G3 Genes | Genomes | Genetics

GENOMIC PREDICTION

A Multiple-Trait Bayesian Variable Selection Regression Method for Integrating Phenotypic Causal Networks in Genome-Wide Association Studies

Zigui Wang,* Deborah Chapman,* Gota Morota,† and Hao Cheng*,[1]

*Department of Animal Science, University of California, Davis, and †Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University

ORCID IDs: 0000-0002-1472-593X (Z.W.); 0000-0002-3567-6911 (G.M.); 0000-0001-5146-7231 (H.C.)

ABSTRACT Bayesian regression methods that incorporate different mixture priors for marker effects are used in multi-trait genomic prediction. These methods can also be extended to genome-wide association studies (GWAS). In multiple-trait GWAS, incorporating the underlying causal structures among traits is essential for comprehensively understanding the relationship between genotypes and traits of interest. Therefore, we develop a GWAS methodology, SEM-Bayesian alphabet, which, by applying the structural equation model (SEM), can be used to incorporate causal structures into multi-trait Bayesian regression methods. SEM-Bayesian alphabet provides a more comprehensive understanding of the genotype-phenotype mapping than multi-trait GWAS by performing GWAS based on indirect, direct and overall marker effects. The superior performance of SEM-Bayesian alphabet was demonstrated by comparing its GWAS results with other similar multi-trait GWAS methods on real and simulated data. The software tool JWAS offers open-source routines to perform these analyses.

KEYWORDS
Structural Equation Models
Bayesian Regression
Variable Selection
GWAS
Genomic Prediction
GenPred
Shared data resources

Genome-wide association studies (GWAS) are widely used to identify associations between single nucleotide polymorphisms (SNPs) and phenotypes (Ozaki *et al.* 2002; Visscher *et al.* 2017; McCarthy *et al.* 2008; Cantor *et al.* 2010). GWAS have successfully mapped quantitative trait loci (QTL) associated with traits of interest, *e.g.*, meat quality and quantity in livestock (Sharma *et al.* 2015), crop yields in plants (Liu and Yan 2018), and diseases in humans (Visscher *et al.* 2017). GWAS are typically based on using linear mixed models to fit one SNP at a time to a single trait (Hackinger and Zeggini 2017). While this allows for a relatively simple statistical model, the interwoven nature of gene expression translates to many traits being correlated with each other (Sodini *et al.* 2018). These correlations can be utilized in multi-trait linear mixed models for GWAS to reduce false positives and increase the statistical power for association mapping (O'Reilly *et al.* 2012; Korte *et al.* 2012).

Conventional multi-trait linear mixed models do not consider the causal relationships between traits. To address this issue, researchers have proposed refining the multi-trait methods with structural equation models (SEM) introduced by Wright (1934) that consider the causal relationship among traits. A model that incorporates causal structures should better reflect underlying genetic mechanisms. Gianola and Sorensen (2004) used SEM to extend conventional multi-trait linear mixed models to accommodate for recursive and simultaneous relationships among traits, which allows traits to be explanatory variables for other traits. Recently, Momen *et al.* (2018, 2019) proposed the

doi: https://doi.org/10.1534/g3.120.401618
Manuscript received July 30, 2020; accepted for publication September 28, 2020; published Early Online October 5, 2020.

[1]Corresponding author: 2139 Meyer Hall, Department of Animal Science, University of California Davis, 1 Shield Avenue, Davis, CA 95616. E-mail: qtlcheng@ucdavis.edu

SEM-based GWAS (SEM-GWAS) methodology by applying SEM to linear mixed models for GWAS. They showed that while conventional GWAS methodology only provides overall SNP effects, SEM-GWAS can capture the complex causal relationships among traits and further decompose the overall SNP effects into direct and indirect effects.

The SEM-GWAS method proposed by Momen *et al.* (2018, 2019) is based on linear mixed models with a fixed substitution effect for the tested SNP and a random effect with covariances defined by a "genomic relationship matrix" computed from genotypes (VanRaden 2008) to account for genetic relatedness. Markers are usually implicitly assumed to affect all traits when the "genomic relationship matrix" is constructed in multi-trait analysis. However, this assumption is not biologically meaningful, especially in multi-trait analyses involving many traits. Cheng *et al.* (2018b) proposed a general class of multi-trait Bayesian variable selection regression methods that use a broad range of mixture priors, *e.g.*, multi-trait BayesCΠ, where each locus can affect any combination of traits, which allows us to more closely model the true biological mechanisms, *e.g.*, pleiotropy (Cheng *et al.* 2018b).

The primary goal of this current research is to develop a multi-trait Bayesian regression GWAS method that more closely resembles the underlying biological mechanisms including pleiotropy and causal structure among traits. In this paper, we develop and implement a new GWAS method called SEM-Bayesian alphabet, which integrates SEM to the multi-trait Bayesian variable selection methods, to incorporate the underlying biological mechanism. The term "Bayesian alphabet" denotes a collection of Bayesian regression models that differ in the priors adopted for marker effects (Gianola 2013). In this paper, we use SEM-BayesCΠ, a Bayesian variable selection method, to show the utility of the SEM-Bayesian alphabet. The performance of our proposed method is studied using real and simulated data.

MATERIALS AND METHODS

Multi-trait Bayesian regression model using mixture priors

Assuming that individuals have all traits measured with a general mean as the only fixed effect, we write the multi-trait model for individual i from n genotyped individuals as:

$$\boldsymbol{y}_i = \boldsymbol{\mu} + \sum_{j=1}^{p} m_{ij}\boldsymbol{\alpha}_j + \boldsymbol{e}_i$$

where $\boldsymbol{y}_i$ is the vector of phenotypes of t traits for individual i, $\boldsymbol{\mu}$ is a vector of overall means for t traits, p is the number of genotyped loci, m_{ij} is the genotype covariate at locus j for individual i (coded as 0,1,2), $\boldsymbol{\alpha}_j$ is the vector of marker effects of t traits for locus j, and $\boldsymbol{e}_i$ is the vector of residuals of t traits for individual i. The fixed effects are assigned flat priors. The residuals, $\boldsymbol{e}_i$, are a priori assumed to be independently and identically distributed multivariate normal vectors with null mean and covariance matrix $\boldsymbol{R}$, which is assumed to have an inverse Wishart prior distribution, $W^{-1}(\boldsymbol{S}_e, \nu_e)$.

Allowing each locus to affect any combination of traits, in a multiple-trait Bayesian variable selection method, *e.g.*, multi-trait BayesCΠ (Cheng *et al.* 2018b), the vector of marker effects at locus j can be written as $\boldsymbol{\alpha}_j = \boldsymbol{D}_j\boldsymbol{\beta}_j$, where $\boldsymbol{D}_j$ is a diagonal matrix whose diagonal element is $\boldsymbol{\delta}_j = (\delta_{j1}, \delta_{j2} \ldots, \delta_{jt})$, where δ_{jk} is the indicator variable indicating whether the marker effect of locus j for trait k is zero or not, and $\boldsymbol{\beta}_j$ is a priori assumed to be independently and identically distributed multivariate normal vectors with null mean and covariance matrix $\boldsymbol{G}$, which is assumed to have an inverse Wishart prior distribution, $W_t^{-1}(\boldsymbol{S}_\beta, \nu_\beta)$. Given that a locus can have an effect on any combination of traits, we use numeric labels $''1'', ''2'', \cdots, ''l''$ to represent all 2^t possible combinations for $\boldsymbol{\delta}_j$, in which case the prior distribution for $\boldsymbol{\delta}_j$ is:

$$p(\boldsymbol{\delta}_j = ''i'') = \Pi_1 I(\boldsymbol{\delta}_j = ''1'') + \Pi_2 I(\boldsymbol{\delta}_j = ''2'') + \ldots + \Pi_l I(\boldsymbol{\delta}_j = ''l'')$$

where Π_i is the prior probability that the vector $\boldsymbol{\delta}_j$ corresponds to the vector labeled "i" and $\sum \Pi_i = 1$. We assume the prior for $\Pi = (\Pi_1, \Pi_2, \ldots \Pi_l)$ is a uniform distribution.

Structural Equation Model

The linear SEM is composed of two parts: the measurement equation analyzing the relationship between the observable variables and latent variables, and the structural equation capturing the connections among latent variables (Anderson and Gerbing 1988). These two equations can be written as:

$$\begin{cases} \boldsymbol{y}_i = \boldsymbol{\Lambda}\begin{pmatrix}\boldsymbol{\eta}_i \\ \boldsymbol{\xi}_i\end{pmatrix} + \boldsymbol{\kappa}_1 & \text{measurement equation} \\ \boldsymbol{\eta}_i = \boldsymbol{\Gamma}_1\boldsymbol{\eta}_i + \boldsymbol{\Gamma}_2\boldsymbol{\xi}_i + \boldsymbol{\kappa}_2 & \text{structural equation} \end{cases}$$

where $\boldsymbol{y}_i$ is the vector of observable variables for individual i, $\boldsymbol{\eta}_i$ is a $q \times 1$ vector of endogenous latent variables, $\boldsymbol{\xi}_i$ is a $r \times 1$ vector with exogenous latent variables, $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ are the matrix of unknown coefficients in structural equation, $\boldsymbol{\Lambda}$ is a $t \times (q + r)$ matrix of unknown structural coefficients, $\boldsymbol{\kappa}_1$ and $\boldsymbol{\kappa}_2$ are $t \times 1$ and $q \times 1$ vectors of residuals. The details of parameter estimation are discussed in Song and Lee (2012).

In our study, no latent variables are assumed and the sole observable variables are phenotypes. Thus only the causal relationship among observable variables, *i.e.*, phenotypes, are fitted in the SEM model (also known as path analysis (Wright 1921)) as:

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}\boldsymbol{y}_i + \boldsymbol{\varepsilon}_i \quad (1)$$

where $\boldsymbol{y}_i$ and $\boldsymbol{\Lambda}$ are defined as above, $\boldsymbol{\varepsilon}_i$ represents everything that is not explained by $\boldsymbol{\Lambda}\boldsymbol{y}_i$, and $\boldsymbol{\Lambda}$ is an $t \times t$ matrix of structural coefficients representing the causal structure recovered from the Inductive Causation (IC) algorithm as described in the next section.

To illustrate, we assume that the phenotypes of three traits for each individual (*i.e.*, y_1, y_2, and y_3 for traits 1, 2, and 3) have the following causal relationship:

$$\begin{cases} y_1 = \varepsilon_1 \\ y_2 = \lambda_{12}y_1 + \varepsilon_2 \\ y_3 = \lambda_{13}y_1 + \lambda_{23}y_2 + \varepsilon_3 \end{cases}$$

where causal coefficient λ_{ij} represents that a 1-unit increase in trait i results in a λ_{ij} unit increase in trait j. Given the causal structure above, the $\boldsymbol{\Lambda}$ can be written as:

$$\boldsymbol{\Lambda} = \begin{pmatrix} 0 & 0 & 0 \\ \lambda_{12} & 0 & 0 \\ \lambda_{13} & \lambda_{23} & 0 \end{pmatrix} \quad (2)$$

Searching causal structure

As described above, fitting the SEM requires the causal structure among all traits to be known before analysis. To explore the wide-range of possible underlying causal structures, we use the method from Valente *et al.* (2010) to discern the causal structure based on the

posterior distribution of the residual covariance matrix. The reason we do not directly apply this method to phenotype data are that the covariance among phenotypes is likely confounded by genetic effects. The process of inferring causal structure is composed of three steps:

1. Fit the multi-trait BayesianCΠ model and obtain the posterior distribution of the residual covariance matrix.
2. Follow Valente *et al.* (2010) to derive the conditional independence relationship among traits based on the posterior distribution of the residual covariance matrix. In detail, we derive the residual partial correlation p(y_i, y_j|h), where h is a set of traits, to test whether trait y_i is conditionally independent from y_j. The highest posterior density (HPD) interval of 0.9 was used to make statistical decisions. If HPD interval of p(y_i, y_j|h) contains zero, y_i and y_j are regarded as conditionally independent on h.
3. Apply the IC algorithm (Pearl 2009) as described in the Appendix to the conditional independence relationship from step 2 to obtain the causal structure.

SEM-BayesCΠ

Assume $\boldsymbol{\varepsilon}_i = \boldsymbol{\mu} + \sum_{j=1}^{p} m_{ij}\boldsymbol{\alpha}_j + \boldsymbol{e}_i$ in equation (1) and follow assumptions in multi-trait BayesCΠ, the SEM-BayesCΠ model can be written as:

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}\boldsymbol{y}_i + \boldsymbol{\mu} + \sum_{j=1}^{p} m_{ij}\boldsymbol{D}_j\boldsymbol{\beta}_j + \boldsymbol{e}_i \tag{3}$$

Move $\boldsymbol{\Lambda}\boldsymbol{y}_i$ from the right side to the left side of equation (3), and define $\boldsymbol{\Lambda}^* = \boldsymbol{I} - \boldsymbol{\Lambda}$, where $\boldsymbol{I}$ is a $t \times t$ identity matrix and $\boldsymbol{\Lambda}$ is a $t \times t$ matrix of structural coefficients based on the discerned causal structure, the model becomes:

$$\boldsymbol{\Lambda}^*\boldsymbol{y}_i = \boldsymbol{\mu} + \sum_{j=1}^{p} m_{ij}\boldsymbol{D}_j\boldsymbol{\beta}_j + \boldsymbol{e}_i \tag{4}$$

To guarantee that the structural coefficient is identifiable, we assume that the residuals for each trait of individual i are independent with each other, which means the residual covariance matrix is diagonal (Wu *et al.* 2010; Momen *et al.* 2018). The vector of all non-zero elements in $\boldsymbol{\Lambda}$, *e.g.*, $\boldsymbol{\lambda} = [\lambda_{12}, \lambda_{13}, \lambda_{23}]$, is assumed to have a prior distribution:

$$\boldsymbol{\lambda}|\lambda_0, \tau^2 \sim N(\mathbf{1}\lambda_0, \boldsymbol{I}\tau^2)$$

where $\mathbf{1}$ is a vector of ones, $\boldsymbol{I}$ is the identity matrix, and λ_0 is a known mean for all elements in $\boldsymbol{\lambda}$. τ^2 is a tuning parameter to adjust the sharpness degree of the prior (Gianola and Sorensen 2004). In this paper, we set $\lambda_0 = 0$ and $\tau^2 = 1$. The priors for the remaining parameters are the same as in the section Multi-trait Bayesian regression model using mixture priors.

Gibbs samplers are used to draw samples for all parameters. The full conditional distribution to draw samples for $\boldsymbol{\lambda}$ is shown below. The derivations of the full conditional distributions of the remaining parameters of interest for Gibbs samplers are in Cheng *et al.* (2018b).

***Full conditional distribution of* $\boldsymbol{\Lambda}$:** We follow Gianola and Sorensen (2004) to obtain the full conditional distribution of $\boldsymbol{\Lambda}$, with the difference between our derivation and Gianola and Sorensen (2004) being that we specify the causal structure with positions of parameters in the $\boldsymbol{\Lambda}$. Let $\boldsymbol{\Omega}$ denote all parameters except $\boldsymbol{\lambda}$ in the SEM-BayesCΠ and use the causal structure $\boldsymbol{\Lambda} = \begin{pmatrix} 0 & 0 & 0 \\ \lambda_{12} & 0 & 0 \\ \lambda_{13} & \lambda_{23} & 0 \end{pmatrix}$ as an example, the left hand side of equation (4), $\boldsymbol{\Lambda}^*\boldsymbol{y}_i$, can be written as:

$$\begin{aligned}
\boldsymbol{\Lambda}^*\boldsymbol{y}_i &= \begin{pmatrix} 1 & 0 & 0 \\ -\lambda_{12} & 1 & 0 \\ -\lambda_{13} & -\lambda_{23} & 1 \end{pmatrix} \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix} \\
&= \begin{pmatrix} y_{i1} \\ y_{i2} - \lambda_{12}y_{i1} \\ y_{i3} - \lambda_{13}y_{i1} - \lambda_{23}y_{i2} \end{pmatrix} \\
&= \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 \\ y_{i1} & 0 & 0 \\ 0 & y_{i1} & y_{i2} \end{pmatrix} \begin{pmatrix} \lambda_{12} \\ \lambda_{13} \\ \lambda_{23} \end{pmatrix} \\
&= \boldsymbol{y}_i - \mathbf{Y}_i\boldsymbol{\lambda}
\end{aligned}$$

The conditional posterior distribution of $\boldsymbol{\lambda}$ can be written as:

$$\begin{aligned}
&p(\boldsymbol{\lambda}|\boldsymbol{\Omega}, \mathbf{y}) \propto \prod_{i=1}^{n} N(\boldsymbol{y}_i|\boldsymbol{\Lambda}^{*-1}(\boldsymbol{\mu} + \sum_{j=1}^{p} m_{ij}\mathbf{D}_j\boldsymbol{\beta}_j), \boldsymbol{\Lambda}^{*-1}\mathbf{R}\boldsymbol{\Lambda}^{*-1'}) \\
&N(\boldsymbol{\lambda}|\mathbf{1}\lambda_0, \mathbf{I}\tau^2) \\
&\propto |\boldsymbol{\Lambda}^*|^{\frac{n}{2}} \prod_{i=1}^{n} N(\boldsymbol{\Lambda}^*\boldsymbol{y}_i|\boldsymbol{\mu} + \sum_{j=1}^{p} m_{ij}\mathbf{D}_j\boldsymbol{\beta}_j, \mathbf{R}) N(\boldsymbol{\lambda}|\mathbf{1}\lambda_0, \mathbf{I}\tau^2) \\
&= |\boldsymbol{\Lambda}^*|^{\frac{n}{2}} exp[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\Lambda}^*\boldsymbol{y}_i - \boldsymbol{\mu} - \sum_{j=1}^{p} m_{ij}\mathbf{D}_j\boldsymbol{\beta}_j)'\mathbf{R}^{-1} \\
&(\boldsymbol{\Lambda}^*\boldsymbol{y}_i - \boldsymbol{\mu} - \sum_{j=1}^{p} m_{ij}\mathbf{D}_j\boldsymbol{\beta}_j)] \\
&\times exp\left[-\frac{1}{2\tau^2}(\boldsymbol{\lambda} - \mathbf{1}\lambda_0)'(\boldsymbol{\lambda} - \mathbf{1}\lambda_0)\right]
\end{aligned} \tag{5}$$

Setting $\boldsymbol{w}_i = \boldsymbol{y}_i - \boldsymbol{\mu} - \sum_{j=1}^{p} m_{ij}\boldsymbol{D}_j\boldsymbol{\beta}_j$, equation (5) can be written as:

$$\begin{aligned}
p(\boldsymbol{\lambda}|\boldsymbol{\Omega}, \mathbf{y}) \sim |\boldsymbol{\Lambda}^*|^{\frac{n}{2}} exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{w}_i - \boldsymbol{Y}_i\boldsymbol{\lambda})'\boldsymbol{R}^{-1}(\boldsymbol{w}_i - \boldsymbol{Y}_i\boldsymbol{\lambda})\right] \\
* exp\left[-\frac{1}{2\tau^2}(\boldsymbol{\lambda} - \mathbf{1}\lambda_0)'(\boldsymbol{\lambda} - \mathbf{1}\lambda_0)\right]
\end{aligned}$$

Following the derivation in Gianola and Sorensen (2004) and the fact that $|\boldsymbol{\Lambda}^*| = 1$ in a recursive system, the full conditional distribution of $\boldsymbol{\lambda}$ is

$$p(\boldsymbol{\lambda}|\boldsymbol{\Omega}, \boldsymbol{y}) \sim N(\hat{\boldsymbol{\lambda}}, V_\lambda),$$

where

$$\begin{aligned}
\hat{\boldsymbol{\lambda}} &= \left(\sum_{i=1}^{n} \boldsymbol{Y}_i'\boldsymbol{R}^{-1}\boldsymbol{Y}_i + \tau^{-2}\boldsymbol{I}\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{Y}_i'\boldsymbol{R}^{-1}\boldsymbol{w}_i + \tau^{-2}\mathbf{1}\lambda_0\right) \\
V_\lambda &= \left(\sum_{i=1}^{n} \boldsymbol{Y}_i'\boldsymbol{R}^{-1}\boldsymbol{Y}_i + \tau^{-2}\boldsymbol{I}\right)^{-1}
\end{aligned}$$

Decomposition of SNP effects

In SEM-BayesCΠ, the marker effect for locus j, $\boldsymbol{\alpha}_j$, is considered as the vector of direct marker effect of t traits. The indirect effect of locus

j of t traits can be calculated as $\sum_{\rho=1}^{t-1}\boldsymbol{\Lambda}^{\rho}\boldsymbol{\alpha}_j$. The overall effect of locus j on t traits is computed as $\sum_{\rho=0}^{t-1}\boldsymbol{\Lambda}^{\rho}\boldsymbol{\alpha}_j$ or $(\boldsymbol{I}-\boldsymbol{\Lambda})^{-1}\boldsymbol{\alpha}_j$, which is the summation of both direct and indirect effect of locus j. For example, given a causal structure $\boldsymbol{\Lambda}=\begin{pmatrix}0 & 0 & 0\\ \lambda_{12} & 0 & 0\\ \lambda_{13} & \lambda_{23} & 0\end{pmatrix}$, the direct effect for locus j on three traits is $\boldsymbol{\alpha}_j=\begin{pmatrix}\alpha_{1j}\\ \alpha_{2j}\\ \alpha_{3j}\end{pmatrix}$, and the indirect effect for locus j on three traits is calculated as

$$\boldsymbol{\Lambda}\boldsymbol{\alpha}_j+\boldsymbol{\Lambda}^2\boldsymbol{\alpha}_j=\begin{pmatrix}0\\ \lambda_{12}\alpha_{1j}\\ (\lambda_{13}+\lambda_{12}\lambda_{23})\alpha_{1j}+\lambda_{23}\alpha_{2j}.\end{pmatrix},$$

and the overall effect of locus j on trait k is $\begin{pmatrix}\alpha_{1j}\\ \lambda_{12}\alpha_{1j}+\alpha_{2j}\\ (\lambda_{13}+\lambda_{12}\lambda_{23})\alpha_{1j}+\lambda_{23}\alpha_{2j}+\alpha_{3j}\end{pmatrix}$.

Inference of association based on genomic windows

Markers in a genomic window are usually highly correlated, indicating that any single marker may not show a strong association with the trait even though a causal variant exists in the window. In this paper, we make an inference of association based on genomic windows, because multiple markers inside a genomic window may jointly capture the signal from the causal variant (Fernando and Garrick 2013; Fernando *et al.* 2017).

To make an inference of association based on genomic windows, posterior distribution for the proportion of the genetic variance explained by markers in genomic window w, q_w, is estimated from MCMC samples of overall, direct, and indirect marker effects as follows. For one MCMC sample of all marker effects on one trait, let $\boldsymbol{\alpha}_{direct}$, $\boldsymbol{\alpha}_{indirect}$, and $\boldsymbol{\alpha}_{overall}$ denote direct, indirect, and overall effects of all markers respectively.

The genetic value that is attributed to genomic window w is calculated as:

$$\begin{aligned}\boldsymbol{a}_{w,direct} &= \boldsymbol{M}_w\boldsymbol{\alpha}_{w,direct}\\ \boldsymbol{a}_{w,indirect} &= \boldsymbol{M}_w\boldsymbol{\alpha}_{w,indirect}\\ \boldsymbol{a}_{w,overall} &= \boldsymbol{M}_w\boldsymbol{\alpha}_{w,overall}\end{aligned}$$

where $\mathbf{M}_w$ is a matrix of marker covariates in window w and $\boldsymbol{\alpha}_{w,direct}$, $\boldsymbol{\alpha}_{w,indirect}$, and $\boldsymbol{\alpha}_{w,overall}$ are the MCMC samples of direct, indirect, and overall marker effects for SNPs in window w. Then the variance explained by the genomic window w is defined as:

$$\sigma^2_{a_{w,direct}}=\frac{\boldsymbol{a}^T_{w,direct}\boldsymbol{a}_{w,direct}}{n}-(\frac{\mathbf{1}_n^T\boldsymbol{a}_{w,direct}}{n})^2$$

$$\sigma^2_{a_{w,indirect}}=\frac{\mathbf{a}^T_{w,indirect}\mathbf{a}_{w,indirect}}{n}-(\frac{\mathbf{1}_n^T\mathbf{a}_{w,indirect}}{n})^2$$

$$\sigma^2_{a_{w,overall}}=\frac{\mathbf{a}^T_{w,overall}\mathbf{a}_{w,overall}}{n}-(\frac{\mathbf{1}_n^T\mathbf{a}_{w,overall}}{n})^2$$

Similarly, the total genetic variance is calculated as:

$$\sigma^2_a=\frac{\boldsymbol{a}^T_{overall}\boldsymbol{a}_{overall}}{n}-\left(\frac{\mathbf{1}_n^T\boldsymbol{a}_{overall}}{n}\right)^2$$

The proportion of the genetic variance explained by direct, indirect, and overall marker effects in the genomic window w is calculated as:

$$\begin{aligned}q_{w,direct} &= \frac{\sigma^2_{a_{w,direct}}}{\sigma^2_a}\\ q_{w,indirect} &= \frac{\sigma^2_{a_{w,indirect}}}{\sigma^2_a}\\ q_{w,overall} &= \frac{\sigma^2_{a_{w,overall}}}{\sigma^2_a}\end{aligned}$$

Given the MCMC samples of q_w, the window posterior probability of association (WPPA) is calculated as the proportion of MCMC samples of q_w that exceed a specific value T (Fernando and Garrick 2013; Chen *et al.* 2017; Lloyd-Jones *et al.* 2017). In this paper, associations are tested for non-overlapping windows of 100 SNPs, and genomic windows that explain over $\frac{1}{N}$ of the total genetic variance were deemed to be of potential interest (*i.e.*, $T=\frac{1}{N}$, where N is the total number of windows).

Data analysis

Real data: The Rice Diversity Panel with 413 *Oryza sativa* individual accessions was used in the analysis. Three traits were considered, including plant height (PH), flowering time in Arkansas (FTA), and panicle number per plant (PN) in our GWAS. After removing the records with missing data for these three traits and genotype with minor allele frequency <0.05, 370 individuals with 33,519 SNPs genotyped were included in our analysis. The phenotypic and genotypic data were publicly available for download from http://www.ricediversity.org/. It has been shown that using a threshold of WPPA $=\alpha$ to declare a significant genomic window restricts the proportion of false positives (PFP) to $<1-\alpha$ (Fernando *et al.* 2017). A previous GWAS (Zhao *et al.* 2011) identified significantly associated SNPs in chromosome 6 for flowering time in Arkansas (FTA) using the same dataset. A threshold of WPPA = 0.8 and p-value = 5×10^{-6} in our GWAS analysis resulted in similarly significant signals. This result suggests that a WPPA of 0.8 and p-value = 5×10^{-6} are reasonable for declaring a significant genomic window.

Simulated data: To compare SEM-BayesCΠ with SEM-GWAS of Momen *et al.* (2018), we simulated data based on real genotypes from the Rice Diversity Panel. The simulation scenarios in Chen *et al.* (2017) were applied to simulate different genetic architectures. The QTL effects were generated from unit-gamma distribution (scale = 1) with three different shape parameters (γ): fewer QTL with large effects ($\gamma=0.18$), fewer QTL with small or large effects and many QTL with intermediate effects ($\gamma=3.0$), and the intermediate case ($\gamma=1.48$). In addition to the distribution of QTL effects, the number of QTL (n_{QTL}) may play an important role in GWAS, thus three numbers of QTL ($n_{QTL}=30,90,300$) combined with the three shape parameters were used to create 9 scenarios. For each scenario, 50 replicated populations were simulated with QTL positions randomly sampled across the genome. Trait 1 was assumed to have a causal effect on trait 2 with causal structural coefficient $\lambda=1.0$. The QTL effects were simulated under two scenarios (Figure 1): the QTL have direct effect on both trait 1 and trait 2 in scenario 1, where QTL only have direct effect on trait 1 in scenario 2. Half of the QTL were simulated following scenario 1, while the remaining followed scenario 2. Phenotypes for two traits were generated based on heritability of 0.5. In the simulated data analysis, the causal structure was assumed known to exclude the bias caused by searching causal structures. The structural coefficients were assumed known in SEM-GWAS.

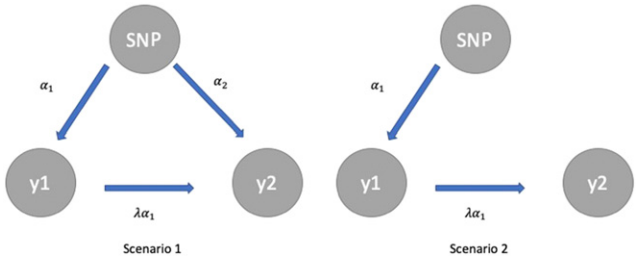

Figure 1 α_1: the direct effect on trait y_1; α_2: the direct effect on trait y_2; $\lambda\alpha_1$: the indirect effect on trait y_2. The graph shows the simulated two QTL effect simulation scenarios. Scenario 1 represents that the QTL has direct effect on both trait 1 and trait 2, whereas scenario 2 represents that the QTL only have direct effect on trait 1.

We have implemented SEM-Bayesian alphabet in JWAS (Cheng *et al.* 2018a), an open-source, publicly available package for single-trait and multi-trait whole-genome analyses written in the freely available Julia language. More details can be found at https://reworkhow.github.io/JWAS.jl/latest/.

RESULTS

Simulated data result

The performances of SEM-BayesCΠ and SEM-GWAS were compared based on the AUC (Area under Receiver Operating Characteristic). Inference of association on genomic windows for SEM-GWAS was based on the minimum p-value (Begum *et al.* 2016), *i.e.*, genomic windows containing at least one significant variant are declared as significant windows. To exclude the irrelevant AUC with low levels of specificity, only the partial area under the curve up until the false positive rate of 5% (pAUC5) (Chen *et al.* 2017; Ma *et al.* 2013) was calculated. For the convenience of comparison, all pAUC5 measurements were rescaled such that the pAUC5 of the random classifier is equal to 1. The R package ROCR (Sing *et al.* 2005) was used to obtain the pAUC5; the paired *t*-tests (p-value < 0.1) were used for comparing both the pAUC5 mean across all scenarios (overall mean comparison) and the pAUC5 mean for each level of n_{QTL} and shape parameters γ (marginal mean comparison).

The GWAS results based on overall, direct and indirect effect were shown in Table 1. For the overall effect result on trait 1, there is no significant difference between SEM-BayesCΠ and SEM-GWAS in both overall mean comparison and marginal mean comparison. The direct effect on trait 1 is the same as the overall effect on trait 1 since the direct effect and overall effect on trait 1 are equal based on the causal structure. For the overall effect on trait 2, the pAUC5 mean of SEM-BayesCΠ is significantly higher than that of SEM-GWAS in the overall mean comparison, and some marginal mean comparisons (*e.g.*, $n_{QTL} = 30$ and $\gamma = 0.18$). For the direct effect on trait 2, though higher overall mean of pAUC5 is usually observed in SEM-BayesCΠ, there is no significant difference (p-value < 0.1) between SEM-BayesCΠ and SEM-GWAS in both overall mean comparison and marginal mean comparison. For the indirect effect result on trait 2, similar to the overall effect result on trait 2, the pAUC5 mean of SEM-BayesCΠ is significantly higher than that of SEM-GWAS in the overall mean comparison, and some marginal mean comparisons (*e.g.*, $n_{QTL} = 30$ and $\gamma = 0.18$).

Real data result

Causal structure and structural coefficients: The causal structure among three traits is inferred by the IC algorithm from the estimated posterior distribution of the residual covariance matrix in the multi-trait BayesCΠ model. Figure 2 shows three potential phenotypic causal structures among traits PH (y_1), FTA (y_2), and PN (y_3) recovered for the 0.9 HPD interval. The causal structure matrices for IC1 ($\mathbf{\Lambda}_1$), IC2 ($\mathbf{\Lambda}_2$), and IC3 ($\mathbf{\Lambda}_3$) are:

Table 1 Overall and marginal mean of rescaled pAUC5 of SEM-BayesCΠ and SEM-GWAS based on overall,direct and indirect effect

		Overall effect		Direct effect		Indirect effect	
	factors	SEMBayesCΠ	SEM-GWAS	SEMBayesCΠ	SEM-GWAS	SEMBayesCΠ	SEM-GWAS
trait 1	n_{QTL}						
	30	4.99	4.66	4.99	4.66	NA	NA
	90	2.00	1.91	2.00	1.91	NA	NA
	300	1.38	1.42	1.38	1.42	NA	NA
	shape (γ)						
	0.18	2.71	2.52	2.71	2.52	NA	NA
	1.48	2.93	2.79	2.93	2.79	NA	NA
	3.00	2.72	2.69	2.72	2.69	NA	NA
	overall	2.79	2.66	2.79	2.66	NA	NA
trait 2	n_{QTL}						
	30	5.49‡	4.74†	4.22	3.91	4.88‡	3.15†
	90	2.09	2.08	1.87	1.87	1.91	1.72
	300	1.34	1.38	1.30	1.33	1.28	1.30
	shape (γ)						
	0.18	3.10‡	2.80†	2.48	2.36	2.79‡	2.29†
	1.48	3.03	2.84	2.56	2.38	2.63	2.46
	3.00	2.84	2.60	2.36	2.38	2.64	2.43
	overall	2.99‡	2.74†	2.46	2.37	2.69‡	2.39†

For both trait 1 and trait 2, comparisons between methods are made for different number of QTL (n_{QTL}) and shape parameters of QTL effects (γ). Estimations are based on 450 simulated data sets including nine scenarios discussed in Simulated data. For each effect, in each row, the values with different symbols have significantly different (p-value <0.1) pAUC5. NA represented pAUC5 was not available because the indirect effect on trait 1 does not exist based on the causal structure. The overall effect result on trait 1 is the same as the direct effect result on trait 1 because the overall effect on trait 1 equals the direct effect on trait 1 based on the causal structure.

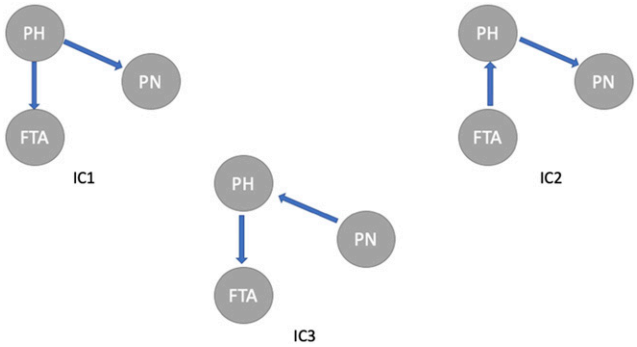

Figure 2 Causal structures among plant height (PH), flowering time in Arkansas (FTA), and panicle number per plant (PN) inferred from the IC algorithms. The edges connecting two traits represent non-null partial correlations as indicated by 0.9 HPD interval. The arrows represent the direction of causal effects.

$$\mathbf{\Lambda}_1 = \begin{pmatrix} 0 & 0 & 0 \\ \lambda_{12} & 0 & 0 \\ \lambda_{13} & 0 & 0 \end{pmatrix}, \mathbf{\Lambda}_2 = \begin{pmatrix} 0 & \lambda_{21} & 0 \\ 0 & 0 & 0 \\ \lambda_{13} & 0 & 0 \end{pmatrix},$$
$$\mathbf{\Lambda}_3 = \begin{pmatrix} 0 & 0 & \lambda_{31} \\ \lambda_{12} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

These three causal structures are fitted in the SEM-BayesCΠ model, and samples from posterior distributions for coefficients in these causal structures are obtained. The 90% credible intervals for structural coefficients in IC1, IC2, and IC3 are shown in Table 2. It is worth noting that the causal structures in Figure 2 provides the same set of marginal and conditional independencies, extra biological knowledge is required to further infer the causal structure. In this paper, the SEM-BayesCΠ model is proposed to incorporate (known) underlying causal structure among traits for GWAS. Thus, for the simplicity of our presentation, the causal structure is assumed known in Real Data Result section, *e.g.*, IC1 is used to demonstrate the performance of SEM-BayesCΠ.

Decomposition of SNP effects: Direct, indirect, and overall SNP effects for all markers are estimated from SEM-BayesCΠ and SEM-GWAS. In SEM-BayesCΠ, direct SNP effects are assigned mixture priors, where each locus can affect any combinations of traits directly; samples from posterior distributions of indirect effects are obtained using joint samples from posterior distributions of $\mathbf{\Lambda}$ and direct SNP effects $\boldsymbol{\alpha}_j$. In IC1, for trait PH, the overall SNP effect is equal to the direct SNP effect, because there is no intermediate trait. For trait FTA, the overall SNP effect is composed of direct SNP effect and indirect SNP effect transmitted from PH. So the overall SNP effect for FTA is given by summing the direct SNP effect and indirect SNP effect. Similarly, for trait PN, the overall SNP effect is obtained by summing the direct SNP effect and indirect SNP effect transmitted from PH.

The results of GWAS from SEM-BayesCΠ and SEM-GWAS incorporating causal structure IC1 are shown in Figure 3. Significant signals are found only for trait FTA. SEM-BayesCΠ adopts a threshold of WPPA = 0.8 to declare a significant genomic window and SEM-GWAS adopts a threshold of p-value = 5×10^{-6}. The overall SNP effects are partitioned into direct and indirect effects, and GWAS are performed for the direct, indirect, and overall SNP effects separately for trait FTA. In Figure 3, the blue points, pink points and green points represents the significant genomic windows located in chromosome 1, chromosome 5 and chromosome 6. Window A contains SNPs from "id1000759" to "id1001229"; window B contains SNPs from "id1023967" to "id1024499"; window C contains SNPs from "id5013234" to "id5013920"; window D contains SNPs from "id6005814" to "id6006470".

For the overall effects, in the SEM-GWAS, window D achieved -log(p-value) 14.21; in the SEM-BayesCΠ, window C achieved WPPA 0.90, window D achieved WPPA 0.88, and window A achieved WPPA 0.82. For the direct effect, in the SEM-GWAS, window D achieved -log(p-value) 12.24; in the SEM-BayesCΠ, window C achieved WPPA 0.90, window D achieved WPPA 0.86, and window A achieved WPPA 0.80. For the indirect effect, in the SEM-GWAS, window B achieved -log(p-value) 14.65; in the SEM-BayesCΠ, although no window is identified as significant in SEM-BayesCΠ, a peak was observed at window B with WPPA 0.52. Further, for all three effects, the results from SEM-BayesCΠ and SEM-GWAS are correlated (the correlation between the WPPA from SEM-BayesCΠ and -log(p-value) from the SEM-GWAS is higher than 0.5). The correlation of indirect effect results from these two methods results achieved 0.70. Also, for both SEM-BayesCΠ and SEM-GWAS, the overall effect is more correlated with direct effect rather than indirect effect. The magnitudes for overall, direct, and indirect SNP effect in SEM-BayesCΠ are also shown in Figure 5. Though most large overall SNP effects consist of a large direct SNP effect and a relatively small indirect SNP effect, the indirect effect of some SNPs play an important role, *e.g.*, the overall effect of SNP "id1024159", as shown in Figure 5, consists of a large indirect SNP effect and relatively small direct effect.

Genetic pleiotropy in SEM-BayesCΠ

As shown in Figure 4, the posterior distribution of the parameter Π is obtained, and markers show different levels of pleiotropy for direct SNP effects. In SEM-BayesCΠ, each SNP can have direct effects on any combination of traits, and the parameter Π is used to estimate the proportion of SNPs having different levels of pleiotropy. Indirect SNP effects on one trait are transmitted from direct SNP effects on other intermediate traits. For example, in causal structure IC1, the indirect SNP effects on FTA is transmitted from direct SNP effects on the trait PH. A SNP having no direct effect on FTA may have overall effects on FTA if its direct effect on trait PH is non-zero. The proportions of markers affecting different combinations of traits through overall SNP effects can also be estimated. This result is shown in Figure 4, and different probabilities are observed for some cases between overall and direct SNP effects. More SNPs have effects on all traits simultaneously when overall SNP effects are considered compared to direct SNP effects (case 1). If only overall SNP effects are considered,

Table 2 The 90% credible interval for causal structural coefficients in the three causal structures

λ	IC1	IC2	IC3
$\lambda_{PH\to FTA}$	(0.30, 0.52)	NA	(0.30, 0.55)
$\lambda_{PH\to PN}$	(-0.21, -0.02)	(-0.20, -0.01)	NA
$\lambda_{FTA\to PH}$	NA	(0.22, 0.44)	NA
$\lambda_{PN\to PH}$	NA	NA	(-0.27, -0.05)

PH, FTA, and PN represent traits plant height, flowering time in Arkansas, panicle number per plant, respectively. $\lambda_{a\to b}$ represents the causal effects of trait *a* on trait *b*. NA denotes structural coefficients those do not exist in the causal structure.

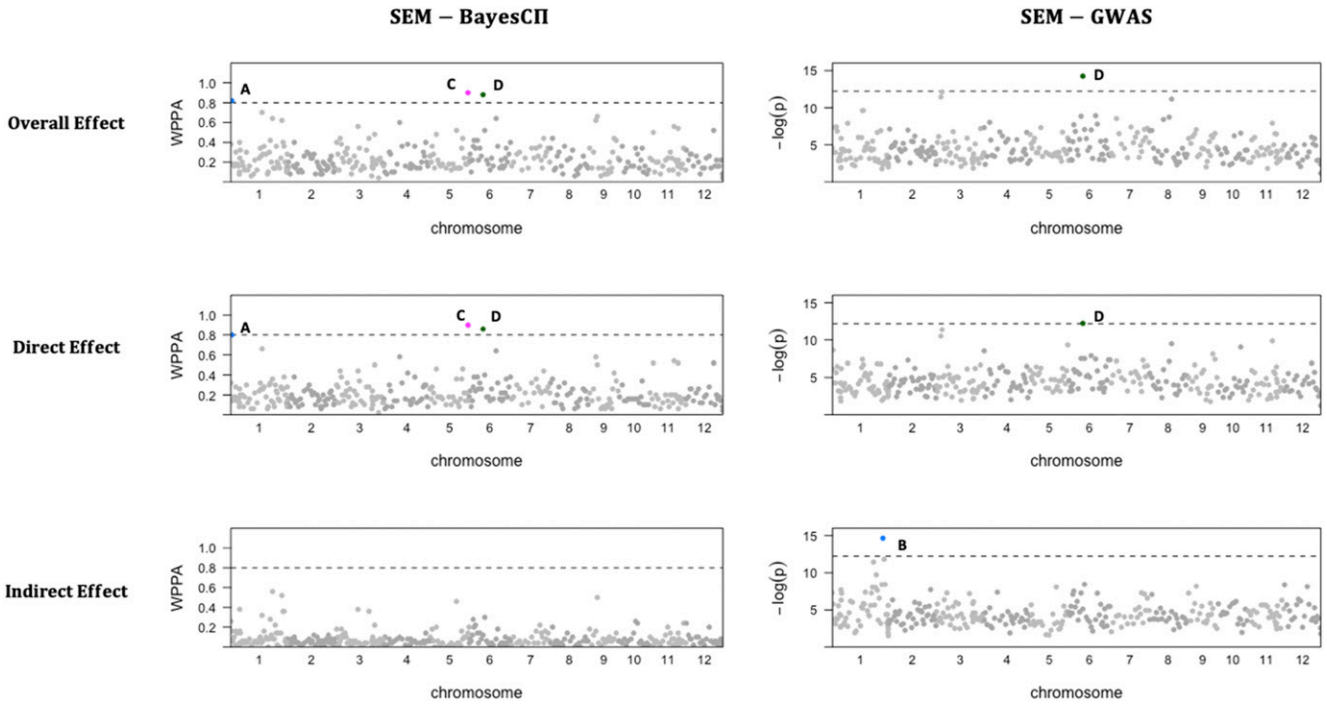

Figure 3 GWAS results based on overall, direct and indirect SNP effects from SEM-BayesCΠ and SEM-GWAS incorporating IC1 causal structure for the trait flowering time at Arkansas (FTA). The horizontal dash line represents the threshold 0.8 or -log(5×10^{-6}). X-axis represents the location of genomic windows along the 12 chromosomes; Y-axis represents window posterior probability of association (WPPA) for SEM-BayesCΠ and negative logarithm of the p-value (-log(p)) for SEM-GWAS. Colored points represent genomic windows with WPPA $\geq$ 0.8 or p-value $\leq 5\times10^{-6}$. The blue points, pink points and green points represents the significant genomic windows located in chromosome 1, chromosome 5 and chromosome 6.

some cases having non-zero probabilities for indirect SNP effects are hidden by the causal relationships among traits (cases 2-4). The same patterns for direct and overall SNP effects are observed in cases 5-8 because there is no causal relationship between trait FTA and PN.

DISCUSSION

The complex causal relationships among multiple traits are usually not considered in conventional multi-trait GWAS. Here we propose the SEM-Bayesian alphabet method to incorporate pre-inferred causal structures among multiple traits into multi-trait Bayesian regression methods. SEM-Bayesian alphabet accounts for causal structures among traits, and has the potential advantage of estimating causal effects, providing genomic window-based inference, as well as providing a comprehensive understanding of the underlying biological mechanism.

GWAS

To show the potential utility of SEM-Bayesian alphabet, simulated data were used to compare SEM-BayesCΠ with SEM-GWAS. A wide variety of potential genomic architectures were constructed by the combination of different levels of skewness of gamma distribution for QTL effects (γ) and different numbers of QTL(n_{QTL}) (Chen *et al.* 2017).

BayesCΠ was also performed to estimate overall SNP effects on the same datasets, and similar results as those in the SEM-BayesCΠ were obtained (not shown in this paper). This is reasonable since the SEM-Bayesian alphabet model can be reduced to a model similar to Bayesian regression by reparameterization, indicating that the joint likelihood functions of SEM-Bayesian alphabet and Bayesian regression are similar. Compared to Bayesian regression, the SEM-Bayesian alphabet provides a more comprehensive understanding of the underlying biological mechanism by decomposing overall SNP effects into direct and indirect SNP effects (Figures 3, 4 and 5).

The comparison between SEM-BayesCΠ and SEM-GWAS using simulated data were shown in Table 1. As shown in our results, SEM-BayesCΠ has relatively the same or higher pAUC5 than SEM-GWAS in all simulation scenarios. In some scenarios, SEM-BayesCΠ has significantly higher pAUC5 than SEM-GWAS. For example, when one trait is affected by few QTL of large effects (*e.g.*, n_{QTL} = 30,*shape*($\gamma = 0.18$)), SEM-BayesCΠ has significantly higher pAUC5 than SEM-GWAS to infer indirect and overall effects. Though significant difference is not observed for direct

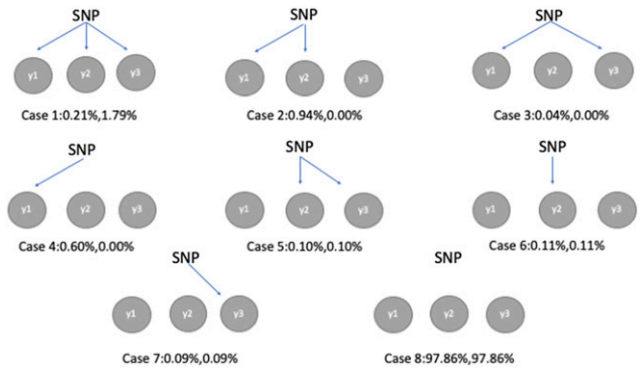

Figure 4 Estimated proportion of markers affecting combinations of traits, Π, from SEM-BayesCΠ incorporating IC1 causal structure. For each scenario, it is estimated for both direct SNP effects (left) and overall SNP effects (right).

effect, higher overall mean of pAUC5 is usually observed in SEM-BayesCΠ.

Causal structure

The causal structure is assumed to be known in SEM-Bayesian alphabet, and it is usually discerned by three types of algorithms: the constraint-based algorithm, the score-based algorithms, and the hybrid algorithms. The IC algorithm (Pearl 2009; Valente *et al.* 2010) used in this paper is a typical constraint-based algorithm, which is based on conditional independence tests. The score-based algorithms apply the heuristic optimization techniques, which set an initial graph structure and assign an initial goodness-of-fit score to it, and then maximize the goodness-of-fit score to obtain the most possible causal structure. The hybrid algorithm is a hybrid of both the constraint-based and the score-based algorithms. It utilizes conditional independence tests to reduce the space of candidate causal structures, and uses network scores to identify the optimal structure among them (Scutari 2014). The causal structures inferred from these algorithms may be different. Note that different evaluation criteria may also result in different outcome causal structures. For example, in this paper, if we choose 0.99 instead of 0.9 HPD interval to search for causal structures, there will be no edge between the traits PH and PN.

Decomposition of SNP effects

In some previous analysis (Mi *et al.* 2010; Momen *et al.* 2018, 2019), the indirect SNP effect of locus j of t traits is obtained by multiplying the estimated $\boldsymbol{\Lambda}$, $\hat{\boldsymbol{\Lambda}}$, and estimated direct SNP effects, $\hat{\boldsymbol{\alpha}}_j$, as $\sum_{\rho=1}^{t-1}\hat{\boldsymbol{\Lambda}}^{\rho}\hat{\boldsymbol{\alpha}}_j$. This is similar to using posterior means of causal structural coefficients and direct SNP effects for calculation of the indirect SNP effects. In our method, indirect SNP effects are estimated using joint samples from posterior distributions of $\boldsymbol{\Lambda}$ and $\boldsymbol{\alpha}_j$. We compared these two approaches for indirect SNP effect estimation on real rice data, and found that the indirect effects estimated from these two approaches are slightly different. The SEM-BayesCΠ approach should be used in indirect SNP effect estimation due to the fact that $\boldsymbol{\Lambda}$ and $\boldsymbol{\alpha}_j$ may be highly dependent.

CONCLUSION

SEM-Bayesian alphabet provides more interpretation into biological mechanisms than Bayesian regression methods by decomposing the overall SNP effects into direct and indirect SNP effects. In SEM-Bayesian alphabet, posterior distributions of the overall, direct, and indirect SNP effects, as well as causal structure coefficients, are obtained, which are used to make inferences about these parameters. Compared to the typical GWAS method incorporating causal structure among multiple traits, such as SEM-GWAS, SEM-Bayesian alphabet obtains the posterior distributions for the proportion of variance attributed to a genomic region to detect causal loci (*i.e.*, the use of WPPA). The level of gene pleiotropy, *e.g.*, proportion of markers affecting different combinations of traits as shown in Figure 4, can also be further dissected into direct and indirect SNP effects. Also, with estimating structural coefficients, SEM-Bayesian alphabet still has relatively same or greater pAUC5 than SEM-GWAS in all scenarios of simulated data. In summary, SEM-Bayesian alphabet offers a more comprehensive understanding of the underlying biological mechanisms including pleiotropy and causal relationships among traits than conventional GWAS, as well as has a potential advantage in the GWAS inference than other GWAS considering complex causal effect among multiple traits.

ACKNOWLEDGMENTS

We want to thank Bruno D. Valente for his explanation of the causal structure searching method and Mehdi Momen for his clarification on the IC algorithm. This work was supported by the United States Department of Agriculture, Agriculture and Food Research Initiative National Institute of Food and Agriculture Competitive grant no. 2018-67015-27957.

LITERATURE CITED

Anderson, J. C., and D. W. Gerbing, 1988 Structural equation modeling in practice: A review and recommended two-step approach. Psychol. Bull. 103: 411–423. https://doi.org/10.1037/0033-2909.103.3.411

Begum, F., M. H. Sharker, S. L. Sherman, G. C. Tseng, and E. Feingold, 2016 Regionally Smoothed Meta-Analysis Methods for GWAS Datasets. Genet. Epidemiol. 40: 154–160. https://doi.org/10.1002/gepi.21949

Cantor, R. M., K. Lange, and J. S. Sinsheimer, 2010 Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. Am. J. Hum. Genet. 86: 6–22. https://doi.org/10.1016/j.ajhg.2009.11.017

Chen, C., J. P. Steibel, and R. J. Tempelman, 2017 Genome-Wide Association Analyses Based on Broadly Different Specifications for Prior Distributions, Genomic Windows, and Estimation Methods. Genetics 206: 1791–1806. https://doi.org/10.1534/genetics.117.202259

Cheng, H., D. J. Garrick, and R. L. Fernando, 2018a JWAS: Julia implementation of whole-genome analysis software. Proceedings of the World Congress on Genetics Applied to Livestock Production **11**: 859.

Cheng, H., K. Kizilkaya, J. Zeng, D. Garrick, and R. Fernando, 2018b Genomic Prediction from Multiple-Trait Bayesian Regression Methods Using Mixture Priors. Genetics 209, 89–103. https://doi.org/10.1534/genetics.118.300650

Chicharro, D., and S. Panzeri, 2014 Algorithms of causal inference for the analysis of effective connectivity among brain regions. Front. Neuroinform. 8: 64. https://doi.org/10.3389/fninf.2014.00064

Fernando, R., A. Toosi, A. Wolc, D. Garrick, and J. Dekkers, 2017 Application of Whole-Genome Prediction Methods for Genome-Wide Association Studies: A Bayesian Approach. J. Agric. Biol. Environ. Stat. 22: 172–193. https://doi.org/10.1007/s13253-017-0277-6

Fernando, R. L., and D. Garrick, 2013 Bayesian Methods Applied to GWAS, pp. 237–274 in *Genome-Wide Association Studies and Genomic Prediction*, Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-447-0_10

Gianola, D., 2013 Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics 194: 573–596. https://doi.org/10.1534/genetics.113.151753

Gianola, D., and D. Sorensen, 2004 Quantitative Genetic Models for Describing Simultaneous and Recursive Relationships Between Phenotypes. Genetics 167: 1407–1424. https://doi.org/10.1534/genetics.103.025734

Hackinger, S., and E. Zeggini, 2017 Statistical methods to detect pleiotropy in human complex traits. Open Biol. 7: 170125. https://doi.org/10.1098/rsob.170125

Korte, A., B. J. Vilhjálmsson, V. Segura, A. Platt, Q. Long *et al.*, 2012 A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat. Genet. 44: 1066–1071. https://doi.org/10.1038/ng.2376

Liu, H.-J., and J. Yan, 2018 Crop genome-wide association study: a harvest of biological relevance. Plant J. 97: 8–18. https://doi.org/10.1111/tpj.14139

Lloyd-Jones, L. R., M. R. Robinson, G. Moser, J. Zeng, S. Beleza *et al.*, 2017 Inference on the Genetic Basis of Eye and Skin Color in an Admixed Population via Bayesian Linear Mixed Models. Genetics 206: 1113–1126. https://doi.org/10.1534/genetics.116.193383

Ma, H., A. I. Bandos, H. E. Rockette, and D. Gur, 2013 On use of partial area under the ROC curve for evaluation of diagnostic performance. Stat. Med. 32: 3449–3458. https://doi.org/10.1002/sim.5777

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little *et al.*, 2008 Genome-wide association studies for complex traits: consensus,

uncertainty and challenges. Nat. Rev. Genet. 9: 356–369. https://doi.org/10.1038/nrg2344

Mi, X., K. Eskridge, D. Wang, P. Stephen Baenziger, 2010 Bayesian mixture structural equation modelling in multiple-trait QTL mapping. Genet. Res. 92: 239–250. https://doi.org/10.1017/S0016672310000236

Momen, M., M. T. Campbell, H. Walia, and G. Morota, 2019 Utilizing trait networks and structural equation models as tools to interpret multi-trait genome-wide association studies. Plant Methods 15: 107. https://doi.org/10.1186/s13007-019-0493-x

Momen, M., A. A. Mehrgardi, M. A. Roudbar, A. Kranis, R. M. Pinto *et al.*, 2018 Including Phenotypic Causal Networks in Genome-Wide Association Studies Using Mixed Effects Structural Equation Models. Front. Genet. 9: 455. https://doi.org/10.3389/fgene.2018.00455

Ozaki, K., Y. Ohnishi, A. Iida, A. Sekine, R. Yamada *et al.*, 2002 Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat. Genet. 32: 650–654. https://doi.org/10.1038/ng1047

O'Reilly, P. F., C. J. Hoggart, Y. Pomyen, F. C. F. Calboli, P. Elliott *et al.*, 2012 MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. PLoS One 7: e34861. https://doi.org/10.1371/journal.pone.0034861

Pearl, J., 2009 Causal inference in statistics: An overview. Stat. Surv. 3: 96–146. https://doi.org/10.1214/09-SS057

Scutari, M., 2014 Bayesian Network Constraint-Based Structure Learning Algorithms: Parallel and Optimised Implementations in the bnlearn R Package.

Sharma, A., J. S. Lee, C. G. Dang, P. Sudrajad, H. C. Kim *et al.*, 2015 Stories and Challenges of Genome Wide Association Studies in Livestock - A Review. Asian-Australas. J. Anim. Sci. 28: 1371–1379. https://doi.org/10.5713/ajas.14.0715

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer, 2005 ROCR: visualizing classifier performance in R. Bioinformatics 21: 3940–3941. https://doi.org/10.1093/bioinformatics/bti623

Sodini, S. M., K. E. Kemper, N. R. Wray, and M. Trzaskowski, 2018 Comparison of Genotypic and Phenotypic Correlations: Cheverud's Conjecture in Humans. Genetics 209: 941–948.

Song, X.-Y., and S.-Y. Lee, 2012 A tutorial on the Bayesian approach for analyzing structural equation models. J. Math. Psychol. 56: 135–148. https://doi.org/10.1016/j.jmp.2012.02.001

Valente, B. D., G. J. M. Rosa, G. Campos, D. Gianola, and M. A. Silva, 2010 Searching for Recursive Causal Structures in Multivariate Quantitative Genetics Mixed Models. Genetics 185: 633–644. https://doi.org/10.1534/genetics.109.112979

VanRaden, P., 2008 Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91: 4414–4423. https://doi.org/10.3168/jds.2007-0980

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy *et al.*, 2017 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. 101: 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Wright, S., 1921 Correlation and Causation. J. Agric. Res. 557–585.

Wright, S., 1934 The Method of Path Coefficients. Ann. Math. Stat. 5: 161–215. https://doi.org/10.1214/aoms/1177732676

Wu, X., B. Heringstad, and D. Gianola, 2010 Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. J. Anim. Breed. Genet. 127: 3–15. https://doi.org/10.1111/j.1439-0388.2009.00835.x

Zhao, K., C.-W. Tung, G. C. Eizenga, M. H. Wright, M. L. Ali *et al.*, 2011 Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat. Commun. 2: 467. https://doi.org/10.1038/ncomms1467

Communicating editor: D.-J. de Koning

APPENDIX

Inductive causation algorithm

We use the IC algorithm (Pearl 2009) to recover the causal structure among a set of traits, denoted as U below, from the conditional independence relationship. The IC algorithm is composed of three steps (Valente *et al.* 2010; Chicharro and Panzeri 2014):

1. For each pair of traits X and Y, search for a set $S_{XY} \subseteq U$ such that $X \perp Y | S_{XY}$ holds. That is, X and Y are independent, conditional on S_{XY}. If there is no such S_{XY}, place an undirected edge between these two traits.

2. If this pair of traits X and Y are non-adjacent (*i.e.*, no un-directed edge between X and Y) with a common neighbor Z (*i.e.*, Z is adjacent to X as well as to Y), and $Z \notin S_{XY}$, place arrowheads pointing to Z, *i.e.*, $X \rightarrow Z \leftarrow Y$.

3. In the partially-oriented graph from step 2, orient as many edges as possible following two requirements:

(a) Any alternative orientation will not yield a new V structure (*i.e.*, $X \rightarrow Z \leftarrow Y$).

(b) Any alternative orientation will not yield a directed cycle.

In summary, we find all the pairs of variables that have a dependent relationship to reconstruct the basic structure of the underlying causal network in step 1. Then we find all the V structures in the network in step 2 and prevent the creation of new V structures or directed cycles in step 3.

Marker effect decomposition

Estimated direct, indirect, and overall SNP effects from SEM-BayesCΠ incorporating the IC1 causal structure for the trait flowering time at Arkansas (FTA) are shown in Figure 5. The SNP "id1024159" has direct effect 0.005 on trait FTA, while its indirect effect transmitted through PH is -0.039. Thus, the overall effect of SNP "id1024159" is mainly determined by the indirect effect.

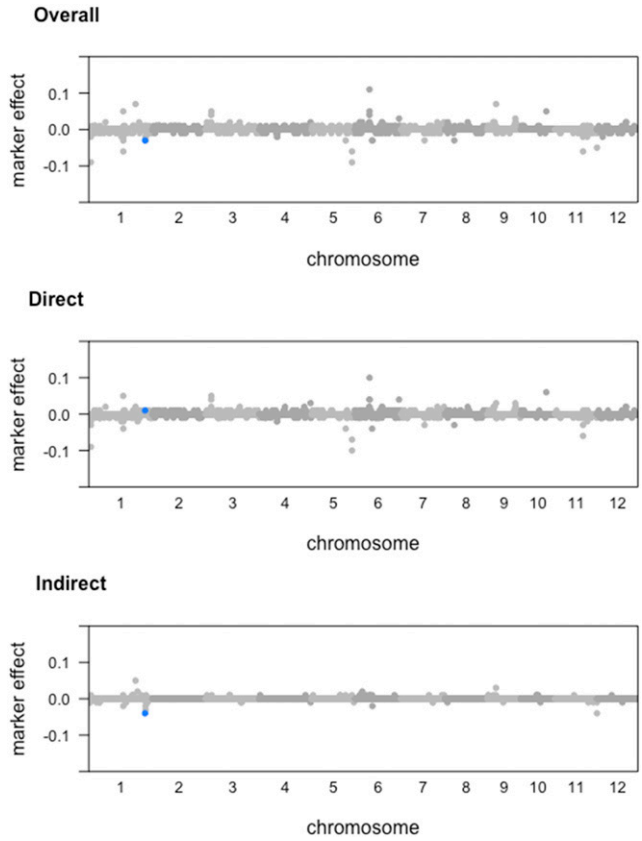

Figure 5 Magnitude of direct, indirect and overall SNP effect from SEM-BayesCΠ incorporating IC1 causal structure for the trait flowering time at Arkansas (FTA). X-axis represents the location of SNPs along the 12 chromosomes. Y-axis represents the magnitude of the marker effects. The blue points represents the SNP "id1024159". For SNP "id1024159", the overall effect is consists of a small direct SNP effect and a relatively large indirect SNP effect.