G3: Genes|Genomes|Genetics Early Online, published on April 16, 2019 as doi:10.1534/g3.119.400154

Genomic Bayesian confirmatory factor analysis and Bayesian network to characterize a wide spectrum of rice phenotypes

Haipeng Yu[1], Malachy T. Campbell[1,2], Qi Zhang[3], Harkamal Walia[2], and Gota Morota[1]

[1]Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061
[2]Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583
[3]Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583

Keywords: Bayesian network, factor analysis, multi-phenotypes, rice

Running title: Network analysis in rice

Corresponding author:
Gota Morota
Department of Animal and Poultry Sciences
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061, USA.
E-mail: morota@vt.edu

Abstract

With the advent of high-throughput phenotyping platforms, plant breeders have a means to assess many traits for large breeding populations. However, understanding the genetic interdependencies among high-dimensional traits in a statistically robust manner remains a major challenge. Since multiple phenotypes likely share mutual relationships, elucidating the interdependencies among economically important traits can better inform breeding decisions and accelerate the genetic improvement of plants. The objective of this study was to leverage confirmatory factor analysis and graphical modeling to elucidate the genetic interdependencies among a diverse agronomic traits in rice. We used a Bayesian network to depict conditional dependencies among phenotypes, which can not be obtained by standard multi-trait analysis. We utilized Bayesian confirmatory factor analysis which hypothesized that 48 observed phenotypes resulted from six latent variables including grain morphology, morphology, flowering time, physiology, yield, and morphological salt response. This was followed by studying the genetics of each latent variable, which is also known as factor, using single nucleotide polymorphisms. Bayesian network structures involving the genomic component of six latent variables were established by fitting four algorithms (i.e., Hill Climbing, Tabu, Max-Min Hill Climbing, and General 2-Phase Restricted Maximization algorithms). Physiological components influenced the flowering time and grain morphology, and morphology and grain morphology influenced yield. In summary, we show the Bayesian network coupled with factor analysis can provide an effective approach to understand the interdependence patterns among phenotypes and to predict the potential influence of external interventions or selection related to target traits in the interrelated complex traits systems.

Introduction

A primary objective in plant breeding is the develop high yielding varieties with specific grain qualities, resilience to pests and abiotic stresses, and superior adaption to the target environment. As a result, plant breeders devote considerable resources to extensive phenotypic evaluation of germplasm and select on multiple traits. These traits are often correlated at a genetic level through common genetic effects (e.g., pleiotropy) or linkage disequilibrium between quantitative trait locus (QTL). Since multiple phenotypes may exhibit mutual relationships, knowledge of the interdependence among agronomically important traits can improve the efficacy of selection and rate of genetic improvement in systems with complex traits.

In a standard quantitative genetic analysis, multivariate phenotypes can be modeled through multi-trait models (MTM) of Henderson and Quaas (1976) or some genomic counterparts (e.g., Calus and Veerkamp 2011; Jia and Jannink 2012) by leveraging genetic or environmental correlations among traits. In particular, MTM has been useful in deriving genetic correlations and enhancing the prediction accuracy of breeding values for traits with low heritability or scarce records via joint modeling with one or more genetically correlated, highly heritable traits (Mrode 2014). Conventional MTM strategies may provide important insight into the genetic relations between agronomically important traits, but they fail to explain how these traits are related. For instance, consider a case where we have three genetically correlated traits: y_1, y_2, and y_3. With MTM, we cannot address whether the relationship between y_1 and y_3 is due to direct effects, or if the relationship is driven by indirect effects mediated by y_2. Bayesian Networks (BN) offer an effective approach to elucidate the underlying network structure in multivariate data and infer network relationships between correlated variables. A BN is a probabilistic graphical model that represents conditional dependencies among a set of variables via a directed acyclic graph (DAG) (Neapolitan *et al.* 2004). In the DAG, the variables are represented by nodes, while their conditional dependencies between nodes are indicated with directed edges. In the context of plant breeding, BN can used to elucidate the interdependencies among traits and inform selection decisions for simultaneously improving multiple traits. For instance in the latter case above ($y_1 \rightarrow y_2 \rightarrow y_3$), selection directly on y_2 will affect the quantity of y_3 without an effect on y_1.

With the advent high-throughput phenotyping (HTP) platforms, plant breeders have been provided with a suite of tools for phenotypic evaluation of large populations (Shakoor *et al.* 2017). These platforms leverage robotics, precise environmental control, and remote sensing techniques to provide accurate, repeatable and high resolution phenotypes for large breeding populations throughout the growing season (Araus and Cairns 2014; Shakoor *et al.* 2017; Araus *et al.* 2018). These data can be used to redefine characteristics underlying superior agronomic performance by quantifying secondary traits associated with seedling vigor, plant architecture, photosynthesis, transpiration, disease resistance, and stress tolerance (Cabrera-Bosquet *et al.* 2016; Sun *et al.* 2017; Crain *et al.* 2018). However given these new approaches, breeders are faced with the new challenge of efficiently utilizing these large multidimesional data sets to improve selection efficiency. The primary challenges associated with multivariate analysis and BN approaches using HTP data is that robust parameter estimates can be untenable because the number of estimated parameters within the model increases with the increasing number of phenotypes. Moreover, even in cases where MTM or BN can be applied, interpreting of interrelationships among a large number of phenotypes can be difficult.

One approach to characterize high-dimensional phenotypes is by using factor analysis (FA). The central idea of FA approaches is to reduce the dimensions of multivariate data sets by constructing unobserved, latent factors, or modules, from correlated phenotypes (de los Campos and Gianola 2007). The biological importance of these latent factors can be interpreted by inspecting the phenotypes that contribute to each factor. Thus, the advantage of FA for large, multivariate data sets is two fold. First, FA provides a means to reduce the dimensions of multivariate data sets thereby providing statistically sound parameter estimates, and easing visualization and interpretation. Secondly, the latent variables/factors themselves may be representative of underlying biological processes that cannot be observed or measured in the population. For instance, several studies have highlighted the effects of plant hormones such as GA on multiple morphological attributes (Wang and Li 2006; Lo *et al.* 2008; Umehara *et al.* 2008; Bhattacharya *et al.* 2010; Brewer *et al.* 2013; Zhou *et al.* 2013). Thus, a latent factor constructed from these morphological traits may provide information on the biosynthesis or sensitivity of these hormones for individuals within the population. If a certain amount of knowledge regarding the biological role of the variables is already known, a varaint of FA, confirmatory factor analysis (CFA), can be used to estimate latent variables based on predetermined biological classes of observed traits (Jöreskog 1969). These latent variables underlie observed phenotypes and can be evaluated for how well the data support the hypothesis. For instance, Peñagaricano *et al.* (2015) performed CFA in swine to derive five latent variables from 19 phenotypic traits and inferred BN structures among those latent variables, thereby demonstrating the potential of this approach.

This study aimed to leverage CFA and graphical modeling to elucidate the genetic interdependencies among traits typically recorded in breeding programs (e.g., yield, plant morphology, phenology, and stress resilience). First, we constructed latent variables, using prior biological knowledge obtained from the literature. Then we connected the observed high-dimensional phenotypes with these to establish latent variables via Bayesian confirmatory factor analysis (BCFA) to reduce the dimensions of the dataset. Further, factor scores computed from BCFA were considered new phenotypes for a Bayesian multivariate analysis to separate breeding values from noise. This was followed by adjustment of breeding values via Cholesky decomposition to eliminate the dependencies introduced by genomic relationships. Finally, the adjusted breeding values were considered inputs to assess the network structure between latent variables by conducting a Gaussian BN analysis. This study is the first, to our knowledge, in rice to characterize various phenotypes with graphical modeling such as BCFA and BN.

Materials and Methods

Phenotypic and genotypic data

The rice dataset comprised $n = 374$ accessions sampled from six subpopulations: temperate japonica (92), tropical japonica (85), indica (77), aus (52), aromatic (12), and admixture of japonica and indica (56) (Zhao *et al.* 2011). The improvement status of each accession was obtained from the USDA-ARS Germplasm Resources Information Network. We used $t = 48$ phenotypes and data regarding 44,000 single-nucleotide polymorphisms (SNP). After removing SNP markers with minor allele frequency less than 0.05, 374 accessions and 33,584 markers were used for further analysis. Of those, 27 phenotypes were reported in Zhao *et al.* (2011) and McCouch *et al.* (2016). These phenotypes can be classified into four categories: flowering time (flowering time at three locations, photoperiod sensitivity), grain morphology (seed length, seed width, seed surface area, seed length to width ratio, seed volume), plant morphology (culm habit/angle, flag leaf length and width, plant height at maturity), and yield traits (panicle fertility, seed number per panicle, number of primary branches on the main panicle, panicle length, and the number of panicles on each plant). Zhao *et al.* (2011) evaluated flowering time-related traits using data from three locations, while the remaining traits were evaluated at one location (Arkansas). The remaining phenotypes were assessed from the salinity stress experiments conducted in Campbell *et al.* (2017). These traits were classified into three categories: morphological salt response, ionic components of salt stress, and plant morphology. The class morphological salt response represents how plant growth is affected by salinity stress and is composed of the ratio of shoot biomass of salt stressed plants to control, the ratio of root biomass of salt stressed plants to control, the ratio of the number of tillers for salt stressed plants to control, and two metrics that represent the ratio of shoot height of salt stressed plants to control. Ionic components of salt stress is composed of traits that quantify ions important for salinity tolerance (Na^+ and K^+) in both root and shoot tissues. Morphology traits are those that describe the growth of the plant in both control and saline conditions (e.g. shoot biomass, root biomass, shoot height, and tiller number). The data used from Campbell *et al.* (2017) were derived from three to six independent greenhouse experiments performed between July and October 2013. Information for all experiments were combined and best linear unbiased estimators were calculated for each line as described in Campbell *et al.* (2017). The detailed descriptions of the phenotypes are summarized in Supplementary Table S1.

Bayesian confirmatory factor analysis

A CFA under the Bayesian framework was performed to model 48 phenotypes. The number of factors and the pattern of phenotype-factor relationships need to be specified in BCFA prior to model fitting. We constructed six latent variables ($q = 6$) from previous reports (Acquaah 2009; Zhao *et al.* 2011; Campbell *et al.* 2017). The six latent variables derived from our analysis represent the grain morphology, morphology, flowering time, ionic components of salt stress, yield, and morphological salt response (Table S1). Each latent variable captures common signals spanning genetic and environmental effects across all its phenotypes. The latent variables, which determine the observed phenotypes can be modeled as

$$\mathbf{T} = \mathbf{\Lambda F} + \mathbf{s},$$

where $\mathbf{T}$ is the $t \times n$ matrix of observed phenotypes, $\mathbf{\Lambda}$ is the $t \times q$ factor loading matrix, $\mathbf{F}$ is the $q \times n$ latent variables matrix, and $\mathbf{s}$ is the $t \times n$ matrix of specific effects. Here, $\mathbf{\Lambda}$ maps latent variables to the observed variables and can be interpreted as the extent of contribution each latent variable to phenotype. This can be derived by solving the following variance-covariance model.

$$var(\mathbf{T}) = \mathbf{\Lambda \Phi \Lambda}' + \mathbf{\Psi},$$

where $\mathbf{\Phi}$ is the variance of latent variables, and $\mathbf{\Psi}$ is the variance of specific effects (Brown 2014). Six latent variables were assumed to account for the covariance in the observed phenotypes. Moreover, latent variables were assumed to be correlated with each other. Prior distributions were assigned to all unknown parameters. The non-zero coefficients within factor loading matrix $\mathbf{\Lambda}$ were assumed to follow a Gaussian distribution with mean of 0 and variance of 0.01. The variance-covariance matrix $\mathbf{\Phi}$ was assigned an inverse Wishart distribution with a 6×6 identity scale matrix $\mathbf{I}_{66}$ and a degree of freedom 7, $\mathbf{\Phi} \sim \mathcal{W}^{-1}(\mathbf{I}_{66}, 7)$ and an inverse Gamma distribution with scale parameter 1 and shape parameter 0.5 was assigned to $\mathbf{\Psi} \sim \Gamma^{-1}(1, 0.5)$.

We employed the blavaan R package (Merkle and Rosseel 2018) jointly with JAGS (Hornik *et al.* 2003) to fit the above BCFA. The blavaan runs the runjags R package (Denwood 2016) to summarize the Markov chain Monte Carlo (MCMC) and samples unknown parameters from the posterior distributions. Three MCMC chains, each of 5,000 samples with 2,000 burn-in, were used to infer the unknown model parameters. The convergence of the parameters was investigated with trace plots and potential scale reduction factor (PSRF) less than 1.2 (Brooks and Gelman 1998). The PSRF computes the difference between estimated variances among multiple Markov chains and estimated variances within the chain. A large difference indicates non-convergence and may require additional Gibbs sampling.

Subsequently, the posterior means of factor scores ($\mathbf{F}$), which reflect the contribution of latent variables to each accession were estimated. Within each draw of Gibbs sampling, $\mathbf{F}$ was sampled from the conditional distribution of $p(\mathbf{F}|\boldsymbol{\theta}, \mathbf{T})$, where $\boldsymbol{\theta}$ refers to the unknown parameters in $\mathbf{\Lambda}$, $\mathbf{\Phi}$, and $\mathbf{\Psi}$. This conditional distribution was derived with data augmentation (Tanner and Wong 1987) assuming $\mathbf{F}$ as missing data (Lee and Song 2012).

Multivariate genomic best linear unbiased prediction

We fitted a Bayesian multivariate genomic best linear unbiased prediction to separate breeding values from population structure and noise in the six factor scores computed previously.

$$\mathbf{F} = \boldsymbol{\mu} + \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\mu}$ is the vector of intercept, $\mathbf{X}$ is the incidence matrix of covariates, $\mathbf{b}$ is the vector of covariate effects, $\mathbf{Z}$ is the incidence matrix relating accessions with additive genetic effects, $\mathbf{u}$ is the vector of additive genetic effects, and $\boldsymbol{\epsilon}$ is the vector of residuals. The incident matrix $\mathbf{X}$ included subpopulation information (temperate japonica, tropical japonica, indica, aus, aromatic, and admixture), as the rice diversity panel used herein shows a clear substructure (Zhao *et al.* 2011).

A flat prior was assigned to $\boldsymbol{\mu}$ and $\mathbf{b}$, and the joint distribution of $\mathbf{u}$ and $\boldsymbol{\epsilon}$ follows multivariate normal

$$\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N \left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma_u} \otimes \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma_\epsilon} \otimes \mathbf{I} \end{pmatrix} \right],$$

where $\mathbf{G}$ represents the second genomic relationship matrix of VanRaden (2008), $\mathbf{I}$ is the identity matrix, $\boldsymbol{\Sigma_u}$ and $\boldsymbol{\Sigma_\epsilon}$ refer to 6×6 dimensional genetic and residual variance-covariance matrices, respectively. An inverse Wishart distribution with a 6×6 identity scale matrix of $\mathbf{I}_{66}$ and a degree of freedom 6 was assigned as prior for $\boldsymbol{\Sigma_u}, \boldsymbol{\Sigma_e} \sim \mathcal{W}^{-1}(\mathbf{I}_{66}, 6)$. These parameters were selected so that relatively uninformative priors were used. The Bayesian multivariate genomic best linear unbiased prediction model was implemented using the MTM R package (https://github.com/QuantGen/MTM). Posterior mean estimates of genomic correlation between latent variables and predicted breeding values ($\hat{\mathbf{u}}$) were then obtained. The convergence of the estimated parameters was verified by trace plots.

Sample independence in the Bayesian network

Theoretically, BN learning algorithms assume sample independence. In the multivariate genomic best linear unbiased prediction, the residuals between phenotypes were assumed independent through $\mathbf{I_{374x374}}$. However, phenotypic dependencies were introduced by the $\mathbf{G}$ matrix for the additive genetic effects, thereby potentially serving as a confounder. Thus, a transformation of $\hat{\mathbf{u}}$ was carried out to derive an adjusted $\hat{\mathbf{u}}^*$ by eliminating the dependencies in $\mathbf{G}$. For a single trait model, the adjusted $\hat{\mathbf{u}}^*$ can be computed by premultiplying $\hat{\mathbf{u}}$ by $\mathbf{L}^{-1}$, where $\mathbf{L}$ is a lower triangular matrix derived from the Choleskey decompostion of $\mathbf{G}$ matrix ($\mathbf{G} = \mathbf{L}\mathbf{L}^{'}$). Since $\mathbf{u} \sim \mathcal{N}(0, \mathbf{G}\sigma_u^2)$, the distribution of $\hat{\mathbf{u}}^*$ follows $\mathcal{N}(0, \mathbf{I}\sigma_u^2)$ (Callanan and Harville 1989; Vazquez *et al.* 2010)

$$\begin{aligned}
Var(\mathbf{u}^*) &= Var(\mathbf{L^{-1}u}) \\
&= \mathbf{L^{-1}}Var(\mathbf{u})(\mathbf{L^{-1}})^{'} \\
&= \mathbf{L^{-1}G}(\mathbf{L^{-1}})^{'}\sigma_u^2 \\
&= \mathbf{L^{-1}LL}^{'}(\mathbf{L}')^{-\mathbf{1}}\sigma_u^2 \\
&= \mathbf{I}\sigma_u^2.
\end{aligned}$$

This transformation can be extended to a multi-traits model by defining $\mathbf{u}^* = \mathbf{M^{-1}u}$, where $\mathbf{M^{-1}} = \mathbf{I_{qq}} \otimes \mathbf{L^{-1}}$ (Töpner *et al.* 2017). Under the multivariate framework, $\mathbf{u}$ follows $\mathcal{N}(0, \boldsymbol{\Sigma_u} \otimes \mathbf{G})$ and the variance of $\mathbf{u}^*$ is

$$\begin{aligned}
Var(\mathbf{u}^*) &= Var(\mathbf{M^{-1}u}) \\
&= (\mathbf{I_{qq}} \otimes \mathbf{L^{-1}})(\boldsymbol{\Sigma_u} \otimes \mathbf{G})(\mathbf{I_{qq}} \otimes \mathbf{L^{-1}})^{'} \\
&= (\mathbf{I_{qq}} \otimes \mathbf{L^{-1}})(\boldsymbol{\Sigma_u} \otimes \mathbf{LL}^{'})(\mathbf{I_{qq}} \otimes \mathbf{L^{-1}})^{'} \\
&= \boldsymbol{\Sigma_u} \otimes \mathbf{I_{nn}},
\end{aligned}$$

where $\mathbf{L^{-1}LL'(L^{-1})'} = \mathbf{I_{nn}}$. This adjusted $\hat{\mathbf{u}}^*$ was used to learn BN structures between predicted breeding values.

Bayesian network

A BN depicts the joint probabilistic distribution of random variables through their conditional independencies (Scutari and Denis 2014)

$$\mathcal{BN} = (\mathcal{G}, X_V),$$

where $\mathcal{G}$ represents a DAG $= (V,\ E)$ with nodes (V) connected by one or more edges (E) conveying the probabilistic relationships and the random vector $X_V = (X_1, ..., X_K)$ is K random variables. The joint probability distribution can be factorized as

$$P(X_V) = P(X_1, ..., X_K) = \prod_{v=1}^{K} P(X_v|Pa(X_v)),$$

where $Pa(X_v)$ denotes a set of parent nodes of child node X_v. The DAG and joint probability distribution are governed by the Markov condition, which states that every random variable is independent of its non-descendants conditioned on its parents. A BN is known as a Gaussian BN, when all variables or phenotypes are defined as marginal or conditional Gaussian distribution as in the present study.

The adjusted breeding values $\hat{\mathbf{u}}^*$ were used to infer a genomic network structure among the aforementioned six latent variables. There are three types of structure-learning algorithms for BN: constraint-based algorithms, score-based algorithms, and a hybrid of these two (Scutari and Denis 2014). The constraint-based algorithms can be originally traced to the inductive causation algorithm (Verma and Pearl 1991), which uses conditional independence tests for network inference. Briefly, the first step is to identify a d-separation set for each pair of nodes and confer an undirected edge between the two if they are not d-separated. The second step is to identify a v-structure for each pair of non-adjacent nodes, where a common neighbor is the outcome of two non-adjacent nodes. In the last step, compelled edges were identified and oriented, where neither cyclic graph nor new v-structures are permitted. The score-based algorithms are based on heuristic approaches, which first assign a goodness-of-fit score for an initial graph structure and then maximize this score by updating the structure (i.e., add, delete, or reverse the edges of initial graph). The hybrid algorithm includes two steps, restrict and maximize, which harness both constraint-based and score-based algorithms to construct a reliable network. In this study, the two score-based (Hill Climbing and Tabu) and two hybrid algorithms (Max-Min Hill Climbing and General 2-Phase Restricted Maximization) were used to perform structure learning. A flow diagram to illustrate the concept of constraint-based Bayesian netwrok structure learning algorithm is shown in Figure 1.

We quantified the strength of edges and uncertainty regarding the direction of networks, using 500 bootstrapping replicates with a size equal to the number of accessions and performed structure learning for each replicate in accordance with Scutari and Denis (2014). Non-parametric bootstrap resampling aimed at reducing the impact of the local optimal structures by computing the probability of the arcs and directions. Subsequently, 500 learned structures were averaged with a strength threshold of 85% or higher to produce a more robust network structure. This process, known as model averaging, returns the final network with arcs present in at least 85% among all 500 networks. Candidate networks were compared on the basis of the Bayesian information criterion (BIC) and Bayesian Gaussian equivalent score (BGe). The BIC accounts for the goodness-of-fit and model complexity, and BGe aims at maximizing the posterior probability of networks per the data. All BN were learned via the bnlearn R package (Scutari 2010). In bnlearn, the BIC score is rescaled by -2, which indicates that the larger BIC refers to a preferred model.

Data availability

Genotypic data regarding the rice accessions can be downloaded from the rice diversity panel website (`http://www.ricediversity.org/`). Phenotypic data used herein are available in Zhao *et al.* (2011), Campbell *et al.* (2017), and Supplementary File S3.

Results

To elucidate the genetic interdependencies among traits typically recorded in breeding programs, we utilized a collection of 48 publicly available phenotypes recorded on a panel of diverse rice accessions (Zhao *et al.* 2011; Campbell *et al.* 2017). The phenotypic data was derived from two independent studies. The first set of phenotypes was recorded from materials grown in two field environments in Arkansas and Faridpur Bangladesh, and in a greenhouse in Aberdeen, UK (Zhao *et al.* 2011). The 34 phenotypes were recorded at maturity and were largely associated with yield (panicle characteristics flowering time, plant morphology (e.g., height and growth habits), and seed morphological traits. The second study consisted of 14 phenotypes were recorded in a greenhouse environment on plants in the active tillering stage (e.g., 30 day-old plants) under control and saline (14 days of 9.5 dS m-2 NaCl stress). The phenotypes from this study can be classified into three categories: morphological traits (e.g., shoot and root biomass, and plant height), morphological responses to salinity (e.g., the ratio of morphological traits in saline conditions to control), and the ionic components of salinity stress (e.g., Na^+, K^+, and Na^+:K^+ in both root and shoot tissues) (Campbell *et al.* 2017). The complete data set provides an in-depth characterization of phenotypic performance at vegetative and reproductive stages in rice using several classes of traits.

Latent variable modeling

The BCFA model grouped the observed phenotypes into the underlying latent variables on the basis of prior biological knowledge, assuming these latent variables determine the observed phenotypes. This allowed us to study the genetics of each latent variable. A measurement model derived from BCFA evaluating the six latent variables is shown in Figure 2. Forty-eight observed phenotypes were hypothesized to result from the six latent variables: 7 for flowering time, 14 for morphology, 5 for yield, 11 for grain morphology, 6 for physiology, and 5 for salt response. The convergence of the parameters was confirmed graphically with the trace plots and a PSRF value less than 1.2 (Brooks and Gelman 1998; Merkle and Rosseel 2018).

The six latent factors showed strong contributions to the 48 observed phenotypes, with standardized regression coefficients ranging from -0.549 to 0.990 for flowering time, -0.349 to 0.925 for morphology, -0.085 to 0.790 for yield, -0.476 to 0.990 for grain morphology, -0.265 to 0.983 for ionic components of salt stress, and -0.022 to 0.939 for salt response. The latent factor flowering time showed a strong positive contribution to flowering time in Arkansas (Fla) and Flowering time in Arkansas in 2007 (Fla7) (0.990 and 0.926, respectively; Table 1), indicating that larger values for the latent factor can be interpreted as a greater number of days from sowing to emergence of the inflorescence. The latent factor morphology showed the largest positive contributions to traits describing height during the vegetative stage (e.g., height to newest ligule in salt (Hls), 0.920; height to newest ligule in control (Hlc), 0.899; height to the tip of first fully expanded leaf in salt (Hfs), 0.907; and height to tip of first fully expanded leaf in control (Hfc), 0.925;) suggesting that this latent factor is an overall representation of plant size. Yield showed large positive contributions to the observed phenotypes primary panicle branch number (Ppn) and seed number per panicle (Snpp) (0.790 and 0.780, respectively), suggesting that larger values for yield indicate a higher degree of branching and seed number. Observed phenotypes describing seed size (e.g., seed volume (Sv) and brown rice volume (Bvl) (0.990 and 0.986, respectively)) were most strongly associated with grain morphology. The latent factor ionic components of salt stress showed strong positive contributions to two observed phenotypes that quantify the ionic components of salt stress (shoot Na^+:K^+ (Ks) and shoot Na^+ (Nas) (0.983 and 0.975, respectively), indicating that higher values for the latent factor result in greater shoot Na^+ and Na^+:K^+. Finally, the latent factor describing morphological salt response showed strong positive contributions to the observed phenotype describing the effect of salt treatment on plant height (ratio of height to tip of newest fully expanded leaf in salt to that of control plants (Hfr) (0.939)), thus larger values for the latent factor may indicate a more tolerant growth response to salinity.

Genomic correlation among latent variables

To understand the genetic relationships between latent variables, genomic correlation analysis was performed. Genomic correlation is due to pleiotropy or linkage disequilibrium between QTL. The genomic correlations among latent variables are shown in Figure 3. Negative correlations were observed between morphological salt response (Msr) and all other five latent variables. In particular, flowering time (-0.5), yield (-0.54), and grain morphology (-0.74) were negatively correlated with morphological salt response. These results suggest that accessions that harbor alleles for more tolerant morphological salt responses may also have alleles associated with longer flowering times, smaller seeds, and low yield. Similarly, a negative correlation was observed between morphology and yield (-0.56) and between morphology and grain morphology (-0.31). Thus, accessions with alleles associated with large plant size may also have alleles that result in low yield, small grain volume, and lower shoot Na^+ and Na^+:K^+. In contrast, a positive correlation was observed between grain morphology and yield (0.49) and between grain morphology and ionic components of salt stress (0.4). Thus, selection for large grain may result in improved yield, and higher shoot Na^+ and Na^+:K^+.

Bayesian network

To infer the possible network structure between latent variables, BN was performed. Prior to BN, the normality of latent variables was assessed using histogram plots combined with density curves as shown in Supplementary Figure S1. Overall, all the six latent variables approximately followed a Gaussian distribution.

The Bayesian networks learned with the score-based and hybrid algorithms are shown in Figure 4. The structures of BN were refined by model averaging with 500 networks from bootstrap resampling to reduce the impact of local optimal structures. The labels of the arcs measure the uncertainty of the arcs, corresponding to strength and direction (in parenthesis). The former measures the frequency of the arc presented among all 500 networks from the bootstrapping replicates and the latter is the frequency of the direction shown conditional on the presence of the arc. We observed minor differences in the structures presented within and across the two types of algorithms used. In general, small differences were observed within algorithm types compared to those across algorithms. The two score-based algorithms produced a greater number of edges than two hybrid algorithms. The Hill Climbing algorithm produced seven directed connections among the six latent variables. Three connections were indicated towards flowering time from morphological salt response, ionic components of salt stress, and morphology, and two edges to yield from morphology and from grain morphology. Other two edges were observed from ionic components of salt stress to grain morphology and from grain morphology to morphological salt response. A similar structure was generated by the Tabu algorithm, except that the connection between salt response and grain morphology presented an opposite direction. The Max-Min Hill Climbing hybrid algorithm yielded six directed edges from morphological salt response to grain morphology, from ionic components of salt stress to grain morphology, from ionic components of salt stress to flowering time, from flowering time to morphology, from morphology to yield, and from grain morphology to yield. An analogous structure with the only difference observed in the directed edge from morphology to flowering time was inferred with the General 2-Phase Restricted Maximization algorithm. Across all four algorithms, there were four common directed edges: from ionic components of salt stress to flowering time and to grain morphology, and from morphology and grain morphology to yield. The most favorable network was considered the one from the Tabu algorithm, which returned the largest network score in terms of BIC (1086.61) and BGe (1080.88). Collectively, these results suggest that there may be a direct genetic influence of morphology and grain morphology on yield, and physiological components of salt tolerance on grain morphology and flowering time.

Discussion

This study is based on the premise that most phenotypes interact to greater or lesser degrees with each other through underlying physiological and molecular pathways. While these physiological pathways are important for the development of agronomically important characteristics, they are often unknown or difficult to assess in large populations. The approach utilized here leverages phenotypes that can be readily assessed in large populations to quantify these underlying unobserved phenotypes, and elucidates the relationships between these variables.

Understanding the behaviors among phenotypes in the complex traits is critical for genetic improvement of agricultural species (Hickey *et al.* 2017). Graphical modeling offers an avenue to decipher bi-directional associations or probabilistic dependencies among variables of interest in plant and animal breeding. For instance, BN and L1-regularized undirected network can be used to model interrelationships of linkage disequilibrium (LD) (Morota *et al.* 2012; Morota and Gianola 2013) or phenotypic, genetic, and environmental interactions (Xavier *et al.* 2017) in a systematic manner. Importantly, MTM elucidates both direct and indirect relationships among phenotypes. Inaccurate interpretation of these relationships may substantially bias selection decisions (Valente *et al.* 2015; Gianola *et al.* 2015). Thus, we applied BCFA to reduce the dimension of the responses by hypothesizing 48 manifest phenotypes originated from the underlying six constructed latent variables as shown in Figure 2 assuming that these latent traits are most important, followed by application of BN to infer the structures among the six biologically relevant latent variables (Figure 4). Note that there are two differences between the approach employed here and a path analysis. A path analysis 1) uses observed variables rather than latent variables and 2) assumes a network structure is known priori. Thus, one advantage of our approach is that it can model a network structure at the level of latent variables and infer a network structure directly from data when prior information is not available from the literature or previous experiments. The BN represents the conditional dependencies between variables. Care must be taken in interpreting these relationships as a causal effect. Although a good BN is expected to describe the underlying causal structure per the data, when the structure is learned solely on the basis of the observed data, it may return multiple equivalent networks that describe the data well. In practice, searching such a causal structure with observed data needs three additional assumptions (Scutari and Denis 2014): 1) each variable is independent of its non-effects (i.e., direct and indirect) conditioned on its direct causes, 2) the probability distribution of variables is supported by a DAG, where the d-separation in DAG provides all dependencies in the probability distribution, and 3) no additional variables influence the variables within the network. Although it may be difficult to meet these assumptions in the observed data, a BN is equipped with suggesting potential causal relationships among latent variables, which can assist in exploring data, making breeding decisions, and improving management strategies in breeding programs (Rosa *et al.* 2011).

Biological meaning of latent variables and their relationships

We performed BCFA to summarize the original 48 phenotypes with the six latent variables. The number of latent variables and which latent variables load onto phenotypes were determined from the literature. The latent variable morphological salt response (Msr) contributed strongly to salt indices for shoot biomass, root biomass, and two indices for plant height (Table 1). Thus, morphological salt response can be interpreted as the morphological responses to salinity stress, with higher values indicating a more tolerant growth response. The latent variable yield is a representation of overall grain productivity, and contributed strongly to the observed phenotypes primary panicle branch number, seed number per panicle, and panicle length. The positive loading scores on these observable phenotypes indicates that more highly branched, productive panicles will have higher values for yield (Table 1). Seed width, seed volume, and seed surface area contributed significantly to the latent variable grain morphology (Grm) (Table 1). Therefore, these results indicate that the grain morphology is a summary of the overall shape of the grain, where high values represent large, round grains, while low values represent small, slender grains. Considering the grain characteristics of rice subpopulations, temperate japonica accessions are expected to have high values for grain morphology, while indica accessions have lower values for grain morphology. Latent variable morphology (Mrp) is a representation of plant biomass during the vegetative stage (28-day-old plants) (Table 1). Shoot biomass, root biomass, and two metrics for plant height contributed largely to morphology, suggesting that accessions with high values for morphology are tall plants with a large biomass.

Genomic correlation analysis among the six latent variables showed meaningful correlations among several pairs. These genetic correlations can either be caused by linkage or pleiotropy. The former is likely to prevail in species with high LD, which is the case in rice where LD ranges from 100 to 200kb (Huang *et al.* 2010). A negative relationship was observed between morphological salt response and three other latent variables (Figure 3). For instance, a negative correlation between morphological salt response and yield indicates that accessions of samples harboring alleles for superior morphological salt responses (e.g., those that are more tolerant) tend to also harbor alleles for poor yield (Figure 3). The rice diversity panel we used is a representative sample of the total genetic diversity within cultivated rice and contains many unimproved traditional varieties (∼12% of lines in the study are landraces and ∼33% classified as cultivars; Supplementary File S2) and modern breeding lines (Eizenga *et al.* 2014). While traditional varieties exhibit superior adaptation to abiotic stresses, they often have very poor agronomic characteristics including low yield, late flowering, and high photoperiod sensitivity (Thomson *et al.* 2009, 2010). Moreover, the indica and japonica subspecies have contrasting salt responses and very different grain morphology. Japonica accessions tend to have short, round seeds and are more sensitive to salt stress, while indica accessions have long, slender grains and often are more salt tolerant

(Zhao *et al.* 2011; Campbell *et al.* 2017). The negative relationship observed between morphological salt response and grain morphology suggests that lines that harbor alleles for high grain morphology (e.g., large, round grains) tend to also harbor alleles for a tolerant growth response to salt stress. However, no studies have yet reported an association between alleles for grain morphology and morphological salt response. Therefore, it remains to be addressed whether this relationship is due to LD or pleitropy.

Genetic correlations observed between other latent variables may suggest a pleiotropic effect among loci. For instance, a negative relationship was observed between morphological salt response and ionic components of salt stress, indicating that accessions harboring alleles associated with superior morphological salt response also tend to harbor alleles for reduced ion content under salt stress (Figure 3). The relationship between salt tolerance, measured in terms of growth or yield, and Na^+ and Na^+:K^+ has been a documented for decades (reviewed by Munns and Tester (2008)). Moreover, natural variation for Na^+ transporters has been utilized to improve growth and yield under saline conditions in rice and other cereals (Ren *et al.* 2005; Byrt *et al.* 2007; Horie *et al.* 2009; Munns *et al.* 2012; Campbell *et al.* 2017). Therefore, the negative genetic relationships observed between morphological salt response and ion content may be due to the pleiotropic effects of some loci.

The genomic relationships among latent variables including morphology, yield, and grain morphology may have resulted from the selection of alleles associated with good agronomic characteristics. A positive relationship was observed between yield and grain morphology, suggesting that alleles that positively contribute to productive panicles also may contribute to large, round grains. Furthermore, the negative genomic correlation observed between morphology and yield indicates that alleles negatively influencing total plant biomass also have a positive contribution to traits for productive panicles. This genomic relationship may reflect the genetics of harvest index, which is defined as the ratio of grain yield to total biomass. Over the past 50 years, rice breeders have selected high harvest index, resulting in plants with short compact morphology and many highly productive panicles (Hay 1995;

Peng *et al.* 2008).

Although BCFA may yield biologically meaningful results, a potential limitation of BCFA is that we assumed each phenotype does not measure more than one latent variable. This assumption may not always strictly concur with the observational data. Therefore, further studies are required to allow each phenotype to potentially load onto multiple factors in the BCFA framework. An alternative approach is to derive the number of latent variables and determine which latent variables load onto phenotypes directly from observed data, using exploratory FA. This approach was not pursued here because accurate estimation of unknown parameters in the exploratory FA requires a large sample size, which was not the case herein (Brown 2014).

Bayesian network of latent variables

The BN is a probabilistic DAG, which represents the conditional dependencies among phenotypes. The genomic correlation among latent variables described in Figure 3 does not inform the flow of genetic signals nor distinguish direct and indirect associations, whereas BN displays directions between latent variables and separate direct and indirect associations. Therefore, the BN describes the possibility that other phenotypes will change if one phenotype is intervened (i.e., selection). However, caution is required to interpret this network as a causal effect, as the causal BN requires more assumptions, which are usually difficult to meet in observational data (Pearl 2009).

Four common edges or consensus subnetworks across the four BN may be the most reliable substructure of latent variables and may describe the dependence between agronomic traits (Figure 4). For example, edges from grain morphology to yield and morphology to yield can be interpreted as final grain productivity is dependant on specific vegetative characteristics as well grain traits. This is because yield, which represents the overall grain productivity of a plant, depends on morphological characteristics such as the degree of tillering, an architecture that allows the plant to efficiently capture light and carbon, and a stature that is resistant to lodging, the degree of panicle branching, as well as specific grain characteristics such as seed volume and shape. Moreover, there is a direct biological linkage between specific vegetative architectural traits such as tillering and plant height, and yield related traits such as panicle branching and number of seeds per panicle. The degree of branching during both vegetative and reproductive development is dependant on the development and initiation of auxiliary meristems. Several genes have been identified in this pathway and have shown to have pleiotropic effects on tillering and panicle branching (reviewed by Liang *et al.* (2014)). For instance, *OsSPL14* has been shown to be an important regulator of auxiliary branching in both vegetative and reproductive stages in rice (Jiao *et al.* 2010; Miura *et al.* 2010). Moreover, other genes such as *OsGhd8* have been reported to regulate other morphological traits such as plant height and yield through increase panicle branching (Yan *et al.* 2011). The biological importance of these dependencies can also be illustrated by viewing them in the context of genetic improvement, as selection for specific architectural traits (represented by the latent variable morphology) and grain characteristics have traditionally been used as traits to improve rice productivity in many conventional breeding programs (Redona and Mackill 1998; Huang *et al.* 2013).

While the above example provides a plausible network structure between latent variables, edges from ionic components of salt stress to flowering time and to grain morphology are an example of instances where caution should be used to infer causation. As mentioned above, there is an inherent difference in salt tolerance and grain morphological traits between the indica and japonica subspecies. The edges observed for these two latent variables (ionic components of salt stress and grain morphology) in BN may be driven by LD between alleles associated with grain morphology and alleles for salt tolerance rather than pleitropy. Thus, given the current data set, genetic effects for grain morphology may still be conditionally dependant on ionic components of salt stress and the BN may be true, even if there is no direct overlap in the genetic mechanisms for the two traits.

We found that there are some uncertain edges among BN in Figure 4. For instance, direction from morphological salt response to grain morphology is supported by 65% (Tabu), 58% (Max-Min Hill Climbing), and 58% (General 2-Phase Restricted Maximization) bootstrap sampling, whereas the opposite direction is supported by 56% bootstrap sampling (Hill Climbing). An analogous uncertainty was also observed between morphology and flowering time, i.e., the path from morphology to flowering time was supported 60% (Hill Climbing), 51% (Tabu), and 52% (General 2-Phase Restricted Maximization), while the reverse direction was supported 51% (Max-Min Hill Climbing) upon bootstrapping. In addition, the two score-based algorithms captured edges between morphological salt response and flowering time with 70% and 76% bootstrapping evidence. However, this connection was not detected in the two hybrid algorithms. In general, inferring the direction of edges was harder than inferring the presence or absence of undirected edges. Finally, the whole structures of BN were evaluated in terms of the BIC score and BGe. Ranking of the networks was consistent across BIC and BGe and the two score-based algorithms produced networks with greater goodness-of-fit than the two hybrid algorithms. The optimal network was produced by the Tabu algorithm. This is consistent with the previous study reporting that the score-based algorithm produced a better fit of networks in data on maize (Töpner *et al.* 2017).

In conclusion, the present results show the utility of CFA and network analysis to characterize various phenotypes in rice. We showed that the joint use of BCFA and BN can be applied to predict the potential influence of external interventions or selection associated with target traits such as yield in the high-dimensional interrelated complex traits system. We contend that the approaches used herein provide greater insights than pairwise-association measures of multiple phenotypes and can be used to analyze the massive amount of diverse image-based phenomics dataset being generated by the automated plant phenomics platforms (e.g., Furbank and Tester 2011). With a large volume of complex traits being collected through phenomics, numerous opportunities to forge new research directions are generated by using network analysis for the growing number of phenotypes.

Acknowledgments

This work was supported by the National Science Foundation under Grant Number 1736192 to QZ, HW, and GM and Virginia Polytechnic Institute and State University startup funds to GM.

References

Acquaah, G., 2009 *Principles of plant genetics and breeding*. John Wiley & Sons.

Araus, J. L. and J. E. Cairns, 2014 Field high-throughput phenotyping: the new crop breeding frontier. Trends in Plant Science **19**: 52–61.

Araus, J. L., S. C. Kefauver, M. Zaman-Allah, M. S. Olsen, and J. E. Cairns, 2018 Translating high-throughput phenotyping into genetic gain. Trends in Plant Science .

Bhattacharya, A., S. Kourmpetli, and M. R. Davey, 2010 Practical applications of manipulating plant architecture by regulating gibberellin metabolism. Journal of Plant Growth Regulation **29**: 249–256.

Brewer, P. B., H. Koltai, and C. A. Beveridge, 2013 Diverse roles of strigolactones in plant development. Molecular Plant **6**: 18–28.

Brooks, S. P. and A. Gelman, 1998 General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics **7**: 434–455.

Brown, T. A., 2014 *Confirmatory factor analysis for applied research*. Guilford Publications.

Byrt, C. S., J. D. Platten, W. Spielmeyer, R. A. James, E. S. Lagudah, *et al.*, 2007 Hkt1; 5-like cation transporters linked to na+ exclusion loci in wheat, nax2 and kna1. Plant Physiology **143**: 1918–1928.

Cabrera-Bosquet, L., C. Fournier, N. Brichet, C. Welcker, B. Suard, *et al.*, 2016 High-throughput estimation of incident light, light interception and radiation-use efficiency of thousands of plants in a phenotyping platform. New Phytologist **212**: 269–281.

Callanan, T. P. and D. A. Harville, 1989 *Some new algorithms for computing maximum likelihood estimates of variance components*. Iowa State University. Department of Statistics. Statistical Laboratory.

Calus, M. P. and R. F. Veerkamp, 2011 Accuracy of multi-trait genomic selection using different methods. Genetics Selection Evolution **43**: 26.

Campbell, M. T., N. Bandillo, F. R. A. Al Shiblawi, S. Sharma, K. Liu, *et al.*, 2017 Allelic variants of oshkt1; 1 underlie the divergence between indica and japonica subspecies of rice (oryza sativa) for root sodium content. PLoS Genetics **13**: e1006823.

Crain, J., S. Mondal, J. Rutkoski, R. P. Singh, and J. Poland, 2018 Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding. The Plant Genome .

de los Campos, G. and D. Gianola, 2007 Factor analysis models for structuring covariance matrices of additive genetic effects: a bayesian implementation. Genetics Selection Evolution **39**: 481.

Denwood, M., 2016 runjags: An r package providing interface utilities, model templates, parallel computing methods and additional distributions for mcmc models in jags. Journal of Statistical Software, Articles **71**: 1–25.

Eizenga, G. C., M. Ali, R. J. Bryant, K. M. Yeater, A. M. McClung, *et al.*, 2014 Registration of the rice diversity panel 1 for genomewide association studies. Journal of Plant Registrations **8**: 109–116.

Furbank, R. T. and M. Tester, 2011 Phenomics-technologies to relieve the phenotyping bottleneck. Trends Plant Sci. **16**: 635–644.

Gianola, D., G. de los Campos, M. A. Toro, H. Naya, C.-C. Schön, *et al.*, 2015 Do molecular markers inform about pleiotropy? Genetics pp. genetics–115.

Hay, R., 1995 Harvest index: a review of its use in plant breeding and crop physiology. Annals of Applied Biology **126**: 197–216.

Henderson, C. and R. Quaas, 1976 Multiple trait evaluation using relatives' records. Journal of Animal Science **43**: 1188–1197.

Hickey, J. M., T. Chiurugwi, I. Mackay, W. Powell, A. Eggen, *et al.*, 2017 Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nature Genetics **49**: 1297.

Horie, T., F. Hauser, and J. I. Schroeder, 2009 Hkt transporter-mediated salinity resistance mechanisms in arabidopsis and monocot crop plants. Trends in Plant Science **14**: 660–668.

Hornik, K., F. Leisch, and A. Zeileis, 2003 Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of DSC*, volume 2, pp. 1–1.

Huang, R., L. Jiang, J. Zheng, T. Wang, H. Wang, *et al.*, 2013 Genetic bases of rice grain shape: so many genes, so little known. Trends in Plant Science **18**: 218–226.

Huang, X., T. Sang, Q. Zhao, Q. Feng, Y. Zhao, *et al.*, 2010 Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genetics **42**: 961.

Jia, Y. and J.-L. Jannink, 2012 Multiple trait genomic selection methods increase genetic value prediction accuracy. Genetics pp. genetics–112.

Jiao, Y., Y. Wang, D. Xue, J. Wang, M. Yan, *et al.*, 2010 Regulation of osspl14 by osmir156 defines ideal plant architecture in rice. Nature Genetics **42**: 541.

Jöreskog, K. G., 1969 A general approach to confirmatory maximum likelihood factor analysis. Psychometrika **34**: 183–202.

Lee, S.-Y. and X.-Y. Song, 2012 *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*. John Wiley & Sons.

Liang, W.-h., F. Shang, Q.-t. Lin, C. Lou, and J. Zhang, 2014 Tillering and panicle branching genes in rice. Gene **537**: 1–5.

Lo, S.-F., S.-Y. Yang, K.-T. Chen, Y.-I. Hsing, J. A. Zeevaart, *et al.*, 2008 A novel class of gibberellin 2-oxidases control semidwarfism, tillering, and root development in rice. The Plant Cell **20**: 2603–2618.

McCouch, S. R., M. H. Wright, C.-W. Tung, L. G. Maron, K. L. McNally, *et al.*, 2016 Open access resources for genome-wide association mapping in rice. Nature Communications **7**: 10532.

Merkle, E. and Y. Rosseel, 2018 blavaan: Bayesian structural equation models via parameter expansion. Journal of Statistical Software, Articles **85**: 1–30.

Miura, K., M. Ikeda, A. Matsubara, X.-J. Song, M. Ito, *et al.*, 2010 Osspl14 promotes panicle branching and higher grain productivity in rice. Nature Genetics **42**: 545.

Morota, G. and D. Gianola, 2013 Evaluation of linkage disequilibrium in wheat with an l1-regularized sparse markov network. Theoretical and Applied Genetics **126**: 1991–2002.

Morota, G., B. Valente, G. Rosa, K. Weigel, and D. Gianola, 2012 An assessment of linkage disequilibrium in holstein cattle using a bayesian network. Journal of Animal Breeding and Genetics **129**: 474–487.

Mrode, R. A., 2014 *Linear models for the prediction of animal breeding values*. Cabi.

Munns, R., R. A. James, B. Xu, A. Athman, S. J. Conn, *et al.*, 2012 Wheat grain yield on saline soils is improved by an ancestral na+ transporter gene. Nature Biotechnology **30**: 360.

Munns, R. and M. Tester, 2008 Mechanisms of salinity tolerance. Annu. Rev. Plant Biol. **59**: 651–681.

Neapolitan, R. E. *et al.*, 2004 *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ.

Pearl, J., 2009 *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, second edition.

Peñagaricano, F., B. Valente, J. Steibel, R. Bates, C. Ernst, *et al.*, 2015 Searching for causal networks involving latent variables in complex traits: application to growth, carcass, and meat quality traits in pigs. Journal of Animal Science **93**: 4617–4623.

Peng, S., G. S. Khush, P. Virk, Q. Tang, and Y. Zou, 2008 Progress in ideotype breeding to increase rice yield potential. Field Crops Research **108**: 32–38.

Redona, E. and D. Mackill, 1998 Quantitative trait locus analysis for rice panicle and grain characteristics. Theoretical and Applied Genetics **96**: 957–963.

Ren, Z.-H., J.-P. Gao, L.-G. Li, X.-L. Cai, W. Huang, *et al.*, 2005 A rice quantitative trait locus for salt tolerance encodes a sodium transporter. Nature Genetics **37**: 1141.

Rosa, G. J., B. D. Valente, G. de los Campos, X.-L. Wu, D. Gianola, *et al.*, 2011 Inferring causal phenotype networks using structural equation models. Genetics Selection Evolution **43**: 6.

Scutari, M., 2010 Learning bayesian networks with the bnlearn r package. Journal of Statistical Software, Articles **35**: 1–22.

Scutari, M. and J.-B. Denis, 2014 *Bayesian networks: with examples in R*. Chapman and Hall/CRC.

Shakoor, N., S. Lee, and T. C. Mockler, 2017 High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. Current Opinion in Plant Biology **38**: 184–192.

Sun, J., J. E. Rutkoski, J. A. Poland, J. Crossa, J.-L. Jannink, *et al.*, 2017 Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. The Plant Genome .

Tanner, M. A. and W. H. Wong, 1987 The calculation of posterior distributions by data augmentation. Journal of the American statistical Association **82**: 528–540.

Thomson, M. J., M. de Ocampo, J. Egdane, M. A. Rahman, A. G. Sajise, *et al.*, 2010 Characterizing the saltol quantitative trait locus for salinity tolerance in rice. Rice **3**: 148–160.

Thomson, M. J., A. M. Ismail, S. R. McCouch, and D. J. Mackill, 2009 Marker assisted breeding. In *Abiotic Stress Adaptation in Plants*, pp. 451–469, Springer.

Töpner, K., G. J. Rosa, D. Gianola, and C.-C. Schön, 2017 Bayesian networks illustrate genomic and residual trait connections in maize (Zea mays L.). G3: Genes, Genomes, Genetics **7**: 2779–2789.

Umehara, M., A. Hanada, S. Yoshida, K. Akiyama, T. Arite, *et al.*, 2008 Inhibition of shoot branching by new terpenoid plant hormones. Nature **455**: 195.

Valente, B. D., G. Morota, F. Peñagaricano, D. Gianola, K. Weigel, *et al.*, 2015 The causal meaning of genomic predictors and how it affects construction and comparison of genome-enabled selection models. Genetics **200**: 483–494.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. **91**: 4414–4423.

Vazquez, A., D. Bates, G. Rosa, D. Gianola, and K. Weigel, 2010 An r package for fitting generalized linear mixed models in animal breeding 1. Journal of animal science **88**: 497–504.

Verma, T. and J. Pearl, 1991 Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pp. 255–270, New York, NY, USA, Elsevier Science Inc.

Wang, Y. and J. Li, 2006 Genes controlling plant architecture. Current Opinion in Biotechnology **17**: 123–129.

Xavier, A., B. Hall, S. Casteel, W. Muir, and K. M. Rainey, 2017 Using unsupervised learning techniques to assess interactions among complex traits in soybeans. Euphytica **213**: 200.

Yan, W.-H., P. Wang, H.-X. Chen, H.-J. Zhou, Q.-P. Li, *et al.*, 2011 A major qtl, ghd8, plays pleiotropic roles in regulating grain productivity, plant height, and heading date in rice. Molecular Plant **4**: 319–330.

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. Nature Communications **2**: 467.

Zhou, F., Q. Lin, L. Zhu, Y. Ren, K. Zhou, *et al.*, 2013 D14–scf d3-dependent degradation of d53 regulates strigolactone signalling. Nature **504**: 406.

Tables

Table 1: Standardized factor loadings obtained from the Bayesian confirmatory factor analysis. PSD refers to the posterior standard deviation of standardized factor loadings.

Latent variable	Observed phenotype	Loading	PSD
Flowering time	Flowering time at Arkansas (Fla)	0.990	0.002
Flowering time	Flowering time at Faridpur (Flf)	0.500	0.045
Flowering time	Flowering time at Aberdeen (Flb)	0.578	0.038
Flowering time	FT ratio of Arkansas/Aberdeen (Flaa)	-0.212	0.053
Flowering time	FT ratio of Faridpur/Aberdeen (Flfa)	-0.549	0.041
Flowering time	Year07 Flowering time at Arkansas (Fla7)	0.926	0.008
Flowering time	Year06 Flowering time at Arkansas (Fla6)	0.886	0.013
Morphology	Culm habit (Cuh)	0.227	0.027
Morphology	Flag leaf length (Fll)	0.116	0.057
Morphology	Flag leaf width (Flw)	-0.044	0.058
Morphology	Plant height (Plh)	0.440	0.047
Morphology	Shoot BM Control (Sbc)	0.534	0.042
Morphology	Shoot BM Salt (Sbs)	0.456	0.048
Morphology	Root BM Control (Rbc)	0.418	0.048
Morphology	Root BM Salt (Rbs)	0.280	0.054
Morphology	Tiller No Salt (Tns)	-0.349	0.051
Morphology	Tiller No Control (Tbc)	-0.318	0.052
Morphology	Ht Lig Salt (Hls)	0.920	0.011
Morphology	Ht Lig Control (Hlc)	0.899	0.014
Morphology	Ht FE Salt (Hfs)	0.907	0.013
Morphology	Ht FE Control (Hfc)	0.925	0.011
Yield	Panicle number per plant (Pnu)	0.190	0.020
Yield	Panicle length (Pal)	0.455	0.057
Yield	Primary panicle branch number (Ppn)	0.790	0.041
Yield	Seed number per panicle (Snpp)	0.780	0.043
Yield	Panicle fertility (Paf)	-0.085	0.081
Grain Morphology	Seed length (Sl)	0.251	0.029
Grain Morphology	Seed width (Sw)	0.876	0.015
Grain Morphology	Seed volume (Sv)	0.990	0.002
Grain Morphology	Seed surface area (Ssa)	0.901	0.012
Grain Morphology	Brown rice seed length (Bsl)	0.158	0.055
Grain Morphology	Brown rice seed width (Bsw)	0.837	0.019
Grain Morphology	Brown rice surface area (Bsa)	0.902	0.012
Grain Morphology	Brown rice volume (Bvl)	0.986	0.002
Grain Morphology	Seed length/width ratio (Slwr)	-0.476	0.045
Grain Morphology	Brown rice length/width ratio (Blwr)	-0.432	0.047
Grain Morphology	Grain length McCouch2016 (Glmc)	0.047	0.064
Ionic components of salt stress	Na K Shoot (Ks)	0.983	0.003
Ionic components of salt stress	Na Shoot (Nas)	0.975	0.004
Ionic components of salt stress	K Shoot Salt (Kss)	-0.265	0.051
Ionic components of salt stress	Na K Root (Kr)	0.061	0.052
Ionic components of salt stress	Na Root (Nar)	0.001	0.053
Ionic components of salt stress	K Root Salt (Krs)	-0.095	0.052
Morphological salt response	Shoot BM Ratio (Sbr)	0.410	0.047
Morphological salt response	Root BM Ratio (Rbr)	0.395	0.051
Morphological salt response	Tiller No Ratio (Tbr)	-0.022	0.057
Morphological salt response	Ht Lig Ratio (Hlr)	0.665	0.036
Morphological salt response	Ht FE Ratio (Hfr)	0.939	0.019

Figures

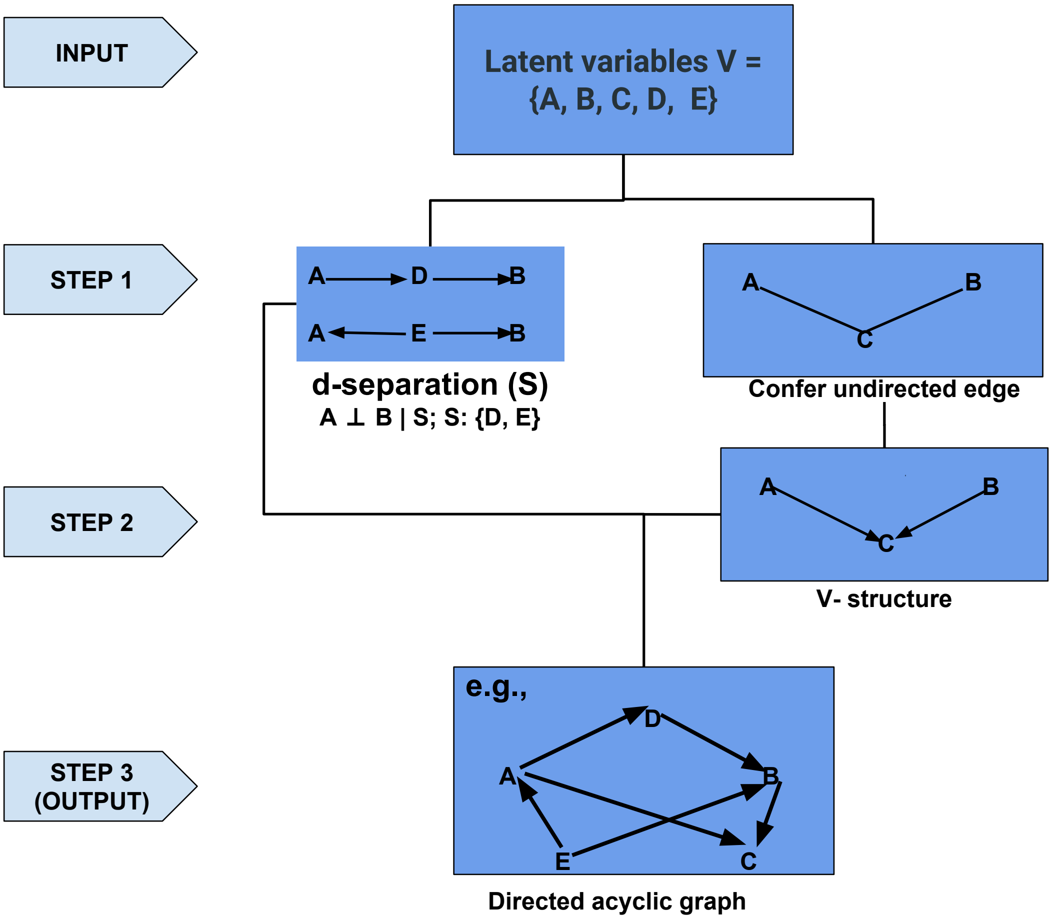

Figure 1: Flow diagram to illustrate the concept of constraint-based structure learning algorithm for a Bayesian network. The A, B, C, D, and E represent five nodes or latent variables. S refers to a set of d-separation. The directed acyclic graph shown in Step 3 is one possible completed partially directed acyclic graph.

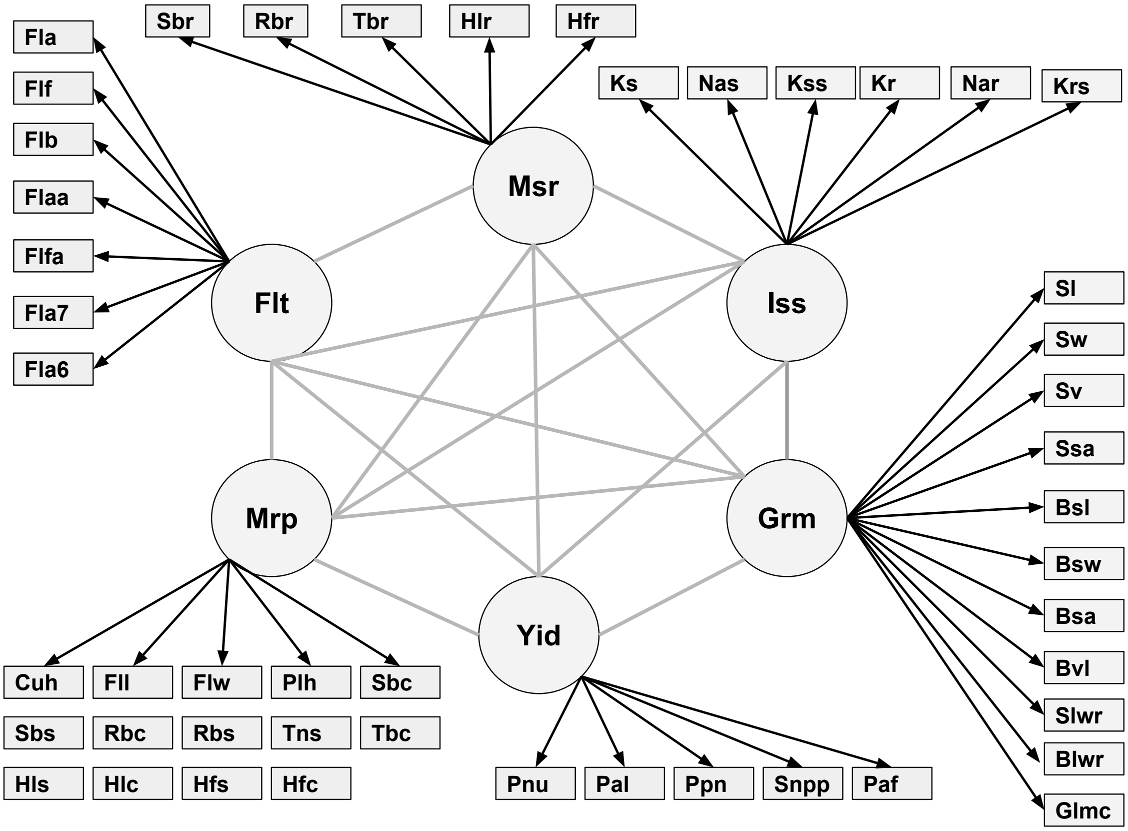

Figure 2: Relationship between six latent variables and observed phenotypes. Msr: morphological salt response; Iss: ionic components of salt stress; Grm: grain morphology; Yid: yield; Mrp: morphology; Flt: flowering time. Abbreviations of observed phenotypes are shown in Table S1.

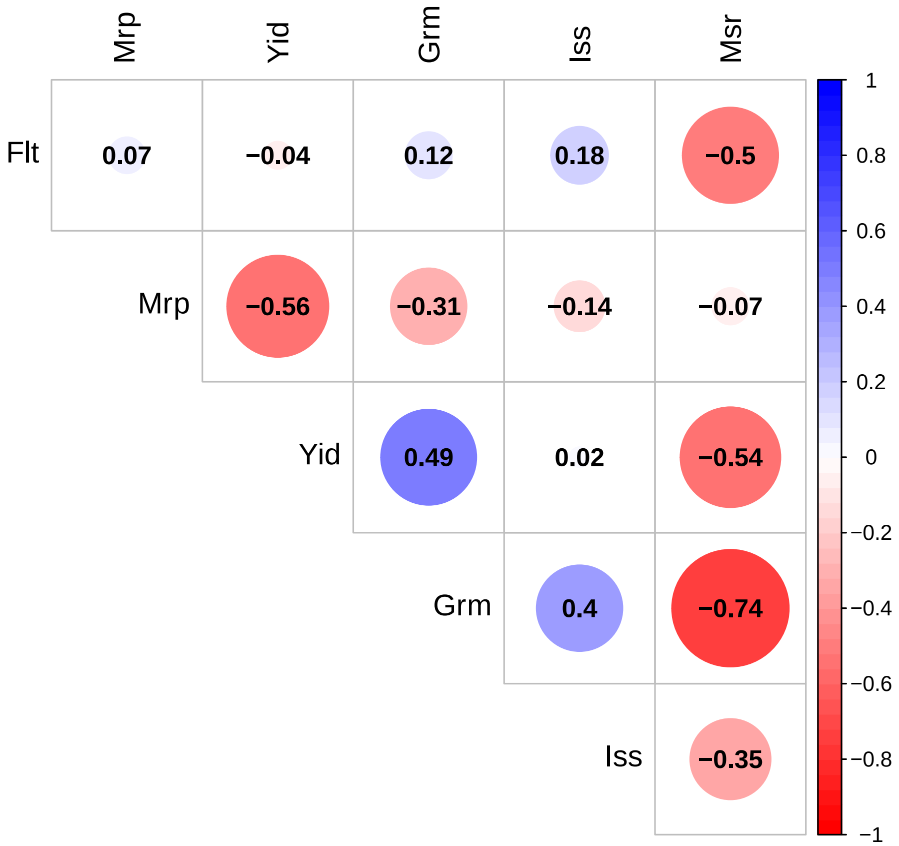

Figure 3: Genomic correlation of six latent variables. The size of each circle, degree of shading, and value reported correspond to the correlation between each pair of latent variables. Msr: morphological salt response; Iss: ionic components of salt stress; Grm: grain morphology; Yid: yield; Mrp: morphology; Flt: flowering time.

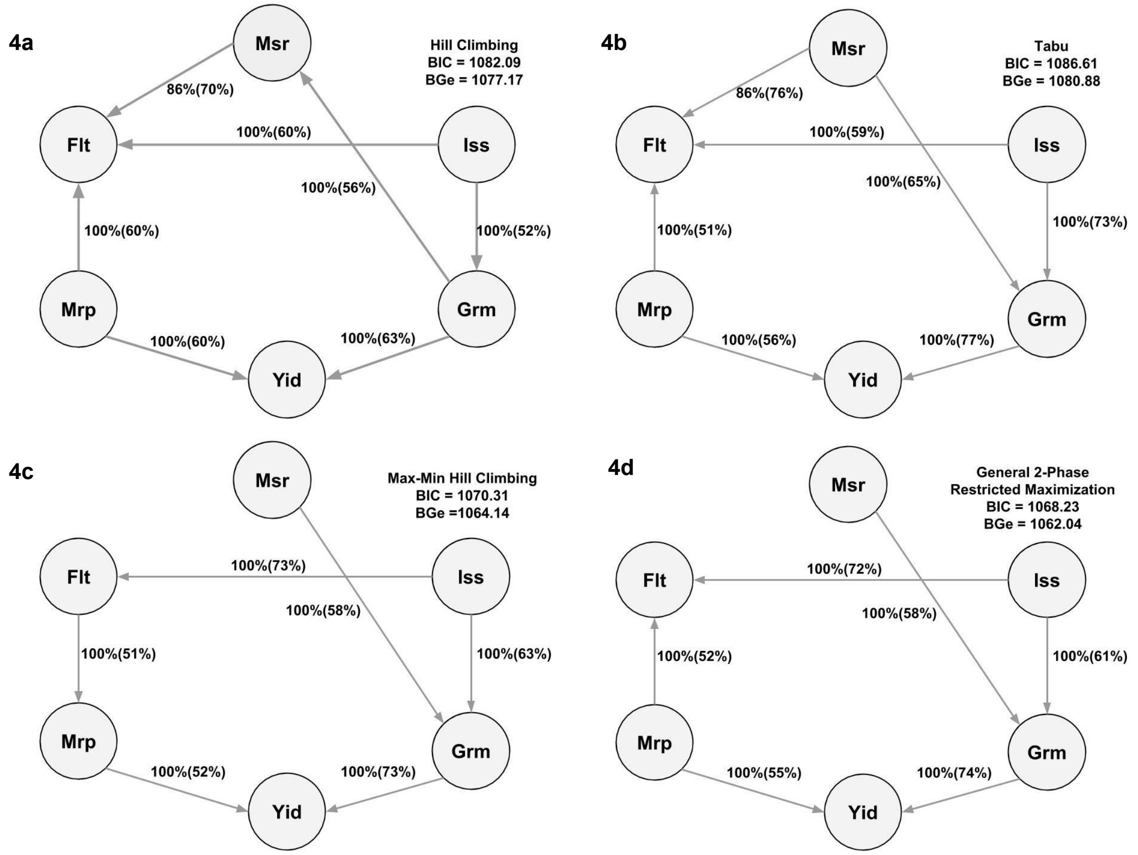

Figure 4: Bayesian networks between six latent variables based on two score-based (4a: Hill Climbing and 4b: Tabu) and two hybrid (4c: Max-Min Hill Climbing and 4d: General 2-Phase Restricted Maximization) algorithms. The quality of the structure was evaluated by bootstrap resampling and model averaging across 500 replications. Labels of the edges refer to the strength and direction (parenthesis) which measure the confidence of the directed edge. The strength indicates the frequency of the edge is present and the direction measures the frequency of the direction conditioned on the presence of edge. BIC: Bayesian information criterion score. BGe: Bayesian Gaussian equivalent score. Msr: morphological salt response; Iss: ionic components of salt stress; Grm: grain morphology; Yid: yield; Mrp: morphology; Flt: flowering time.