SEMs and the genetic effects estimated from the traditional simple regressions and multiple regressions. Finally, the sparse SEMs are applied to exome sequence data from the NHLBI's Exome Sequencing Project (ESP) with 11 phenotypes. A program implementing the developed sparse SEMs for quantitative genetic analysis with multiple phenotypes will be published as an R package.

Methods

Multivariate quantitative trait association analysis can be investigated by phenotype-genotype networks, which can be represented as a graph. Phenotypes, covariates such as age, sex, race, and SNPs are variables. Variables are represented as nodes in the graph. We assume that causal relationships among phenotypes exist. Therefore, a phenotype network is represented by a directed graph. A directed edge between two nodes indicates the causal relationship between them. Since SNPs do not have causal relationships among them, a genotype network is represented as an undirected graph. An edge between two nodes in the genotype network indicates their correlation. Since all SNPs and covariates may cause changes in phenotypes, the phenotype network and genotype network are connected by edges directed from covariates and SNP to the phenotypes. The phenotypes and connections between phenotypes, covariates and SNPs can be modeled by structural equations. The genotype network can be leant by graphical LASSO (GLASSO) [39], here we didn't focus on genotype network in this paper. An example of phenotype-genotype network is shown in Additional file 1: Figure S1.

SEMs for multivariate association analysis

The SEMs offer a general statistical framework for inferring phenotype networks and connections between genotypes and phenotypes. Assume that n individuals are sampled. We consider M phenotypes that are referred to as endogenous variables. The endogenous variables are jointly determined in the model and are also influenced by the variables outside the model. We denote the n observations on the M endogenous variables by the matrix $Y = [y_1, y_2, \ldots, y_M]$, where $y_i = [y_{1i}, \ldots, y_{ni}]^T$ is a vector of collecting n observation of the endogenous variable i. Covariates, genetic variants as exogenous or predetermined variables are denoted by $X = [x_1, \ldots, x_K]$ where $x_i = [x_{1i,} \ldots, x_{ni}]^T$. Similar to independent variables in the regression, the exogenous variables are outside the models and are not influenced by the variables in the model. Similarly, random errors are denoted by $E = [e_1, \ldots, e_M]$, where we assume $E[e_i] = 0$ and $E[e_i e_i^T] = \sigma_i^2 I_n$ for $i = 1, \ldots, M$. Recall that the relationships between the phenotypes and genotypes are traditionally described by the regressions where the phenotypes are taken as dependent variables and genotypes are taken as independent variables are predictors. In the regression models, the dependence relationships among dependent variables or phenotypes cannot be explicitly expressed. Therefore, the regression models cannot be used to determine which phenotypes cause the variations of which phenotypes. To overcome this limitation, we introduce linear structural equations. The linear structural equations for modeling relationships among phenotypes and genotypes can be written as [38].

$$\begin{array}{l} y_1\gamma_{11} + y_2\gamma_{21} + \ldots + y_M\gamma_{M1} + x_1\beta_{11} + x_2\beta_{21} + \ldots + x_K\beta_{K1} + e_1 = 0 \\ \quad\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ y_1\gamma_{1M} + y_2\gamma_{2M} + \ldots + y_M\gamma_{MM} + x_1\beta_{1M} + x_2\beta_{2M} + \ldots + x_K\beta_{KM} + e_M = 0 \end{array} \tag{1}$$

where the γ's and β's are the structural parameters of the system that are unknown.

Variables in the SEMs can be classified into two basic types of variables: observed variables that can be measured and the residual error variables that cannot be measured and represent all other unmodeled causes of the variables. Most observed variables (e. g. phenotypes such as BMI, blood pressure, high density lipoprotein, low density lipoprotein) are random. Some observed variables may be nonrandom or control variables (e. g. genotypes, drug dosages) whose values remain the same in repeated random sampling or might manipulated by the experimenter. The observed variables will be further classified into exogenous variables (e.g. genotypes, age, sex, race), which lie outside the model, and endogenous variables (e.g. phenotypes), whose values are determined through joint interaction with other variables within the system. All nonrandom variables can be viewed as exogenous variables. Phenotypes are viewed as endogenous variables. The terms exogenous and endogenous are model specific. It may be that an exogenous variable in one model is endogenous in another. The structural parameters γ describe the relationships between phenotypes and parameters β measure the direct genetic effects of the genotypes on the phenotypes.

In matrix notation the SEMs (1) can be rewritten as

$$Y\Gamma + XB + E = 0, \tag{2}$$

where $\Gamma = [\Gamma_1, \ldots, \Gamma_M]$, $\Gamma_i = [\gamma_{1i}, \ldots, \gamma_{Mi}]^T$, $B = [B_1, \ldots, B_M]$, $B_i = [\beta_{1i}, \ldots, \beta_{Ki}]^T$.

We assume that the random errors in the structural equations are independent and uncorrelated with exogenous variables. We apply the sparsity penalty to each equation to ensure that the sparse SEMs are identifiable.

Two-stage least square estimates of the parameters in the SEMs

The ordinary least squares estimator is biased and inconsistent for the parameters of structural equations. To ensure the consistent estimates of the parameters in the SEMs, we use a generalized least square method that can be interpreted as a two-stage least square estimate method to estimate the parameters in the SEMs [38].

Recalling that y_i is the vector of observations of the variable i, let Y_{-i} be the observation matrix Y after removing y_i from it and γ_{-i} be the parameter vector Γ_i after removing the parameter γ_{ii}. The ith equation:

$$Y\Gamma_i + XB_i + e_i = 0$$

can be rewritten as

$$\begin{aligned} y_i &= Y_{-i}\gamma_{-i} + XB_i + e_i \qquad (3) \\ &= W_i\Delta_i + e_i, \end{aligned}$$

where $W_i = [\,Y_{-i}\, X], \Delta_i = \left[\gamma_{-i}^T\ B_i^T\right]^T$.

Multiplying by the matrix X^T on both sides of eq. (3), we obtain

$$\begin{aligned} X^T y_i &= X^T Y_{-i}\gamma_{-i} + \left(X^T X\right)B_i + X^T e_i \\ &= X^T W_i\Delta_i + X^T e_i. \end{aligned} \tag{4}$$

It is known that

$$\mathrm{cov}\left(X^T e_i, X^T e_i\right) = X^T X\sigma_i^2.$$

The generalized least square estimate $\hat{\Delta}_i$ is given by

$$\hat{\Delta}_i = \left[W_i^T X\left(X^T X\right)^{-1} X^T W_i\right]^{-1} W_i^T X\left(X^T X\right)^{-1} X^T y_i. \tag{5}$$

The generalized least square estimate $\hat{\Delta}_i$ can be interpreted as a two-stage least square estimate [38].

Suppose that in the first stage, Y_{-i} is regressed on X to obtain

$$\hat{\Pi}_i = \left(X^T X\right)^{-1} X^T Y_{-i} \text{ and } \hat{Y}_{-i} = X\hat{\Pi}_i.$$

Then,

$$\begin{aligned} \hat{W}_i &= \left[\,\hat{Y}_{-i}\ \ \mathrm{X}\right] \\ &= X\left(X^T X\right)^{-1} X^T W_i. \end{aligned}$$

Eq. (5) can be reduced to

$$\hat{\Delta}_i = \left(\hat{W}_i^T \hat{W}_i\right)^{-1} \hat{W}_i^T y_i. \tag{6}$$

Therefore, if W_i in eq. (3) is replaced by $\hat{W}_i$, eq. (6) can be interpreted as that in the second stage, y_i is regressed on $\hat{Y}_i$ and X to obtain estimate $\hat{\Delta}_i$.

Sparse SEMs and alternating direction method of multipliers

In general, the genotype-phenotype networks are sparse. Therefore, Γ and B are sparse matrices. In order to obtain sparse estimates of Γ and B, the natural approach is the l_1 -norm penalized regression of eq. (4). Using weighted least square and l_1 -norm penalization, we can form the following optimization problem:

$$\begin{aligned} &\min_{\Delta_i}\ f(\Delta_i) + \lambda||\Delta_i||_1 \\ &\text{where } f(\Delta_i) = \left(X^T y_i – X^T W_i\Delta_i\right)^T \left(X^T X\right)^{-1} \left(X^T y_i – X^T W_i\Delta_i\right). \end{aligned} \tag{7}$$

The size of the genotype-phenotype network may be large. The efficient ADMM [37] algorithm is used to solve the optimization problem (7). The procedure for implementing ADMM is given below (more detailed descriptions are provided in Appendix 1).

```
Algorithm:

For $i = 1,...,M$

Step 1. Initialization

$u^0 := 0$
$\Delta_i^0 := [W_i^T X(X^T X)^{-1} X^T W_i + \rho I]^{-1} W_i^T X(X^T X)^{-1} X^T y_i$
$Z_i^0 := \Delta_i^0.$

Carry out steps 2,3 and 4 until convergence

Step 2.

$\Delta_i^{(k+1)} := [\frac{1}{\rho} I - \frac{1}{\rho} W_i^T X(\rho X^T X + X^T W_i W_i^T X)^{-1} X^T W_i][W_i^T X(X^T X)^{-1} X^T y_i + \rho(Z_i^k - u^k)]$

Step 3.

$Z_i^{(k+1)} := \mathrm{sgn}(\Delta_i^{k+1} + u^k)(|\Delta_i^{k+1} + u^k| - \frac{\lambda}{\rho})_+,$

where

$|x|_+ = \begin{cases} x & x \geq 0 \\ 0 & x < 0. \end{cases}$

Step 4.

$u^{(k+)} :== u^{(k)} + (\Delta_i^{(k+1)} - Z_i^{(k+1)}).$
```

Under some assumptions convergence of ADMM can be proved [37]. In practice, although it can be slow to converge to high accuracy, ADMM converges to modest accuracy within a few tens of iterations. When large-scale problems and parameter estimation problems are considered, modest accuracy is sufficient. Therefore, ADMM may work very well for structure and parameter estimation in the genotype-phenotype networks.

Most of the elements of matrices Γ and B are equal to zero. The l_1 – regularized Lasso for the two stage least squares approach and ADMM algorithms are expected to shrink most of the coefficient matrices Γ and B toward zero, yielding sparse network structures. The sparsity-controlling parameter λ will be estimated via cross validation or set by users to get reasonable results. We abbreviate this sparse two stage least square estimation of SEMs as S2SEMs.

Sparse functional structural equation models for phenotype and genotype networks

Fast and cheaper next generation sequencing (NGS) technologies will generate unprecedentedly massive and highly-dimensional genomic variation data. Despite their promise, next generation sequencing platforms also have three specific features: high error rates, enrichment of rare variants and large proportion of missing values. Available causal analysis platforms for genetic studies which are mainly designed for common variants provide useful tools for single marker-based pleiotropic genetic analysis, but have limitations in analyzing thousands of sequences collected for very large population-based studies of humans. To address the critical barrier in causal genetic analysis with NGS data, we extend the multivariate SEMs to functional SEMs where exogenous genotype profiles across a genomic region are represented as a function of the genomic position. To effectively reduce the dimension of the data, we use genetic variant profiles which will recognize information contained in the physical location of the SNP as a major data form. The densely distributed genetic variants across the genomes in large samples can be viewed as realizations of a Poisson process. The densely typed genetic variants in a genomic region for each individual are so close that these genetic variant profiles can be treated as observed data taken from curves. The genetic variant profiles are called functional.

Large simulations have shown that combining information across multiple variants in a genomic region of analysis will greatly enhance power to detect association of rare variants [9]. To jointly utilize multi-locus genetic information and reduce the dimension of the NGS data, we propose to use a genomic region or a gene as a unit in multiple trait association analysis and develop sparse functional structural equation models (FSEMs) for construction and analysis of the phenotype and genotype networks. The FSEMs collectively analyze the contribution of multiple variants to the traits, reduce the errors in the NGS data via data reduction techniques and can effectively deal with missing data through the smooth mechanism of the function curves of the data.

Let t be a genomic position. Define a genotype profile $x_i(t)$ of the i-th individual as

$$x_i(\mathrm{t}) = \begin{cases} 2P_q(\mathrm{t}), & \mathrm{QQ} \\ P_q(\mathrm{t})\text{-}\mathrm{P_Q}(\mathrm{t}), & \mathrm{Qq} \\ -2P_Q(\mathrm{t}), & \mathrm{qq} \end{cases}$$

where Q and q are two alleles of the marker at the genomic position t, $P_Q(t)$ and $P_q(t)$ are the frequencies of the alleles Q and q, respectively. Suppose that we are interested in k genomic regions or genes $[a_j, b_j]$, denoted as $T_j, j = 1, 2, \ldots, k$. We consider the following functional structural equation models:

$$\begin{array}{l} \mathrm{y}_1\gamma_{11} + y_2\gamma_{21} + \ldots + y_M\gamma_{M1} + \int_{T_1} x_1(t)\beta_{11}(t)dt + \ldots + \int_{T_k} x_k(t)\beta_{k1}(t)dt + e_1 = 0 \\ \mathrm{y}_1\gamma_{12} + y_2\gamma_{22} + \ldots + y_M\gamma_{M2} + \int_{T_1} x_1(t)\beta_{12}(t)dt + \ldots + \int_{T_k} x_k(t)\beta_{k2}(t)dt + e_2 = 0 \\ \vdots \qquad\qquad\qquad\qquad \vdots \qquad\qquad\qquad\qquad \vdots \\ \mathrm{y}_1\gamma_{1M} + y_2\gamma_{2M} + \ldots + y_M\gamma_{MM} + \int_{T_1} x_1(t)\beta_{1M}(t)dt + \ldots + \int_{T_k} x_k(t)\beta_{kM}(t)dt + e_M = 0 \end{array} \tag{8}$$

where $\beta_{ij}(t)$ are genetic effect functions.

Functional principal components (FPCs) are efficient summary statistics. The FPCs simultaneously employs genetic information of the individual variants and correlation information (linkage disequilibrium) among all variants. The FPCs view the genetic variation across the genomic region as a function of its genomic location and uses intrinsic functional dependence structure of the data and all available genetic information of the variants in the genomic region. The neighboring genetic variants are linked. The genotypes at one SNP are dependent on the genotypes at nearby SNPs. The FPCs account for the space-ordering of the genetic variation data. Expanding the genotype functions in terms of a few orthogonal FPCs will substantially reduce the dimensions of the genetic variation data while preserving the intrinsic correlation structure and the space-ordering of the data. Specifically, For each genomic region or gene, we use functional principal component analysis to calculate principal component function [14]. We expand $x_{nj}(t), n = 1, \ldots, N, j = 1, 2, \ldots, k$ in each genomic region in terms of orthogonal principal component functions:

$$x_{ij}(t) = \sum_{l=1}^{L_j} \eta_{ijl}\phi_{jl}(t), j = 1, \ldots, k,$$

where $\phi_{jl}(t), j = 1, \ldots, k, l = 1, \ldots, L_j$ are the l-th principal component function in the j-th genomic region or gene and η_{ijl} are the functional principal component scores of the i-th individual.

Let η be a matrix collection of all functional principal component scores, the parameter matrix B can be defined as that in Appendix 2, matrices Y and Γ can be defined as that in the previous section. The structural functional equations can be reduced in terms of functional principal component scores (Appendix 2):

$$Y\Gamma_i + \eta B_i + e_i = 0,$$

which can be rewritten as

$$y_i = W_i\Delta_i + e_i,$$

where $W_i = [\,Y_{-i}\ \eta], \Delta_i = \begin{bmatrix} \gamma_{-i}^T & B_i^T \end{bmatrix}^T$.

Then, the sparse FSEMs are transformed to

$$\min_{\Delta_i}\ f(\Delta_i) + \lambda||\Delta_i||_1$$
$$\text{where} f(\Delta_i) = \left(\eta^T y_i - \eta^T W_i\Delta_i\right)^T \left(\eta^T\eta\right)^{-1}\left(\eta^T y_i - \eta^T W_i\Delta_i\right). \tag{9}$$

The ADMM algorithms for solving the sparse FSEMs are the same as that in the previous section if the matrix X is replaced by a functional principal component score matrix η (Appendix 2).

The functional SEMs can efficiently combine both common and rare genetic variants across the gene region and are suitable for NGS data [14]. This model extends the single variant-based network analysis to gene-based analysis, which can deal with hundreds of genes that may include tens of thousands SNPs. However, due to the computational limitation, we cannot directly handle the whole genome sequencing data. To construct whole genome genotype-phenotype networks, the network construction consists of two stages. At the first stage, we can group the genes based on the metabolic pathways or cluster analysis, with each group having at most hundreds of genes, and then apply the functional SEMs to each group of genes to find the set of genes significantly connected to the phenotypes. At the second stage, adding the sets of significantly connected genes identified at the first stage together to form a new of set of genes for network construction. The functional SEMs are again applied to the new set of genes to construct the final genotype-phenotype network.

Effect decomposition and estimation

To make this paper self-contained, we introduce basic concepts and methods for decomposition and estimation of the effects. In the genotype-phenotype network analysis we are interested in estimation of effects of genetic variants on phenotypes, which is referred to as genetic effects and effects of treatment on phenotypes. All genetic effects and treatment effects can be decomposed as total (causal), direct effects and indirect effects. Distinction between total, direct and indirect effects are of great practical importance in genetic association analysis [40]. The total effect measures the changes of response variable Y (phenotype) would take on the value y when variable X is set to x by external intervention. Direct effect is defined as sensitivity of Y to changes in X while all other variables in the model are held fixed. Indirect effect is to measure the portion of the effect which can be explained by mediation alone, while inhibiting the capacity of Y to respond to X [41]. The total effect is equal to the summation of direct and indirect effects.

Given a directed graph model G, one can compute total effects using intervention calculus [34, 42]. Suppose that the expected value of a response variable Y, after X is assigned value x by intervention is denoted by $E[Y|do(X = x)]$. The total effect is defined as

$$\frac{\partial}{\partial x}E[Y|do(X = x)]. \tag{10}$$

Note X_j is called a parent of X in G if there is a directed edge $X_j \rightarrow X$. Let pa_x denote the set of all parents of X in G. In the linear SEMs, we assume that $E[Y|X, \text{pa}_x]$ is linear in X and pa_x:

$$E[Y|X,\ \mathrm{pa}_x] = \alpha + \beta X + \gamma^T \mathrm{pa}_x. \tag{11}$$

Then,

$$\frac{\partial}{\partial x} E[Y|do(X = x)] = \beta.$$

When a directed graph is given, it is easy to calculate total effect [42]. Assume that there are k directed paths from X to Y and p_i are the product of the path coefficients along the i-th path. The total effect of X on Y is then defined as $\sum_{i=1}^{k} p_i$. As shown in Additional file 1: Figure S2, the total effect of X on Y is $ag + bdh + acdh$. By its definition, direct effect measures the sensitivity of Y to changes in X while all other variables in the model are held fixed. In other words, all links from X to Y other than the direct link will be blocked. As a consequences, the direct effect is equal to the path coefficient from X to Y. In the linear SEMs, the indirect effect of X on Y mediated by M is equal to the sum of the products associated with directed paths from X to Y through M [42]. In Additional file 1: Figure S2, there is no direct effect from X to Y. The indirect effect of X on Y which is mediated by B and D is equal to bdh.

In the SEMs for genotype-phenotype networks, since all SNPs only form undirected graph and there are no directed links between SNPs although we can observe linkage (or correlation) between SNPs; SNPs in the genotype-phenotype networks do not have parents. The total effect of SNP X on Y is the regression coefficient β of the following linear regression:

$$E[Y|do\ (X = x)] = \alpha + \beta x,$$

which is a simple regression of Y on X. This indicates that the traditional simple regression for association studies captures the total effect of a genetic variant on a phenotype.

If we include environments and risk factors such as smoking and obesity in the model and want to evaluate the effects of the environments and risk factors on the phenotype, these variables play mediating roles and will also be taken as phenotypes. We denote these mediating phenotypes by Y_{ME}. Since genetic variants, and other risk factors and phenotypes will affect the mediating phenotypes, the mediating phenotypes in the graphics may have parents. Their parents are denoted by S. Total effect of the mediation phenotype on the target phenotype is calculated by

$$E\left[Y|do\ \left(Y_{ME} = y_{ME}, X_{pa} = x_{pa}\right)\right] = \alpha + \beta y_{ME} + \gamma^T x_{pa}, \tag{12}$$

where β is the total effect of the mediation phenotype Y_{ME} on the target phenotype Y. In this case, a simple regression of Y on Y_{ME} can no longer be used to measure the total effect of the mediation phenotype Y_{ME} on the target phenotype Y. To observe this, we simulated 1000 individuals with the SEM as shown in Additional file 1: Figure S3. Each variable has a noise term distributed as $N(0, 1)$. The total effect of the mediation phenotype Y_{ME} on the target phenotype Y is 3.5. We obtain the simple regression:

$$Y = 1.39 + 5.85 Y_{ME}.$$

It is clear that the coefficient of the simple regression is 5.85. This value is far away from the total effect 3.5. However, using eq. (12) we obtain

$$Y = 3.54 Y_{ME} + 5.85 X,$$

where the regression coefficient 3.54 measured the total effect of the mediation phenotype Y_{ME} on the target phenotype Y.

Test statistics for path coefficients

Testing connection between the j-th gene and the i-th phenotype in the genotype-phenotype network, we formally investigate the problem of testing the coefficient of the path directed from the j- th gene to the i-th phenotype:

$$H_0 : \beta_{ji}(t) = 0\ ,\ \ \forall t \in \left[0, T_j\right] \tag{13}$$

against

$$H_a : \beta_{ji}(t) \neq 0\ .$$

If the coefficient function of path or genetic effect function $\beta_{ji}(t)$ is expanded in terms of the principal component functions:

$$\beta_{ji}(t) = \sum\nolimits_{g=1}^{G} b_{jig} \phi_{jg}(t),$$

then testing the null hypothesis H_0 in Eq. (13) is equivalent to testing the hypothesis:

$$H_0 : b_{jig} = 0\ ,\ \forall \mathrm{g}. \tag{14}$$

The path coefficients b_{jig} can be estimated by solving problems (8) and (9). Let $\hat{b}_{ji} = \left[b_{ji1}, \dots, b_{jiG}\right]^T$. The covariance matrix of the vector of the estimators of path coefficients for the i -th equation is given by [38]

$$\hat{\Sigma}_i = \sigma_{ii}\left[W_i^T \eta\left(\eta^T \eta\right)^{-1} \eta^T W_i\right]^{-1}, \tag{15}$$

where

$$\sigma_{ii} = \left(y_i - W_i \hat{\Delta}_i\right)^T \left(y_i - W_i \hat{\Delta}_i\right)/n. \tag{16}$$

Let Λ_i be the submatrix that corresponds to b_{ji} in the matrix $\hat{\Sigma}_i$. Define the statistic for testing the directed connection from the j-th gene to the i-th phenotype as

$$T_g = \hat{b}_{ji}^T \Lambda_i^{-1} \hat{b}_{ji}. \qquad (17)$$

Under the null hypothesis of no association $H_0 : b_{ji} = 0$, T_g is asymptotically distributed as a central $\chi^2_{(G)}$ distribution where G is the number of functional principal components in the expansion of $\beta_{ji}(t)$.

For testing a single parameter or single variant's path coefficient in the SEMs, the l-th parameter of the i-th equation, the statistic is given by

$$T_c = \frac{\hat{\Delta}_{il}^2}{\operatorname{var}\left(\hat{\Delta}_{il}\right)}, \qquad (18)$$

where $\operatorname{var}\left(\hat{\Delta}_{il}\right)$ is the l-th diagonal element of the matrix $\hat{\Sigma}_i$. Under the null hypothesis $H_0 : \Delta_{il} = 0$, T_c is asymptotically distributed as a central $\chi^2_{(1)}$ distribution.

Testing for the path coefficients within the network results in multiple testing problems. Both false discovery rate approach and Bonferroni correction can be used to adjust for multiple testing [43, 44].

Results

Model evaluation by simulations

We evaluated the performance of the sparse SEM approach for genetic analysis of multiple quantitative traits in simulation studies of a genotype-phenotype network where SNP-based simulations and gene-based simulations were considered. The simulations were carried out for common variants, rare variants and half common and half rare variants. The genotype data were selected from the NHLBI's Exome Sequencing Project (ESP) with 3248 individuals of European origin, which were then used to generate a population of 1,000,000 individuals.

We first study the SNP-based simulations. The genotype-phenotype network consisted of two parts. The first part was the phenotype network that was modeled by a DAG. The second part was the connections between the genotypes and phenotypes in which the genotypes were directed to the phenotypes. We randomly generated a genotype-phenotype network structure (see an example in Additional file 1: Figure S4). The parameters Γ_{ij} in the SEMs for modeling phenotype sub-network were generated from a uniformly distributed random variable over the interval (0.5, 1) or (−1,-0.5) if an edge from node j to node i was presented in the phenotype sub-network; otherwise Γ_{ij} =0. Similarly, the parameters B_{ij} in the SEMs for modeling the direction from the genotype (SNP) node j to the phenotype node i were generated from a uniformly distributed random variable over the interval (0, 1) or (−1,0) if an edge from node j to node i was presented in the genotype-phenotype network, otherwise $B_{ij} = 0$. The indicator variables for coding genotypes of the SNP were as previously described. Using the randomly generated network structure and parameters in the structural equations, we produced the phenotypes by the model: $Y = -XB\Gamma^{-1} + \varepsilon\Gamma^{-1}$, where $\varepsilon \sim N(0, 0.01 \times I)$, and X is a matrix of indicator variables for coding genotypes. For the randomly generated phenotype network, the expected number of degrees per node is three. Simulations were repeated 100 times. Five-fold cross validation was used to determine the penalty parameter λ that was then employed to infer the network while running power simulations. Two measures: the power of detection (PD) and the false discovery rate (FDR) were used to evaluate the performance of the algorithms for identification of the network structures. Specifically, let N_t be the total number of edges among 1000 replicates of the network and $\hat{N}_t$ be the total number of edges detected by the inference algorithm, N_{true} be the total number of true edges detected among simulated network and N_{False} be the false edges detected among $\hat{N}_t$. Now, the power of detection (PD) is defined by $\frac{N_{True}}{\hat{N}_t}$ and false discovery rate (FDR) is defined by $\frac{N_{False}}{\hat{N}_t}$.

In the SNP-based simulations we first compared the S2SEM with ADMM algorithms with the sparse maximum likelihood SEMs (SML) with coordinate ascent algorithms [36]. The SML method assumes each phenotype has one priori known QTL, and only focus on the inference of phenotype network. So in this comparison we only calculate PD and FDR for phenotype network.

We first compare the power and FDR of the S2SEM and SML under the assumption that each QTL had only connection with one phenotype, and no pleiotropic effects were present. We considered two scenarios: 10 phenotypes and 10 SNPs, 30 phenotypes and 30 SNPs. Results were shown in Additional file 1: Figure S5. We observed that if the variants were common variants, the SML had a higher power and a lower FDR than the S2SEM. However, once rare variants were included the SML substantially lost power and increased FDR.

Now we compare the power and FDR of the S2SEM and SML under the assumption of the presence of pleiotropic genetic effects. We considered three scenarios: ten phenotypes and SNPs, 30 phenotypes and SNPs, and 100 phenotypes and SNPs. The simulation results were shown in Fig. 1. We observed that even though SML still showed very high power for common variants, its FDR was large when the genotype had pleiotropic effects. Again, in the presence of rare variants the S2SEM had a higher power and a lower FDR than the SML. Even if in the presence of only common variants, we also observed an interesting feature that when the number of phenotypes and SNPs exceeds some threshold, the power of the S2SEM became higher than the SML.

Next we compare the computational time of the S2SEM and SML methods. Table 1 showed their

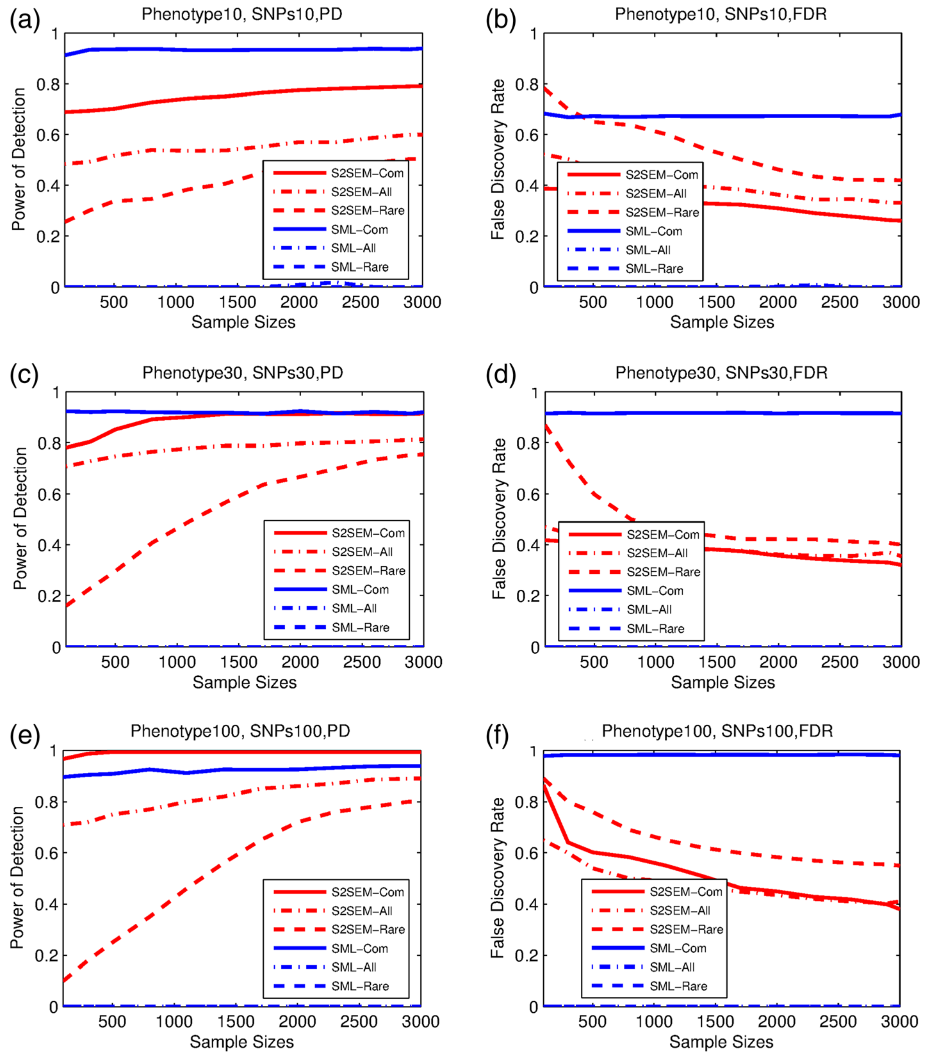

Fig. 1 Performance of S2SEM and SML. The power and FDR of the two methods for phenotype networks inference when the phenotype and genotype number is 10 (**a**, **b**), 30 (**c**, **d**) and 100 (**e**, **f**) respectively, each genotype has pleiotropic effect for another phenotype

Table 1 The computation time for S2SEM and SML methods of one replicate for the simulations of common variants

Time(s)/replicate	Phenotypes 10, Common SNPs 10		Phenotypes 30, Common SNPs 30		Phenotypes 100, Common SNPs 100	
Sample	S2SEM	SML	S2SEM	SML	S2SEM	SML
100	0.712	0.825	5.148	16.386	273.873	353.402
300	0.926	1.598	7.058	37.006	123.936	873.426
500	1.228	2.560	9.037	58.360	143.593	1535.544
800	1.850	4.230	11.967	94.564	170.799	2470.602
1100	2.463	6.315	15.541	130.116	309.795	4098.201
1400	2.972	8.755	18.856	169.985	218.933	5269.562
1700	3.736	12.032	23.879	212.400	285.213	4989.240
2000	4.515	15.473	27.664	252.855	311.480	5973.181
2300	5.517	19.388	33.685	297.721	375.167	7268.095
2600	6.488	23.515	38.413	347.531	625.959	8950.179
2900	7.646	28.127	45.708	393.149	724.829	10455.262

program running time for one replicate of simulations in the presence of the common variants. The computer CPU was Intel (R) Xeon E7-4870. We can see that our S2SEM method are much faster than SML method.

When we study the general genotype-phenotype networks, construction of large genotype-phenotype requires heavy computations. The SML methods are not suitable for the general genotype-phenotype network inference due to its large computational time and have not be applied to general genotype-phenotype network estimation. Next we compared the S2SEM with the QTLnet algorithm [22] which can be used for joint inference of causal network and genetic architecture for correlated phenotypes. We considered three scenarios: ten phenotypes with 30 SNPs, 30 phenotypes with 100 SNPs and 100 phenotypes and 1000 SNPs. The procedures for randomly generating genotype-phenotype networks were described as in the previous section. We assumed that on the average, each phenotype was affected by three genetic variants.

Figure 2a and c showed the power of two methods: S2SEM and QTLnet for detecting the structure of the genotype (common variants, rare variants and both common and rare variants)– phenotype network as a function of sample size.

We observed three features. The first, the power of S2SEM in all three cases was higher than QTLnet method. Second, the power of the two methods to detect the structure of the networks with the common variants was the highest, followed by the half common and half rare variants. The power of two methods to detect the structure of the network with the rare variants was the

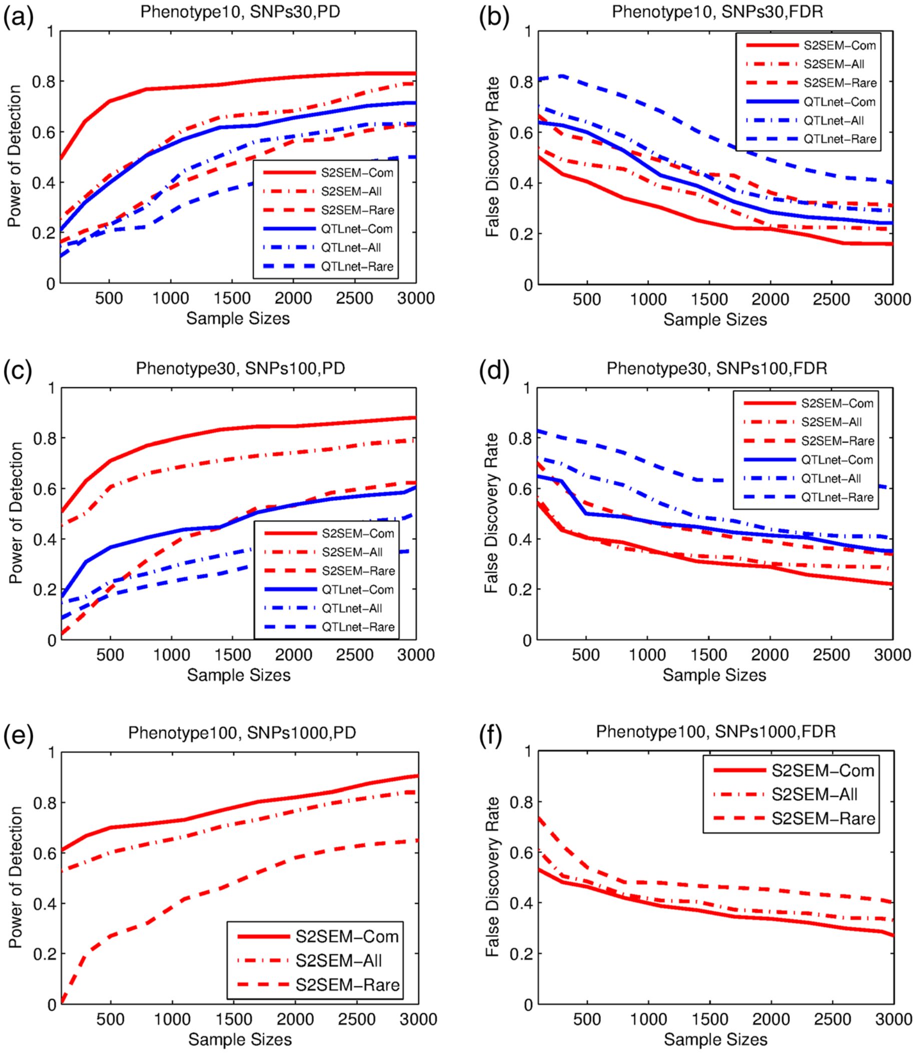

Fig. 2 Performance of S2SEM and QTLnet. The power and FDR of the two methods for genotype-phenotype networks inference in three different settings, (**a**,**b**) is results for 10 phenotypes and 30 SNPs, (**c**,**d**) is results for 30 phenotypes and 100 SNPs. QTLnet method is too time consuming to obtain results for 100 phenotypes, (**e**,**f**) only gives results of S2SEM for different SNPs frequencies

lowest. Third, in general, the power increased when the sample sizes increased. To fully evaluate the performance of the two methods, we also presented the FDR for detection of the structure of the networks as a function of sample sizes in Fig. 2 (b) and (d). It was clear that the FDR of the S2SEM in all three cases was lower than QTLnet method. The FDR of two methods to detect the structure of the networks with the common variants was the lowest, followed by the both common and rare variants. The FDR of two methods to detect the structure of the network with the rare variants was the highest. However, the false discovery rates for these two methods and in three cases were larger than 0.1 even the sample sizes reached 3000, and it is larger than 0.3 in the rare variants case.

Finally, the simulation results for the third scenario: 100 phenotypes and 1000 SNPs were shown in Fig. 2(e) and (f). The power and FDR patterns of the S2SEM and QTLnet were the same as that for the previous two scenarios. We also observed that the sample sizes increased as the sizes of the network increased. However, when the sample sizes exceeded 2000 the impact of the sizes of the network on the power became small. Table 2 showed the required computational times for network construction using the S2SEM and QTLnet. It was clear that the S2SEM still can estimate the large genotype-phenotype networks in a short time. However, the QTLnet method could not estimate such large genotype-phenotype networks in a reasonable time.

Figures 1 and 2 showed that the power of the variant by variant tests for identifying the network structure with the rare variants was low. To increase the power and reduce data dimensions, we develop functional SEMs (FSEMs) for network analysis using a genomic region or gene as a unit of analysis. To evaluate this strategy, we presented Fig. 3 to compare the power and FDR of the gene-based FSEMs and the SNP-based SEMs for detection of the network structures. Since the original papers for QTLnet [22] did not develop the gene-based statistics, in Fig. 3 we did not present the results of QTLnet algorithm. Simulation were conducted for two settings: ten phenotypes with ten genes (ten SNPs for each gene), and 30 phenotypes with 100 genes (ten SNPs for each gene). We observed that in all three cases: common, rare and both common and rare variants, the gene-based FSEM had much higher power and smaller FDR than the SNP-based SEMs. It is interesting to observe that even if for the rare variants the gene-based method can reach the power as high as 85 % when sample sizes were larger than 3000.

Application to real data examples

To evaluate its performance, we applied the sparse functional SEMs with a gene as a unit of analysis to a sample of 1011 European-Americans (EA) with complete exome sequencing (total of 1,861,447 common and rare variants, 18,025 genes, of which, 5288 genes were mapped to 259 pathways downloaded from the KEGG database) and 11 phenotypes: high density lipoprotein cholesterol (HDL), low density lipoprotein cholesterol (LDL), triglyceride (Trig) and total cholesterol (TotChol), fast glucose, systolic blood pressure (SBP), diastolic blood pressure, body mass index (BMI), fastinsulin, Fibrinogen, and platelet count (PLATELET) (no missing phenotype data). Inverse rank normal transformation of the phenotypes was used in the analysis.

The analysis consisted of two stages. At the first stage, the sparse functional SEMs were applied to each of the 259 KEGG pathways and 11 phenotypes to infer genotype-phenotype networks. The remaining 12,737 genes which were not mapped to KEGG pathways were divided into 100 groups according to the order of

Table 2 The computation time for S2SEM and QTLnet methods of one replicate for the simulations of common variants

Time(s)/replicate	Phenotypes 10, Common SNPs 30		Phenotypes 30, Common SNPs 100		Phenotypes 100, Common SNPs 1000
Sample	S2SEM	QTLnet	S2SEM	QTLnet	S2SEM
100	1.36	1320.652	27.7625	12027.462	273.873
300	1.81063	1398.773	14.7055	12700.457	123.936
500	2.37817	1476.835	19.29666	13373.882	143.593
800	3.45897	1554.466	26.23715	14047.312	170.799
1100	4.70929	1632.869	36.27526	14720.764	309.795
1400	5.89862	1710.212	43.54272	15394.201	218.933
1700	7.49375	1788.496	55.96197	16067.649	285.213
2000	9.03683	1866.751	64.34067	16741.084	311.480
2300	11.21737	1944.063	78.7097	17414.523	375.167
2600	12.93615	2022.158	87.91563	18087.964	625.959
2900	15.51082	2100.947	105.14222	18761.402	724.829

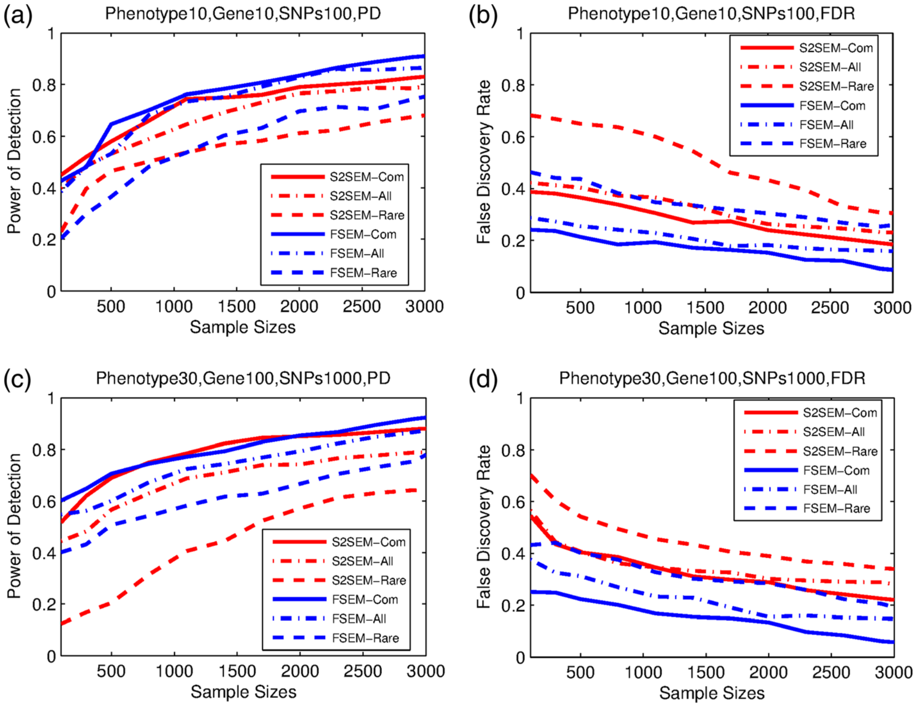

Fig. 3 Performance of gene-based S2SEM and FSEM. The power and FDR of the two methods for inference of genotype-phenotype networks in two different settings, (**a**,**b**) is results for 10 phenoytypes and 10 genes which include 100 SNPs, (**c**,**d**) is results for 30 phenotypes and 100 genes which include 1000 SNPs

chromosomes. Again, the sparse functional SEMs were applied to each group of genes and 11 phenotypes. We identified 1789 genes with *P*-values for testing path coefficients < 0.05 from the analysis at the first stage. To dissect pleiotropic genetic structure, at the second stage, we select 142 genes that were connected with more than one phenotype for further analysis. The sparse functional SEMs were applied to the selected 142 genes and 11 phenotypes to infer genotype-phenotype networks. To improve the accuracy of estimation, a stability selection procedure was used to infer the structure of the network. In other words, we randomly resampled data and estimated the genotype-phenotype networks 100 times. We only selected arrows when their *P*-values for testing the path-coefficients were less than 0.05 and they were present in the estimated network more than 80 times, i.e., the probability for each arrow to be selected was more than 0.8. We identified a genotype-phenotype network with 137 genes directly connected to 11 phenotypes and 341 edges. One hundred fourteen genes out of 137 genes showed pleiotropic genetic effects. The results were presented in Fig. 4. Additional file 2: Table S1 shows the path coefficients and *p*-values according to network in Fig. 4. We observed that the most causal relationships among phenotypes had *P*-values < 10^{-7} and stability ~1. This showed that the inference about phenotype sub-network is highly reliable. We also observed that large proportion of the edges in the phenotype network had two directions. This demonstrated that the SEMs had limitations for inferring causal networks.

We observed that *PIK3R5* directly affected seven phenotypes, *HS1BP3* directly affected five phenotypes, 11 genes directly affected four phenotypes, and 33 genes directly affected three phenotypes and remaining 102 genes directly affected two phenotypes. To assess the roles of path analysis in detecting genetic pleiotropic effects, we presented Table 3 that summarizes the *P*-values of three genes that affecting more than four phenotypes for path coefficients, the marginal effects of single and multiple traits (simple regression and multiple regression), and the minimum of *P*-values derived from principal component analysis (PCA) based regression. Table 3 showed that the most *P*-values for path coefficient were less than that for the marginal effect of corresponding single trait. Estimation of the marginal genetic effect of single trait only explores information of the target trait and genetic variants. However, estimation of the path coefficient uses information of all the relevant traits and genetic variants. This implies that path analysis has higher power to detect genetic risk variants than the traditional marginal analysis. From Table 3 we also observed that in general, each gene had at least one path with *P*-value for path coefficient was less or close to that

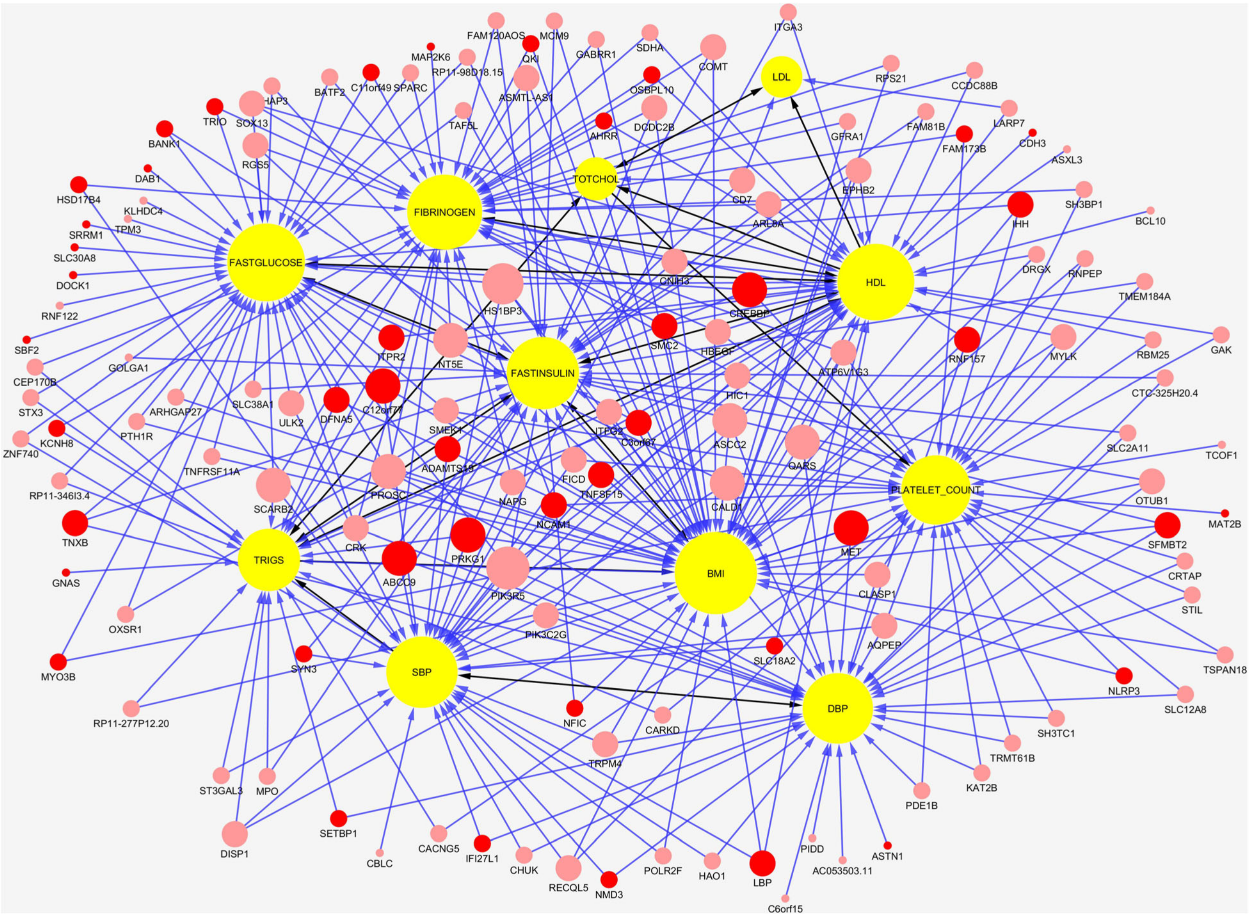

Fig. 4 A genotype-phenotype network consisted of 137 genes and 11 phenotypes. One hundred fourteen genes of all the gene nodes showed pleiotropic genetic effect. The nodes in yellow color represented the phenotypes, the nodes in light red color represented genes influencing phenotype variation, the nodes in the red color represented genes from our network were reported to be associated with 11 phenotypes or cardiovascular diseases phenotypes, the black arrows indicated the causal relations between phenotypes, the blue arrows indicted the contribution of the gene to one phenotype

for marginal effects of multiple relevant traits or their PCA analysis. Additional file 3: Table S2 summarizes the results for all the 13 genes that connected to more than 4 phenotypes.

SEMs provide a powerful tool to distinguish four types of effects: direct, indirect, total and marginal (estimated by a simple regression) effects. Additional file 4: Table S3 summarized the direct, indirect, total and marginal effects of one variable (phenotype or gene) that was referred to as the causal on another variable (phenotype) that was referred to as outcome for all 11 phenotypes and 137 genes in the Fig. 4. Investigating each type of effect allows a more comprehensive understanding of the relationship between variables.

Additional file 4: Table S3 listed the total 1414 pairs of causal relations between variables. We observed 343 (24.3 %) pairs of relations with direct effects, 1283 (90.7 %) pairs of relations with indirect effects and 212 (15 %) pairs of relations with both direct and indirect effects (Table 4 showed examples of pairs with both direct and indirect effects). This implied that the most effects are indirect effects due to mediation. In the quantitative trait locus (QTL) analysis, we often identify QTL by testing association of the marginal effect with the single trait. The SEMs provide complimentary information about path coefficients. In Table 5 we listed 25 tests in which the P-values for testing path coefficients were smaller than that for testing the marginal effects (coefficient of simple regression model, SRG) and 25 tests in which the P-values for testing marginal effects were smaller than that for testing the path coefficients. This showed that using SEMs for path analysis will discover additional QTLs that may be missed by marginal association analysis. In theory, the total effect of the causal X on outcome Y is equal to the summation of the product of the path coefficients along all possible paths between X and Y [34]. In the previous section, the total effect is defined as the coefficient

Table 3 *P*-values for four different effect tests

Outcome	Causal	Stability	*P*-value for path coefficient	*P*-value (Single trait marginal)	Min (*P*-value) PCA	*P*-value (Multiple traits marginal)
					1.44E-03	1.57E-03
SBP	PIK3R5	1	8.38E-05	9.25E-02		
DBP	PIK3R5	1	1.20E-03	7.62E-02		
TRIGS	PIK3R5	0.9	5.59E-03	1.07E-01		
FASTGLUCOSE	PIK3R5	0.9	1.31E-02	2.01E-01		
BMI	PIK3R5	0.88	1.63E-02	2.61E-01		
FASTINSULIN	PIK3R5	0.92	3.00E-02	7.94E-01		
HDL	PIK3R5	0.81	3.84E-02	2.39E-01		
					2.17E-04	2.43E-04
DBP	HS1BP3	0.96	1.21E-04	1.58E-01		
HDL	HS1BP3	0.98	6.78E-04	8.15E-04		
SBP	HS1BP3	0.87	6.08E-03	9.10E-01		
BMI	HS1BP3	0.94	1.71E-02	1.91E-01		
FASTGLUCOSE	HS1BP3	0.91	1.99E-02	5.95E-02		
					4.79E-04	5.24E-04
DBP	ABCC9	1	5.30E-06	6.49E-02		
SBP	ABCC9	0.94	2.56E-04	4.63E-01		
FASTINSULIN	ABCC9	0.95	2.71E-03	2.10E-03		
FIBRINOGEN	ABCC9	0.96	5.52E-03	7.71E-02		

The tested *p*-value for the path coefficient, mariginal effects of single trait and multiple traits, and minimum of *P*-values from PCA analysis (example of three genes that connected to more than four phenotypes)

Table 4 An example of 20 pairs of variables that had both direct and indirect effects

Outcome	Causal	Direct effect	Indirect effect	Total effect	Marginal effect
BMI	ATP6V1G3	−0.5701	0.0107	−0.5594	−0.5360
BMI	C12orf77	−1.3704	−0.2745	−1.6449	−1.7371
BMI	EPHB2	−0.2627	0.0544	−0.2083	−0.2017
BMI	PIK3R5	0.0631	−0.0119	0.0512	0.0541
BMI	RPS21	0.7630	0.0123	0.7754	0.8282
DBP	CHUK	0.0553	−0.0355	0.0199	0.0213
FASTGLUCOSE	C11orf49	0.6186	0.0146	0.6332	0.5846
FASTGLUCOSE	C12orf77	−1.2863	−0.2366	−1.5228	−1.5121
FASTGLUCOSE	PIK3R5	0.0674	−0.0133	0.0541	0.0530
FASTGLUCOSE	TAF5L	0.2911	0.0196	0.3107	0.2942
FASTINSULIN	SFMBT2	0.2874	−0.0898	0.1975	0.1930
FIBRINOGEN	FAM120AOS	−0.1550	0.0033	−0.1516	−0.1597
HDL	CREBBP	0.1168	−0.0176	0.0992	0.1097
HDL	ITPR2	0.0946	−0.0542	0.0404	0.0434
LDL	TOTCHOL	0.9458	0.0161	0.9619	0.9398
SBP	AQPEP	−0.0425	0.0282	−0.0143	−0.0156
SBP	CHUK	−0.0595	0.0335	−0.0261	−0.0272
SBP	MET	−0.0596	0.0376	−0.0220	−0.0212

Table 5 Twenty-five pairs of *P*-values for testing path coefficients and marginal effects, respectively

		P-value				*P*-value	
Outcome	Causal	Path coeff	Marginal effect	Outcome	Causal	Path coeff	Marginal effect
DBP	ABCC9	5.30E-06	6.49E-02	HDL	ITFG2	4.42E-04	1.98E-05
DBP	TRPM4	9.51E-06	7.24E-02	TRIGS	ST3GAL3	2.21E-04	1.16E-04
BMI	POLR2F	1.14E-05	3.93E-04	PLATELET	CRTAP	3.79E-04	1.58E-04
PLATELET	TRMT61B	1.17E-05	1.02E-04	HDL	ASCC2	5.58E-04	1.99E-04
SBP	CHUK	1.44E-05	1.76E-01	FASTGLUCOSE	PTH1R	5.72E-03	2.00E-04
DBP	QARS	2.00E-05	8.82E-03	PLATELET	PDE1B	4.61E-04	3.02E-04
PLATELET	TCOF1	2.10E-05	1.36E-02	HDL	HS1BP3	6.78E-04	3.56E-04
FIBRINOGEN	SMC2	2.33E-05	1.06E-03	TRIGS	ITFG2	7.32E-03	3.81E-04
HDL	DRGX	3.20E-05	3.22E-04	HDL	NT5E	3.26E-02	5.86E-04
DBP	CHUK	3.33E-05	2.88E-01	HDL	EPHB2	4.30E-03	6.33E-04
FIBRINOGEN	C11orf49	4.58E-05	5.10E-05	FIBRINOGEN	CEP170B	4.71E-03	7.04E-04
LDL	LARP7	4.58E-05	3.33E-01	DBP	PIDD	1.51E-03	7.53E-04
LDL	CD7	5.16E-05	7.68E-02	HDL	HBEGF	3.43E-03	8.44E-04
FIBRINOGEN	CRK	8.66E-05	1.45E-03	FASTGLUCOSE	RNF122	4.29E-02	9.17E-04
PLATELET	MYLK	9.80E-05	1.14E-03	DBP	AC053503.11	5.62E-03	1.08E-03
SBP	STX3	1.07E-04	1.60E-03	FIBRINOGEN	CNIH3	1.07E-02	1.13E-03
FASTINSULIN	MPO	1.09E-04	3.22E-03	BMI	NFIC	1.74E-03	1.15E-03
FIBRINOGEN	OSBPL10	1.18E-04	3.99E-03	FASTGLUCOSE	IHH	4.66E-02	1.17E-03
SBP	ST3GAL3	1.56E-04	1.99E-03	FASTGLUCOSE	ARHGAP27	1.87E-03	1.30E-03
HDL	LARP7	1.60E-04	3.31E-02	SBP	SLC18A2	1.42E-02	1.44E-03
HDL	SOX13	1.60E-04	6.78E-03	BMI	QKI	1.93E-03	1.61E-03
PLATELET	SH3TC1	2.18E-04	1.25E-02	FASTGLUCOSE	SLC38A1	1.44E-02	1.65E-03
SBP	ABCC9	2.56E-04	4.63E-01	SBP	ZNF740	4.68E-03	1.81E-03
BMI	DCDC2B	2.64E-04	8.70E-02	TRIGS	ADAMTS19	2.31E-02	1.82E-03
SBP	TRPM4	2.71E-04	2.19E-01	PLATELET	GAK	1.54E-02	2.02E-03

$\beta_{YX.Parent_X}$ of X in the linear regression of Y on X and its parent set. Let $Z = Parent_x$. The total effect $\beta_{YX.Z}$ can be expressed by [34]

$$\beta_{YX.Z} = \beta_{YX} \frac{\frac{1-\rho_{YZ}\rho_{ZX}}{\rho_{YX}}}{1-\rho_{XZ}{}^2}. \quad (19)$$

Since causal relations between SNPs do not exist, any SNP does not have its parent, i.e., the set $Z = \phi$ is empty. Therefore, for the SNP or gene X, we have $\beta_{YX.Z} = \beta_{YX}$. For example, the estimator of the direct effect of gene *MET* on the phenotype SBP was −0.0596. *MET* also had path *MET* → DBP → SBP. The indirect effect of *MET* on SBP was 0.0621 × 0.605 = 0.0376. Thus, the total effect of *MET* on SBP was −0.022. The marginal effect β_{YX} of *MET* on SBP estimated by SRG was −0.0212. The total effect of *MET* on SBP estimated from Fig. 4 was close to the marginal effect of *MET* on SBP. This example showed that if the causal relationships among the variables were completely captured by a DAG, the total effect and marginal effect were almost equal. Therefore, in the genotype-phenotype estimation process, we can use the relationship between the total and marginal effects to check whether the causal relationship modeled by a DAG is complete.

Multiple SNPs within a gene jointly have significant genetic effects, but individually each SNP make mild contributions to the phenotype variation. Table 6 listed *P*-values of 22 SNPs in seven genes for testing the path coefficients. We observed that single SNP made only a mild contribution to the direct effect, the multiple SNPs made significant contributions to the phenotype variation. This showed that the gene-based genotype-phenotype inference had higher power than the single SNP-based genotype-phenotype inference.

Since the most existing methods for genotype-phenotype network estimation only take a single SNP as a variable (unite of analysis) and cannot take a gene as a unite of analysis, next we illustrate the application of S2SEMs

Table 6 *P*-values of 22 SNPs in seven genes for testing path coefficients

Phenotype				*P*-value testing path coef	
	Gene	Chr	SNP position	Gene	SNP
FASTGLUCOSE	SEMA3B	3	50310922	5.98E-06	6.41E-05
FASTGLUCOSE	DNAJC16	1	15873386	1.09E-05	5.62E-03
FASTGLUCOSE	DNAJC16	1	15874961		3.61E-03
FASTGLUCOSE	DNAJC16	1	15905501		3.94E-02
DBP	OBSCN	1	228404668	1.48E-05	8.81E-02
DBP	OBSCN	1	228461187		9.96E-02
DBP	OBSCN	1	228482028		9.68E-02
DBP	OBSCN	1	228496066		8.61E-02
DBP	OBSCN	1	228503711		6.52E-04
DBP	OBSCN	1	228565208		2.66E-03
DBP	OBSCN	1	228565445		8.28E-02
HDL	SOX13	1	204085609	4.31E-05	2.45E-02
HDL	SOX13	1	204092129		9.14E-04
HDL	SOX13	1	204094963		9.40E-02
HDL	SOX13	1	204095220		3.56E-02
HDL	SOX13	1	204095280		3.79E-02
HDL	SRRM5	19	44099538	4.74E-05	6.37E-02
HDL	SRRM5	19	44111890		2.21E-05
FIBRINOGEN	SLC45A4	8	142225990	1.63E-04	3.97E-02
FIBRINOGEN	SLC45A4	8	142226108		1.61E-02
FIBRINOGEN	SLC45A4	8	142228909		8.87E-04
FIBRINOGEN	LHFPL2	5	77784738	5.95E-06	4.21E-06

for inference of genotype-phenotype network using SNPs and compared their results with that of QTL-driven phenotype network method (QTLnet) [22]. The number of SNPs in 137 genes was 5482. Due to the limitation of the size of the genotype-phenotype network which the sparse multivariate SEMs can estimate, from 137 genes in Fig. 4 we selected 45 genes that were reported to be associated with the 11 phenotypes in the analysis or other cardiovascular disease (CVD) related phenotypes in the literature. A total of 1993 SNPs in the 45 genes (248 common and 1745 rare SNPs) were included in the analysis.

The gene-based genotype-phenotype network with 55 nodes (11 phenotypes and 44 genes) and 110 edges estimated using the selected 45 genes and the FSEM method was shown in Fig. 5. S2SEM can also be used to estimate gene-based genotype-phenotype network. The procedures were as follows. At the first stage, S2SEM method and all 1993 SNPs were used to estimate the SNP-phenotype network (Fig. 6) where a gene was connected to a phenotype if the minimum of *P*-values for the coefficients of all the paths connecting SNPs within a gene and a phenotype was less than 0.05. At the second stage, we used Bonferroni correction to adjust *P*-values for multiple tests. In Fig. 7, we plotted the estimated gene-phenotype network with 17 nodes (11 phenotypes and six genes) and 22 edges using the selected 45 genes and the gene-based S2SEM method where a gene was connected to a phenotype if the Bonferroni correction adjusted *P*-values for path coefficients connecting gene and phenotype was less than 0.05. Figures 5 and 7 showed that the gene-based FSEM method can identify much more genes influencing phenotypes than the gene-based S2SEM method.

Next we study the SNP-based genotype-phenotype network estimation using the S2SEM method. In other words, we connected genes to the phenotypes using the minimum of *P*-values for the coefficients of all the paths that connect SNPs within a gene and a phenotype without Bonferroni correction. Figure 6 plotted the estimated genotype-phenotype network with 42 nodes (11 phenotypes and 31 genes) and 78 edges using S2SEM method and 1993 SNPs in the 45 genes. The path coefficients and *P*-values (<0.05) for the path coefficients of the edges connecting the SNPs in the gene to the phenotypes were summarized in Additional file 5: Table S4. In Additional file 1: Figure S5, we plotted the estimated genotype-phenotype network with 13 nodes (ten connected phenotypes, one isolated phenotype and two genes) and 20 edges using QTLnet method. In Additional file 6: Table S5, we listed the edges of the estimated network which connect the genes and phenotypes using QTLnet method. While the QTLnet method only identified two genes: *LBP* connected to the phenotypes TRIGS and TOTCHOL, and *DOCK1* connected to the phenotype HDL, the SNP-based and gene-based S2SEM method, respectively, discovered 31 and six genes connected to phenotypes. These results showed that all proposed SEM methods including FSEM, gene-based and SNP-based S2SEM methods outperform the QTLnet method.

Similar to the gene-based FSEM method, we observed several remarkable features from these results obtained by the S2SEM method. First, we observed three SNPs that showed pleiotropic genetic effects (rs138251768 in the gene *ADAMTS19* effected SBP and DBP, rs116623954 in the gene *CNIH3* affected FASTINSULIN and FIBRINOGEN, rs13223756 in the gene *MET* affected SBP and DBP). Second, multiple SNPs in the same gene affected the same phenotype. Three SNPs: rs754555, rs754554 and rs754553 in the gene *DFNA5* jointly affected BMI, two SNPs: rs11017658 and rs61758438 in the gene *DOCK1* jointly affected SBP. Third, the pleiotropic effects of the gene were due to different SNPs. The SNPs: rs564665 and rs141647150 in the gene *DAB1* affected phenotypes DBP and FASTGLUCOSE, respectively; the SNPs:

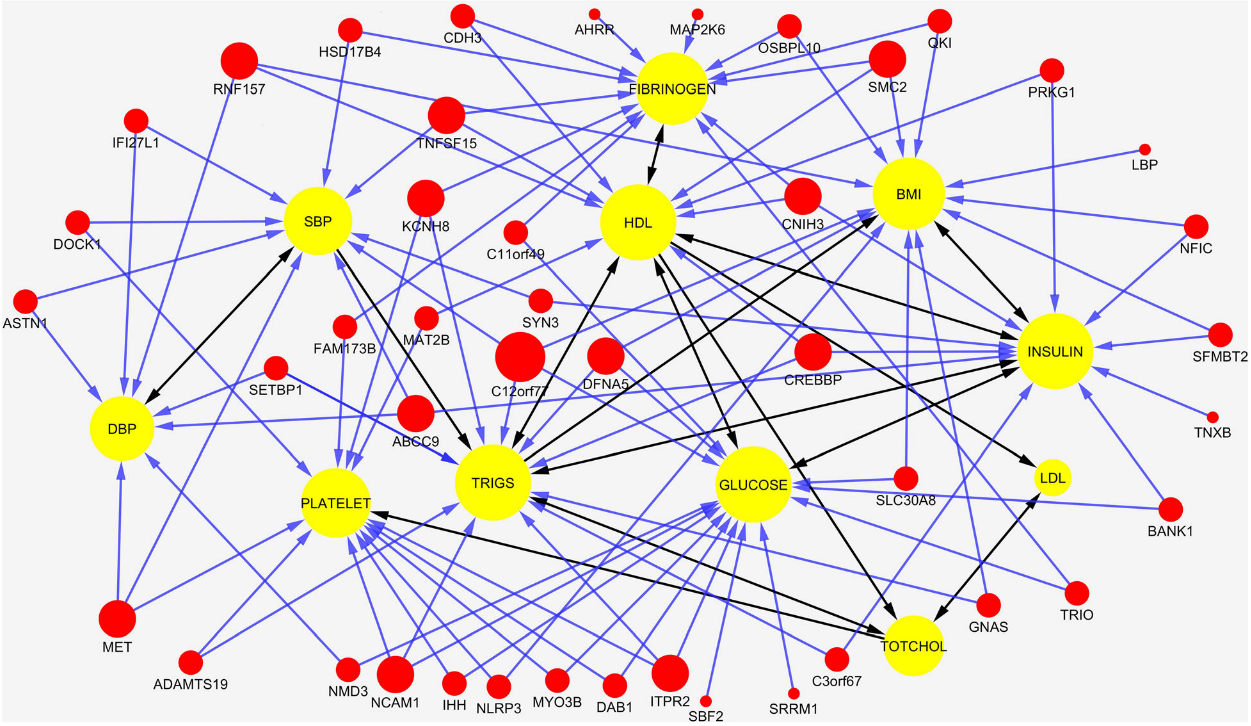

Fig. 5 A genotype-phenotype network consisted of 44 genes and 11 phenotypes. The network was constructed using FSEM from 45 genes. Nodes and edges are the same as described in Fig. 4

rs376043577 and rs3731878 in the gene *IHH* affected BMI and PLATELET COUNT, respectively; SNPs rs2232585 and rs2232605 in the gene *LBP* affected FIBRINOGEN and PLATELET COUNT, respectively. SNPs rs2305610, rs372123385 and rs17027957 in the gene *OSBPL10* affected BMI, DBP and FIBRINOGEN, respectively; three SNPs: rs144082896, rs140962261 and rs11547635 in the gene *SYN3* affected TRIGS, SBP and FASTINSULIN, respectively. You can find more examples from Additional file 5: Table S4 and Additional file 6: Table S5. Due to space limitation, they are omitted here.

In summary, we jointly estimated genetic architecture and phenotype network with 137 genes that were significantly connected to phenotypes. A total of 45 genes out of 137 genes were reported to be associated with 11 phenotypes or CVD related phenotypes, Additional file 7: Table S6 summarized the results of the reported 45 genes and their associated phenotypes. For the reported phenotypes, 6 phenotypes are from the analyzed 11 phenotypes. According to Fig. 4, Gene *SMC2* was connected with phenotypes: BMI, HDL and FIBRINOGEN. It was reported associated with HDL and BMI [45, 46], and also related with respiratory function and Echocardiography [47, 48]. Gene *RNF157* was connected with HDL, and it was reported associated with blood pressure [49] and HDL [45]. The other pairs of association for these six phenotypes were found through indirect paths from Fig. 4. For example, gene *DAB1*, *DFNA5* and *DOCK1* were reported associated with LDL [46], and there are indirect path from these genes to LDL according to Fig. 4. From these results we can summarized that our gene-based FSEMs has a rather high power to detect genetic pleiotropic effects, and it also provide a tool to decompose the effects into direct and indirect effects.

Discussion

Alternative to the standard marginal models for genetic association analysis of multiple correlated phenotypes, we have developed sparse SEMs and sparse FSEMs as a statistical framework for joint analysis of genetic architecture and causal phenotype network, which may emerge as a new generation of genetic analysis of multiple phenotypes exploring the causal network structures of the phenotypes. To facilitate using SEMs as a new paradigm for genetic analysis of multiple phenotypes, several issues have been addressed in this paper.

The first issue is to develop a unified framework for joint analysis of genetic architecture and causal phenotype network with both GWAS and the NGS data. The traditional multivariate SEMs can be applied to infer genotype-phenotype network with common variants, but are difficult to deal with rare variants. To overcome this limitation, we extend the multivariate SEMs to

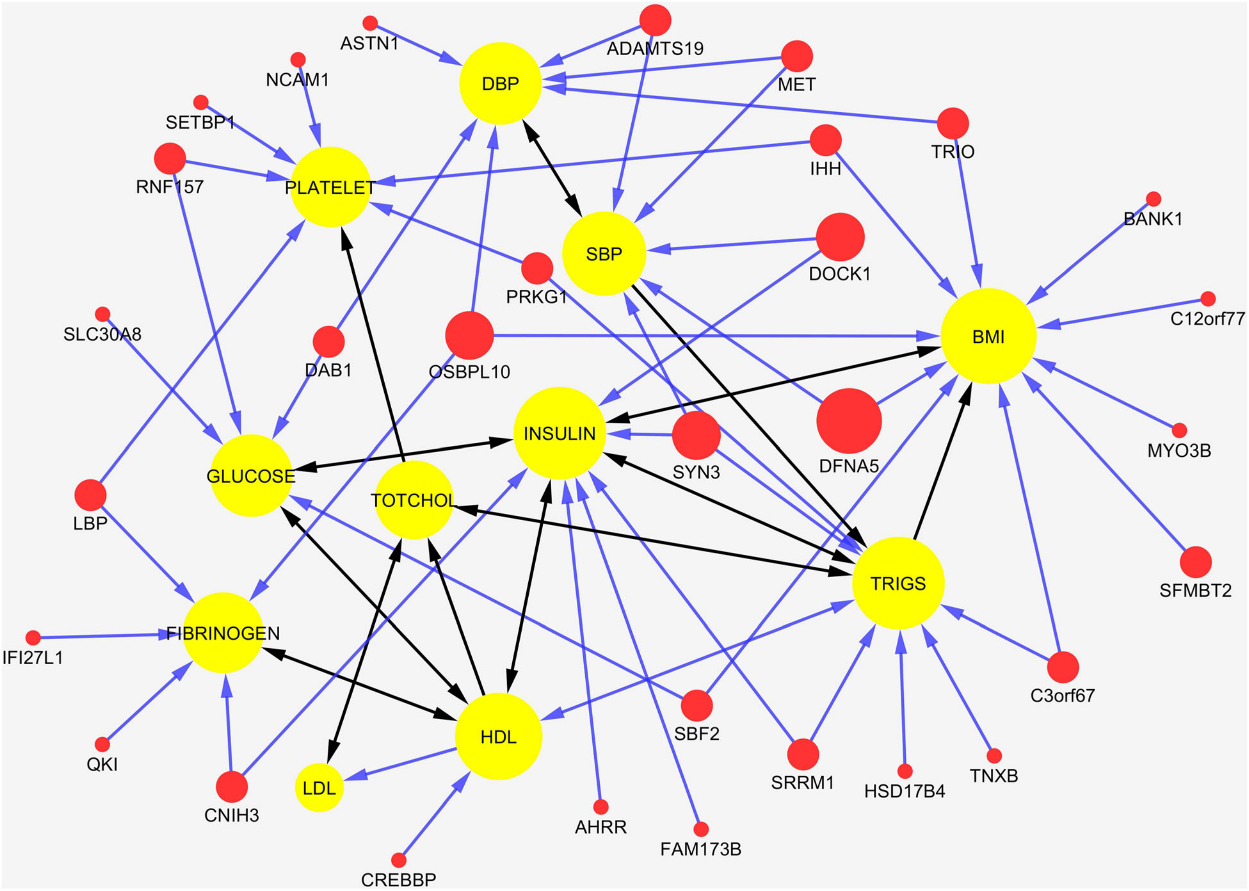

Fig. 6 A genotype-phenotype network consisted of 31 genes and 11 phenotypes. The network was constructed using SNP-based S2SEM method from 1993 SNPs of 45 genes. Nodes and edges are the same as described in Fig. 4

functional SEMs where exogenous genotype profiles across a genomic region or a gene are represented as a function of the genomic position for genetic analysis of multiple quantitative traits. In other words, we extend the variant-based genotype-phenotype network analysis to gene-based genotype-phenotype network analysis.

The second issue is how to develop statistical methods for jointly inferring genetic architecture and casual phenotype network structure. There is increasing consensus that the structure of the network in nature is sparse. However, the traditional estimation methods for the SEMs do not take the sparsity presented in the network into account. To solve this problem, we developed sparse SEMs and sparse functional SEMs to automatically incorporate the sparse condition into the estimation process. The widely used estimation method for the SEMs is the maximum likelihood method. However, the penalized maximum likelihood method and coordinate descent algorithms are not scalable to SEMs of high dimension. To overcome this limitation, we develop the ADMM-based sparse two-stage least square estimation method for the structure and parameter estimation of the SEMs. Our experience showed that the newly developed ADMM-based sparse two-stage least square estimation methods can infer networks with hundreds of nodes.

The third issue is the true structure discovery. An essential problem for the genotype-phenotype network analysis is to accurately estimate the network structure. By large scale simulations we showed that the true network structure can be accurately recovered with high probability. We also compared the performance of the sparse two-stage least square estimate methods with the QTLnet method. We demonstrated that for all the three cases (common, rare and both common and rare variants) our sparse two-stage SEMs (S2SEM) outperformed QTLnet method. Since the gene-based version of QTLnet method has not been developed we only compared the power and false discovery rates of the variant-based SEMs and gene-based functional SEMs. We found that for all spectrums of allele frequencies (common, rare and both common and rare variants) the gene-

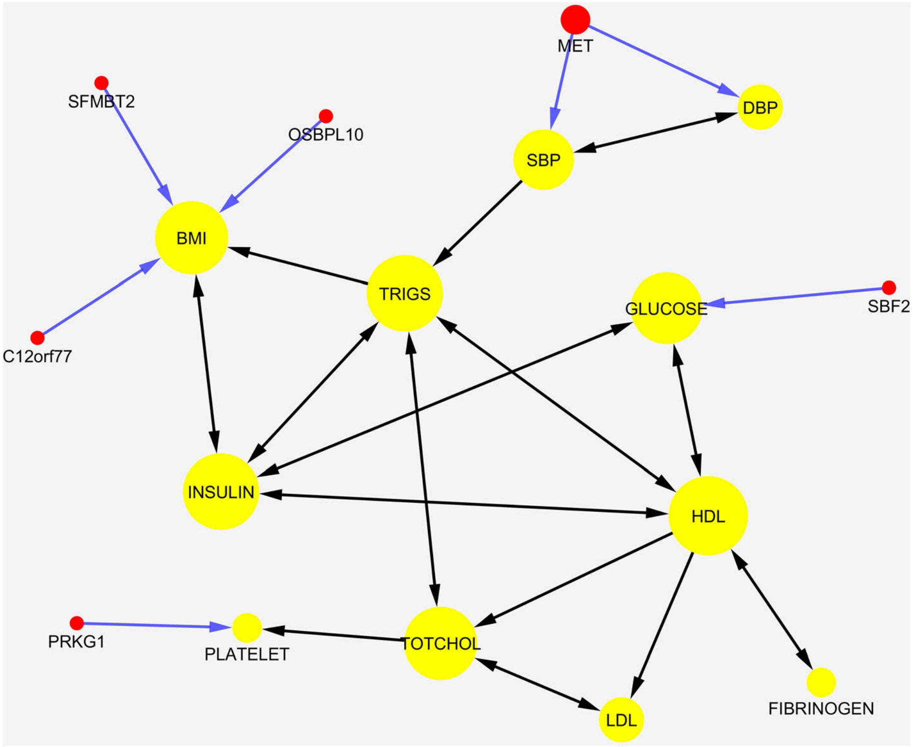

Fig. 7 A genotype-phenotype network consisted of six genes and 11 phenotypes. The network was constructed using the gene-based S2SEM method from 1993 SNPs of 45 genes where a gene was connected to a phenotype if the Bonferroni correction adjusted *P*-values for path coefficients connecting gene and phenotype was less than 0.05. Nodes and edges are the same as described in Fig. 4

based functional SEMs substantially outperformed the variant-based multivariate SEMs.

The fourth issue is how to distinguish four types of effects: direct, indirect, total and marginal effects. The current paradigm for genetic association analysis of multiple phenotypes is genetic marginal analysis in which the effects of the genetic variants on the phenotypes are estimated by regressing phenotypes on the genetic variants. This paradigm is unable to unravel the structure of the genotype-phenotype network and to estimate direct, indirect and total effects of the genetic variants on the phenotypes. The direct, indirect and total genetic effects provide valuable information for dissecting genetic structure of complex traits. We developed sparse SEMs and FSEMs as a causal inference tool to estimate direct, indirect and total genetic effects in addition to estimating marginal genetic effects. We observed that the most effects were indirect effects due to mediation. In traditional QTL analysis, we often identify QTL by testing association of the marginal effect with the single trait. The FSEMs and SEMs provide complimentary information about path coefficients. Interestingly, we found that many *P*-values for testing path coefficients were smaller than that for testing the marginal effects. This demonstrated that only using marginal association analysis we might miss identification of many significant QTLs.

The fifth issue is how to solve the large genotype-phenotype networks with up to hundreds of nodes or genes. A key to the large network inference is computation efficiency of the algorithms. Two strategies were employed to solve this problem. The first strategy was to reduce the dimension of data using functional data analysis. We first expand the genotype profiles in a genomic region (gene) in terms of orthonormal eigenfunctions. Genetic information across all variants in the genomic region including all single variant variation and their linkage disequilibrium is compressed into functional principal component scores. We use genetic information compressed into functional principal component scores to infer genotype-phenotype networks. The second strategy is to use ADMM algorithms to optimally solve the sparse SEM problem. The widely used algorithms for sparse SEMs are coordinate descent algorithms borrowed from the lasso originally designed for the sparse linear regression. The ADMM algorithms are parallel and efficient. Their convergence rates are fast. The ADMM algorithms allow inferring networks with hundreds or even thousands of nodes.

Major limitation of the SEMs for joint inference of genetic architecture and causal phenotype networks is the presence of two directions associated with one edge in the estimated network, which leads to a cyclic graph. To remove the cycles from the graph we need to strictly enforce the global constraint that the graph structure has to be acyclic. Such problems are often casted into a combinatorial optimization problem. We rank graph structures via a scoring metric that measure how well the DAG models fit the data. Combinatorial optimization algorithms are then used to search the optimal DAG with the best score [50].

Although their application to genome-wide genotype-phenotype network construction is difficult due to computational limitations, the SEMs are suitable to the phenome-wide association studies where starting phenomics, defined as the unbiased study of a large number of phenotypes in a population. We study the complex networks between multiple expressed phenotypes and genetic variants. Since the number of genetic variants in the phenome-wide association is quite limited and hence the size of the genotype-phenotype network is limited, the required computational time of construction of genotype-phenotype networks using SEMs is in the range the current computer system can reach. Advances in biosensors and sequencing technologies generate large amounts of phenotype and genetic data. SEMs and causal inference may emerge as a new paradigm of genetic studies of complex traits. The main purpose of this paper is to stimulate discussions about what are the optimal strategies to facilitate the development of a new generation of genetic analysis. We hope that our results will greatly increase the confidence in joint inference of genetic architecture and causal phenotype networks.

Conclusions

We have developed sparse SEMs and sparse FSEMs as a statistical framework for joint analysis of genetic architecture and causal phenotype network, which may emerge as a new generation of genetic analysis of multiple phenotypes. Our proposed sparse functional SEMs can incorporate both common and rare variants into the analysis and the ADMM algorithm can efficiently solve the penalized SEMs. Using this model we can jointly infer genetic architecture and casual phenotype network structure, and decompose the genetic effect into direct, indirect and total effect. Using large scale simulations we showed that the proposed methods have higher power to detect true causal genetic pleiotropic structure than other existing methods.

Appendix 1

Alternating direction method of multipliers for sparse SEMs

The optimization problem (7) can be further reduced to

$$\begin{aligned} &\min \quad f(\Delta_i) + \lambda ||Z_i||_1 \\ &\text{subject to} \quad \Delta_i - Z_i = 0. \end{aligned} \tag{20}$$

To solve the optimization problem (20), we form the augmented Lagrangian

$$\begin{aligned} L_\rho(\Delta_{i,} Z_i, \mu) = & f(\Delta_i) + \lambda ||Z_i||_1 + \mu^T(\Delta_i - Z_i) \\ & + \frac{\rho}{2}||\Delta_i - Z_i||_2^2. \end{aligned} \tag{21}$$

The alternating direction method of multipliers (ADMM) consists of the iterations:

$$\Delta_i{}^{(k+1)} :== \underset{\Delta_i}{\arg\min} \; L_\rho(\Delta_i, Z_i{}^{(k)}, \mu^{(k)}) \tag{22}$$

$$Z_i^{(k+1)} :== \underset{Z_i}{\arg\min} \; L_\rho\left(\Delta_i^{(k+1),}, Z_i, \mu^{(k)}\right) \tag{23}$$

$$\mu^{(k+1)} :== \mu^{(k+1)} + \rho\left(\Delta_i^{(k+1)} - Z_i^{(k+1)}\right), \tag{24}$$

where $\rho > 0$. Let $u = \frac{\mu}{\rho}$. Eq. (20, 21 and 22) can be reduced to

$$\Delta_i^{(k+1)} :== \underset{\Delta_i}{\arg\min} \left(f(\Delta_i) + \frac{\rho}{2}||\Delta_i - Z_i^{(k)} + u^{(k)}||_2^2\right) \tag{25}$$

$$Z_i^{(k+1)} :== \underset{Z_i}{\arg\min} \, (\lambda ||Z_i||_1 + \frac{\rho}{2}||\Delta_i^{(k+1)} - Z_i + u^{(k)}||_2^2) \tag{26}$$

$$u^{(k+)} :== u^{(k)} + \left(\Delta_i^{(k+1)} - Z_i^{(k+1)}\right). \tag{27}$$

Solving minimization problem (25), we obtain

$$\begin{aligned} \Delta_i^{(k+1)} = & \left[W_i^T X\left(X^T X\right)^{-1} X^T W_i + \rho I\right]^{-1} \\ & \left[W_i^T X\left(X^T X\right)^{-1} X^T y_i + \rho\left(Z_i^k - u^k\right)\right], \end{aligned}$$

which can be reduced to

$$\begin{aligned} \Delta_i^{(k+1)} = & \left[\frac{1}{\rho} I - \frac{1}{\rho} W_i^T X\left(\rho X^T X + X^T W_i W_i^T X\right)^{-1} X^T W_i\right] \\ & \left[W_i^T X\left(X^T X\right)^{-1} X^T y_i + \rho\left(Z_i^k - u^k\right)\right] \end{aligned} \tag{28}$$

The optimization problem (26) is non-differentiable. Although the first term in (26) is not differentiable, we still can obtain a simple closed-form solution to the problem (26) using subdiffenrential calculus [37]. Let Γ_j be a generalized derivative of the j-th component Z_i^j of the vector Z_i and $\Gamma = [\Gamma_1, \ldots, \Gamma_{M+K-1}]^T$ where

$$\Gamma_j = \begin{cases} 1 & Z_i^j > 0 \\ [-1, 1] & Z_i^j = 0 \\ -1 & Z_i^j < 0 \end{cases}$$

Then, we have

$$\frac{\lambda}{\rho}\Gamma + Z_i = \Delta_{i^{k+1}} + u^k,$$

which implies that

$$Z_i^{(k+1)} = \operatorname{sgn}\left(\Delta_i^{k+1} + u^k\right)\left(\left|\Delta_i^{k+1} + u^k\right| - \frac{\lambda}{\rho}\right)_+, \tag{29}$$

Where

$$|x|_+ = \begin{cases} x & x \geq 0 \\ 0 & x < 0. \end{cases}$$

Appendix 2

Estimation of parameters in the sparse structural functional equation models for the genotype-phenotype networks

Assume that the sparse SFEMs are given by

$$\begin{aligned}
&y_1\gamma_{11} + y_2\gamma_{21} + \ldots + y_M\gamma_{M1} + \int_{T_1} x_1(t)\beta_{11}(t)dt + \ldots + \int_{T_k} x_k(t)\beta_{k1}(t)dt + e_1 = 0 \\
&y1\gamma_{12} + y_2\gamma_{22} + \ldots + y_M\gamma_{M2} + \int_{T_1} x_1(t)\beta_{12}(t)dt + \ldots + \int_{T_k} x_k(t)\beta_{k2}(t)dt + e_2 = 0 \\
&\vdots \qquad\qquad\qquad\qquad\qquad \vdots \qquad\qquad\qquad\qquad\qquad \vdots \\
&y_1\gamma_{1M} + y_2\gamma_{2M} + \ldots + y_M\gamma_{MM} + \int_{T_1} x_1(t)\beta_{1M}(t)dt + \ldots + \int_{T_k} x_k(t)\beta_{kM}(t)dt + e_M = 0
\end{aligned} \tag{30}$$

For each genomic region or gene, we use functional principal component analysis to calculate principal component function [14]. We expand $x_{ij}(t), j = 1, 2, \ldots, k$ in each genomic region in terms of orthogonal principal component functions:

$$x_{ij}(t) = \sum_{l=1}^{L_j} \eta_{ijl}\phi_{jl}(t), j = 1, \ldots, k, \tag{31}$$

where $\phi_{jl}(t), j = 1, \ldots, k, l = 1, \ldots, L_j$ are the l-th principal component function in the j-th genomic region and η_{ijl} are the functional principal component scores of the i-th individual. Using the functional principal component expansion of $x_{ij}(t)$, we obtain

$$\begin{aligned}
\int_T x_{ij}(t)\beta_{jm}(t)dt &= \int_T \sum_{l=1}^{L_j} \eta_{ijl}\phi_{jl}(t)\beta_{jm}(t)dt \\
&= \sum_{l=1}^{L_j} \eta_{ijl}b_{jlm}, i = 1, \ldots, n, j \\
&= 1, \ldots, k, m = 1, \ldots, M.
\end{aligned} \tag{32}$$

Let $x_j(t) = [x_{1j}(t), \ldots, x_{nj}(t)]^T, \eta_{jl} = [\eta_{1jl}, \ldots, \eta_{njl}]^T$. Substituting eq. (32) into eq. (30), we obtain

$$\begin{aligned}
&y_1r_{11} + y_2r_{21} + \cdots + y_Mr_{M1} + \sum_{l=1}^{L_1}\eta_{1l}b_{1l1} + \cdots + \sum_{l=1}^{L_k}\eta_{kl}b_{kl1} + e_1 = 0 \\
&\vdots \qquad\qquad\qquad\qquad\qquad \vdots \qquad\qquad\qquad\qquad\qquad \vdots \\
&y_1r_{1M} + y_2r_{2M} + \cdots + y_Mr_{MM} + \sum_{l=1}^{L_1}\eta_{1l}b_{1lM} + \cdots + \sum_{l=1}^{L_k}\eta_{kl}b_{klM} \\
&\qquad + e_M = 0
\end{aligned} \tag{33}$$

Let $\eta = \left[\eta_{11}, \ldots, \eta_{1L_{1,\ldots,}}\eta_{k1}, \ldots, \eta_{kL_k}\right]$, $B = \begin{bmatrix} b_{111} & \cdots & b_{11M} \\ \vdots & \vdots & \vdots \\ b_{1L_11} & \cdots & b_{1L_1M} \\ \vdots & \vdots & \vdots \\ b_{k11} & \cdots & b_{k1M} \\ \vdots & \vdots & \vdots \\ b_{kL_k1} & \vdots & b_{kL_kM} \end{bmatrix}$ and Y, Γ and E be defined as before.

In matrix form, eq. (33) can be rewritten as

$$Y\Gamma + \eta B + E = 0, \tag{34}$$

which has the same form as the Eq. (2) has.

If we consider only one genomic region or gene, the matrices η and B will be reduced to $\eta = [\eta_1, \ldots, \eta_L]$ and $B = \begin{bmatrix} b_{11} \cdots b_{1M} \\ \vdots \quad \vdots \quad \vdots \\ b_{L1} \cdots b_{LM} \end{bmatrix}$.

If we take functional principal component scores as predictors, the models and algorithms for network structure and parameter estimation will be similar to that discussed in Appendix 1. Specifically, the i-th equation is given by

$$Y\Gamma_i + \eta B_i + e_i = 0,$$

which can be rewritten as

$$y_i = W_i\Delta_i + e_i, \tag{35}$$

where $W_i = [\,Y_{-i}\,\eta], \Delta_i = [\,\gamma_{-i}\,B_i]$.

Then, the sparse SFEMs are transformed to

$$\min_{\Delta_i} \; f(\Delta_i) + \lambda||\Delta_i||_1$$
$$\text{where}\, f(\Delta_i) = \left(\eta^T y_i - \eta^T W_i \Delta_i\right)^T \left(\eta^T \eta\right)^{-1} \left(\eta^T y_i - \eta^T W_i \Delta_i\right). \tag{36}$$

Finally, ADMM algorithms are given by

Algorithm:

For $i = 1,...,M$

Step 1. Initialization

$u^0 := 0$
$\Delta_i^0 := [W_i^T \eta(\eta^T \eta)^{-1} \eta^T W_i]^{-1} W_i^T \eta(\eta^T \eta)^{-1} \eta^T y_i$
$Z_i^0 := \Delta_i^0.$

Carry out steps 2,3 and 4 until convergence

Step 2.

$$\Delta_i^{(k+1)} := [\frac{1}{\rho} I - \frac{1}{\rho} W_i^T \eta(\rho\eta^T \eta + \eta^T W_i W_i^T \eta)^{-1} \eta^T W_i][W_i^T \eta(\eta^T \eta)^{-1} \eta^T y_i + \rho(Z_i^k - u^k)]$$

Step 3.

$$Z_i^{(k+1)} := \text{sgn}(\Delta_i^{k+1} + u^k)(|\,\Delta_i^{k+1} + u^k\,| - \frac{\lambda}{\rho})_+.$$

Step 4.

$$u^{(k+)} :== u^{(k)} + (\Delta_i^{(k+1)} - Z_i^{(k+1)}).$$

Additional files

Additional file 1: Figure S1. A scheme of genotype-phenotype network. **Figure S2**. Diagram associated with effect decomposition. **Figure S3**. Diagram of a simulation example for illustrating equation (12). **Figure S4**. An example for the simulated genotype-phenotype network. The network consisted of ten phenotype nodes and 30 genotype (SNP) nodes. **Figure S5**. Performance of S2SEM and SML for phenotype network inference. The power and FDR of the two methods for inference of phenotype networks when the phenotype and genotype number is 10 and 30 respectively. **Figure S6**. A genotype-phenotype network consisted of two genes that were reported to be associated with phenotypes in the analysis or other CVD related phenotypes in the literatures and ten phenotypes (one isolated phenotype didn't appear) estimated using QTLnet method. The nodes in yellow color represented the phenotypes, the nodes in the red color represented genes, the black arrows indicated the causal relations between phenotypes and the blue arrows indicted the contribution of the gene to one phenotype. (DOCX 4101 kb)

Additional file 2: Table S1. Path coefficients and *P*-values in Fig. 4. (XLSX 40 kb)

Additional file 3: Table S2. *P*-values for the path coefficient, mariginal effects of single trait and multiple traits, and minimum of *P*-values from PCA analysis for 13 genes that are connected to more than 4 phenotypes in Fig. 4. (XLSX 17 kb)

Additional file 4: Table S3. Direct, indirect, total and marginal effects for Fig. 4. (XLSX 94 kb)

Additional file 5: Table S4. *P*-value for the path coefficient (SNPs to Phenotype) of the genotype-phenotype networks with 31 genes and 11 phenotypes estimated using S2SEM method. (XLSX 13 kb)

Additional file 6: Table S5. SNPs to phenotype edges in the genotype-phenotype network estimated by QTL net method. (XLSX 8 kb)

Additional file 7: Table S6. Reported 45 genes out of 137 genes that are associated with the phenotypes in the analysis or other CVD related phenotypes. (XLSX 11 kb)

Abbreviations

ADMM: Alternative direction methods of multiplier, an algorithm that solves convex optimization problems; BMI: Body mass index; CVD: Cardiovascular disease; DAGs: Directed acyclic graphs; DBP: Diastolic blood pressure; ESP: Exome Sequencing Project; FDR: The false discovery rate; FPCs: Functional principal components; FSEMs: Functional structure equation models; HDL: High density lipoprotein cholesterol; LDL: Low density lipoprotein cholesterol; NGS: Next generation sequencing; PCA: Principal component analysis; PD: The power of detection; PLATELET: Platelet count; QTL: Quantitative traits locus; S2SEM: Sparse two-stage structure equation models; SBP: Systolic blood pressure; SML: Sparse maximum likelihood SEMs; SRG: Simple regression model; TotChol: Total cholesterol; Trig: Triglyceride; VLDL: Very low density lipoproteins

Acknowledgments

Mr. Xiong was supported by Grant 1R01AR057120–01 and 1R01HL106034-01, from the National Institutes of Health and NHLBI. Ms. Wang and Mr. Jin were supported by Grant 31330038 from the National Science Foundation of China and grant B13016 from the Programme of Introducing Talents of Discipline to Universities (111 Project). Ms. Wang was also supported by China Scholarship Council.
The authors thank two anonymous reviewers for the thorough reading of the manuscript and thoughtful suggestions that improved the manuscript.
The authors wish to acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff and study participants in creating this resource for biomedical research. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO) and RC2 HL-102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO).

Funding

NIH, National Heart, Lung, and Blood Institute (NHLBI), National Science Foundation of China, Programme of Introducing Talents of Discipline to Universities (111 Project). These funding agencies supported the design of the study, analysis, interpretation of data and writing the manuscript.

Availability of data and materials

Exome sequence data were generated from the NHLBI's Exome Sequencing Project (ESP) and have been deposited in dbGaP as part of the ESP cohort data. The data can be downloaded from dbGaP (https://esp.gs.washington.edu/drupal/dbGaP_Releases).

Authors' contributions

MX and PW proposed the model and drafted the manuscript. PW implemented the model and carried out most of the real data analysis. PW and MR completed the simulations together. LJ, MX and PW designed the analysis for real data application. All authors read and approved the final manuscript.

Competing interest

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details
[1]Human Genetics Center, Department of Biostatistics, University of Texas School of Public Health, Houston, TX 77030, USA. [2]State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China. [3]Human Genetics Center, The University of Texas Health Science Center at Houston, P.O. Box 20186, Houston, TX 77225, USA.

Received: 10 December 2015 Accepted: 18 October 2016
Published online: 07 November 2016

References

1. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010;34(2):188–93.
2. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83(3):311–21.
3. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5(2):e1000384.
4. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei L-J, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010;86(6):832–8.
5. Li Y, Byrnes AE, Li M. To identify associations with rare variants, just WHaIT: weighted haplotype and imputation-based tests. Am J Hum Genet. 2010;87(5):728–35.
6. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82–93.
7. Yi N, Zhi D. Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol. 2011;35(1):57–69.
8. Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010;70(1):42–54.
9. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. PLoS Genet. 2011;7(3):e1001322.
10. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet. 2011;7(2):e1001289.
11. Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet. 2010;6(10):e1001156.
12. Luo L, Boerwinkle E, Xiong M. Association studies for next-generation sequencing. Genome Res. 2011;21(7):1099–108.
13. Fan R, Wang Y, Mills JL, Wilson AF, Bailey-Wilson JE, Xiong M. Functional linear models for association analysis of quantitative traits. Genet Epidemiol. 2013;37(7):726–42.
14. Luo L, Zhu Y, Xiong M. Quantitative trait locus analysis for next-generation sequencing with the functional linear models. J Med Genet. 2012;49(8):513–24.
15. Stephens M. A unified framework for association analysis with multiple related phenotypes. PLoS ONE. 2013;8(7):e65245.
16. Aschard H, Vilhjálmsson BJ, Greliche N, Morange P-E, Trégouët D-A, Kraft P. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. Am J Hum Genet. 2014;94(5):662–76.
17. Schifano ED, Li L, Christiani DC, Lin X. Genome-wide association analysis for multiple continuous secondary phenotypes. Am J Hum Genet. 2013;92(5):744–59.
18. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014;11(4):407–9.
19. Bollen KA. Structural equations with latent variables. New York: John Wiley & Sons; 2014.
20. Lawson HA, Cady JE, Partridge C, Wolf JB, Semenkovich CF, Cheverud JM. Genetic effects at pleiotropic loci are context-dependent with consequences for the maintenance of genetic variation in populations. PLoS Genet. 2011;7(9):e1002256.
21. Rosa GJ, Valente BD, de Los Campos G, Wu X-L, Gianola D, Silva MA. Inferring causal phenotype networks using structural equation models. Genet Sel Evol. 2011;43(6). doi:10.1186/1297-9686-43-6
22. Neto EC, Keller MP, Attie AD, Yandell BS. Causal Graphical Models in Systems Genetics: A Unified Framework for Joint Inference of Causal Network and Genetic Architecture for Correlated Phenotypes. Ann Appl Stat. 2010;4(1):320–39.
23. Neto EC, Ferrara CT, Attie AD, Yandell BS. Inferring causal phenotype networks from segregating populations. Genetics. 2008;179(2):1089–100.
24. Rockman MV. Reverse engineering the genotype-phenotype map with natural genetic variation. Nature. 2008;456(7223):738–44.
25. Winrow CJ, Williams DL, Kasarskis A, Millstein J, Laposky AD, Yang HS, Mrazek K, Zhou L, Owens JR, Radzicki D, et al. Uncovering the genetic landscape for multiple sleep-wake traits. PLoS ONE. 2009;4(4):e5161.
26. Hageman RS, Leduc MS, Korstanje R, Paigen B, Churchill GA. A Bayesian framework for inference of the genotype-phenotype map for segregating populations. Genetics. 2011;187(4):1163–70.
27. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008;6(5):e107.
28. Li Y, Tesson BM, Churchill GA, Jansen RC. Critical reasoning on causal inference in genome-wide linkage and association studies. Trends Genet. 2010;26(12):493–8.
29. Duarte CW, Zeng ZB. High-confidence discovery of genetic network regulators in expression quantitative trait loci data. Genetics. 2011;187(3):955–64.
30. Mi X, Eskridge K, Wang D, Baenziger PS, Campbell BT, Gill KS, Dweikat I, Bovaird J. Regression-based multi-trait QTL mapping using a structural equation model. Stat Appl Genet Mol. 2010;9(1):1–23.
31. Valente BD, Rosa GJ, de Los CG, Gianola D, Silva MA. Searching for recursive causal structures in multivariate quantitative genetics mixed models. Genetics. 2010;185(2):633–44.
32. Li R, Tsaih S-W, Shockley K, Stylianou IM, Wergedal J, Paigen B, Churchill GA. Structural model analysis of multiple quantitative traits. PLoS Genet. 2006;2(7):e114.
33. Hauser A, Bühlmann P. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. J Mach Learn Res. 2012;13(1):2409–64.
34. Maathuis MH, Kalisch M, Bühlmann P. Estimating high-dimensional intervention effects from observational data. Ann Stat. 2009;37(6A):3133–64.
35. Kalisch M, Bühlmann P. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J Mach Learn Res. 2007;8:613–36.
36. Cai X, Bazerque JA, Giannakis GB. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. PLoS Comp Biol. 2013;9(5):e1003068.
37. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trend Mach Learn. 2011;3(1):1–122.
38. Judge GG, Hill RC, Griffiths WE, Lutkepohl H, Lee T-C. Introduction to the Theory and Practice of Econometrics. New York: Wiley; 1982.
39. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9(3):432–41.
40. Pearl J. Direct and indirect effects. In: Proceedings of the seventeenth conference on uncertainty in artificial intelligence: 2001. San Francisco: Morgan Kaufmann Publishers Inc; 2001. p. 411–20.
41. Pearl J. The deductive approach to causal inference. J Causal Infer. 2014;2(2):115–29.
42. Chen B, Pearl J. Graphical tools for linear structural equation modeling. In.: DTIC Document; 2014.
43. Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J Clin Epidemiol. 2014;67(8):850–7.
44. Veazie PJ. When to combine hypotheses and adjust for multiple tests. Health Serv Res. 2006;41(3p1):804–18.
45. Kathiresan S, Manning AK, Demissie S, D'Agostino RB, Surti A, Guiducci C, Gianniny L, Burtt NP, Melander O, Orho-Melander M, et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. BMC Med Genet. 2007;8 Suppl 1:S17.
46. Fox CS, Heard-Costa N, Cupples LA, Dupuis J, Vasan RS, Atwood LD. Genome-wide association to body mass index and waist circumference: the Framingham Heart Study 100K project. BMC Med Genet. 2007;8 Suppl 1:S18.
47. Wilk JB, Walter RE, Laramie JM, Gottlieb DJ, O'Connor GT. Framingham Heart Study genome-wide association: results for pulmonary function measures. BMC Med Genet. 2007;8 Suppl 1:S8.
48. Vasan RS, Larson MG, Aragam J, Wang TJ, Mitchell GF, Kathiresan S, Newton-Cheh C, Vita JA, Keyes MJ, O'Donnell CJ, et al. Genome-wide association of echocardiographic dimensions, brachial artery endothelial

function and treadmill exercise responses in the Framingham Heart Study. BMC Med Genet. 2007;8 Suppl 1:S2.

49. O'Donnell CJ, Cupples LA, D'Agostino RB, Fox CS, Hoffmann U, Hwang SJ, Ingellson E, Liu C, Murabito JM, Polak JF, et al. Genome-wide association study for subclinical atherosclerosis in major arterial territories in the NHLBI's Framingham Heart Study. BMC Med Genet. 2007;8 Suppl 1:S4.
50. Loh P-L, Bühlmann P. High-dimensional learning of linear causal networks via inverse covariance estimation. J Mach Learn Res. 2014;15(1):3065–105.