Harnessing phenotypic networks and structural equation models to improve genome-wide association analysis

Mehdi Momen[1], Malachy T. Campbell[1,2], Harkamal Walia[2], and Gota Morota[1*]

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

Keywords: structural equation modeling, Bayesian network, genome-wide association, multiple-traits

Running title: network analysis in rice

ORCID: 0000-0002-2562-2741 (MM), 0000-0002-8257-3595 (MTC), 0000-0002-9712-5824 (HW), and 0000-0002-3567-6911 (GM).

* Corresponding author:

Gota Morota
Department of Animal and Poultry Sciences
Virginia Polytechnic Institute and State University
175 West Campus Drive
Blacksburg, Virginia 24061 USA.
E-mail: morota@vt.edu

Abstract

Plant breeders and breeders alike seek to develop cultivars with maximal agronomic value. The merit of breeding material is often assessed using many, often genetically correlated traits. As intervention on one trait will affect the value of another, breeding decisions should consider the relationships between traits. With the proliferation of multi-trait genome-wide association studies (MTM-GWAS), we can infer putative genetic signals at the multivariate scale. However, a standard MTM-GWAS does not accommodate the network structure of phenotypes, and therefore does not address how the traits are interrelated. We extended the scope of MTM-GWAS by incorporating phenotypic network structures into GWAS using structural equation models (SEM-GWAS). In this network GWAS model, one or more phenotypes appear in the equations for other phenotypes as explanatory variables. A salient feature of SEM-GWAS is that it can partition the total single nucleotide polymorphism (SNP) effects into direct and indirect effects. In this paper, we illustrate the utility of SEM-GWAS using biomass, root biomass, water use, and water use efficiency in rice. We found that water use efficiency is directly impacted by biomass and water use and indirectly by biomass and root biomass. In addition, SEM-GWAS partitioned significant SNP effects influencing water use efficiency into direct and indirect effects as a function of biomass, root biomass, and water use efficiency, providing further biological insights. These results suggest that the use of SEM may enhance our understanding of complex relationships between GWAS traits.

Background

Elite high-yielding crop varieties are the result of generations of targeted selection for multiple characteristics. In many cases, plant and animal breeders alike seek to improve many, often correlated, phenotypes simultaneously. Thus, breeders must consider the interaction between traits during selection. For instance, genetic selection for one trait may increase or decrease the expression of another trait, depending on the genetic correlation between the two. While consideration of the genetic correlation between traits is essential in this respect, modeling recursive interactions between phenotypes provides important insights for developing breeding and management strategies for crops that cannot be realized with conventional multivariate approaches alone.. In particular, inferring the structure of phenotypic networks from observational data is critical for our understanding of the interdependence of multiple phenotypes (Valente et al., 2010; Wang and van Eeuwijk, 2014; Yu et al., 2018).

Genome-wide association studies (GWAS) have become increasingly popular approaches for the elucidation of the genetic basis of economically important traits. They have been successful in identifying single nucleotide polymorphism (SNPs) associated with a wide spectrum of phenotypes, including yield, abiotic and biotic stresses, and morphology in plants (Huang and Han, 2014). For many studies, multiple, often correlated, traits are recorded on the same material, and association mapping is preformed for each trait separately. While such approaches may yield powerful, biologically meaningful results, they fail to adequately capture the genetic interdependancy among traits and impose limitations on understanding the genetic mechanisms underlying a complex system of traits. When multiple phenotypes possess correlated structures, multi-trait GWAS (MTM-GWAS), which is the application of mutli-trait models (MTM) (Henderson and Quaas, 1976) to GWAS, is a standard approach. The rationale behind this is to leverage genetic correlations among phenotypes to increase statistical power for the detection of quantitative trait loci, particularly for traits that have low heritability or are scarcely recorded.

While MTM-GWAS is a powerful approach to capture the genetic correlations between traits for genetic inference, it fails to address how the traits are interrelated, or elucidate the mechanisms that give rise to the observed correlation. The early work of Sewall Wright sought to infer causative relations between correlated variables though path analysis (Wright, 1921). This seminal work gave rise to structural equation models (SEM), which assesses the nature and magnitude of direct and indirect effects of multiple interacting variables. Although SEM remains a powerful approach to model the relationships among variables in complex systems, its use has been limited in biology.

Recently, Momen et al. (2018) proposed the SEM-GWAS framework by incorporating phenotypic networks and SNPs into MTM-GWAS through SEM (Wright, 1921; Haavelmo, 1943). In contrast to standard multivariate statistical techniques, the SEM framework opens up a multivariate modeling strategy that accounts for recursive (an effect from one phenotype is passed onto another phenotype) and simultaneous (reciprocal) structures among its variables (Goldberger, 1972; Bielby and Hauser, 1977). Momen et al. (2018) showed that SEM-GWAS can supplement MTM-GWAS, and is capable of partitioning the source of the SNP effects into direct and indirect effects, which helps to provide a better understanding of the relevant biological mechanisms. In contrast, MTM-GWAS, which does not take the network structure between phenotypes into account, estimates overall SNP effects that are mediated by other phenotypes, and combines direct and indirect SNP effects.

Current climate projections predict an increase in the incidence of drought events and elevated temperatures throughout the growing season (Wehner et al., 2017). These elevated temperatures will drive higher evapotranspirational demands, and combined with the inconsistency of rainfall events, will increase the rate of drought onset and intensity, and impact on crop growth and productivity (Challinor et al., 2014; Mann and Gleick, 2015; Otkin et al., 2017; Zampieri et al., 2017; Zhao et al., 2017). To counter the effects of climate change on agricultural productivity, drought-resilient crops must be developed. However, progress towards this goal is often hindered by the complex biological basis of drought tolerance (Tuberosa and Salvi, 2006; Sinclair, 2011; Mir et al., 2012; Passioura, 2012). The ability to maintain productivity under limited water availability involves a suite of morphological and physiological traits (Passioura, 2012). Among these is the ability to access available water and utilize it for growth. Thus, studying traits associated with water capture (e.g. root biomass and architecture) and utilization (e.g. water-use efficiency) is essential. However, of equal importance is a robust statistical framework that allows these complex traits to be analyzed jointly and causal relationships among traits to be inferred.

In this study, we applied SEM-GWAS and MTM-GWAS to incorporate the phenotypic network structures related to shoot and root biomass and to drought responses in rice (*Oryza sativa* L.) from a graphical modeling perspective. Graphical modeling offers statistical inferences regarding complex associations among multivariate phenotypes. Plant biomass and drought stress responses are considered to be interconnected through physiological pathways that may be related to each other, requiring the specification of recursive effects using SEM. We combined GWAS with two graphical modeling approaches: a Bayesian network was used to infer how each SNP affects a focal phenotype directly or indirectly through other phenotypes, and SEM was applied to represent the interrelationships among SNPs and multiple phenotypes in the form of equations and path diagrams.

Materials and Methods

Experimental data set

The plant material used in our analysis consisted of a rice diversity panel of $n = 357$ inbred accessions of *O. sativa* collected from a diverse range of regions, which are expected to capture much of the genetic diversity within cultivated rice (Zhao et al., 2011). All lines were genotyped with 700,000 SNPs using the high-density rice array from Affymetrix (Santa Clara, CA, USA) such that there was approximately 1 SNP every 0.54Kb across the rice genome (Zhao et al., 2011; McCouch et al., 2016). We used PLINK v1.9 software (Purcell et al., 2007) to remove SNPs with a call rate ≤ 0.95 and a minor allele frequency ≤ 0.05. Missing genotypes were imputed using Beagle software version 3.3.2 (Browning and Browning, 2007). Finally, 411,066 SNPs were retained for further analysis.

Phenotypic data

We analyzed four biologically important traits for drought responses in rice: projected shoot area (PSA), root biomass (RB), water use (WU), and water use efficiency (WUE). These phenotypes are derived from two separate studies (Campbell et al., 2017a, 2018). The aim of the first study was to evaluate the effects of drought on shoot growth (Campbell et al., 2018). Here, the diversity panel was phenotyped using an automated phenotyping platform in Adelaide, SA, Australia. This new phenotyping technology enables us to produce high-resolution spatial and temporal image-derived phenotypes, which can be used to capture dynamic growth, development, and stress responses (Berger et al., 2010; Golzarian et al., 2011; Campbell et al., 2015, 2017b).

The plants were phenotyped over a period of 20 days, starting at 13 days after they were transplanted into soil and ending at 33 days. Each day, the plants were watered to a specific target weight to ensure the soil was completely saturated. The plants were then imaged from three angles. These pictures were processed to remove all background objects, leaving just pixels for the green shoot tissue. We summed the pixels from each picture to obtain an estimate of the shoot biomass. We refer to this metric as PSA. With this system, we also obtained the weights, prior to watering and after watering, for each pot on each day. From this data, we estimated the amount of water that is used by each plant. WU was calculated as Pot $\text{Weight}_{(r-1)} - \text{Pot Weight}_{(r)}$, where r is time, and WUE is the ratio of PSA to WU. Although this data has not yet been published, a description of the phenotyping system and insight into the experimental design can be found in Campbell et al. (2015).

The aim of the second study was to assess salinity tolerance in the rice diversity panel. The plants were grown in a hydroponics system in a greenhouse. Salt stress was imposed for two weeks, and destructive phenotyping performed at 28 days after transplantation. A number of traits were recorded, including RB. The experimental design of this study is fully described in (Campbell et al., 2017a). All the aforementioned phenotypes were measured under controlled conditions. The 15th day of imaging was selected for analysis of PSA, WU, and WUE, which is equivalent to 28 days after transplantation, so that it matched the age at which RB was recorded.

Multi-trait genomic best linear unbiased prediction

A Bayesian multi-trait genomic best linear unbiased prediction (MT-GBLUP) model was used for four traits to obtain posterior means of model residuals as inputs for inferring a phenotypic network.

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{g} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is the vector observations for $t = 4$ traits, $\boldsymbol{\mu}$ 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, $\mathbf{g}$ is the vector of additive genetic effects, and $\boldsymbol{\epsilon}$ is the vector of residuals. The incident matrix $\mathbf{X}$ only included intercepts for the four traits examined in this study. Under the infinitesimal model of inheritance, the $\mathbf{g}$ and $\boldsymbol{\epsilon}$ were assumed to follow a multivariate Gaussian distribution $\mathbf{g} \sim N(0, \sum_g \otimes \mathbf{G})$ and $\boldsymbol{\epsilon} \sim N(0, \sum_\epsilon \otimes \mathbf{I})$, respectively, where $\mathbf{G}$ is the $n \times n$ genomic relationship matrix for genetic effects, $\mathbf{I}$ is the identify matrix for residuals, $\sum_g$ and $\sum_\epsilon$ are the $t \times t$ variance-covariance matrices of genetic effects and residuals, respectively, and $\otimes$ denotes the Kronecker product. The $\mathbf{G}$ matrix was computed as $\mathbf{W}\mathbf{W}'/2\sum_{j=1}^{m} p_j(1-p_j)$, where $\mathbf{W}$ is the centered marker incidence matrix taking values of $0-2p_j$ for zero copies of the reference allele, $1-2p_j$ for one copy of the reference allele, and $2-2p_j$ for two copies of the reference allele (VanRaden, 2008). Here, p_j is the allele frequency at SNP $j = 1, \cdots, m$. We assigned flat priors for the intercept and the vector of fixed effects. The vectors of random additive genetic effects and residual effects were assigned independent multivariate normal priors with null mean and inverse Wishart distributions for the covariance matrices.

A Markov chain Monte Carlo (MCMC) approach based on Gibbs sampler was used to explore posterior distributions. We used a burn-in of 25,000 MCMC samples followed by an additional 150,000 MCMC samples. The MCMC samples were thinned with a factor of two, resulting in 75,000 MCMC samples for inference. Posterior means were then calculated for estimating model parameters. The MTM R package was used to fit the above regression model (https://github.com/QuantGen/MTM).

Learning structures using Bayesian network

Networks or graphs can be used to model interactions. Bayesian networks describe conditional independence relationships among multivariate phenotypes. Each phenotype is connected by an edge to another phenotype if they directly affect each other given the rest of the phenotypes, whereas the absence of edge implies conditional independence given the rest of phenotypes. Several algorithms have been proposed to infer plausible structures in Bayesian networks, assuming independence among the realization of random variables (Scutari, 2010). The estimated residuals from MT-GBLUP were used as inputs, and we applied the Max-Min

Parents and Children (MMPC) algorithm from the constraint-based structure learning category to infer the network structure among the four traits examined in this study (Scutari et al., 2018). We selected this algorithm because it was suggested in a recent study, Töpner et al. (2017), which showed that the constraint-based algorithms performed better for the construction of networks than score-based counterparts. This algorithm is similar to the inductive causation algorithm (Tsamardinos et al., 2003) that was first used in Valente et al. (2010) to infer a phenotypic network. The bnlearn R package was used to learn the Bayesian phenotypic network throughout this analysis with mutual information as the test, and the statistically significant level set at $\alpha = 0.01$ (Scutari, 2010). We computed the Bayesian information criterion (BIC) score of a network and estimated the strength and uncertainty of direction of each edge probabilistically by bootstrapping as described in Scutari and Denis (2014). In addition, the strength of the edge was assessed by computing the change in the BIC score when that particular edge was removed from the network, while keeping the rest of the network intact.

Multi-trait GWAS

We used the following MTM-GWAS that does not account for the inferred network structure by extending the single trait GWAS counterpart of Kennedy et al. (1992) and Yu et al. (2006). For ease of presentation, it is assumed that each phenotype has null mean.

$$.\mathbf{y} = \mathbf{w}\mathbf{s} + \mathbf{Z}\mathbf{g} + \boldsymbol{\epsilon},$$

where $\mathbf{w}$ is the jth SNP being tested, $\mathbf{s}$ represents the vector of fixed jth SNP effect, and $\mathbf{g}$ is the vector of additive polygenic effect. The aforementioned variance-covariance structures were assumed for $\mathbf{g}$ and $\boldsymbol{\epsilon}$. The MTM-GWAS was fitted individually for each SNP, where the output is a vector of marker effect estimates for each trait, i.e. $\hat{\mathbf{s}} = [\hat{s}_{\text{PSA}}, \hat{s}_{\text{WU}}, \hat{s}_{\text{WUE}}, \hat{s}_{\text{RB}}]$.

Structural equation model for GWAS

A structural equation model is capable of conveying directed network relationships among multivariate phenotypes involving recursive effects. The SEM described in Gianola and Sorensen (2004) in the context of linear mixed models was extended for GWAS, according to Momen et al. (2018).

$$\mathbf{y} = \mathbf{\Lambda y} + \mathbf{ws} + \mathbf{Zg} + \boldsymbol{\epsilon}$$

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{y}_3 \\ \boldsymbol{y}_4 \end{pmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{I}_1\lambda_{\text{PSA}\rightarrow\text{RB}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{I}_1\lambda_{\text{PSA}\rightarrow\text{WU}} & \mathbf{I}_2\lambda_{\text{RB}\rightarrow\text{WU}} & \mathbf{0} & \mathbf{0} \\ \mathbf{I}_1\lambda_{\text{PSA}\rightarrow\text{WUE}} & \mathbf{I}_2\lambda_{\text{RB}\rightarrow\text{WUE}} & \mathbf{I}_3\lambda_{\text{WU}\rightarrow\text{WUE}} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{y}_3 \\ \boldsymbol{y}_4 \end{bmatrix}$$

$$+ \begin{bmatrix} \mathbf{w}_{j1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_{j2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{w}_{j3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{w}_{j4} \end{bmatrix} \begin{bmatrix} s_{j1} \\ s_{j2} \\ s_{j3} \\ s_{j4} \end{bmatrix}$$

$$+ \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_4 \end{bmatrix} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \mathbf{g}_3 \\ \mathbf{g}_4 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \boldsymbol{\epsilon}_3 \\ \boldsymbol{\epsilon}_4 \end{bmatrix}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{\Lambda}$ is the lower triangular matrix of regression coefficients or structural coefficients based on the learned network structure from the Bayesian network, and the other terms are as defined earlier.

Note that the structural coefficients $\mathbf{\Lambda}$ determine that the phenotypes which appear in the left-hand side also appear in the right-hand side, and represent the edge effect size from phenotype to phenotype in Bayesian networks. If all elements of $\mathbf{\Lambda}$ are equal to 0, then this model is equivalent to MTM-GWAS. Gianola and Sorensen (2004) showed that the reduction and re-parameterization of a SEM mixed model can yield the same joint probability distribution of observation as MTM, suggesting that the expected likelihoods of MTM and SEM are the same (Varona et al., 2007). For example, we can rewrite the SEM-GWAS model as

$$\begin{aligned}\mathbf{y} &= (\mathbf{I}-\mathbf{\Lambda})^{-1}\mathbf{ws} + (\mathbf{I}-\mathbf{\Lambda})^{-1}\mathbf{Zg} + (\mathbf{I}-\mathbf{\Lambda})^{-1}\boldsymbol{\epsilon} \\ &= \boldsymbol{\theta}^* + \mathbf{g}^* + \boldsymbol{\epsilon}^*\end{aligned}$$

where $\text{Var}(\mathbf{g}^*) \sim (\mathbf{I}-\mathbf{\Lambda})^{-1}\mathbf{G}(\mathbf{I}-\mathbf{\Lambda})^{'-1}$ and $\text{Var}(\boldsymbol{\epsilon}^*) \sim (\mathbf{I}-\mathbf{\Lambda})^{-1}\mathbf{R}(\mathbf{I}-\mathbf{\Lambda})^{'-1}$. This transformation changes SEM-GWAS into MTM-GWAS, which ignores the network relationships among traits (Gianola and Sorensen, 2004; Varona et al., 2007). However, Valente et al. (2013) stated that SEM allows for the prediction of the effects of external interventions, which can be useful for making selection decisions that are not possible with MTM. We used SNP Snappy software to perform MTM-GWAS and SEM-GWAS (Meyer and Tier, 2012). To identify candidate SNPs that may explain direct (in the absence of mediation by other traits) and indirect (with intervention and mediation by other traits) effects for each trait, the SNPs from MTM-GWAS were ranked according to p-values for each trait. The 20 most significant SNPs were then selected, and all genes within 200 kb were considered to be potential candidate genes.

Results

Trait correlations and network structure

Multi-phenotypes were split into genetic values and residuals by fitting the MT-GBLUP. The estimates of genomic and residual correlations among the four traits measured in this study are shown in Table 1. Correlations between all traits ranged from 0.48 to 0.92 for genomics and -0.13 to 0.83 for residuals. The estimated genomic correlations can arise from pleiotropy or linkage disequilibrium (LD). Although pleiotropy is the most durable and stable source of genetic correlations, LD is considered to be less important than pleiotropy because alleles at two linked loci may become non-randomly associated by chance and be distorted through recombination (Gianola et al., 2015; Momen et al., 2017).

We postulated that the learned networks can provide a deeper insight into relationships among traits than simple correlations or covariances. Figure 1 shows a network structure inferred using the MMPC algorithm. This is a fully recursive structure because there is at least one incoming or outgoing edge for each node. Unlike the MTM-GWAS model, the inferred graph structure explains how the phenotypes may be related to each other either directly or indirectly mediated by one or more variables. We found a direct dependency between PSA and WUE, which can also be mediated by WU. A direct connection was also found between RB and WU, and WU and WUE.

Measuring the strength of probabilistic dependence for each arc is crucial in Bayesian network learning (Scutari and Denis, 2014). As shown in Figure 1, the strength of each arc was assessed with 2,500 bootstrap samples with a significance level at $\alpha = 0.01$. The labels on the edges indicate the proportion of bootstrap samples supporting the presence of the edge and the proportion supporting the direction of the edges are provided in parentheses. Learned structures were averaged with a strength threshold of 85% or higher to produce a more robust network structure. Edges that did not meet this threshold were removed from the networks. In addition, we used BIC as goodness-of-fit statistics measuring how well the paths mirror the dependence structure of the data (Table 2). The BIC assign higher scores to any path that fit the data better. The BIC score reports the importance of each arc by its removal from the learned structure. We found that removing PSA → WUE resulted in the largest decrease in the BIC score, suggesting that this path is playing the most important role in the network structure. This was followed by WU → WUE, RB → WU, and PSA → WU.

Structural equation coefficients

The inferred Bayesian network among PSA, RB, WU, and WUE in Figure 1 was modeled using a set of structural equations to estimate SEM parameters and SNP effects, as shown in Figure 2, which can be statistically expressed as

$$
\begin{aligned}
\mathbf{y}_{1_{\mathrm{PSA}}} &= \mathbf{w}_j s_{j(y_{1_{\mathrm{PSA}}})} + \mathbf{Z}_1\mathbf{g}_1 + \boldsymbol{\epsilon}_1 \\
\mathbf{y}_{2_{\mathrm{RB}}} &= \mathbf{w}_j s_{j(y_{2_{\mathrm{RB}}})} + \mathbf{Z}_2\mathbf{g}_2 + \boldsymbol{\epsilon}_2 \\
\mathbf{y}_{3_{\mathrm{WU}}} &= \lambda_{13}\mathbf{y}_{1_{\mathrm{PSA}}} + \lambda_{23}\mathbf{y}_{2_{\mathrm{RB}}} + \mathbf{w}_j s_{j(y_{3_{\mathrm{WU}}})} + \mathbf{Z}_3\mathbf{g}_3 + \boldsymbol{\epsilon}_3 \\
&= \lambda_{13}[\mathbf{w}_j s_{j(y_{1_{\mathrm{PSA}}})} + \mathbf{Z}_1\mathbf{g}_1 + \boldsymbol{\epsilon}_1] + \lambda_{23}[\mathbf{w}_j s_{j(y_{2_{\mathrm{RB}}})} + \mathbf{Z}_2\mathbf{g}_2 + \boldsymbol{\epsilon}_2] + \mathbf{w}_j s_{j(y_{3_{\mathrm{WU}}})} + \mathbf{Z}_3\mathbf{g}_3 + \boldsymbol{\epsilon}_3 \\
\mathbf{y}_{4_{\mathrm{WUE}}} &= \lambda_{14}\mathbf{y}_{1_{\mathrm{PSA}}} + \lambda_{34}\mathbf{y}_{3_{\mathrm{WU}}} + \mathbf{w}_j s_{j(y_{4_{\mathrm{WUE}}})} + \mathbf{Z}\mathbf{g} + \boldsymbol{\epsilon}_4 \\
&= \lambda_{14}[\mathbf{w}_j s_{j(y_{1_{\mathrm{PSA}}})} + \mathbf{Z}_1\mathbf{g}_1 + \boldsymbol{\epsilon}_1] \\
&+ \lambda_{34}\{\lambda_{13}[\mathbf{w}_j s_{j(y_{1_{\mathrm{PSA}}})} + \mathbf{Z}_1\mathbf{g}_1 + \boldsymbol{\epsilon}_1] + \lambda_{23}[\mathbf{w}_j s_{j(y_{2_{\mathrm{RB}}})} + \mathbf{Z}_2\mathbf{g}_2 + \boldsymbol{\epsilon}_2] + \mathbf{w}_j s_{j(y_{3_{\mathrm{WU}}})} + \mathbf{Z}_3\mathbf{g}_3 + \boldsymbol{\epsilon}_3\} \\
&+ \mathbf{w}_j s_{j(y_{4_{\mathrm{WUE}}})} + \mathbf{Z}\mathbf{g} + \boldsymbol{\epsilon}_4.
\end{aligned}
$$

The corresponding estimated $\boldsymbol{\Lambda}$ matrix is

$$
\boldsymbol{\Lambda} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \lambda_{13_{\mathrm{PSA}\to\mathrm{WU}}} & \lambda_{23_{\mathrm{RB}\to\mathrm{WU}}} & 0 & 0 \\ \lambda_{14_{\mathrm{PSA}\to\mathrm{WUE}}} & 0 & \lambda_{34_{\mathrm{WU}\to\mathrm{WUE}}} & 0 \end{bmatrix}.
$$

Table 3 represents the magnitude of estimated structural path coefficients: λ_{13}, λ_{23}, λ_{14}, and λ_{34} for PSA on WU, RB on WU, PSA on WUE, and WU on WUE, respectively. The structural coefficients ($\lambda_{ii'}$) describe the rate of change of trait i with respect to trait $i^{'}$. The largest magnitude of the structural coefficient was 1.339, which was estimated for PSA→WUE, whereas the lowest was 0.005, which was estimated for RB→ WU. The WU→ WUE relationship has a negative path coefficient, whereas the remainder were all positive.

Interpretation of SNP effects

We implemented SEM-GWAS as an extension of the MTM-GWAS method for analysis of the joint genetic architecture of the four measured traits, to partition SNP effects into direct and indirect (Alwin and Hauser, 1975). The results of the decomposition of SNP effects are discussed for each trait separately below. Because the network only revealed indirect effects for WU and WUE, we focused on these traits for candidate gene discovery.

Projected Shoot Area (**PSA**): Figure 3 shows a Manhattan plot of SNP effects on the PSA. According to the path diagram, there is no intervening trait or any mediator variable for PSA (Figure 2). It is possible that the PSA architecture is only influenced by the direct SNP effects, and is not affected by any other mediators or pathways. Hence, the total effect of jth SNP on PSA is equal to its direct effects.

$$\begin{aligned}
\text{Direct}_{s_j \rightarrow y_{1_{\text{PSA}}}} &= s_{j(y_{1_{\text{PSA}}})} \\
\text{Total}_{s_j \rightarrow y_{1_{\text{PSA}}}} &= \text{Direct}_{s_j \rightarrow y_{1_{\text{PSA}}}} \\
&= s_{j(y_{1_{\text{PSA}}})}
\end{aligned}$$

Root Biomass (**RB**): No incoming edges were detected for RB, resulting in a similar pattern to PSA, which suggests that SNP effects on RB were not mediated by other phenotypes. As shown in Figure 4, a Manhattan plot for RB consists of direct and total effects.

$$\begin{aligned}\text{Direct}_{s_j \rightarrow y_{2_{\text{RB}}}} &= s_{j(y_{2_{\text{RB}}})}\\ \text{Total}_{s_j \rightarrow y_{2_{\text{RB}}}} &= \text{Direct}_{s_j \rightarrow y_{2_{\text{RB}}}}\\ &= s_{j(y_{2_{\text{RB}}})}\end{aligned}$$

Water use (**WU**): Based on Figure 2, a total single SNP effect on WU is attributable to two mediators, as it has two incoming edges: PSA and RB. Thus, the SNP effects transmitted from PSA and RB also contribute to the total SNP effects on WU. Under these conditions, the estimated total SNP effects for WU cannot be simply described as the direct effect of a given SNP, since the indirect effects of PSA and RB must also be considered. This is different to MTM-GWAS, which does not distinguish between the effects mediated by mediator phenotypes, and only captures the overall SNP effects. Here it should be noted that the extent of SNP effects on PSA and RB are controlled by the structural equation coefficients λ_{13} and λ_{23}. Figure 5 shows a Manhattan plot of SNP effects on WU. We found that the indirect RB → WU path had the least impact on overall effects, whereas indirect PSA → WU path had almost the same contribution as the direct SNP effects.

$$\begin{aligned}\text{Direct}_{s_j \rightarrow y_{3_{\text{WU}}}} &= s_{j(y_{3_{\text{WU}}})}\\ \text{Indirect}(1)_{s_j \rightarrow y_{3_{\text{WU}}}} &= \lambda_{13} s_{j(y_{1_{\text{PSA}}})}\\ \text{Indirect}(2)_{s_j \rightarrow y_{3_{\text{WU}}}} &= \lambda_{23} s_{j(y_{2_{\text{RB}}})}\\ \text{Total}_{s_j \rightarrow y_{2_{\text{WU}}}} &= \text{Direct}_{s_j \rightarrow y_{2_{\text{WU}}}} + \text{Indirect}(1)_{s_j \rightarrow y_{3_{\text{WU}}}} + \text{Indirect}(2)_{s_j \rightarrow y_{3_{\text{WU}}}}\\ &= s_{j(y_{3_{\text{WU}}})} + \lambda_{13} s_{j(y_{1_{\text{PSA}}})} + \lambda_{23} s_{j(y_{2_{\text{RB}}})}\end{aligned}$$

Water Usage Efficiency (**WUE**): The overall SNP effects for WUE can be partitioned into one direct and four indirect genetic signals (Figure 2). WUE is the only phenotype trait that does not have any outgoing path to other traits. According to Figure 6, the extents of the SNP effects among the four indirect paths were 1) RB → WUE mediated by WU, 2) PSA → WUE mediated by WU, 3) WU → WUE, and 4) PSA → WUE, in increasing order. We found that the SNP effect transmitted through RB had the smallest effect on the WUE, suggesting that modifying the size of the QTL effect for RB may not have a noticeable effect on WUE, whereas a change in PSA had a noticeable effect on WUE. The magnitude of the relationship between RB and WUE is proportional to the product of structural coefficients $\lambda_{23} \times \lambda_{34} = 0.005 \times -0.455$. PSA influenced WUE via two indirect paths, and strongly depends on the structural coefficients $\lambda_{14} = 1.339$ and $\lambda_{13}\lambda_{34} = 0.767 \times -0.455$ for PSA → WUE and PSA → WU → WUE, respectively. It should be noted that the indirect effect transmitted through PSA → WUE was greater than the direct effects of a given SNP on WUE. This is because the structural coefficient between WU and WUE has a negative sign, resulting in transmitted indirect SNP effects that can change the sign and magnitude of the total effect on WUE, even from positive values to negative values. However, this indicates that the modification and selection of plants for WU may impact WUE, even for the opposite direction.

The direct and indirect effects are summarized with the following equation:

$$\begin{aligned}
\text{Direct}_{s_j \to y_{4\text{WUE}}} &= s_{j(y_{4\text{WUE}})} \\
\text{Indirect}(1)_{s_j \to y_{4\text{WUE}}} &= \lambda_{14} s_{j(y_{1\text{PSA}})} \\
\text{Indirect}(2)_{s_j \to y_{4\text{WUE}}} &= \lambda_{34} s_{j(y_{3\text{WU}})} \\
\text{Indirect}(3)_{s_j \to y_{4\text{WUE}}} &= \lambda_{13}\lambda_{34} s_{j(y_{1\text{PSA}})} \\
\text{Indirect}(4)_{s_j \to y_{4\text{WUE}}} &= \lambda_{23}\lambda_{34} s_{j(y_{2\text{RB}})} \\
\text{Total}_{S_j \to y_{4\text{WUE}}} &= \text{Direct}_{s_j \to y_{4\text{WUE}}} + \text{Indirect}(1)_{s_j \to y_{4\text{WUE}}} + \text{Indirect}(2)_{s_j \to y_{4\text{WUE}}} \\
&\quad + \text{Indirect}(3)_{s_j \to y_{4\text{WUE}}} + \text{Indirect}(4)_{s_j \to y_{4\text{WUE}}} \\
&= s_{j(y_{4\text{WUE}})} + \lambda_{14} s_{j(y_{1\text{PSA}})} + \lambda_{34} s_{j(y_{3\text{WU}})} + \lambda_{13}\lambda_{34} s_{j(y_{1\text{PSA}})} + \lambda_{23}\lambda_{34} s_{j(y_{2\text{RB}})}
\end{aligned}$$

The indirect and direct SNP effects across all possible paths with the total effect for WU and WUE are compared in Supplementary Figures 1 and 2. The results showed a positive agreement for PSA→WU and direct effect with total effect on WU, whereas the RB→WU showed less association with total effect (Supplementary Figure 1). A positive association between direct and indirect effects was also observed for WU. When the paths to WUE were mediated by WU, all transmitted indirect effects have negative associations with total SNP effects (Supplementary Figure 2). PSA→WU→WUE showed a greater association with total SNP effects than that of RB→WU→WUE and WU→WUE. The strongest positive association with total effect was observed for PSA→WUE. The positive association between total effects with direct effect, and direct with indirect, were also relatively high. Supplementary Figure 3 shows that the agreement between the total SNP signals derived from MTM-GWAS and SEM-GWAS. We found that PSA and RB presented a stronger agreement between MTM-GWAS and SEM-GWAS, probably because the direct effect is equivalent to the total effect for these phenotypes, and does not require the estimation of additional parameters. The only discrepancy that may arise is that there might be some differences in the inferred effects, due to the methods used for inference. In contrast, the association between MTM-GWAS and SEM-GWAS was slightly weaker for WU and WUE, due to uncertainty regarding the additional estimated structural coefficients associated with the indirect effects included in the computation of total effects, especially given that our model is not fully recursive.

Trade offs between MTM- and SEM-GWAS models suggest enrichment of candidate genes for the traits

Nineteen of the top 20 SNPs showed a direct effect on WU ($P_{direct} < 0.01$), while for WUE all SNPs showed an indirect effect ($P_{indirect} \geq 0.01$). Interestingly, for both traits, all indirect effects at these loci could be attributed to PSA, indicating that alleles that influence shoot biomass may have an effect on WU and WUE. The positive relationship between dry matter production and WU is widely documented across multiple crops, and is simply because larger plants have a greater water demand than small plants (Ehdaie, 1995; Hubick et al., 1986; Ismail and Hall, 1992). Moreover, in this study the plants were grown under simulated paddy conditions (i.e., with water-saturated soil); thus; there was sufficient water to meet these demands and sustain shoot growth in larger plants. In conditions where water is limited such relationships may not hold true.

Several candidate genes associated with plant growth were identified in close proximity to SNPs with indirect effects. For instance, two genes with known roles in the regulation of organ size and plant growth, *SMOS1* and *OVP1*, were identified for WU and WUE, respectively. *OVP1* was located near the most significant SNP identified for WUE, and SEM-GWAS showed that this SNP influences WUE indirectly through PSA. *OVP1* is known to influence abiotic stress responses in rice, as well as growth and development in Arabidopsis (Zhang et al., 2011; Khadilkar et al., 2015; Schilling et al., 2014). In rice, ectopic expression of *OVP1* led to increased cell membrane integrity and accumulation of proline during cold stress (Zhang et al., 2011). The production of proline is important for the maintenance of cell water relations during water deficits. High proline levels are often observed during osmotic stresses, and effectively reduce the osmotic potential of the cell, which restores turgor pressure and facilitates cell growth. While Zhang et al. (2011) demonstrated a role for *OVP1* in cold tolerance, the mechanisms that lead to the observed improvement in cold tolerance remain to be elucidated. However, the Arabidopsis ortholog of *OVP1*, *AVP1*, has been widely characterized and has been shown to be involved with the partitioning of photosynthates into the phloem and transport to the roots (Khadilkar et al., 2015). Khadilkar et al. (2015) showed that higher expression of *AVP1* led to increased phloem loading of photosynthates, and resulted in a larger overall shoot and root biomass. Moreover, Schilling et al. (2014) showed similar effects in barley plants, which over expressed *AVP1*, further indicating that this gene may influence plant growth (Schilling et al., 2014).

SMOS1 is located at $\sim$ 18.81 Mb on chromosome 5, and encodes an AP2 transcription factor. Initially identified through a mutant screen, *SMOS1* knockout plants exhibit nearly normal vegetative and reproductive development; however the leaf blade, leaf sheath, roots, flowers, and seeds are significantly reduced in the mutant lines (Aya et al., 2014). The shorter length of these organs was attributed to a reduction in cell size, indicating that this gene is involved in the regulation of cell growth. These observations were further supported by Aya et al. (2014) and Hirano et al. (2017), who showed that *SMOS1* binds to the promoter of the cell expansion gene, phosphate-induced protein 1 (*PHI1*). While the effect of *OVP1* and *SMOS1* on shoot growth and water use efficiency remain to be elucidated in rice, the known functions of these genes, as well as their presence in close proximity to SNPs with indirect effects on WUE through PSA, are encouraging and warrant further investigation.

Two notable genes were identified in close proximity to SNPs with direct effects on WU that have been shown to participate in ABA-induced stomatal closure. The stomatal aperture is controlled by a cascade of events that involve ABA as an upstream signal and reactive oxygen species (ROS) as an intermediate signal. The first gene, *PYL11*, encodes an ABA receptor. Kim et al. (2011) determined that *PYL11* plays a role in seed germination and early growth, and showed that over-expression of *PYL11* led to hypersensitivity to ABA. However, in a recent study, Miao et al. (2018) generated multiple high-order PYL knockout mutants in rice, and characterized several traits in field conditions (Miao et al., 2018). After ABA treatment, a greater proportion of stomates remained open in *pyl11* compared to WT, indicating that stomatal closure is impaired in the *pyl11* mutants. However, it was also shown that the total stomatal aperture of *pyl11* was still greater than other *pyl* mutants, suggesting that other genes may have a stronger effect on stomatal responses to ABA.

The second gene,*RADICAL-INDUCED CELL DEATH1 (RCD1)*, is located at ~ 35.87 Mb on chromosome 3, and encodes a WWE-domain containing protein. *RCD1* has been well characterized in Arabidopsis for hormonal responses and *ROS* homeostasis (Ahlfors et al., 2004). Interestingly, *RCD1* and other members of the Similar to *RCD* One (SRO) family have been shown to be involved with the regulation of the stomatal aperture and water loss. For example, Ahlfors et al. (2004) showed that *rcd1* mutants exhibit greater stomatal conductance and greater water loss than the WT (Ahlfors et al., 2004). The over-expression of a *RCD1* ortholog in rice, *OsSRO1c*, resulted in the opposite phenotype being observed, with a decreased stomatal aperture and reduced water loss compared with the WT (You et al., 2012). The *ROS* H_2O_2 has been shown to act downstream of ABA and to result in stomatal closure. Members of the SRO family are involved in the regulation of ROS homeostasis; thus, the stomata and water loss phenotypes exhibited by mis-regulation of SRO or *RCD1* may be due to the inability to properly regulate H_2O_2 levels (You et al., 2012).

Discussion

The relationship between biomass and WU in rice may involve complex network pathways with recursive effects. These network relationships cannot be modeled using a standard MTM-GWAS model. In this study, we incorporated the network structure between four phenotypes, PSA, RB, WU, and WUE, into a multivariate GWAS model using SEM. In GWAS, a distinction between undirected edges and directed edges is crucial, because often biologists and breeders are interested in studying and improving a suite of traits rather than a single trait in isolation. Moreover, intervention on one trait often influences the expression of another (Shipley, 2016). As highlighted in Alwin and Hauser (1975), one of the advantages of SEM is that it is capable of splitting the total effects into direct and indirect effects. In regards to genetic studies, SEM enables the researcher to elucidate the underlying mechanism by which an intervention trait may influence phenotypes using a network relationship (Wu et al., 2010; Onogi et al., 2016).

Detecting putative causal genes is of considerable interest for determining which traits will be affected by specific loci from a biological perspective, and consequently partitioning the genetic signals according to the paths determined. Although the parameter interpretations of SEM as applied to QTL mapping (Li et al., 2006; Mi et al., 2010), expression QTL (Liu et al., 2008), or genetic selection (Valente et al., 2013) have been actively pursued, the work of Momen et al. (2018) marks one of the first studies to pay particular attention at the level of individual SNP effect in genome-wide SEM analyses. The SEM embeds a flexible framework for performing such network analysis in a GWAS context, and the current study demonstrates its the first application in crops. We assumed that modeling a system of four traits in rice simultaneously may help us to examine the sources of SNP effects in GWAS in greater depth. Therefore, we compared two GWAS methodologies that have the ability to embed multiple traits jointly, so that the estimated SNP effects from both models have different meanings. The main significance of SEM-GWAS, relative to MTM-GWAS, is to include the relationship between SNPs and measured phenotypes, coupled with relationships that are potentially meditated by other phenotypes (mediator traits). This advances GWAS, and consequently the information obtained from phenotypic networks describing such interrelationships can be used to predict the behavior of complex systems (Momen et al., 2018). Although we analyzed the observed phenotypes in the current study, the factor analysis component of SEM can be added to SEM-GWAS by deriving latent factors from multiple phenotypes (e.g., Verhulst et al., 2017; Leal-Gutiérrez et al., 2018). The inference of a phenotypic network structure was carried out using a Bayesian network, which has applications in genetics ranging from modeling linkage disequilibrium (Morota et al., 2012) to epistasis (Han et al., 2012).

Effective water use and water capture are essential for the growth of plants in arid environments, where water is a limiting factor. These processes are tightly intertwined, and therefore must be studied in a holistic manner. In the current study, we sought to understand the genetic basis of water use, water capture, and growth by examining PSA, RB, WU, and WUE in a diverse panel of rice accessions. The identification of several genes that have been reported to regulate one or more of these processes highlights the interconnectedness of PSA, RB, WU, and WUE (Ho et al., 2005; Zhang et al., 2011; Schilling et al., 2014). We used SEM analysis to observe significant interactions between intermediate variables and independent variables in each of the four phenotypes studied. The two most significant QTL identified harbored two genes that are known to regulate *OVP1* (which is located near the most significant SNP identified for WUE) and *SMOS1*, for WUE and WU, respectively. As discussed above, the effect of *OVP1* and *SMOS1* on shoot growth and water use efficiency remain to be elucidated in rice; their known functions, as well as their presence in close proximity to SNPs with indirect effects on WUE through PSA, are encouraging and warrant further investigation. We also found two important genes in close proximity to SNPs that have direct effects on WU, and have been shown to participate in ABA-induced stomatal closure. The first gene, *PYL11*, encodes an ABA receptor and the second gene, *RCD1*, is located at 35.87 Mb on chromosome 3 and encodes a WWE-domain containing protein. The identification of these genes within this QTL interval suggests that these genes may have an impact on RB and WU. These findings highlight the significant potential and importance of mediator relationship inclusion in the association between other variables in the inferred graph.

A deep understanding of the complex relationship between effective water use and water capture, and its impact on plant growth in arid environments, is critical as we continue to develop germplasm that is resilient to challenging future climates. As with the significant recent advances in phenotyping and remote sensing technologies, tomorrow's plant breeders will have a new suite of tools to quantify morphological, physiological, and environmental variables at high resolutions. To fully harness these emerging technologies and leverage these multi-dimensional datasets for crop improvement, new analytical approaches must be developed that integrate genomic and phenomic data in a biologically meaningful framework. This study examined multiple phenotypes determined using a Bayesian network that may serve as potential factors to allow intervention in complex trait GWAS. The SEM-GWAS seems to provide enhanced statistical analysis of MTM-GWAS by accounting for phenotypic network structures.

Acknowledgments

This work was partially supported by the National Science Foundation under Grant Number 1736192 to HW and GM.

Author contribution statement

MTC and HW designed and conducted the experiments. MM and MTC analyzed the data. MM and GM conceived the idea and wrote the manuscript. MTC and HW discussed results and revised the manuscript. GM supervised and directed the study. All authors read and approved the manuscript.

Tables

Table 1: Genomic (upper triangular), residual (lower triangular) correlations and genomic heritablities (diagonals) of four traits in the rice with posterior standard deviations in parentheses. Projected shoot area (PSA), root biomass (RB), water use (WU), and water use efficiency (WUE).

	PSA	WU	WUE	RB
PSA	0.677 (0.092)	0.846 (0.043)	0.920 (0.018)	0.515 (0.102)
WU	0.443 (0.152)	0.643 (0.097)	0.744 (0.076)	0.479 (0.114)
WUE	0.829 (0.052)	0.106 (0.182)	0.576 (0.092)	0.517 (0.107)
RB	0.030 (0.218)	-0.134 (0.216)	0.111 (0.195)	0.733 (0.083)

Table 2: Bayesian information criterion (BIC) for the network learned using the Max-Min Parents and Children (MMPC) algorithm. BIC denote BIC scores for pairs of nodes and reports the change in the score caused by an arc removal relative to the entire network score. Projected shoot area (PSA), root biomass (RB), water use (WU), and water use efficiency (WUE).

Algorithm	from	to	BIC
MMPC	PSA	WUE	-311.039
	PSA	WU	-2.680
	WU	WUE	-108.154
	RB	WU	-24.284

Table 3: Structural coefficients (λ) estimates derived from the structural equation models. Projected shoot area (PSA), root biomass (RB), water use (WU), and water use efficiency (WUE).

Path	λ	Structural coefficient
PSA → WU	λ_{13}	0.767
RB → WU	λ_{23}	0.005
PSA → WUE	λ_{14}	1.339
WU → WUE	λ_{34}	-0.455

Figures

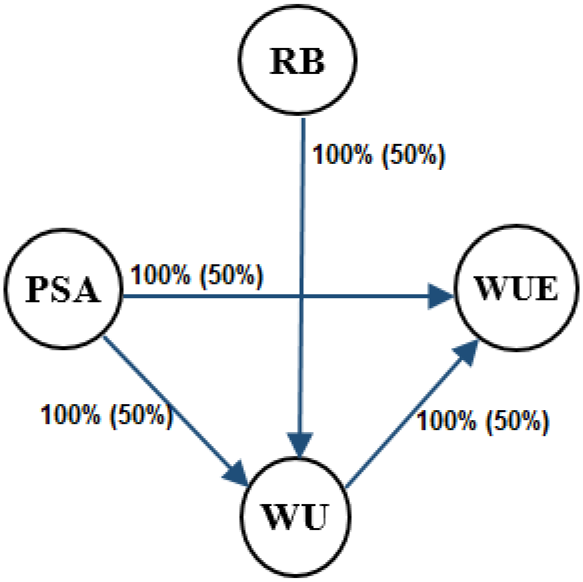

Figure 1: Scheme of inferred network structure using the Max-Min Parents and Children (MMPC) algorithm. Structure learning test was performed with 2,500 bootstrap samples with mutual information as the test statistic with a significance level at $\alpha = 0.01$. Labels of the edges refer to the strength 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. PSA: projected shoot area; RB: root biomass; WU: water use; WUE: water use efficiency.

454
Figures

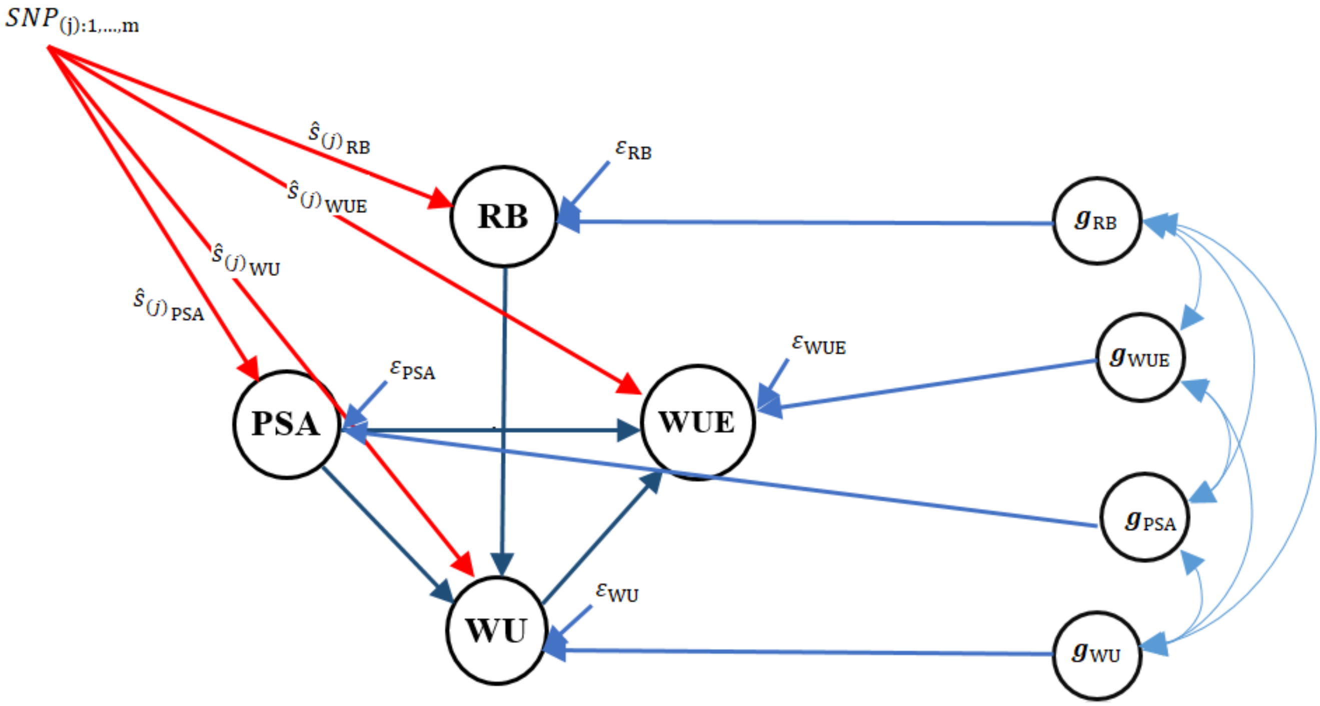

Figure 2: Pictorial representation of phenotypic network and SNP effects ($\hat{s}$) using the structural equation model for four traits. Unidirectional arrows indicate the direction of effects and bidirectional arrows represent genetic correlations (g) among phenotypes. PSA: projected shoot area; RB: root biomass; WU: water use; WUE: water use efficiency; ϵ: residual.

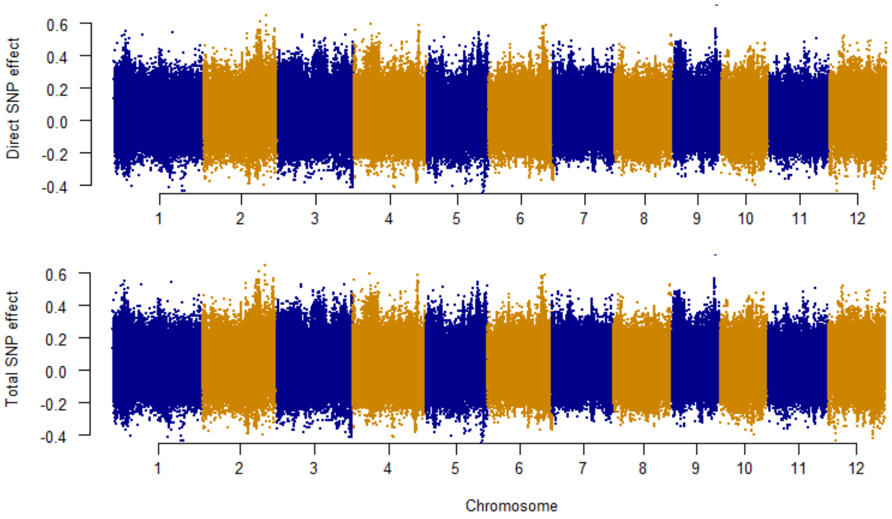

Figure 3: Manhattan plots of direct (affecting each trait without any mediation) and total (sum of all direct and indirect) SNP effects on projected shoot area (PSA) using SEM-GWAS based on the network learned by the MMPC algorithm. Each point represents a SNP and the height of the SNP represents the extent of its association with PSA.

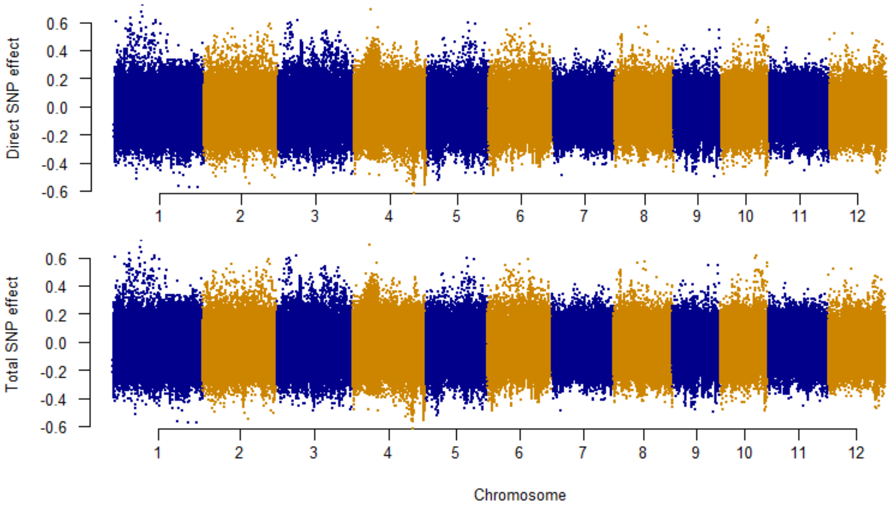

Figure 4: Manhattan plots of direct (affecting each trait without any mediation) and total (sum of all direct and indirect) SNP effects on root biomass (RB) using SEM-GWAS based on the network learned by the MMPC algorithm. Each point represents a SNP and the height of the SNP represents the extent of its association with RB.

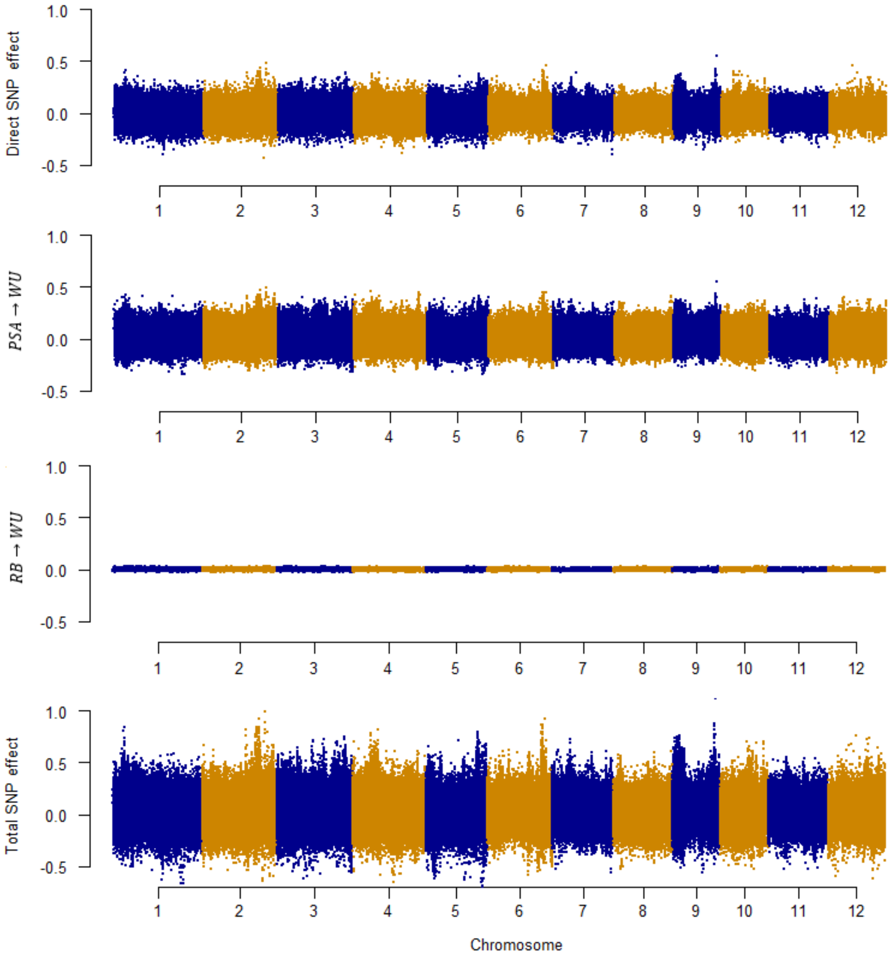

Figure 5: Manhattan plot of direct (affecting each trait without any mediation), indirect (mediated by other phenotypes), and total (sum of all direct and indirect) SNP effects on water use (WU) using SEM-GWAS based on the network learned by the MMPC algorithm. Each point represents a SNP and the height of the SNP represents the extent of its association with WU.

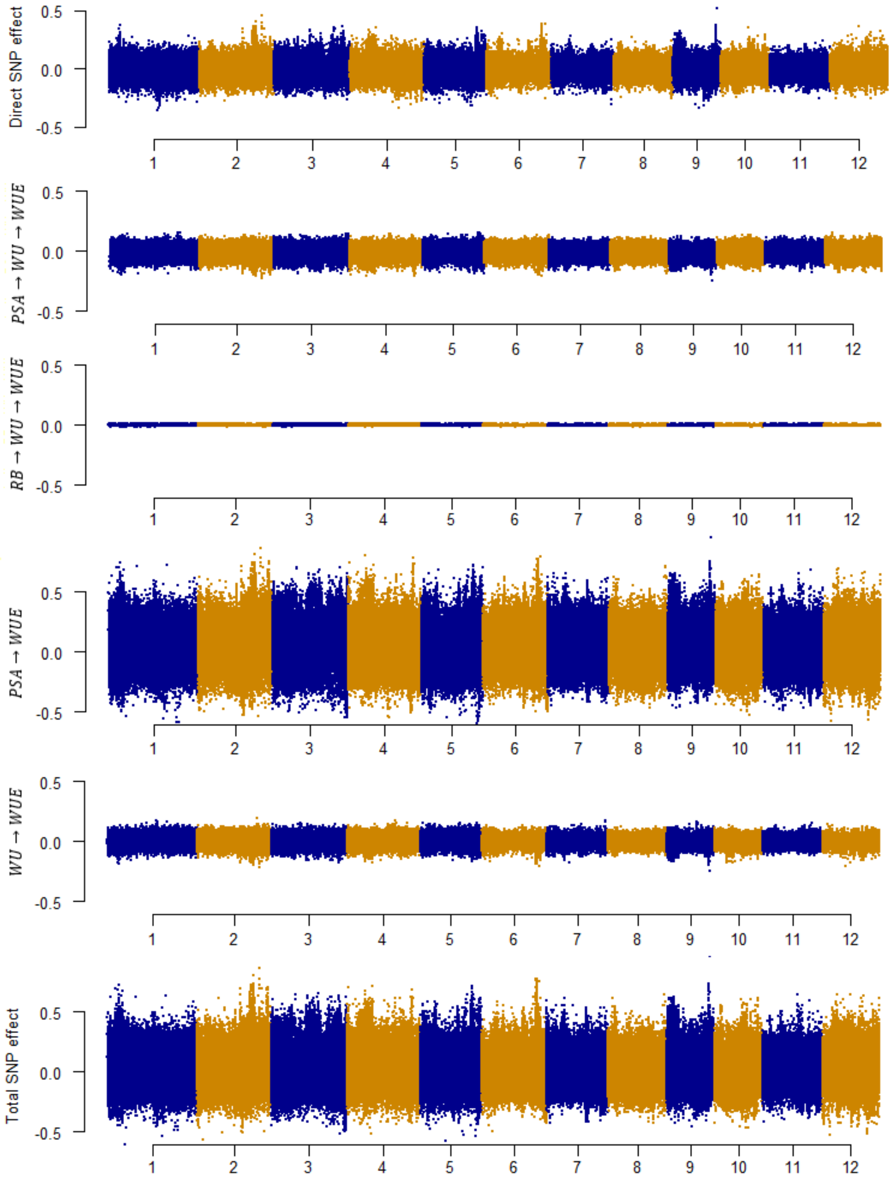

Figure 6: Manhattan plot of direct (affecting each trait without any mediation), indirect (mediated by other phenotypes), and total (sum of all direct and indirect) SNP effects on water use efficiency (WUE)using SEM-GWAS based on the network learned by the MMPC algorithm. Each point represents a SNP and the height of the SNP represents the extent of its association with WUE.

References

Ahlfors, R., Lång, S., Overmyer, K., Jaspers, P., Brosché, M., Tauriainen, A., Kollist, H., Tuominen, H., Belles-Boix, E., Piippo, M., et al. (2004). Arabidopsis radical-induced cell death1 belongs to the wwe protein–protein interaction domain protein family and modulates abscisic acid, ethylene, and methyl jasmonate responses. *The Plant Cell*, 16(7):1925–1937.

Alwin, D. F. and Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, pages 37–47.

Aya, K., Hobo, T., Sato-Izawa, K., Ueguchi-Tanaka, M., Kitano, H., and Matsuoka, M. (2014). A novel ap2-type transcription factor, small organ size1, controls organ size downstream of an auxin signaling pathway. *Plant and Cell Physiology*, 55(5):897–912.

Berger, B., Parent, B., and Tester, M. (2010). High-throughput shoot imaging to study drought responses. *Journal of Experimental Botany*, 61(13):3519–3528.

Bielby, W. T. and Hauser, R. M. (1977). Structural equation models. *Annual Review of Sociology*, 3(1):137–161.

Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097.

Campbell, M., Walia, H., and Morota, G. (2018). Utilizing random regression models for genomic prediction of a longitudinal trait derived from high-throughput phenotyping. *Plant Direct*, 2(9):e00080.

Campbell, M. T., Bandillo, N., Al Shiblawi, F. R. A., Sharma, S., Liu, K., Du, Q., Schmitz, A. J., Zhang, C., Véry, A.-A., Lorenz, A. J., et al. (2017a). Allelic variants of oshkt1; 1

underlie the divergence between indica and japonica subspecies of rice (oryza sativa) for root sodium content. *PLoS Genetics*, 13(6):e1006823.

Campbell, M. T., Du, Q., Liu, K., Brien, C. J., Berger, B., Zhang, C., and Walia, H. (2017b). A comprehensive image-based phenomic analysis reveals the complex genetic architecture of shoot growth dynamics in rice. *The Plant Genome*, 10(2).

Campbell, T. M., Avi, C. K., Berger, B., Chris, J. B., Wang, D., and Walia, H. (2015). Integrating image-based phenomics and association analysis to dissect the genetic architecture of temporal salinity responses in rice. *Plant physiology*, pages pp–00450.

Challinor, A. J., Watson, J., Lobell, D. B., Howden, S., Smith, D., and Chhetri, N. (2014). A meta-analysis of crop yield under climate change and adaptation. *Nature Climate Change*, 4(4):287.

Ehdaie, B. (1995). Variation in water-use efficiency and its components in wheat: II. pot and field experiments. *Crop Science*, 35(6):1617–1626.

Gianola, D., de los Campos, G., Toro, M. A., Naya, H., Schön, C.-C., and Sorensen, D. (2015). Do molecular markers inform about pleiotropy? *Genetics*, 201(1):23–29.

Gianola, D. and Sorensen, D. (2004). Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics*, 167(3):1407–1424.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001.

Golzarian, M. R., Frick, R. A., Rajendran, K., Berger, B., Roy, S., Tester, M., and Lun, D. S. (2011). Accurate inference of shoot biomass from high-throughput images of cereal plants. *Plant Methods*, 7(1):2.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, pages 1–12.

Han, B., Chen, X.-w., Talebizadeh, Z., and Xu, H. (2012). Genetic studies of complex human diseases: characterizing snp-disease associations using bayesian networks. *BMC Systems Biology*, 6(3):S14.

Henderson, C. and Quaas, R. (1976). Multiple trait evaluation using relatives' records. *Journal of Animal Science*, 43(6):1188–1197.

Hirano, K., Yoshida, H., Aya, K., Kawamura, M., Hayashi, M., Hobo, T., Sato-Izawa, K., Kitano, H., Ueguchi-Tanaka, M., and Matsuoka, M. (2017). Small organ size 1 and small organ size 2/dwarf and low-tillering form a complex to integrate auxin and brassinosteroid signaling in rice. *Molecular Plant*, 10(4):590–604.

Ho, M. D., Rosas, J. C., Brown, K. M., and Lynch, J. P. (2005). Root architectural tradeoffs for water and phosphorus acquisition. *Functional Plant Biology*, 32(8):737–748.

Huang, X. and Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annual review of plant biology*, 65:531–551.

Hubick, K., Farquhar, G., and Shorter, R. (1986). Correlation between water-use efficiency and carbon isotope discrimination in diverse peanut (*Arachis*) germplasm. *Functional Plant Biology*, 13(6):803–816.

Ismail, A. M. and Hall, A. (1992). Correlation between water-use efficiency and carbon isotope discrimination in diverse cowpea genotypes and isogenic lines. *Crop Science*, 32(1):7–12.

Kennedy, B., Quinton, M., and Van Arendonk, J. (1992). Estimation of effects of single genes on quantitative traits. *Journal of Animal Science*, 70(7):2000–2012.

Khadilkar, A. S., Yadav, U. P., Salazar, C., Shulaev, V., Paez-Valencia, J., Pizzio, G. A., Gaxiola, R. A., and Ayre, B. G. (2015). Constitutive and companion cell-specific overexpression of AVP1, encoding a proton-pumping pyrophosphatase, enhances biomass accumulation, phloem loading and long-distance transport. *Plant Physiology*, pages pp–01409.

Kim, H., Hwang, H., Hong, J.-W., Lee, Y.-N., Ahn, I. P., Yoon, I. S., Yoo, S.-D., Lee, S., Lee, S. C., and Kim, B.-G. (2011). A rice orthologue of the aba receptor, ospyl/rcar5, is a positive regulator of the aba signal transduction pathway in seed germination and early seedling growth. *Journal of Experimental Botany*, 63(2):1013–1024.

Leal-Gutiérrez, J. D., Rezende, F. M., Elzo, M. A., Johnson, D., Peñagaricano, F., and Mateescu, R. G. (2018). Structural equation modeling and whole-genome scans uncover chromosome regions and enriched pathways for carcass and meat quality in beef. *Frontiers in Genetics*, 9.

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

Liu, B., de La Fuente, A., and Hoeschele, I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*.

Mann, M. E. and Gleick, P. H. (2015). Climate change and california drought in the 21st century. *Proceedings of the National Academy of Sciences*, 112(13):3858–3859.

McCouch, S. R., Wright, M. H., Tung, C.-W., Maron, L. G., McNally, K. L., Fitzgerald, M., Singh, N., DeClerck, G., Agosto-Perez, F., Korniliev, P., et al. (2016). Open access resources for genome-wide association mapping in rice. *Nature communications*, 7:10532.

Meyer, K. and Tier, B. (2012). snp snappy: A strategy for fast genome-wide association studies fitting a full mixed model. *Genetics*, 190(1):275–277.

Mi, X., Eskridge, K., Wang, D., Baenziger, P. S., Campbell, B. T., Gill, K. S., and Dweikat,

I. (2010). Bayesian mixture structural equation modelling in multiple-trait qtl mapping. *Genetics Research*, 92(3):239–250.

Miao, C., Xiao, L., Hua, K., Zou, C., Zhao, Y., Bressan, R. A., and Zhu, J.-K. (2018). Mutations in a subfamily of abscisic acid receptor genes promote rice growth and productivity. *Proceedings of the National Academy of Sciences*, page 201804774.

Mir, R. R., Zaman-Allah, M., Sreenivasulu, N., Trethowan, R., and Varshney, R. K. (2012). Integrated genomics, physiology and breeding approaches for improving drought tolerance in crops. *Theoretical and Applied Genetics*, 125(4):625–645.

Momen, M., Mehrgardi, A. A., Roudbar, M. A., Kranis, A., Pinto, R. M., Valente, B. D., Morota, G., Rosa, G. J., and Gianola, D. (2018). Including phenotypic causal networks in genome-wide association studies using mixed effects structural equation models. *bioRxiv*, page 251421.

Momen, M., Mehrgardi, A. A., Sheikhy, A., Esmailizadeh, A., Fozi, M. A., Kranis, A., Valente, B. D., Rosa, G. J., and Gianola, D. (2017). A predictive assessment of genetic correlations between traits in chickens using markers. *Genetics Selection Evolution*, 49(1):16.

Morota, G., Valente, B., Rosa, G., Weigel, K., and Gianola, D. (2012). An assessment of linkage disequilibrium in holstein cattle using a bayesian network. *Journal of Animal Breeding and Genetics*, 129(6):474–487.

Onogi, A., Ideta, O., Yoshioka, T., Ebana, K., Yamasaki, M., and Iwata, H. (2016). Uncovering a nuisance influence of a phenological trait of plants using a nonlinear structural equation: Application to days to heading and culm length in asian cultivated rice (oryza sativa l.). *PloS One*, 11(2):e0148609.

Otkin, J. A., Svoboda, M., Hunt, E. D., Ford, T. W., Anderson, M. C., Hain, C., and Basara, J. B. (2017). Flash droughts: A review and assessment of the challenges imposed by rapid onset droughts in the united states. *Bulletin of the American Meteorological Society*, (2017).

Passioura, J. (2012). Phenotyping for drought tolerance in grain crops: when is it useful to breeders? *Functional Plant Biology*, 39(11):851–859.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

Schilling, R. K., Marschner, P., Shavrukov, Y., Berger, B., Tester, M., Roy, S. J., and Plett, D. C. (2014). Expression of the arabidopsis vacuolar h+-pyrophosphatase gene (AVP1) improves the shoot biomass of transgenic barley and increases grain yield in a saline field. *Plant Biotechnology Journal*, 12(3):378–386.

Scutari, M. (2010). Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software, Articles*, 35(3):1–22.

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

Scutari, M., Graafland, C. E., and Gutiérrez, J. M. (2018). Who learns better bayesian network structures: Constraint-based, score-based or hybrid algorithms? *arXiv preprint arXiv:1805.11908*.

Shipley, B. (2016). *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R*. Cambridge University Press.

Sinclair, T. R. (2011). Challenges in breeding for yield increase for drought. *Trends in plant science*, 16(6):289–293.

Töpner, K., Rosa, G. J., Gianola, D., and Schön, C.-C. (2017). Bayesian networks illustrate genomic and residual trait connections in maize (zea mays l.). *G3: Genes, Genomes, Genetics*, 7(8):2779–2789.

Tsamardinos, I., Aliferis, C. F., and Statnikov, A. (2003). Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678. ACM.

Tuberosa, R. and Salvi, S. (2006). Genomics-based approaches to improve drought tolerance of crops. *Trends in Plant Science*, 11(8):405–412.

Valente, B. D., Rosa, G. J., Gianola, D., Wu, X.-L., and Weigel, K. A. (2013). Is structural equation modeling advantageous for the genetic improvement of multiple traits? *Genetics*, 194(3):561–572.

Valente, B. D., Rosa, G. J., Gustavo, A., Gianola, D., and Silva, M. A. (2010). Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics*.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423.

Varona, L., Sorensen, D., and Thompson, R. (2007). Analysis of litter size and average litter weight in pigs using a recursive model. *Genetics*, 177(3):1791–1799.

Verhulst, B., Maes, H. H., and Neale, M. C. (2017). GW-SEM: A statistical package to conduct genome-wide structural equation modeling. *Behavior Genetics*, 47(3):345–359.

Wang, H. and van Eeuwijk, F. A. (2014). A new method to infer causal phenotype networks using qtl and phenotypic information. *PloS One*, 9(8):e103997.

Wehner, M. F., Arnold, J. R., Knutson, T., Kunkel, K. E., and N, L. A. (2017). Droughts, Floods, and Wildfires. In *Climate Science Special Report: Fourth National Climate Assessment, Volume I.*, pages 231–256.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585.

Wu, X.-L., Heringstad, B., and Gianola, D. (2010). Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. *Journal of Animal Breeding and Genetics*, 127(1):3–15.

You, J., Zong, W., Li, X., Ning, J., Hu, H., Li, X., Xiao, J., and Xiong, L. (2012). The snac1-targeted gene ossro1c modulates stomatal closure and oxidative stress tolerance by regulating hydrogen peroxide in rice. *Journal of Experimental Botany*, 64(2):569–583.

Yu, H., Campbell, M. T., Zhang, Q., Walia, H., and Morota, G. (2018). Genomic bayesian confirmatory factor analysis and bayesian network to characterize a wide spectrum of rice phenotypes. *bioRxiv*, page 435792.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203.

Zampieri, M., Ceglar, A., Dentener, F., and Toreti, A. (2017). Wheat yield loss attributable to heat waves, drought and water excess at the global, national and subnational scales. *Environmental Research Letters*, 12(6):064008.

Zhang, J., Li, J., Wang, X., and Chen, J. (2011). Ovp1, a vacuolar h+-translocating inorganic pyrophosphatase (v-ppase), overexpression improved rice cold tolerance. *Plant Physiology and Biochemistry*, 49(1):33–38.

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P., et al. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proceedings of the National Academy of Sciences*, 114(35):9326–9331.

Zhao, K., Tung, C.-W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., Norton, G. J., Islam, M. R., Reynolds, A., Mezey, J., et al. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in oryza sativa. *Nature Communications*, 2:467.