



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

12

13 Running title: network analysis in rice

14

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

17

18 \* Corresponding author:

19

20 Gota Morota

21 Department of Animal and Poultry Sciences

22 Virginia Polytechnic Institute and State University

23 175 West Campus Drive

24 Blacksburg, Virginia 24061 USA.

25 E-mail: [morota@vt.edu](mailto:morota@vt.edu)

26

## 27 **Abstract**

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

## 47 **Background**

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

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

73 While MTM-GWAS is a powerful approach to capture the genetic correlations between

74 traits for genetic inference, it fails to address how the traits are interrelated, or elucidate  
75 the mechanisms that give rise to the observed correlation. The early work of Sewall Wright  
76 sought to infer causative relations between correlated variables through path analysis (Wright,  
77 1921). This seminal work gave rise to structural equation models (SEM), which assesses  
78 the nature and magnitude of direct and indirect effects of multiple interacting variables.  
79 Although SEM remains a powerful approach to model the relationships among variables in  
80 complex systems, its use has been limited in biology.

81 Recently, Momen et al. (2018) proposed the SEM-GWAS framework by incorporating  
82 phenotypic networks and SNPs into MTM-GWAS through SEM (Wright, 1921; Haavelmo,  
83 1943). In contrast to standard multivariate statistical techniques, the SEM framework opens  
84 up a multivariate modeling strategy that accounts for recursive (an effect from one pheno-  
85 type is passed onto another phenotype) and simultaneous (reciprocal) structures among its  
86 variables (Goldberger, 1972; Bielby and Hauser, 1977). Momen et al. (2018) showed that  
87 SEM-GWAS can supplement MTM-GWAS, and is capable of partitioning the source of the  
88 SNP effects into direct and indirect effects, which helps to provide a better understanding  
89 of the relevant biological mechanisms. In contrast, MTM-GWAS, which does not take the  
90 network structure between phenotypes into account, estimates overall SNP effects that are  
91 mediated by other phenotypes, and combines direct and indirect SNP effects.

92 Current climate projections predict an increase in the incidence of drought events and  
93 elevated temperatures throughout the growing season (Wehner et al., 2017). These elevated  
94 temperatures will drive higher evapotranspirational demands, and combined with the incon-  
95 sistency of rainfall events, will increase the rate of drought onset and intensity, and impact on  
96 crop growth and productivity (Challinor et al., 2014; Mann and Gleick, 2015; Otkin et al.,  
97 2017; Zampieri et al., 2017; Zhao et al., 2017). To counter the effects of climate change  
98 on agricultural productivity, drought-resilient crops must be developed. However, progress  
99 towards this goal is often hindered by the complex biological basis of drought tolerance  
100 (Tuberosa and Salvi, 2006; Sinclair, 2011; Mir et al., 2012; Passioura, 2012). The ability to

101 maintain productivity under limited water availability involves a suite of morphological and  
102 physiological traits (Passioura, 2012). Among these is the ability to access available water  
103 and utilize it for growth. Thus, studying traits associated with water capture (e.g. root  
104 biomass and architecture) and utilization (e.g. water-use efficiency) is essential. However,  
105 of equal importance is a robust statistical framework that allows these complex traits to be  
106 analyzed jointly and causal relationships among traits to be inferred.

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

## 117 **Materials and Methods**

### 118 **Experimental data set**

119 The plant material used in our analysis consisted of a rice diversity panel of  $n = 357$  inbred  
120 accessions of *O. sativa* collected from a diverse range of regions, which are expected to  
121 capture much of the genetic diversity within cultivated rice (Zhao et al., 2011). All lines were  
122 genotyped with 700,000 SNPs using the high-density rice array from Affymetrix (Santa Clara,  
123 CA, USA) such that there was approximately 1 SNP every 0.54Kb across the rice genome  
124 (Zhao et al., 2011; McCouch et al., 2016). We used PLINK v1.9 software (Purcell et al.,  
125 2007) to remove SNPs with a call rate  $\leq 0.95$  and a minor allele frequency  $\leq 0.05$ . Missing  
126 genotypes were imputed using Beagle software version 3.3.2 (Browning and Browning, 2007).  
127 Finally, 411,066 SNPs were retained for further analysis.

### 128 **Phenotypic data**

129 We analyzed four biologically important traits for drought responses in rice: projected shoot  
130 area (PSA), root biomass (RB), water use (WU), and water use efficiency (WUE). These  
131 phenotypes are derived from two separate studies (Campbell et al., 2017a, 2018). The aim  
132 of the first study was to evaluate the effects of drought on shoot growth (Campbell et al.,  
133 2018). Here, the diversity panel was phenotyped using an automated phenotyping platform  
134 in Adelaide, SA, Australia. This new phenotyping technology enables us to produce high-  
135 resolution spatial and temporal image-derived phenotypes, which can be used to capture  
136 dynamic growth, development, and stress responses (Berger et al., 2010; Golzarian et al.,  
137 2011; Campbell et al., 2015, 2017b).

138 The plants were phenotyped over a period of 20 days, starting at 13 days after they were  
139 transplanted into soil and ending at 33 days. Each day, the plants were watered to a specific  
140 target weight to ensure the soil was completely saturated. The plants were then imaged  
141 from three angles. These pictures were processed to remove all background objects, leaving

142 just pixels for the green shoot tissue. We summed the pixels from each picture to obtain an  
143 estimate of the shoot biomass. We refer to this metric as PSA. With this system, we also  
144 obtained the weights, prior to watering and after watering, for each pot on each day. From  
145 this data, we estimated the amount of water that is used by each plant. WU was calculated  
146 as  $\text{Pot Weight}_{(r-1)} - \text{Pot Weight}_{(r)}$ , where  $r$  is time, and WUE is the ratio of PSA to WU.  
147 Although this data has not yet been published, a description of the phenotyping system and  
148 insight into the experimental design can be found in Campbell et al. (2015).

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

## 157 **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.

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

158 where  $\mathbf{y}$  is the vector observations for  $t = 4$  traits,  $\boldsymbol{\mu}$  is the vector of intercept,  $\mathbf{X}$  is the  
159 incidence matrix of covariates,  $\mathbf{b}$  is the vector of covariate effects,  $\mathbf{Z}$  is the incidence matrix  
160 relating accessions with additive genetic effects,  $\mathbf{g}$  is the vector of additive genetic effects,  
161 and  $\boldsymbol{\epsilon}$  is the vector of residuals. The incident matrix  $\mathbf{X}$  only included intercepts for the



162 four traits examined in this study. Under the infinitesimal model of inheritance, the  $\mathbf{g}$   
163 and  $\boldsymbol{\epsilon}$  were assumed to follow a multivariate Gaussian distribution  $\mathbf{g} \sim N(0, \sum_g \otimes \mathbf{G})$  and  
164  $\boldsymbol{\epsilon} \sim N(0, \sum_\epsilon \otimes \mathbf{I})$ , respectively, where  $\mathbf{G}$  is the  $n \times n$  genomic relationship matrix for genetic  
165 effects,  $\mathbf{I}$  is the identity matrix for residuals,  $\sum_g$  and  $\sum_\epsilon$  are the  $t \times t$  variance-covariance  
166 matrices of genetic effects and residuals, respectively, and  $\otimes$  denotes the Kronecker product.  
167 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  
168 incidence matrix taking values of  $0 - 2p_j$  for zero copies of the reference allele,  $1 - 2p_j$  for  
169 one copy of the reference allele, and  $2 - 2p_j$  for two copies of the reference allele (VanRaden,  
170 2008). Here,  $p_j$  is the allele frequency at SNP  $j = 1, \dots, m$ . We assigned flat priors for the  
171 intercept and the vector of fixed effects. The vectors of random additive genetic effects and  
172 residual effects were assigned independent multivariate normal priors with null mean and  
173 inverse Wishart distributions for the covariance matrices.

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

## 180 **Learning structures using Bayesian network**

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

188 Parents and Children (MMPC) algorithm from the constraint-based structure learning cat-  
189 egory to infer the network structure among the four traits examined in this study (Scutari  
190 et al., 2018). We selected this algorithm because it was suggested in a recent study, Töpner  
191 et al. (2017), which showed that the constraint-based algorithms performed better for the  
192 construction of networks than score-based counterparts. This algorithm is similar to the  
193 inductive causation algorithm (Tsamardinos et al., 2003) that was first used in Valente et al.  
194 (2010) to infer a phenotypic network. The bnlearn R package was used to learn the Bayesian  
195 phenotypic network throughout this analysis with mutual information as the test, and the  
196 statistically significant level set at  $\alpha = 0.01$  (Scutari, 2010). We computed the Bayesian  
197 information criterion (BIC) score of a network and estimated the strength and uncertainty  
198 of direction of each edge probabilistically by bootstrapping as described in Scutari and Denis  
199 (2014). In addition, the strength of the edge was assessed by computing the change in the  
200 BIC score when that particular edge was removed from the network, while keeping the rest  
201 of the network intact.

## 202 **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.

$$.y = \mathbf{w}s + \mathbf{Zg} + \epsilon,$$

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

## 207 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).

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

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \mathbf{y}_4 \end{pmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{I}_1\lambda_{\text{PSA} \rightarrow \text{RB}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{I}_1\lambda_{\text{PSA} \rightarrow \text{WU}} & \mathbf{I}_2\lambda_{\text{RB} \rightarrow \text{WU}} & \mathbf{0} & \mathbf{0} \\ \mathbf{I}_1\lambda_{\text{PSA} \rightarrow \text{WUE}} & \mathbf{I}_2\lambda_{\text{RB} \rightarrow \text{WUE}} & \mathbf{I}_3\lambda_{\text{WU} \rightarrow \text{WUE}} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \mathbf{y}_4 \end{pmatrix} \\ + \begin{bmatrix} \mathbf{w}_{j1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_{j2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{w}_{j3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{w}_{j4} \end{bmatrix} \begin{pmatrix} s_{j1} \\ s_{j2} \\ s_{j3} \\ s_{j4} \end{pmatrix} \\ + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_4 \end{bmatrix} \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \mathbf{g}_3 \\ \mathbf{g}_4 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \boldsymbol{\epsilon}_3 \\ \boldsymbol{\epsilon}_4 \end{pmatrix}$$

208 where  $\mathbf{I}$  is the identity matrix,  $\mathbf{\Lambda}$  is the lower triangular matrix of regression coefficients  
 209 or structural coefficients based on the learned network structure from the Bayesian network,  
 210 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  $\Lambda$  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

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

211 where  $\text{Var}(\mathbf{g}^*) \sim (\mathbf{I} - \Lambda)^{-1} \mathbf{G} (\mathbf{I} - \Lambda)^{\prime -1}$  and  $\text{Var}(\boldsymbol{\epsilon}^*) \sim (\mathbf{I} - \Lambda)^{-1} \mathbf{R} (\mathbf{I} - \Lambda)^{\prime -1}$ . This trans-  
212 formation changes SEM-GWAS into MTM-GWAS, which ignores the network relationships  
213 among traits (Gianola and Sorensen, 2004; Varona et al., 2007). However, Valente et al.  
214 (2013) stated that SEM allows for the prediction of the effects of external interventions,  
215 which can be useful for making selection decisions that are not possible with MTM. We  
216 used SNP Snappy software to perform MTM-GWAS and SEM-GWAS (Meyer and Tier,  
217 2012). To identify candidate SNPs that may explain direct (in the absence of mediation by  
218 other traits) and indirect (with intervention and mediation by other traits) effects for each  
219 trait, the SNPs from MTM-GWAS were ranked according to  $p$ -values for each trait. The 20  
220 most significant SNPs were then selected, and all genes within 200 kb were considered to be  
221 potential candidate genes.

## 222 Results

### 223 Trait correlations and network structure

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

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

240 Measuring the strength of probabilistic dependence for each arc is crucial in Bayesian  
241 network learning (Scutari and Denis, 2014). As shown in Figure 1, the strength of each arc  
242 was assessed with 2,500 bootstrap samples with a significance level at  $\alpha = 0.01$ . The labels  
243 on the edges indicate the proportion of bootstrap samples supporting the presence of the  
244 edge and the proportion supporting the direction of the edges are provided in parentheses.  
245 Learned structures were averaged with a strength threshold of 85% or higher to produce a  
246 more robust network structure. Edges that did not meet this threshold were removed from  
247 the networks. In addition, we used BIC as goodness-of-fit statistics measuring how well the

248 paths mirror the dependence structure of the data (Table 2). The BIC assign higher scores  
 249 to any path that fit the data better. The BIC score reports the importance of each arc by its  
 250 removal from the learned structure. We found that removing PSA  $\rightarrow$  WUE resulted in the  
 251 largest decrease in the BIC score, suggesting that this path is playing the most important  
 252 role in the network structure. This was followed by WU  $\rightarrow$  WUE, RB  $\rightarrow$  WU, and PSA  $\rightarrow$   
 253 WU.

## 254 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

$$\begin{aligned}
 \mathbf{y}_{1\text{PSA}} &= \mathbf{w}_j s_j(y_{1\text{PSA}}) + \mathbf{Z}_1 \mathbf{g}_1 + \epsilon_1 \\
 \mathbf{y}_{2\text{RB}} &= \mathbf{w}_j s_j(y_{2\text{RB}}) + \mathbf{Z}_2 \mathbf{g}_2 + \epsilon_2 \\
 \mathbf{y}_{3\text{WU}} &= \lambda_{13} \mathbf{y}_{1\text{PSA}} + \lambda_{23} \mathbf{y}_{2\text{RB}} + \mathbf{w}_j s_j(y_{3\text{WU}}) + \mathbf{Z}_3 \mathbf{g}_3 + \epsilon_3 \\
 &= \lambda_{13} [\mathbf{w}_j s_j(y_{1\text{PSA}}) + \mathbf{Z}_1 \mathbf{g}_1 + \epsilon_1] + \lambda_{23} [\mathbf{w}_j s_j(y_{2\text{RB}}) + \mathbf{Z}_2 \mathbf{g}_2 + \epsilon_2] + \mathbf{w}_j s_j(y_{3\text{WU}}) + \mathbf{Z}_3 \mathbf{g}_3 + \epsilon_3 \\
 \mathbf{y}_{4\text{WUE}} &= \lambda_{14} \mathbf{y}_{1\text{PSA}} + \lambda_{34} \mathbf{y}_{3\text{WU}} + \mathbf{w}_j s_j(y_{4\text{WUE}}) + \mathbf{Z} \mathbf{g} + \epsilon_4 \\
 &= \lambda_{14} [\mathbf{w}_j s_j(y_{1\text{PSA}}) + \mathbf{Z}_1 \mathbf{g}_1 + \epsilon_1] \\
 &\quad + \lambda_{34} \{ \lambda_{13} [\mathbf{w}_j s_j(y_{1\text{PSA}}) + \mathbf{Z}_1 \mathbf{g}_1 + \epsilon_1] + \lambda_{23} [\mathbf{w}_j s_j(y_{2\text{RB}}) + \mathbf{Z}_2 \mathbf{g}_2 + \epsilon_2] + \mathbf{w}_j s_j(y_{3\text{WU}}) + \mathbf{Z}_3 \mathbf{g}_3 + \epsilon_3 \} \\
 &\quad + \mathbf{w}_j s_j(y_{4\text{WUE}}) + \mathbf{Z} \mathbf{g} + \epsilon_4.
 \end{aligned}$$

255 The corresponding estimated  $\mathbf{\Lambda}$  matrix is

$$\mathbf{\Lambda} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \lambda_{13\text{PSA} \rightarrow \text{WU}} & \lambda_{23\text{RB} \rightarrow \text{WU}} & 0 & 0 \\ \lambda_{14\text{PSA} \rightarrow \text{WUE}} & 0 & \lambda_{34\text{WU} \rightarrow \text{WUE}} & 0 \end{bmatrix}.$$

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

## 262 Interpretation of SNP effects

263 We implemented SEM-GWAS as an extension of the MTM-GWAS method for analysis of  
264 the joint genetic architecture of the four measured traits, to partition SNP effects into direct  
265 and indirect (Alwin and Hauser, 1975). The results of the decomposition of SNP effects are  
266 discussed for each trait separately below. Because the network only revealed indirect effects  
267 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  $j$ th 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  $\lambda_{13}$  and  $\lambda_{23}$ . Figure 5 shows a Manhattan plot of SNP effects on WU. We found that the indirect RB  $\rightarrow$  WU path had the least impact on overall effects, whereas indirect PSA  $\rightarrow$  WU path had almost the same contribution as the direct SNP effects.

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

268 Water Usage Efficiency (**WUE**): The overall SNP effects for WUE can be partitioned into  
 269 one direct and four indirect genetic signals (Figure 2). WUE is the only phenotype trait  
 270 that does not have any outgoing path to other traits. According to Figure 6, the extents of



271 the SNP effects among the four indirect paths were 1) RB  $\rightarrow$  WUE mediated by WU, 2)  
 272 PSA  $\rightarrow$  WUE mediated by WU, 3) WU  $\rightarrow$  WUE, and 4) PSA  $\rightarrow$  WUE, in increasing order.  
 273 We found that the SNP effect transmitted through RB had the smallest effect on the WUE,  
 274 suggesting that modifying the size of the QTL effect for RB may not have a noticeable effect  
 275 on WUE, whereas a change in PSA had a noticeable effect on WUE. The magnitude of the  
 276 relationship between RB and WUE is proportional to the product of structural coefficients  
 277  $\lambda_{23} \times \lambda_{34} = 0.005 \times -0.455$ . PSA influenced WUE via two indirect paths, and strongly  
 278 depends on the structural coefficients  $\lambda_{14} = 1.339$  and  $\lambda_{13}\lambda_{34} = 0.767 \times -0.455$  for PSA  $\rightarrow$   
 279 WUE and PSA  $\rightarrow$  WU  $\rightarrow$  WUE, respectively. It should be noted that the indirect effect  
 280 transmitted through PSA  $\rightarrow$  WUE was greater than the direct effects of a given SNP on  
 281 WUE. This is because the structural coefficient between WU and WUE has a negative sign,  
 282 resulting in transmitted indirect SNP effects that can change the sign and magnitude of the  
 283 total effect on WUE, even from positive values to negative values. However, this indicates  
 284 that the modification and selection of plants for WU may impact WUE, even for the opposite  
 285 direction.

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

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

286 The indirect and direct SNP effects across all possible paths with the total effect for WU

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

### 306 **Trade offs between MTM- and SEM-GWAS models suggest enrich-** 307 **ment of candidate genes for the traits**

308 Nineteen of the top 20 SNPs showed a direct effect on WU ( $P_{direct} < 0.01$ ), while for WUE  
309 all SNPs showed an indirect effect ( $P_{indirect} \geq 0.01$ ). Interestingly, for both traits, all indirect  
310 effects at these loci could be attributed to PSA, indicating that alleles that influence shoot  
311 biomass may have an effect on WU and WUE. The positive relationship between dry matter  
312 production and WU is widely documented across multiple crops, and is simply because larger

313 plants have a greater water demand than small plants (Ehdaie, 1995; Hubick et al., 1986;  
314 Ismail and Hall, 1992). Moreover, in this study the plants were grown under simulated  
315 paddy conditions (i.e., with water-saturated soil); thus; there was sufficient water to meet  
316 these demands and sustain shoot growth in larger plants. In conditions where water is limited  
317 such relationships may not hold true.

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

338 *SMOS1* is located at  $\sim 18.81$  Mb on chromosome 5, and encodes an AP2 transcription  
339 factor. Initially identified through a mutant screen, *SMOS1* knockout plants exhibit nearly

340 normal vegetative and reproductive development; however the leaf blade, leaf sheath, roots,  
341 flowers, and seeds are significantly reduced in the mutant lines (Aya et al., 2014). The shorter  
342 length of these organs was attributed to a reduction in cell size, indicating that this gene  
343 is involved in the regulation of cell growth. These observations were further supported by  
344 Aya et al. (2014) and Hirano et al. (2017), who showed that *SMOS1* binds to the promoter  
345 of the cell expansion gene, phosphate-induced protein 1 (*PHI1*). While the effect of *OVP1*  
346 and *SMOS1* on shoot growth and water use efficiency remain to be elucidated in rice, the  
347 known functions of these genes, as well as their presence in close proximity to SNPs with  
348 indirect effects on WUE through PSA, are encouraging and warrant further investigation.

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

361 The second gene, *RADICAL-INDUCED CELL DEATH1 (RCD1)*, is located at  $\sim 35.87$   
362 Mb on chromosome 3, and encodes a WWE-domain containing protein. *RCD1* has been well  
363 characterized in Arabidopsis for hormonal responses and ROS homeostasis (Ahlfors et al.,  
364 2004). Interestingly, *RCD1* and other members of the Similar to *RCD* One (SRO) family  
365 have been shown to be involved with the regulation of the stomatal aperture and water  
366 loss. For example, Ahlfors et al. (2004) showed that *rcd1* mutants exhibit greater stomatal

367 conductance and greater water loss than the WT (Ahlfors et al., 2004). The over-expression  
368 of a *RCD1* ortholog in rice, *OsSRO1c*, resulted in the opposite phenotype being observed,  
369 with a decreased stomatal aperture and reduced water loss compared with the WT (You  
370 et al., 2012). The *ROS H<sub>2</sub>O<sub>2</sub>* has been shown to act downstream of ABA and to result  
371 in stomatal closure. Members of the SRO family are involved in the regulation of ROS  
372 homeostasis; thus, the stomata and water loss phenotypes exhibited by mis-regulation of  
373 SRO or *RCD1* may be due to the inability to properly regulate *H<sub>2</sub>O<sub>2</sub>* levels (You et al.,  
374 2012).

## 375 Discussion

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

388 Detecting putative causal genes is of considerable interest for determining which traits will  
389 be affected by specific loci from a biological perspective, and consequently partitioning the  
390 genetic signals according to the paths determined. Although the parameter interpretations  
391 of SEM as applied to QTL mapping (Li et al., 2006; Mi et al., 2010), expression QTL (Liu  
392 et al., 2008), or genetic selection (Valente et al., 2013) have been actively pursued, the work  
393 of Momen et al. (2018) marks one of the first studies to pay particular attention at the level of  
394 individual SNP effect in genome-wide SEM analyses. The SEM embeds a flexible framework  
395 for performing such network analysis in a GWAS context, and the current study demonstrates  
396 its the first application in crops. We assumed that modeling a system of four traits in rice  
397 simultaneously may help us to examine the sources of SNP effects in GWAS in greater depth.  
398 Therefore, we compared two GWAS methodologies that have the ability to embed multiple  
399 traits jointly, so that the estimated SNP effects from both models have different meanings.  
400 The main significance of SEM-GWAS, relative to MTM-GWAS, is to include the relationship  
401 between SNPs and measured phenotypes, coupled with relationships that are potentially

402 meditated by other phenotypes (mediator traits). This advances GWAS, and consequently  
403 the information obtained from phenotypic networks describing such interrelationships can be  
404 used to predict the behavior of complex systems (Momen et al., 2018). Although we analyzed  
405 the observed phenotypes in the current study, the factor analysis component of SEM can  
406 be added to SEM-GWAS by deriving latent factors from multiple phenotypes (e.g., Verhulst  
407 et al., 2017; Leal-Gutiérrez et al., 2018). The inference of a phenotypic network structure  
408 was carried out using a Bayesian network, which has applications in genetics ranging from  
409 modeling linkage disequilibrium (Morota et al., 2012) to epistasis (Han et al., 2012).

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

429 impact on RB and WU. These findings highlight the significant potential and importance  
430 of mediator relationship inclusion in the association between other variables in the inferred  
431 graph.

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



## 444 **Acknowledgments**

445 This work was partially supported by the National Science Foundation under Grant Number  
446 1736192 to HW and GM.

## 447 **Author contribution statement**

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

## 452 Tables

Table 1: Genomic (upper triangular), residual (lower triangular) correlations and genomic heritabilities (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 ( $\lambda$ ) estimates derived from the structural equation models. Projected shoot area (PSA), root biomass (RB), water use (WU), and water use efficiency (WUE).

Path	$\lambda$	Structural coefficient
PSA $\rightarrow$ WU	$\lambda_{13}$	0.767
RB $\rightarrow$ WU	$\lambda_{23}$	0.005
PSA $\rightarrow$ WUE	$\lambda_{14}$	1.339
WU $\rightarrow$ WUE	$\lambda_{34}$	-0.455

## 453 Figures

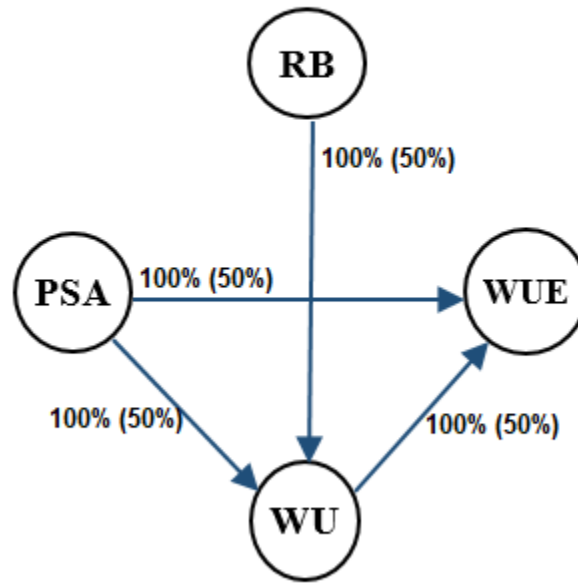


Figure 1: Scheme of inferred network structure using the Max-Min Parents and Children (MMPC) algorithm. Structure learning test was performed with 2,500 bootstrap samples with mutual information as the test statistic with a significance level at  $\alpha = 0.01$ . Labels of the edges refer to the strength 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**

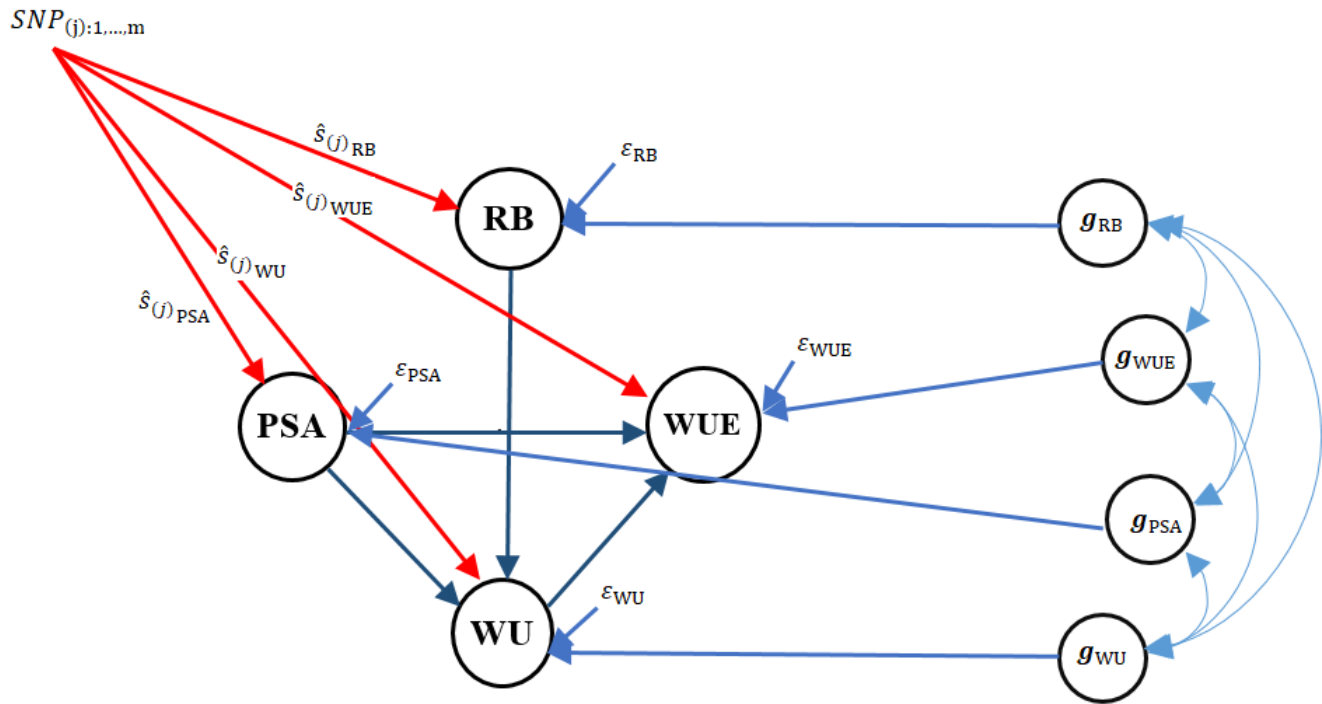


Figure 2: Pictorial representation of phenotypic network and SNP effects ( $\hat{s}$ ) using the structural equation model for four traits. Unidirectional arrows indicate the direction of effects and bidirectional arrows represent genetic correlations ( $g$ ) among phenotypes. PSA: projected shoot area;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

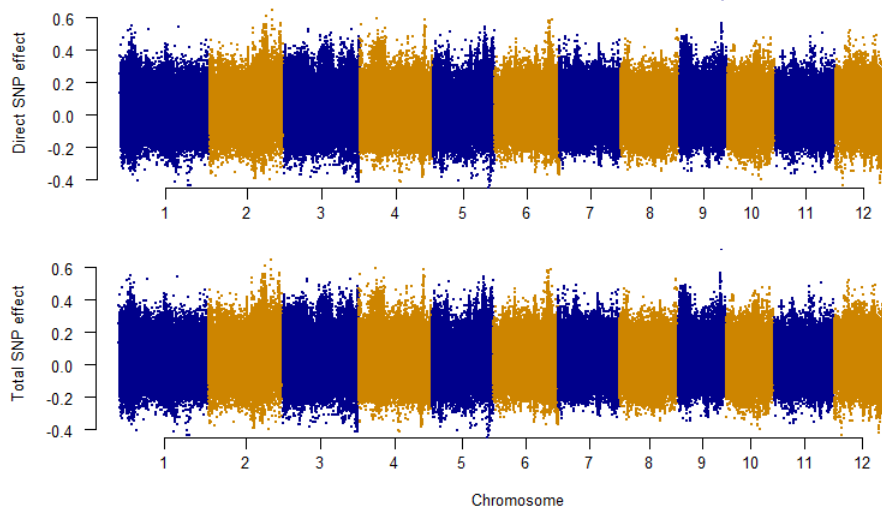


Figure 3: Manhattan plots of direct (affecting each trait without any mediation) and total (sum of all direct and indirect) SNP effects on projected shoot area (PSA) using SEM-GWAS based on the network learned by the MMPC algorithm. Each point represents a SNP and the height of the SNP represents the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

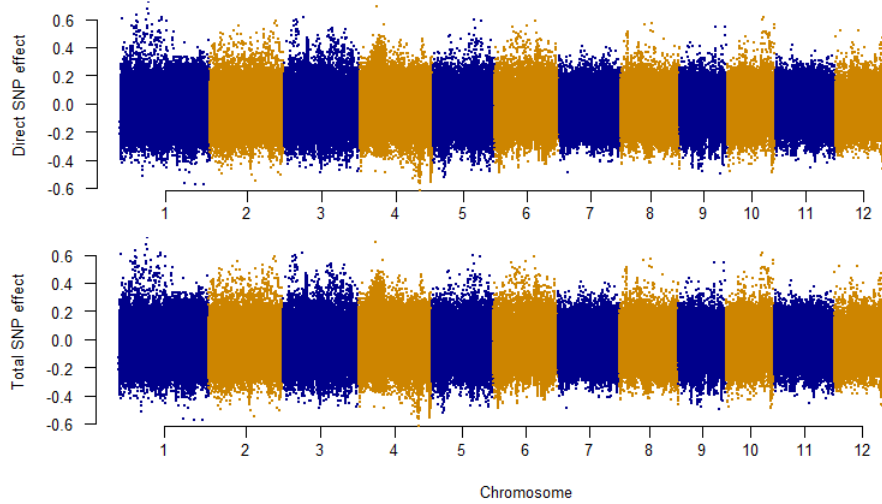


Figure 4: Manhattan plots of direct (affecting each trait without any mediation) and total (sum of all direct and indirect) SNP effects on root biomass (RB) using SEM-GWAS based on the network learned by the MMPC algorithm. Each point represents a SNP and the height of the SNP represents the extent of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



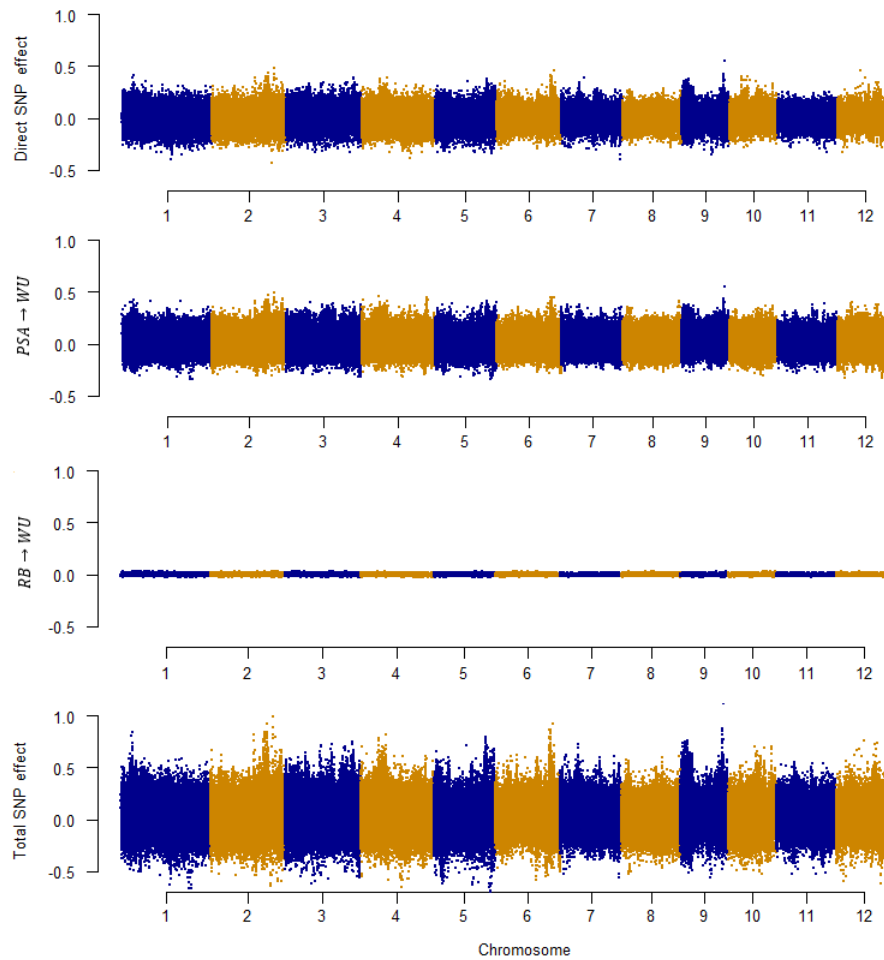


Figure 5: Manhattan plot of direct (affecting each trait without any mediation), indirect (mediated by other phenotypes), and total (sum of all direct and indirect) SNP effects on water use (WU) using SEM-GWAS based on the network learned by the MMPC algorithm. Each point represents a SNP and the height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

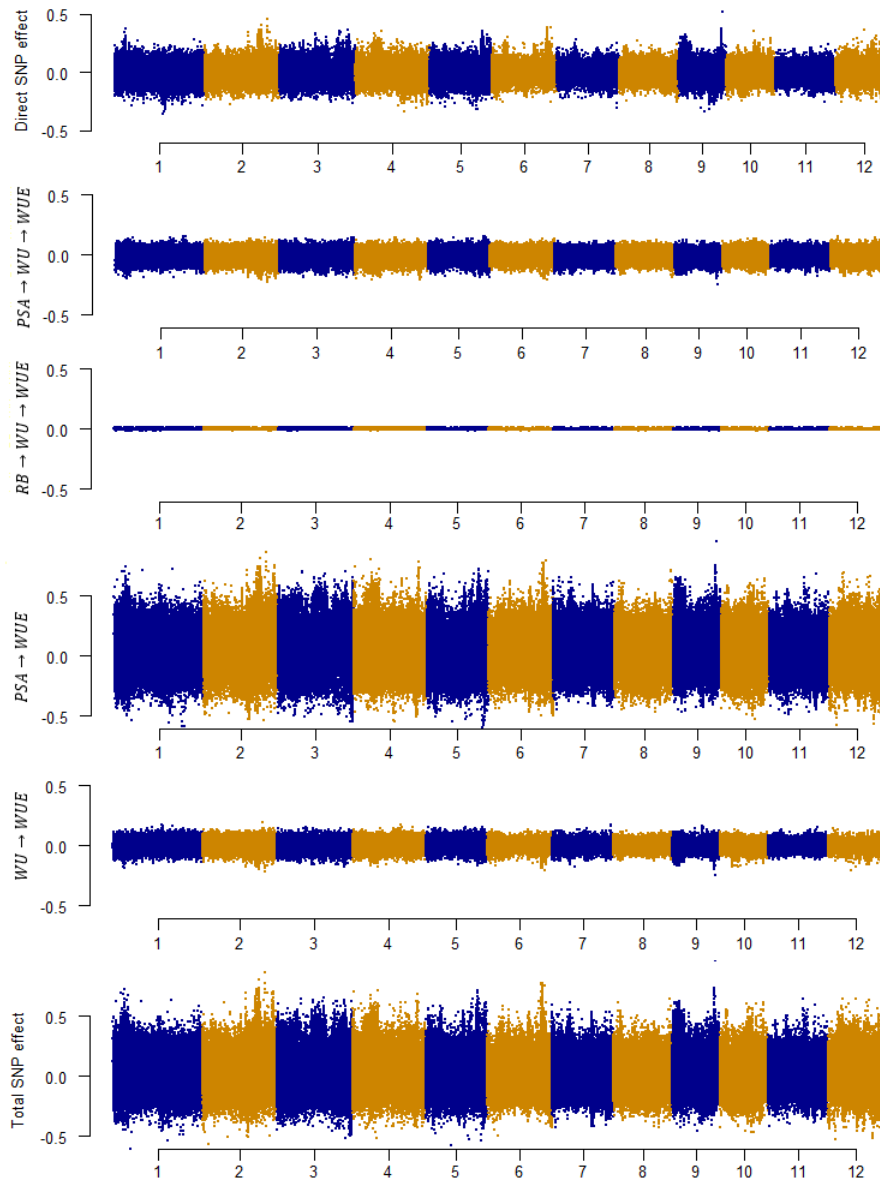


Figure 6: Manhattan plot of direct (affecting each trait without any mediation), indirect (mediated by other phenotypes), and total (sum of all direct and indirect) SNP effects on water use efficiency (WUE) using SEM-GWAS based on the network learned by the MMPC algorithm. Each point represents a SNP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

## 455 References

- 456 Ahlfors, R., Lång, S., Overmyer, K., Jaspers, P., Brosché, M., Tauriainen, A., Kollist, H.,  
457 Tuominen, H., Belles-Boix, E., Piippo, M., et al. (2004). Arabidopsis radical-induced cell  
458 death1 belongs to the wwe protein–protein interaction domain protein family and modu-  
459 lates abscisic acid, ethylene, and methyl jasmonate responses. *The Plant Cell*, 16(7):1925–  
460 1937.
- 461 Alwin, D. F. and Hauser, R. M. (1975). The decomposition of effects in path analysis.  
462 *American Sociological Review*, pages 37–47.
- 463 Aya, K., Hobo, T., Sato-Izawa, K., Ueguchi-Tanaka, M., Kitano, H., and Matsuoka, M.  
464 (2014). A novel ap2-type transcription factor, small organ size1, controls organ size down-  
465 stream of an auxin signaling pathway. *Plant and Cell Physiology*, 55(5):897–912.
- 466 Berger, B., Parent, B., and Tester, M. (2010). High-throughput shoot imaging to study  
467 drought responses. *Journal of Experimental Botany*, 61(13):3519–3528.
- 468 Bielby, W. T. and Hauser, R. M. (1977). Structural equation models. *Annual Review of*  
469 *Sociology*, 3(1):137–161.
- 470 Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and  
471 missing-data inference for whole-genome association studies by use of localized haplotype  
472 clustering. *The American Journal of Human Genetics*, 81(5):1084–1097.
- 473 Campbell, M., Walia, H., and Morota, G. (2018). Utilizing random regression models for ge-  
474 nomic prediction of a longitudinal trait derived from high-throughput phenotyping. *Plant*  
475 *Direct*, 2(9):e00080.
- 476 Campbell, M. T., Bandillo, N., Al Shiblawi, F. R. A., Sharma, S., Liu, K., Du, Q., Schmitz,  
477 A. J., Zhang, C., Véry, A.-A., Lorenz, A. J., et al. (2017a). Allelic variants of oshkt1; 1

- 478 underlie the divergence between indica and japonica subspecies of rice (*Oryza sativa*) for  
479 root sodium content. *PLoS Genetics*, 13(6):e1006823.
- 480 Campbell, M. T., Du, Q., Liu, K., Brien, C. J., Berger, B., Zhang, C., and Walia, H. (2017b).  
481 A comprehensive image-based phenomic analysis reveals the complex genetic architecture  
482 of shoot growth dynamics in rice. *The Plant Genome*, 10(2).
- 483 Campbell, T. M., Avi, C. K., Berger, B., Chris, J. B., Wang, D., and Walia, H. (2015). Inte-  
484 grating image-based phenomics and association analysis to dissect the genetic architecture  
485 of temporal salinity responses in rice. *Plant physiology*, pages pp-00450.
- 486 Challinor, A. J., Watson, J., Lobell, D. B., Howden, S., Smith, D., and Chhetri, N. (2014). A  
487 meta-analysis of crop yield under climate change and adaptation. *Nature Climate Change*,  
488 4(4):287.
- 489 Ehdaie, B. (1995). Variation in water-use efficiency and its components in wheat: II. pot  
490 and field experiments. *Crop Science*, 35(6):1617–1626.
- 491 Gianola, D., de los Campos, G., Toro, M. A., Naya, H., Schön, C.-C., and Sorensen, D.  
492 (2015). Do molecular markers inform about pleiotropy? *Genetics*, 201(1):23–29.
- 493 Gianola, D. and Sorensen, D. (2004). Quantitative genetic models for describing simultaneous  
494 and recursive relationships between phenotypes. *Genetics*, 167(3):1407–1424.
- 495 Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica*:  
496 *Journal of the Econometric Society*, pages 979–1001.
- 497 Golzarian, M. R., Frick, R. A., Rajendran, K., Berger, B., Roy, S., Tester, M., and Lun,  
498 D. S. (2011). Accurate inference of shoot biomass from high-throughput images of cereal  
499 plants. *Plant Methods*, 7(1):2.
- 500 Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations.  
501 *Econometrica, Journal of the Econometric Society*, pages 1–12.

- 502 Han, B., Chen, X.-w., Talebizadeh, Z., and Xu, H. (2012). Genetic studies of complex human  
503 diseases: characterizing snp-disease associations using bayesian networks. *BMC Systems*  
504 *Biology*, 6(3):S14.
- 505 Henderson, C. and Quaas, R. (1976). Multiple trait evaluation using relatives' records.  
506 *Journal of Animal Science*, 43(6):1188–1197.
- 507 Hirano, K., Yoshida, H., Aya, K., Kawamura, M., Hayashi, M., Hobo, T., Sato-Izawa, K.,  
508 Kitano, H., Ueguchi-Tanaka, M., and Matsuoka, M. (2017). Small organ size 1 and small  
509 organ size 2/dwarf and low-tillering form a complex to integrate auxin and brassinosteroid  
510 signaling in rice. *Molecular Plant*, 10(4):590–604.
- 511 Ho, M. D., Rosas, J. C., Brown, K. M., and Lynch, J. P. (2005). Root architectural tradeoffs  
512 for water and phosphorus acquisition. *Functional Plant Biology*, 32(8):737–748.
- 513 Huang, X. and Han, B. (2014). Natural variations and genome-wide association studies in  
514 crop plants. *Annual review of plant biology*, 65:531–551.
- 515 Hubick, K., Farquhar, G., and Shorter, R. (1986). Correlation between water-use efficiency  
516 and carbon isotope discrimination in diverse peanut (*Arachis*) germplasm. *Functional*  
517 *Plant Biology*, 13(6):803–816.
- 518 Ismail, A. M. and Hall, A. (1992). Correlation between water-use efficiency and carbon iso-  
519 tope discrimination in diverse cowpea genotypes and isogenic lines. *Crop Science*, 32(1):7–  
520 12.
- 521 Kennedy, B., Quinton, M., and Van Arendonk, J. (1992). Estimation of effects of single  
522 genes on quantitative traits. *Journal of Animal Science*, 70(7):2000–2012.
- 523 Khadilkar, A. S., Yadav, U. P., Salazar, C., Shulaev, V., Paez-Valencia, J., Pizzio, G. A.,  
524 Gaxiola, R. A., and Ayre, B. G. (2015). Constitutive and companion cell-specific overex-

- 525    pression of AVP1, encoding a proton-pumping pyrophosphatase, enhances biomass accu-  
526    mulation, phloem loading and long-distance transport. *Plant Physiology*, pages pp–01409.
- 527 Kim, H., Hwang, H., Hong, J.-W., Lee, Y.-N., Ahn, I. P., Yoon, I. S., Yoo, S.-D., Lee, S.,  
528 Lee, S. C., and Kim, B.-G. (2011). A rice orthologue of the aba receptor, ospyl/rcar5, is  
529 a positive regulator of the aba signal transduction pathway in seed germination and early  
530 seedling growth. *Journal of Experimental Botany*, 63(2):1013–1024.
- 531 Leal-Gutiérrez, J. D., Rezende, F. M., Elzo, M. A., Johnson, D., Peñagaricano, F., and  
532 Mateescu, R. G. (2018). Structural equation modeling and whole-genome scans uncover  
533 chromosome regions and enriched pathways for carcass and meat quality in beef. *Frontiers*  
534 *in Genetics*, 9.
- 535 Li, R., Tsaih, S.-W., Shockley, K., Stylianou, I. M., Wergedal, J., Paigen, B., and Churchill,  
536 G. A. (2006). Structural model analysis of multiple quantitative traits. *PLoS Genetics*,  
537 2(7):e114.
- 538 Liu, B., de La Fuente, A., and Hoeschele, I. (2008). Gene network inference via structural  
539 equation modeling in genetical genomics experiments. *Genetics*.
- 540 Mann, M. E. and Gleick, P. H. (2015). Climate change and california drought in the 21st  
541 century. *Proceedings of the National Academy of Sciences*, 112(13):3858–3859.
- 542 McCouch, S. R., Wright, M. H., Tung, C.-W., Maron, L. G., McNally, K. L., Fitzgerald,  
543 M., Singh, N., DeClerck, G., Agosto-Perez, F., Korniliev, P., et al. (2016). Open access  
544 resources for genome-wide association mapping in rice. *Nature communications*, 7:10532.
- 545 Meyer, K. and Tier, B. (2012). snp snappy: A strategy for fast genome-wide association  
546 studies fitting a full mixed model. *Genetics*, 190(1):275–277.
- 547 Mi, X., Eskridge, K., Wang, D., Baenziger, P. S., Campbell, B. T., Gill, K. S., and Dweikat,

- 548 I. (2010). Bayesian mixture structural equation modelling in multiple-trait qtl mapping.  
549 *Genetics Research*, 92(3):239–250.
- 550 Miao, C., Xiao, L., Hua, K., Zou, C., Zhao, Y., Bressan, R. A., and Zhu, J.-K. (2018). Muta-  
551 tions in a subfamily of abscisic acid receptor genes promote rice growth and productivity.  
552 *Proceedings of the National Academy of Sciences*, page 201804774.
- 553 Mir, R. R., Zaman-Allah, M., Sreenivasulu, N., Trethowan, R., and Varshney, R. K. (2012).  
554 Integrated genomics, physiology and breeding approaches for improving drought tolerance  
555 in crops. *Theoretical and Applied Genetics*, 125(4):625–645.
- 556 Momen, M., Mehrgardi, A. A., Roudbar, M. A., Kranis, A., Pinto, R. M., Valente, B. D.,  
557 Morota, G., Rosa, G. J., and Gianola, D. (2018). Including phenotypic causal networks in  
558 genome-wide association studies using mixed effects structural equation models. *bioRxiv*,  
559 page 251421.
- 560 Momen, M., Mehrgardi, A. A., Sheikhy, A., Esmailizadeh, A., Fozi, M. A., Kranis, A.,  
561 Valente, B. D., Rosa, G. J., and Gianola, D. (2017). A predictive assessment of ge-  
562 netic correlations between traits in chickens using markers. *Genetics Selection Evolution*,  
563 49(1):16.
- 564 Morota, G., Valente, B., Rosa, G., Weigel, K., and Gianola, D. (2012). An assessment  
565 of linkage disequilibrium in holstein cattle using a bayesian network. *Journal of Animal*  
566 *Breeding and Genetics*, 129(6):474–487.
- 567 Onogi, A., Ideta, O., Yoshioka, T., Ebana, K., Yamasaki, M., and Iwata, H. (2016). Un-  
568 covering a nuisance influence of a phenological trait of plants using a nonlinear structural  
569 equation: Application to days to heading and culm length in asian cultivated rice (*oryza*  
570 *sativa* l.). *PloS One*, 11(2):e0148609.
- 571 Otkin, J. A., Svoboda, M., Hunt, E. D., Ford, T. W., Anderson, M. C., Hain, C., and  
572 Basara, J. B. (2017). Flash droughts: A review and assessment of the challenges imposed

- 573 by rapid onset droughts in the united states. *Bulletin of the American Meteorological*  
574 *Society*, (2017).
- 575 Passioura, J. (2012). Phenotyping for drought tolerance in grain crops: when is it useful to  
576 breeders? *Functional Plant Biology*, 39(11):851–859.
- 577 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller,  
578 J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-  
579 genome association and population-based linkage analyses. *The American Journal of*  
580 *Human Genetics*, 81(3):559–575.
- 581 Schilling, R. K., Marschner, P., Shavrukov, Y., Berger, B., Tester, M., Roy, S. J., and Plett,  
582 D. C. (2014). Expression of the arabidopsis vacuolar h<sup>+</sup>-pyrophosphatase gene (AVP1)  
583 improves the shoot biomass of transgenic barley and increases grain yield in a saline field.  
584 *Plant Biotechnology Journal*, 12(3):378–386.
- 585 Scutari, M. (2010). Learning bayesian networks with the bnlearn r package. *Journal of*  
586 *Statistical Software, Articles*, 35(3):1–22.
- 587 Scutari, M. and Denis, J.-B. (2014). *Bayesian networks: with examples in R*. Chapman and  
588 Hall/CRC.
- 589 Scutari, M., Graafland, C. E., and Gutiérrez, J. M. (2018). Who learns better bayesian  
590 network structures: Constraint-based, score-based or hybrid algorithms? *arXiv preprint*  
591 *arXiv:1805.11908*.
- 592 Shipley, B. (2016). *Cause and correlation in biology: a user's guide to path analysis, struc-*  
593 *tural equations and causal inference with R*. Cambridge University Press.
- 594 Sinclair, T. R. (2011). Challenges in breeding for yield increase for drought. *Trends in plant*  
595 *science*, 16(6):289–293.



- 596 Töpner, K., Rosa, G. J., Gianola, D., and Schön, C.-C. (2017). Bayesian networks illustrate  
597 genomic and residual trait connections in maize (*zea mays* l.). *G3: Genes, Genomes,*  
598 *Genetics*, 7(8):2779–2789.
- 599 Tsamardinos, I., Aliferis, C. F., and Statnikov, A. (2003). Time and sample efficient discovery  
600 of markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD*  
601 *international conference on Knowledge discovery and data mining*, pages 673–678. ACM.
- 602 Tuberosa, R. and Salvi, S. (2006). Genomics-based approaches to improve drought tolerance  
603 of crops. *Trends in Plant Science*, 11(8):405–412.
- 604 Valente, B. D., Rosa, G. J., Gianola, D., Wu, X.-L., and Weigel, K. A. (2013). Is structural  
605 equation modeling advantageous for the genetic improvement of multiple traits? *Genetics*,  
606 194(3):561–572.
- 607 Valente, B. D., Rosa, G. J., Gustavo, A., Gianola, D., and Silva, M. A. (2010). Searching for  
608 recursive causal structures in multivariate quantitative genetics mixed models. *Genetics*.
- 609 VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of*  
610 *Dairy Science*, 91(11):4414–4423.
- 611 Varona, L., Sorensen, D., and Thompson, R. (2007). Analysis of litter size and average litter  
612 weight in pigs using a recursive model. *Genetics*, 177(3):1791–1799.
- 613 Verhulst, B., Maes, H. H., and Neale, M. C. (2017). GW-SEM: A statistical package to  
614 conduct genome-wide structural equation modeling. *Behavior Genetics*, 47(3):345–359.
- 615 Wang, H. and van Eeuwijk, F. A. (2014). A new method to infer causal phenotype networks  
616 using qtl and phenotypic information. *PloS One*, 9(8):e103997.
- 617 Wehner, M. F., Arnold, J. R., Knutson, T., Kunkel, K. E., and N, L. A. (2017). Droughts,  
618 Floods, and Wildfires. In *Climate Science Special Report: Fourth National Climate As-*  
619 *essment, Volume I.*, pages 231–256.

- 620 Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20(7):557–  
621 585.
- 622 Wu, X.-L., Heringstad, B., and Gianola, D. (2010). Bayesian structural equation models for  
623 inferring relationships between phenotypes: a review of methodology, identifiability, and  
624 applications. *Journal of Animal Breeding and Genetics*, 127(1):3–15.
- 625 You, J., Zong, W., Li, X., Ning, J., Hu, H., Li, X., Xiao, J., and Xiong, L. (2012). The  
626 *snac1*-targeted gene *ossro1c* modulates stomatal closure and oxidative stress tolerance by  
627 regulating hydrogen peroxide in rice. *Journal of Experimental Botany*, 64(2):569–583.
- 628 Yu, H., Campbell, M. T., Zhang, Q., Walia, H., and Morota, G. (2018). Genomic bayesian  
629 confirmatory factor analysis and bayesian network to characterize a wide spectrum of rice  
630 phenotypes. *bioRxiv*, page 435792.
- 631 Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D.,  
632 Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method  
633 for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*,  
634 38(2):203.
- 635 Zampieri, M., Ceglar, A., Dentener, F., and Toreti, A. (2017). Wheat yield loss attributable  
636 to heat waves, drought and water excess at the global, national and subnational scales.  
637 *Environmental Research Letters*, 12(6):064008.
- 638 Zhang, J., Li, J., Wang, X., and Chen, J. (2011). *Ovp1*, a vacuolar h<sup>+</sup>-translocating inorganic  
639 pyrophosphatase (*v-ppase*), overexpression improved rice cold tolerance. *Plant Physiology*  
640 *and Biochemistry*, 49(1):33–38.
- 641 Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y.,  
642 Bassu, S., Ciaia, P., et al. (2017). Temperature increase reduces global yields of major  
643 crops in four independent estimates. *Proceedings of the National Academy of Sciences*,  
644 114(35):9326–9331.

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