Modeling multiple phenotypes in wheat using data-driven genomic exploratory factor analysis and Bayesian network learning

Mehdi Momen[1], Madhav Bhatta[2], Waseem Hussain[3], Haipeng Yu[1], and Gota Morota[1*]

[1]Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA 24061
[2]Department of Agronomy, University of Wisconsin-Madison, Madison, WI 53706, USA
[3]International Rice Research Institute, Los Banos, Philippines

* 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

Running title: Factor analysis for wheat

Keywords: Bayesian network, confirmatory factor analysis, exploratory factor analysis, multi-trait, wheat

ORCID: 0000-0002-2562-2741 (MM), 0000-0002-4959-0481 (MB), 0000-0002-6861-0193 (WH), 0000-0002-8923-9733 (HY), and 0000-0002-3567-6911 (GM).

Abstract

Inferring trait networks from a large volume of genetically correlated diverse phenotypes such as yield, architecture, and disease resistance can provide information on the manner in which complex phenotypes are interrelated. However, studies on statistical methods tailored to multi-dimensional phenotypes are limited, whereas numerous methods are available for evaluating the massive number of genetic markers. Factor analysis operates at the level of latent variables predicted to generate observed responses. The objectives of this study were to illustrate the manner in which data-driven exploratory factor analysis can map observed phenotypes into a smaller number of latent variables and infer a genomic latent factor network using 45 agro-morphological, disease, and grain mineral phenotypes measured in synthetic hexaploid wheat lines (*Triticum Aestivum L.*). In total, eight latent factors including grain yield, architecture, flag leaf-related traits, grain minerals, yellow rust, two types of stem rust, and leaf rust were identified as common sources of the observed phenotypes. The genetic component of the factor scores for each latent variable was fed into a Bayesian network to obtain a trait structure reflecting the genetic interdependency among traits. Three directed paths were consistently identified by two Bayesian network algorithms. Flag leaf-related traits influenced leaf rust, and yellow rust and stem rust influenced grain yield. Additional paths that were identified included flag leaf-related traits to minerals and minerals to architecture. This study shows that data-driven exploratory factor analysis can reveal smaller dimensional common latent phenotypes that are likely to give rise to numerous observed field phenotypes without relying on prior biological knowledge. The inferred genomic latent factor structure from the Bayesian network provides insights for plant breeding to simultaneously improve multiple traits, as an intervention on one trait will affect the values of focal phenotypes in an interrelated complex trait system.

Background

With the development of high-throughput phenotyping technologies, phenomics has been generating plant measurements at a greater level of resolution and dimensionality (Araus and Cairns, 2014; Watanabe et al., 2017). Integrating these diverse and heterogeneous data to improve the biological understanding of plant systems and interpret the underlying interrelationships among phenotypes remains challenging (Morota et al., 2019). One approach is to model each measurement as a different trait using a multi-trait model (Henderson and Quaas, 1976). However, in a high-dimensional specification, where the number of traits measured per genotype can reach hundreds or thousands, this approach leads to dramatic increases in the computational burden or difficulties in interpreting the results. Recently, Yu et al. (2019) showed that factor analysis can be used to reduce the dimension of response variables by assuming latent factors that give rise to observed phenotypes in rice. They used confirmatory factor analysis (CFA), which requires knowledge of the phenotype-factor category before data analysis. However, reliable phenotype-factor patterns are not always known in advance. Alternatively, exploratory factor analysis (EFA) can be used to perform latent variable analysis by estimating patterns from data when a latent structure cannot be determined a priori. EFA identifies underlying latent factors to represent observed measurements, which is useful when the exact number and meaning of latent factors are unknown (Jöreskog, 1967; Hoyle and Duvall, 2004).

The first objective of this study was to illustrate the utility of EFA for revealing the underlying genomic latent structure of agronomic or agro-morphological phenotypes for synthetic hexaploid wheat lines (*T. aestivum L*). Grain yield in wheat is influenced by several agro-morphological traits. However, successfully incorporating yield-promoting agro-morphological traits in breeding programs to improve genetic gains requires detailed knowledge of the interrelationships between and among traits. The second objective was to determine a trait network structure among the genomic latent factors using a Bayesian network. This is an essential task because breeding programs often aim to improve multiple corre-

lated traits concurrently. Knowledge of directed trait networks accounting for the genetic interdependency among traits can improve the understanding of the manner in which the selection of one phenotype may increase or decrease the observation of another phenotype, providing additional insight beyond associations (Valente et al., 2015). The current study demonstrates the advantages of the joint application of factor analysis and Bayesian network as a data-driven approach to discover interrelationships between a set of many correlated traits in wheat.

Materials and Methods

Plant materials

A diversity panel of $n = 123$ synthetic hexaploid wheat lines, derived from an interspecific cross between wild accessions of goat grass (*Aegilops tauschii L.*) and diverse accessions of cultivated durum wheat (*Triticum turgidum L.*), was used in this study. These plant materials were shared by the International Winter Wheat Improvement Program in Turkey and are available at http://www.iwwip.org. Pedigree information and other details on these lines were reported previously (Bhatta et al., 2018a,c,d). Briefly, the lines originated from two breeding programs. The first group of synthetics comprises 14 lines developed by Kyoto University, Japan, from 1 Langdon durum parent crossed with 14 different accessions of *Ae. tauschii*. The second group consists of 109 lines developed by the International Maize and Wheat Improvement Center from crosses between 6 winter durum wheats and 11 different *Ae. tauschii* accessions. The synthetic lines used in this study are unique; they were recently developed (F8–F9 generations) and tested for multiple traits for use in a breeding program.

Phenotypic and genotypic data

We analyzed 16 agronomic-, 16 grain mineral-, and 13 wheat rust-related phenotypes in the current study. Agronomic traits including grain yield (GY), harvest index (HI), biomass weight (BMWT), grain volume weight (GVWT), flag leaf length (FLL), flag leaf width (FLW), flag leaf area (FLA), rachis break (RB), sterile spikelet (SP), spike length (SL), seeds per spike (SPS), spikelet number (SN), fertile spikelet (FS), spike weight (SW), grain weight per spike (GPS), and spike harvest index (SHI) were measured using previously described standard procedures (Bhatta et al., 2018a; Morgounov et al., 2018; Hussain et al., 2017). Grain minerals including arsenic (As), calcium (Ca), cadmium (Cd), cobalt (Co), copper (Cu), iron (Fe), potassium (K), lithium (Li), magnesium (Mg), manganese (Mn), molybdenum (Mo), nickel (Ni), phosphorous (P), sulfur (S), titanium (Ti), and zinc (Zn)

were measured via inductively-coupled plasma mass spectrometry (ICP-MS, Agilent 7500cx, Agilent Technologies, Santa Clara, CA, USA) at the University of Nebraska Redox Biology Center, Proteomics and Metabolomics Core (Guttieri et al., 2015; Bhatta et al., 2018a). The wheat rust (leaf stem and yellow rusts) disease severity, coefficient of infection, and infection type were tested under field conditions as previously described (Peterson et al., 1948; Morgounov et al., 2018; Bhatta et al., 2018d). Wheat rust traits collected from several locations in Turkey and one location in Kenya included the leaf rust coefficient of infection (LRCI), leaf rust infection type (LRIT), leaf rust severity (LRS), stem rust coefficient of infection at Haymana (SRCIH), stem rust infection type at Haymana (SRITH), stem rust severity at Haymana (SRSH), stem rust coefficient of infection at Kastamonu (SRCIK), stem rust infection type at Kastamonu (SRITK), stem rust severity at Kastamonu (SRSK), yellow rust coefficient of infection at Haymana (YRCIH), yellow rust infection type at Haymana (YRIH), yellow rust severity at Haymana (YRSH), and yellow rust severity at Kastamonu (YRSK). All lines were genotyped with the genotyping by sequencing technology (Bhatta et al., 2018c). After setting a minor allele frequency threshold of 0.05, 35,648 markers remained for analysis.

Experimental design and analysis

The experiments were conducted across several locations in Turkey and one location in Kenya in 2017. The experimental design was an alpha lattice design with two replications (Barreto et al., 1996). A linear mixed model coupled with restricted maximum likelihood implemented in the PROC MIXED procedure in SAS 9.4 (SAS Institute, Inc., Cary, NC, USA) was used to obtain the adjusted means for each trait from the following model (Bhatta et al., 2018b).

$$y_{ijkl} = \mu + r_i + b(r)_{ji} + c_k + g_{l(ji)} + \epsilon_{ijkl},$$

where y_{ijk} is the trait of interest; μ is the overall mean; r_i is the effect of ith replication; $b(r)_{ji}$ is the effect of the jth block within the ith replication; c_k is the kth check; g_{lji} (new variable, where check is coded as 0 and entry is coded as 1, and the genotype is considered a new variable $\times$ entry) is the effect of the lth genotype within the jth incomplete block of the ith replication; and ϵ_{ijkl} is the residual.

Exploratory factor analysis

Exploratory factor analysis can reveal the latent structure among phenotypes when no hypotheses about the nature of the underlying factor can be assumed *a priori*. This section closely follows the work of Yu et al. (2020). The aforementioned $t = 45$ phenotypes were analyzed using EFA by fitting

$$\mathbf{Y} = \mathbf{\Lambda F} + \mathbf{U}, \tag{1}$$

where $\mathbf{Y}$ is the $t \times n$ phenotypic matrix; $\mathbf{\Lambda}$ is the $t \times q$ matrix of factor loading indicating the relation between phenotypes and latent common factors; $\mathbf{F}$ is the $q \times n$ matrix of latent factor scores; and $\mathbf{U}$ is the $t \times n$ vector of unique effects that is not explained by q underlying common factors. The variance-covariance matrix of $\mathbf{Y}$ is

$$\mathbf{\Sigma} = \mathbf{\Lambda \Phi \Lambda'} + \mathbf{\Psi}, \tag{2}$$

where $\mathbf{\Sigma}$ is the $t \times t$ variance-covariance matrix of phenotypes, $\mathbf{\Phi}$ is the variance of factor scores, and $\mathbf{\Psi}$ is a $t \times t$ diagonal matrix of unique variance. The elements of $\mathbf{\Lambda}$, $\mathbf{\Phi}$, and $\mathbf{\Psi}$ are parameters of the model to be estimated from the data. We assumed $\mathbf{\Phi} = \mathbf{I}$ yielding factors each with unit variance (Jöreskog, 1967; Anderson, 2003). With the assumption of $\mathbf{F} \sim \mathscr{N}(\mathbf{0}, \mathbf{I})$, parameters $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ were estimated by maximizing the log-likelihood of $\mathscr{L}(\mathbf{\Lambda}, \mathbf{\Psi}|\mathbf{Y})$ using the R package psych (Revelle, 2018) along with a varimax rotation (Kaiser, 1958). A threshold of $\lambda > |0.3|$ was first applied to screen out factor loading values. Then each phenotype was assigned to only one of the factors based on its largest loading.

Parallel analysis was performed to estimate the optimum number of factors from data in EFA (Horn, 1965; Hayton et al., 2004). This is conducted by generating simulated data from the observed data. Next, the eigenvalues were extracted until the observed data had a smaller eigenvalue than the simulated data. The number of eigenvalues was used as the number of optimum factors.

The factor ability of the data set was also assessed by estimating the Kaiser-Meyer-Olkin measure of sampling adequacy (Cerny and Kaiser, 1977). This criterion measures the adequacy of the dataset for factor analysis by investigating the correlation and partial correlation matrices of the phenotypes. The measure of sampling adequacy ranges between 0 to 1, and values closer to 1 are preferred. When the measure of sampling adequacy is less than 0.5, the dataset is not recommended for factor analysis (Cerny and Kaiser, 1977).

Confirmatory factor analysis

Once the phenotype-factor pattern was established by EFA, Bayesian CFA was used to obtain factor scores. Although EFA and CFA are similar, there are also clear differences. In general, EFA is used to find a latent structure in data, whereas CFA requires the phenotype-latent variable category to be known before analysis and is often used to estimate factor scores based on the structure from EFA. The differences between EFA and CFA are shown in Figure 1. In a Bayesian setting, all unknowns in equations (1) and (2) were assigned priors. The assignment of priors was performed according to Yu et al. (2019, 2020) using the default priors in the blavaan R package (Merkle and Rosseel, 2018). A Gaussian distribution with a mean of zero and variance of 100 was assigned to the factor loading term. The variance-covariance matrix of the latent factors followed an inverse Wishart distribution with a scale matrix of an 8×8 identity matrix and degree of freedom of 8. Each error variance followed an inverse Gamma distribution with a shape parameter of 1 and scale parameter of 0.5. The factor scores of latent variables ($\mathbf{F}$) were sampled from the conditional distribution of $p(\mathbf{F}|\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{\Psi}, \mathbf{Y})$ (Lee and Song, 2012) using a data augmentation technique (Tanner and Wong, 1987). The

posterior mean of $\mathbf{F}$ was considered a new phenotype in subsequent analysis. Convergence was diagnosed by the potential scale reduction factor (PSRF) (Gelman et al., 1992; Brown, 2014). This criterion utilizes at least two Markov chains, which are considered to be mixed to a stationary status if the ratio of between the chain variance to within the chain variance is close to 1. In total, two chains, each consisting of 5,000 Markov chain Monte Carlo samples after 2,000 burn-in samples, were collected to derive the posterior means.

Multi-trait genomic best linear unbiased prediction

A Bayesian multi-trait genomic best linear unbiased prediction model was applied to partition inferred latent variables into genetic and environmental components.

$$\mathbf{F} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{g} + \mathbf{e},$$

where $\mathbf{F}$ is the vector of estimated factor scores, $\mathbf{X}$ is the incidence matrix of covariates including the intercept and the top three principal components accounting for population structure, $\mathbf{b}$ is the vector of covariate effects, $\mathbf{Z}$ is the incidence matrix relating the factor scores of each latent variable to additive genetic effect, $\mathbf{g}$ is a vector of additive genetic effect, and $\mathbf{e}$ is the vector of residuals. Under the infinitesimal model of inheritance, $\mathbf{g}$ and $\mathbf{e}$ were assumed to follow a multivariate Gaussian distribution of $\mathbf{g} \sim N(0, \Sigma_g \otimes \mathbf{G})$ and $\mathbf{e} \sim N(0, \Sigma_e \otimes \mathbf{I})$, respectively. Here, $\mathbf{G}$ is a $n \times n$ genomic relationship matrix, $\mathbf{I}$ is a $n \times n$ identity matrix, Σ_g and Σ_e are variance-covariance matrices of additive genetic effect and residuals, respectively, and $\otimes$ is the Kronecker product. The $\mathbf{G}$ matrix was set 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 the values of $0 - 2p_j$ for zero copies of the reference allele, $1 - 2p_j$ for one copy of the reference allele, $2 - 2p_j$ for two copies of the reference allele, and p_j is the allele frequency at marker $j = 1, \cdots, m$ (VanRaden, 2008). The prior distribution specifications followed those of Momen et al. (2019). A flat prior was assigned for $\mathbf{b}$. The vectors of additive

genetic and residual effects were assigned independent multivariate Gaussian priors with null mean and inverse Wishart distributions for the covariance matrices Σ_g and Σ_e. A Gibbs sampler was used to obtain posterior distributions. A burn-in of 10,000 samples followed by an additional 90,000 samples, thinned by a factor of two, resulted in 45,000 available samples for posterior mean inferences. The MTM R package was used to fit the model (https://github.com/QuantGen/MTM).

Bayesian network structure learning

The posterior means of genetic values of latent variables obtained from the Bayesian multi-trait genomic best linear unbiased prediction model were used to examine the manner in which the traits are interrelated using a Bayesian network. A Bayesian network is a graphical representation of the conditional independence among random variables based on a directed acyclic graph (Heckerman et al., 1995). For example, if an arrow arises from phenotype A to phenotype B, phenotype A is considered to impact phenotype B directly conditional on the remaining phenotypes, whereas the absence of an edge implies conditional independence given the remaining phenotypes. In this study, the Tabu search (Tabu) and Max-Min Hill-Climbing (MMHC) algorithms were applied to learn the underlying trait network structure of latent variables at the genetic level using the bnlearn R package (Scutari and Denis, 2014). These two algorithms were chosen because they yielded a reasonable result in a recent study (Yu et al., 2019). The Bayesian information criterion (BIC) score was calculated for the whole network and for each edge. A higher BIC score leads to greater model fit because the BIC score is rescaled by -2 in the bnlearn package. Additionally, the strength and uncertainty of the direction of each edge were estimated probabilistically by bootstrapping (Scutari and Denis, 2014). Before fitting the Bayesian network structure learning algorithms, genetic values of latent variables were transformed to be uncorrelated to meet the primary assumption of a Bayesian network (Töpner et al., 2017; Yu et al., 2019).

Data availability

The data are available from the previously published studies. The agronomic, grain minerals, and rust related phenotypic data are available from Bhatta et al. (2018a,d, 2019) and the marker data are available from Bhatta et al. (2018d).

Results

Assessing factorability and factor selection

Figure 2 shows the Pearsons correlation coefficients among all observed variables represented in a heat map. Moderate to high correlations were observed within the spike-, mineral-, and rust- related traits. Because the objective of factor analysis is to model the interrelationships between observed traits with a smaller subset of latent variables, the presence of some block structures in the heat map suggests that our dataset is suited for factor analysis. This observation was supported by the overall Kaiser-Meyer-Olkin measure of sampling adequacy, which was estimated as 0.7, indicating that the factorability of the dataset was sufficient. Parallel analysis was performed to determine the appropriate number of latent variables. The first eight eigenvalues extracted from the original data were larger than the first eight eigenvalues obtained from simulated random data. Thus, eight underlying latent variables were examined in subsequent analysis.

Factor loading from EFA

Factor analysis was performed to understand the biological meaning of the eight latent factors by investigating the co-variation among measured observations using EFA. Figure 3 summarizes the degree of the contributions of unobserved factors to the observed phenotypes. Because EFA allows the cross-loading of phenotypes, an additional step is required so that each phenotype loads only on one factor. A heat map of the estimated factor loading values for each phenotype is shown in Figure 3A. The results showed that each variable had some nonzero loadings on several factors. Figure 3B shows the phenotype-latent variable pattern after selecting the largest loading for each phenotype and imposing a threshold of $> |0.30|$. This resulted in each phenotype loading on only one factor except for GVWT, RB, SP, and YRSK, which did not load on to any factors. The results showed that all mineral-related traits including As, Ca, Cd, Co, Cu, Fe, K, Li, Mg, Mn, Mo, Ni, P, S, Ti, and Zn were loaded

on the first factor (F1) ranging from 0.34 to 0.98. Seven agronomic traits including FS, SL, SN, SPS, SW, GPS, and SHI were placed on the second factor (F2) and biologically all appear to be related to the plant structure. In this category, the lowest loading was estimated for the SHI (0.44) and the largest for GPS (0.91). The 12 disease-related phenotypes were distributed among 4 factors (F3, F4, F5, and F6) with a loading of at least 0.8 in their categories. FLL, FLW, and FLA traits with 0.84, 0.73, and 0.98 loadings, respectively, were placed on the seventh factor (F7). Finally, GY, HI, and BM loaded on the eighth factor (F8).

Figure 4 shows the overall inferred latent structure of the data. The biological meanings attached to the eight factors according to the EFA analysis were GYL: grain yield; ARC: plant architecture; FL: flag and leaf, MIN: minerals; YRD: yellow rust disease; SRDK: stem rust disease at Kastamonu; SRDH: stem rust disease at Haymana; and LRD: leaf rust disease. These estimated latent factors were subsequently evaluated to determine their genetic interrelationships.

Confirmatory factor analysis

Table 1 shows the posterior means and their posterior standard deviations of the standardized loadings, PSRF, and R^2 statistics from the Bayesian CFA. Convergence was diagnosed from the PSRF of each observed phenotype. The estimated PSRF values for all phenotypes were close to 1, suggesting that they converged to a stationary status. The result showed that the eight latent factors strongly contributed to the observed phenotypes. For the latent factor GYL, the lowest and highest loading values were obtained for HI and GY, respectively. For the FL latent factor, all three phenotypes presented a loading of at least 0.77. In ARC, the factor loading values varied from SHI to FS in ascending order. The MIN latent factor was associated with the 16 observed phenotypes, which was the largest factor. The lowest and highest loading values were obtained for Ti and Mg, respectively. The remaining four latent factors including LRD, SRHD, SRKD, and YRD, which are relevant to diseases, showed

that the data fit well with >0.8 loading. The extent of R^2 values mostly agreed with the estimated loadings with a correlation of 0.99.

Bayesian network among genomic latent factors

The Bayesian network was used to investigate the interrelationships among the genetic components of latent factors. Because SRDH and SRDK capture the same set of phenotypes with a high correlation (Figure 3) but were collected at different locations, only SRDH was used for trait network structure analysis. As shown in Figure 5, Tabu yielded six directed edges from FL to LRD and MIN, from YRD to LRD and GYL, from MIN to ARC, and from SRDH to GYL. However, MMHC only produced three directed edges that were a subset of the Tabu network. Thus, the consensus network has common directed edges from FL and LRD, from YRD to GYL, and SRDH to GYL. These results suggest that there is stronger evidence that FL, YRD, and SRDH directly influence LRD, GYL, and GYL, respectively. In both networks, the bootstrapping results revealed that confidence was always higher regarding the presence or absence of edges compared to the directions of edges. The goodness-of-fit statistics measured by BIC is shown in Table 2. This table shows how well the paths mirror the dependence structure of the data. According to the BIC values, Tabu yielded a larger BIC score than the MMHC algorithms for the entire network (-423.61 vs. -437.39). For each specific path, removing SRDH $\rightarrow$ GYL resulted in the largest decrease in the BIC score, suggesting that this path plays the most important role in the network structure. This was followed by YRD $\rightarrow$ GYL and FL $\rightarrow$ LRD. The top three most influential paths in Tabu formed the network structure of MMHC.

Discussion

Data-driven latent variable analysis

With the availability of large volumes of measured observations per individual because of recent advances in phenomics, it is critical to develop a phenotype-centric statistical approach. Factor analysis is an effective method for handling many response variables in a quantitative genetic framework (Runcie and Mukherjee, 2013; Peñagaricano et al., 2015; Rocha et al., 2018; Yu et al., 2019, 2020). The central idea behind factor analysis is to model the observed phenotypes through unobserved latent factors by maximizing the common variance between correlated phenotypes. In the current study, latent factors were directly inferred from the field data of physiological and morphological phenotypes in wheat using EFA followed by estimating their factor scores by CFA. This allowed the analysis of the lower dimensional data because the number of latent factors was less than the number of observed phenotypes. The combination of EFA and CFA enabled the evaluation of the genetics of latent factors that were predicted to give rise to the observed phenotypes. Our results demonstrate that a data-driven approach for estimating latent factors using EFA is useful because the observed traits were uniquely assigned to one of the factors with biological interpretations. This contrasts with the results of a recent study by Yu et al. (2019), in which observed phenotypes were classified into factors based on prior biological knowledge. However, in most scenarios, the phenotype-latent variable pattern may be unknown. In contrast, EFA can be used to perform latent variable analysis by estimating latent factors from data when the latent structure cannot be determined *a priori*.

The interrelationships among latent variables were investigated at the genomic level using Tabu and MMHC. Based on the BIC values, Tabu resulted in a better fit than MMHC. This agrees with the findings of recent studies using Bayesian networks (Töpner et al., 2017; Scutari et al., 2018; Yu et al., 2019). The trait network structure inferred from MMHC was a subset of that of MMHC. Additionally, the three directed paths identified from MMHC

were the top three most important paths in Tabu according to BIC. This suggests that the networks structures were consistent between Tabu and MMHC. Thus, the trait network derived from MMHC can be considered the consensus network that is more reliable. The network structures from Tabu and MMHC may become aligned by increasing the sample size. Inferring a trait network from observational data is an emerging topic in quantitative genetics (Valente et al., 2010). Because breeders are often interested in the impact of external intervention or the selection of one trait over other traits, distinguishing undirected edges from directed edges is important. The trait network learned in this study can also be integrated into SEM-GWAS, which is a framework to perform multi-trait genome-wide association analysis derived from structural equation models (Momen et al., 2018, 2019). The combination of data-driven EFA and Bayesian network approaches is particularly useful for analyzing image-based high-throughput phenotyping data, where relationships within image-based phenotypes and between classical phenotypes and image-based phenotypes may not always be obvious.

Biological meaning of the inferred relationships

Previous studies revealed the negative genetic associations of yellow and stem rust traits with grain yield traits. Wheat rust diseases are foliar fungal diseases whose infection on the flag leaf close to the grain filling period causes a decline in the photosynthetic ability of the plant, drastically decreasing the grain filling process and reducing the biomass yield, thousand kernel weight, and harvest index (He et al., 2019; Bhatta et al., 2018a; Herrera-Foessel et al., 2006). Thus, the reduction of these important traits results in a reduction in the final grain yield (SRDH $\rightarrow$ GYL and YRD $\rightarrow$ GYL). Wheat leaf rust may be affected by flag leaf traits such as FLL, FLW, and FLA (FL $\rightarrow$ LRD). As the flag leaf area increases, the surface also becomes greater, increasing the risk of disease infection on the wider and longer leaves.

Flag leaf traits play important roles in the synthesis, translocation, and remobilization

of photo-assimilates and minerals to the grains (Sperotto et al., 2013). A recent study on *Triticum sps.* showed that the flag leaf contains two- to three-fold higher concentrations of Fe and Zn than the grain mineral concentrations (Hu et al., 2017). They also found strong positive correlations between leaf and grain Fe and Zn concentrations. Another study used more than 120 hexaploid wheat lines and reported a significant positive correlation of flag leaf N concentrations at anthesis with grain Fe, Mn, and Cu (SHI et al., 2013). These results suggest that flag leaf traits play an important role in determining the grain mineral concentration, which agrees with our results indicating a direct link from FL to MN.

Foliar diseases such as yellow rust, caused by *Puccinia striiformis f. sp. tritici (Pst)*, is an important foliar fungal disease of wheat that causes major yield loss (Bhatta et al., 2019). This disease produces rust pustules on leaves and reduces the process of photosynthesis and translocation of photosynthate to grain yield traits, which in turn inhibit grain filling, possibly resulting in a significant reduction in grain weight and ultimately reducing grain yield (Ye et al., 2019; Murray and Murray, 2005). A recent study on winter wheat germplasm showed that yellow rust infection seriously damaged the photosynthetic function of leaves at an earlier stage of grain filling, leading to biomass loss (He et al., 2019). Additionally, the presence of foliar diseases in wheat is associated with a reduction in the biomass weight and harvest index by reducing the healthy leaf area and affecting healthy spike growth (Gooding et al., 2000; Dimmock and Gooding, 2002), indicating that yellow rust traits affected grain yield-related traits (YRD $\rightarrow$ GYL).

Several studies have reported negative associations between grain minerals and architecture-related traits. A larger number of seeds per spike and kernel size in wheat is associated with lower grain mineral accumulation in the grain, which is mainly attributed to the grain mineral dilution effect (Bhatta et al., 2018a; Guttieri et al., 2015). Similarly, the nitrogen concentration in the grains depends on their position within the spike Calderini and Ortiz-Monasterio (2003); Herzog and Stamp (1983), suggesting that spike architecture traits have important impacts on grain mineral traits (MIN $\rightarrow$ ARC).

Conclusions

This study demonstrates that data-driven latent variable analysis can reveal the underlying structure of phenotypes on a smaller dimensional scale. Thus, determining the genetic effects of correlated traits by factor analysis is an efficient approach for learning the minimum set of core factors contributing to high-dimensional observed phenotypes. Additionally, by reconstructing a more general structure of genomic latent factors from observed phenotypes using a Bayesian network, a clearer picture of trait interdependency can be obtained, which is useful for developing breeding and management strategies for crops such as wheat.

Acknowledgements

This work was supported in part by Virginia Polytechnic Institute and State University startup funds to GM.

Author contributions

MM, MB, WH, and GM conceived the study. MM and HY analyzed the data. MM drafted the manuscript. WH, MB, HY, and GM revised the manuscript. GM supervised and directed the study. All authors read and approved the manuscript.

Conflict of interests

The authors declare that they have no competing interests.

References

Anderson, T. (2003). An introduction to multivariate statistical analysis (wiley series in probability and statistics). *July 11*.

Araus, J. L. and Cairns, J. E. (2014). Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science*, 19(1):52–61.

Barreto, H., Edmeades, G., Chapman, S., and Crossa, J. (1996). The alpha lattice design in plant breeding and agronomy: generation and analysis. *Drought-and Low N-Tolerant Maize*, page 544.

Bhatta, M., Baenziger, P., Waters, B., Poudel, R., Belamkar, V., Poland, J., and Morgounov, A. (2018a). Genome-wide association study reveals novel genomic regions associated with 10 grain minerals in synthetic hexaploid wheat. *International Journal of Molecular Sciences*, 19(10):3237.

Bhatta, M., Morgounov, A., Belamkar, V., and Baenziger, P. S. (2018b). Genome-wide association study reveals novel genomic regions for grain yield and yield-related traits in drought-stressed synthetic hexaploid wheat. *International journal of molecular sciences*, 19(10):3011.

Bhatta, M., Morgounov, A., Belamkar, V., Poland, J., and Baenziger, P. S. (2018c). Unlocking the novel genetic diversity and population structure of synthetic hexaploid wheat. *BMC Genomics*, 19(1):591.

Bhatta, M., Morgounov, A., Belamkar, V., Wegulo, S. N., Dababat, A. A., Erginbas-Orakci, G., Bouhssini, M. E., Gautam, P., Poland, J., Akci, N., et al. (2019). Genome-wide association study for multiple biotic stress resistance in synthetic hexaploid wheat. *International Journal of Molecular Sciences*, 20(15):3667.

Bhatta, M., Morgounov, A., Belamkar, V., Yorgancılar, A., and Baenziger, P. S. (2018d). Genome-wide association study reveals favorable alleles associated with common bunt resistance in synthetic hexaploid wheat. *Euphytica*, 214(11):200.

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

Calderini, D. F. and Ortiz-Monasterio, I. (2003). Grain position affects grain macronutrient and micronutrient concentrations in wheat. *Crop Science*, 43(1):141–151.

Cerny, B. A. and Kaiser, H. F. (1977). A study of a measure of sampling adequacy for factor-analytic correlation matrices. *Multivariate Behavioral Research*, 12(1):43–47.

Dimmock, J. and Gooding, M. (2002). The effects of fungicides on rate and duration of grain filling in winter wheat in relation to maintenance of flag leaf green area. *The Journal of Agricultural Science*, 138(1):1–16.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Gooding, M., Dimmock, J., France, J., and Jones, S. (2000). Green leaf area decline of wheat flag leaves: the influence of fungicides and relationships with mean grain weight and grain yield. *Annals of applied Biology*, 136(1):77–84.

Guttieri, M. J., Baenziger, P. S., Frels, K., Carver, B., Arnall, B., and Waters, B. M. (2015). Variation for grain mineral concentration in a diversity panel of current and historical great plains hard winter wheat germplasm. *Crop Science*, 55(3):1035–1052.

Hayton, J. C., Allen, D. G., and Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, 7(2):191–205.

He, C., Zhang, Y., Zhou, W., Guo, Q., Bai, B., Shen, S., and Huang, G. (2019). Study on stripe rust (puccinia striiformis) effect on grain filling and seed morphology building of special winter wheat germplasm huixianhong. *PlOS ONE*, 14(5):e0215066.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243.

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

Herrera-Foessel, S., Singh, R., Huerta-Espino, J., Crossa, J., Yuen, J., and Djurle, A. (2006). Effect of leaf rust on grain yield and yield traits of durum wheats with race-specific and slow-rusting resistance to leaf rust. *Plant Disease*, 90(8):1065–1072.

Herzog, H. and Stamp, P. (1983). Dry matter and nitrogen accumulation in grains at different ear positions in gigas, semidwarf and normal spring wheats. *Euphytica*, 32(2):511–520.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

Hoyle, R. H. and Duvall, J. L. (2004). Determining the number of factors in exploratory and confirmatory factor analysis. *Handbook of quantitative methodology for the social sciences*, pages 301–315.

Hu, X., Liu, J., Zhang, L., Wu, B., Hu, J., Liu, D., and Zheng, Y. (2017). Zn and fe concentration variations of grain and flag leaf and the relationship with nam-g1 gene in triticum timopheevii (zhuk.) zhuk. ssp. timopheevii. *Cereal Research Communications*, 45(3):421–431.

Hussain, W., Baenziger, P. S., Belamkar, V., Guttieri, M. J., Venegas, J. P., Easterly, A., Sallam, A., and Poland, J. (2017). Genotyping-by-sequencing derived high-density linkage

map and its application to qtl mapping of flag leaf traits in bread wheat. *Scientific Reports*, 7(1):16394.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32(4):443–482.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

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

Merkle, E. C. and Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, 85(4):1–30.

Momen, M., Ayatollahi Mehrgardi, A., Amiri Roudbar, M., Kranis, A., Mercuri Pinto, R., 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. *Frontiers in genetics*, 9:455.

Momen, M., Campbell, M. T., Walia, H., and Morota, G. (2019). Utilizing trait networks and structural equation models as tools to interpret multi-trait genome-wide association studies. *Plant methods*, 15(1):107.

Morgounov, A., Abugalieva, A., Akan, K., Akın, B., Baenziger, S., Bhatta, M., Dababat, A. A., Demir, L., Dutbayev, Y., El Bouhssini, M., et al. (2018). High-yielding winter synthetic hexaploid wheats resistant to multiple diseases and pests. *Plant Genetic Resources*, 16(3):273–278.

Morota, G., Jarquin, D., Campbell, M. T., and Iwata, H. (2019). Statistical methods for

the quantitative genetic analysis of high-throughput phenotyping data. *arXiv preprint arXiv:1904.12341*.

Murray, D. G. and Murray, G. M. (2005). *Stripe rust: Understanding the disease in wheat*. NSW Department of Primary Industries.

Peñagaricano, F., Valente, B., Steibel, J., Bates, R., Ernst, C., Khatib, H., and Rosa, G. (2015). Searching for causal networks involving latent variables in complex traits: Application to growth, carcass, and meat quality traits in pigs. *Journal of Animal Science*, 93(10):4617–4623.

Peterson, R. F., Campbell, A., and Hannah, A. (1948). A diagrammatic scale for estimating rust intensity on leaves and stems of cereals. *Canadian Journal of Research*, 26(5):496–500.

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.8.12.

Rocha, J. R. d. A. S. d. C., Machado, J. C., and Carneiro, P. C. S. (2018). Multitrait index based on factor analysis and ideotype-design: proposal and application on elephant grass breeding for bioenergy. *GCB Bioenergy*, 10(1):52–60.

Runcie, D. E. and Mukherjee, S. (2013). Dissecting high-dimensional phenotypes with bayesian sparse factor analysis of genetic covariance matrices. *Genetics*, 194(3):753–767.

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*.

SHI, R.-l., TONG, Y.-p., JING, R.-l., ZHANG, F.-s., and ZOU, C.-q. (2013). Characterization of quantitative trait loci for grain minerals in hexaploid wheat (triticum aestivum l.). *Journal of Integrative Agriculture*, 12(9):1512–1521.

Sperotto, R., Ricachenevsky, F., de A Waldow, V., Müller, A., Dressler, V., and Fett, J. (2013). Rice grain fe, mn and zn accumulation: How important are flag leaves and seed number? *Plant, Soil and Environment*, 59(6):262–266.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

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.

Valente, B. D., Morota, G., Peñagaricano, F., Gianola, D., Weigel, K., and Rosa, G. J. (2015). The causal meaning of genomic predictors and how it affects construction and comparison of genome-enabled selection models. *Genetics*, 200(2):483–494.

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.

Watanabe, K., Guo, W., Arai, K., Takanashi, H., Kajiya-Kanegae, H., Kobayashi, M., Yano, K., Tokunaga, T., Fujiwara, T., Tsutsumi, N., et al. (2017). High-throughput phenotyping of sorghum plant height using an unmanned aerial vehicle and its application to genomic prediction modeling. *Frontiers in plant science*, 8:421.

Ye, X., Li, J., Cheng, Y., Yao, F., Long, L., Wang, Y., Wu, Y., Li, J., Wang, J., Jiang, Q.,

et al. (2019). Genome-wide association study reveals new loci for yield-related traits in sichuan wheat germplasm under stripe rust stress. *BMC Genomics*, 20(1):1–17.

Yu, H., Campbell, M. T., Zhang, Q., Walia, H., and Morota, G. (2019). Genomic Bayesian confirmatory factor analysis and bayesian network to characterize a wide spectrum of rice phenotypes. *G3: Genes, Genomes, Genetics*, 9(6):1975–1986.

Yu, H., Morota, G., Celestino, E. F., Dahlen, C. R., Wagner, S. A., Riley, D. G., and Hanna, L. L. H. (2020). Deciphering cattle temperament measures derived from a four-platform standing scale using genetic factor analytic modeling. *Frontiers in Genetics*, 11:599.

Tables

Table 1: Factor loading values from the Bayesian confirmatory factor analysis. PSD: posterior standard deviation, PSRF: potential scale reduction factor, GYL: grain yield, ARC: plant architecture, FL: flag and leaf, MIN: mineral-related traits, YRD: yellow rust, SRDK: stem rust at Kastamonu, SRDH: stem rust at Haymana, LRD: leaf rust, and R^2: coefficient of determination.

Latent factor	Phenotype	Loading	PSD	PSRF	R^2
GYL:					
	grain yield	0.998	0.071	1.000	0.996
	harvest index	0.571	0.090	1.000	0.327
	biomass weight	0.823	0.081	1.000	0.677
FL:					
	flag leaf length	0.849	0.080	1.002	0.720
	flag leaf width	0.771	0.082	1.002	0.594
	flag leaf area	0.999	0.071	1.005	0.998
ARC:					
	fertile spikelet	0.867	0.098	1.006	0.752
	spike length	0.543	0.097	1.001	0.295
	spikelet number	0.776	0.099	1.005	0.602
	seeds per spike	0.796	0.088	1.001	0.633
	spike weight	0.740	0.110	1.003	0.548
	grain weight per spike	0.854	0.108	1.005	0.730
	spike harvest index	0.462	0.107	1.001	0.214
MIN:					
	arsenic	0.483	0.098	1.001	0.234
	calcium	0.884	0.086	1.005	0.782
	cadmium	0.767	0.091	1.003	0.588
	colbalt	0.468	0.101	1.001	0.219
	copper	0.940	0.083	1.005	0.883
	iron	0.927	0.084	1.005	0.858
	potassium	0.773	0.091	1.003	0.598
	lithium	0.379	0.102	1.000	0.144
	magnesium	0.984	0.078	1.007	0.968
	manganese	0.928	0.084	1.006	0.861
	molybdenum	0.757	0.091	1.002	0.573
	nickel	0.531	0.098	1.001	0.282
	phosphorous	0.974	0.080	1.007	0.949
	sulphur	0.750	0.091	1.002	0.563
	titanium	0.365	0.100	1.000	0.133
	zinc	0.817	0.089	1.003	0.667
LRD:					
	leaf rust severity	0.996	0.071	1.016	0.992
	leaf rust infection type	0.813	0.081	1.006	0.662
	leaf rust coefficient of infection	0.998	0.071	1.015	0.997
SRDH:					
	stem rust severity at Haymana	0.955	0.074	1.002	0.912
	stem rust infection type at Haymana	0.872	0.078	1.001	0.760
	stem rust coefficient of infection at Haymana	0.998	0.071	1.003	0.997
SRDK:					
	stem rust severity at Kastamanu	0.968	0.073	1.012	0.937
	stem rust infection type at Kastamonu	0.912	0.076	1.009	0.832
	stem rust coefficient of infection at Kastamonu	0.991	0.071	1.012	0.982
YRD:					
	yellow rust coefficient of infection at Haymana	0.999	0.071	1.002	0.999
	yellow rust infection type at Haymana	0.824	0.080	1.001	0.680
	yellow rust severity at Haymana	0.973	0.073	1.002	0.946

Table 2: Bayesian information criterion (BIC) scores for pairs of nodes reporting the change in the score caused by an arc removal relative to the entire network score. Tabu: Tabu Search, MMHC: Max-Min Hill-Climbing, GYL: grain yield traits, FL: flag and leaf traits, MIN: mineral traits, ARC: architecture traits, LRD: leaf rust disease, SRDH: steam rust disease at Haymana, and YRD: yellow rust disease.

Algorithm	from	to	BIC
Tabu	FL	MIN	-2.074
	FL	LRD	-9.648
	MIN	ARC	-5.884
	SRDH	GYL	-32.297
	YRD	GYL	-16.399
	YRD	ARC	-5.916
MMHC	FL	LRD	-5.989
	SRDH	GYL	-32.297
	YRD	GYL	-16.3997

Figures

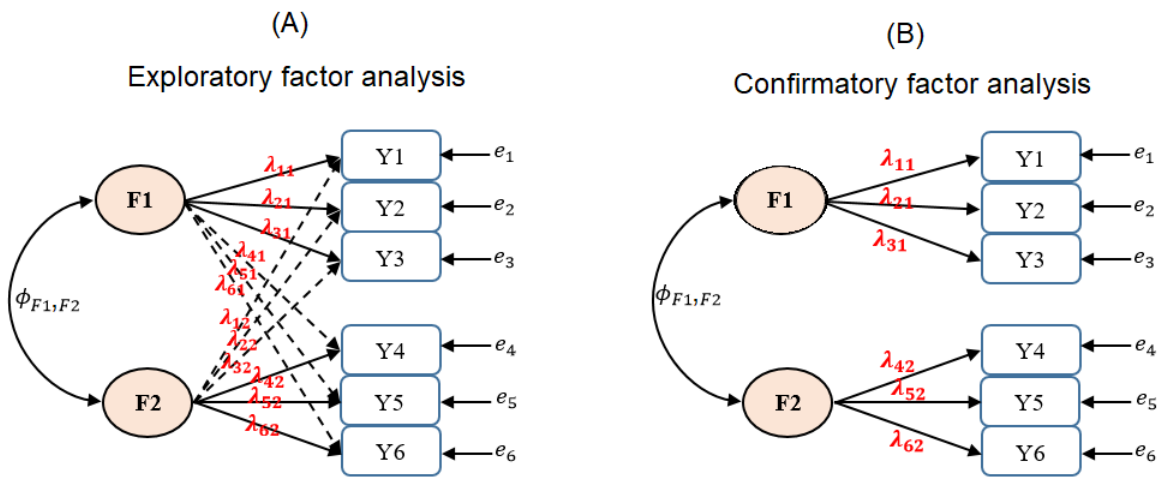

Figure 1: A graphical representation of exploratory factor analysis (panel A) and confirmatory factor analysis (panel B) assuming that there are hypothetical six observed phenotypes ($\mathrm{Y1}, \mathrm{Y2}, \cdots, \mathrm{Y6}$) and two unobserved latent factors (F1 and F2). The double headed arrow is the covariance between the two latent factors ($\Phi_{F1,F2}$). $e_1, e_2, \ldots, e_6$ represent the residuals. Exploratory factor analysis estimates the phenotype-factor relationship from the data by allowing cross-loading. By choosing the largest factor loading value for each phenotype, phenotypes can be uniquely assigned to one of the two factors. In this example, Y1, Y2, and Y3 loaded on the F1 (with loadings of λ_{11}, λ_{21}, and λ_{31}) and Y4, Y5, and Y6 loaded on F2 (with loadings of λ_{42}, λ_{52}, and λ_{62}). Confirmatory factor analysis assumes that this relationship is known *a priori*.

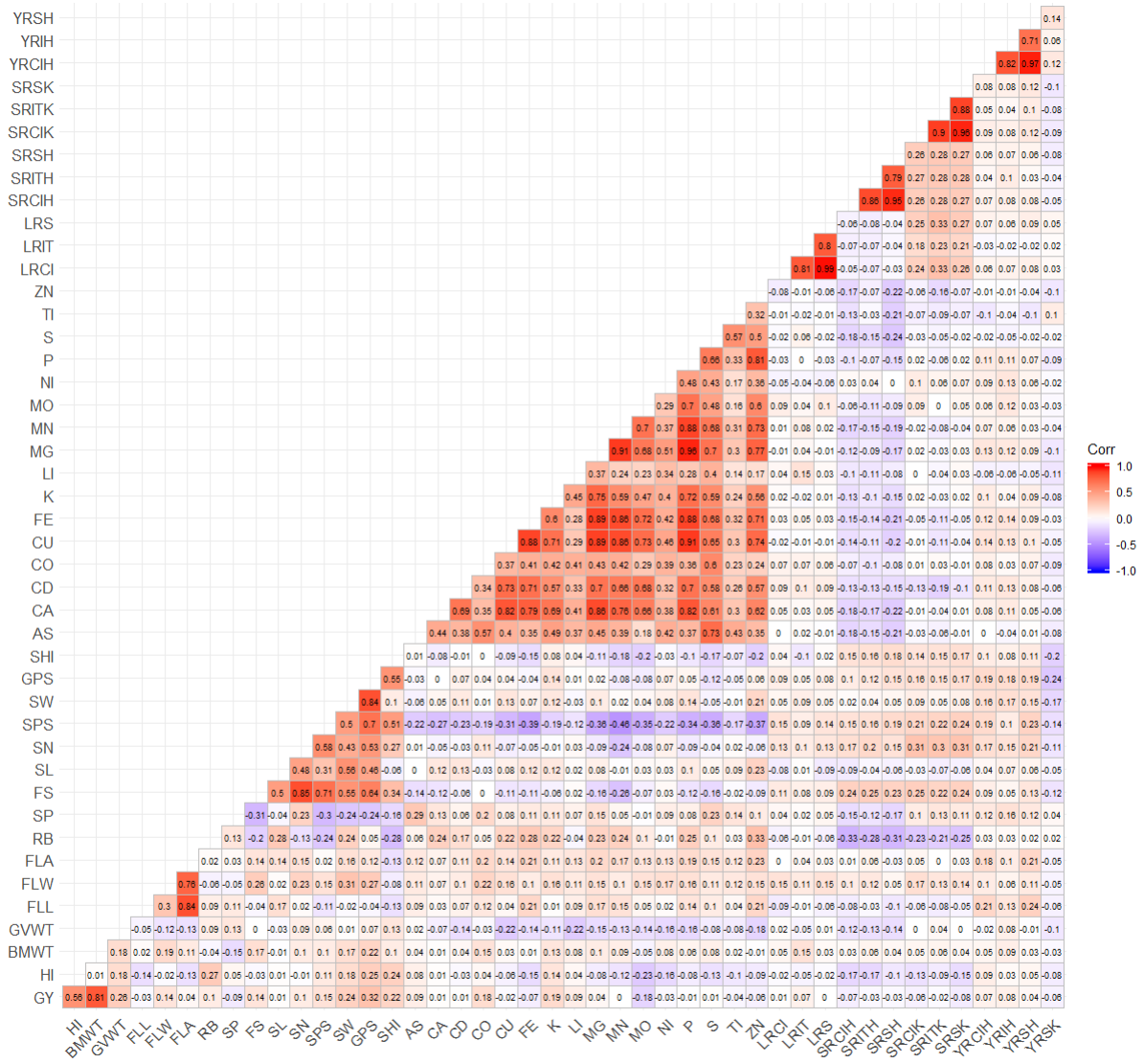

Figure 2: Pairwise Pearson's correlations between 45 phenotypes. GY: grain yield, HI: harvest index, BWT: biomass weight, GVWT: grain volume weight, FLL: flag leaf length, FLW: flag leaf width, FLA: flag leaf area, SL: spike length, SN: spikelet number,SP: sterile spikelet, FS: fertile spikelet, RB: rachis break, SPS: seeds per spike, SW: spike weight, GPS: grain weight per spike, SHI: spike harvest index, AS: arsenic, CA: calcium, CD: cadmium, CO: colbalt, CU: copper, FE: iron, K: potassium, LI: lithium, MG: magnesium, MN: manganese, MO: molybdenum, NI: nickel, P: phosphorous, S: sulphur, TI: titanium, ZN: zinc, LRCI: leaf rust coefficient of infection, LRIT: leaf rust infection type, LRS: leaf rust severity, SRCIH: steam rust coefficient of infection at Haymana, SRITH: stem rust infection type at Haymana, SRSH: stem rust severity at Haymana, SRCIK: stem rust coefficient of infection at Kastamonu, SRITK: stem rust infection type at Kastamonu, SRSK: stem rust severity at Kastamanu, YRCIH: yellow rust coefficient of infection at Haymana, YRIH: yellow rust infection type at Haymana, YRSH: yellow rust severity at Haymana, YRSK: yellow rust severity at Kastamonu.

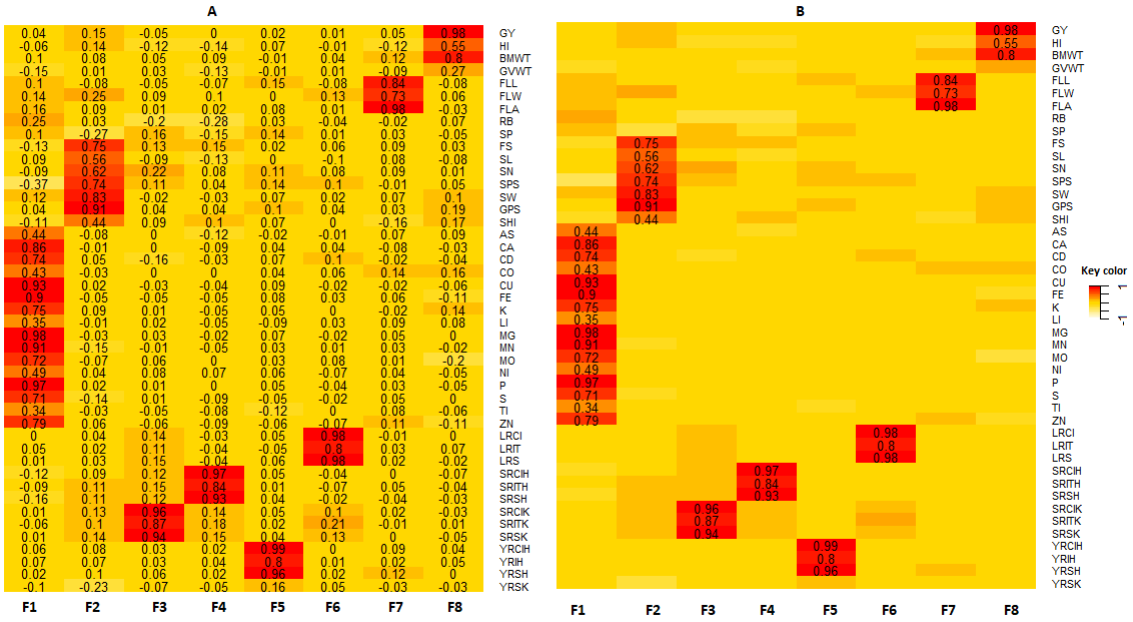

Figure 3: Panel A: heat map of factor loading values. Panel B: heat map of factor loading values after removing cross-loading by setting a cut-off value of $|\lambda| > 0.30$. The rows of each panel correspond to the observed phenotypes and the columns correspond to the eight factors (F1 to F8). Abbreviations of observed phenotypes are shown in Figure 2.

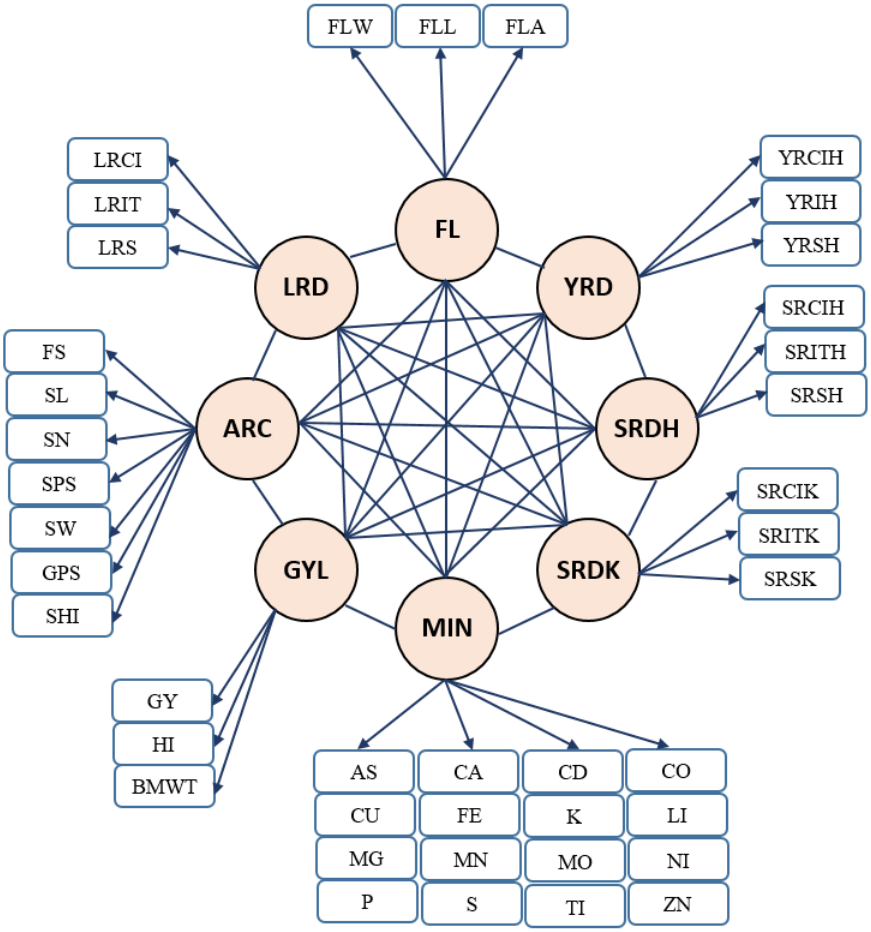

Figure 4: Relationship between eight latent variables and observed phenotypes based on exploratory factor analysis. GYL: grain yield related traits, ARC: architecture related trait, FL: flag and leaf related traits, MIN: mineral-related traits, YRD: yellow rust related traits, SRDK: stem rust related traits at Kastamonu, SRDH: stem rust related traits at Haymana, LRD: leaf rust related traits. The eight latent factors were assumed to be correlated. Abbreviations of observed phenotypes are shown in Figure 2.

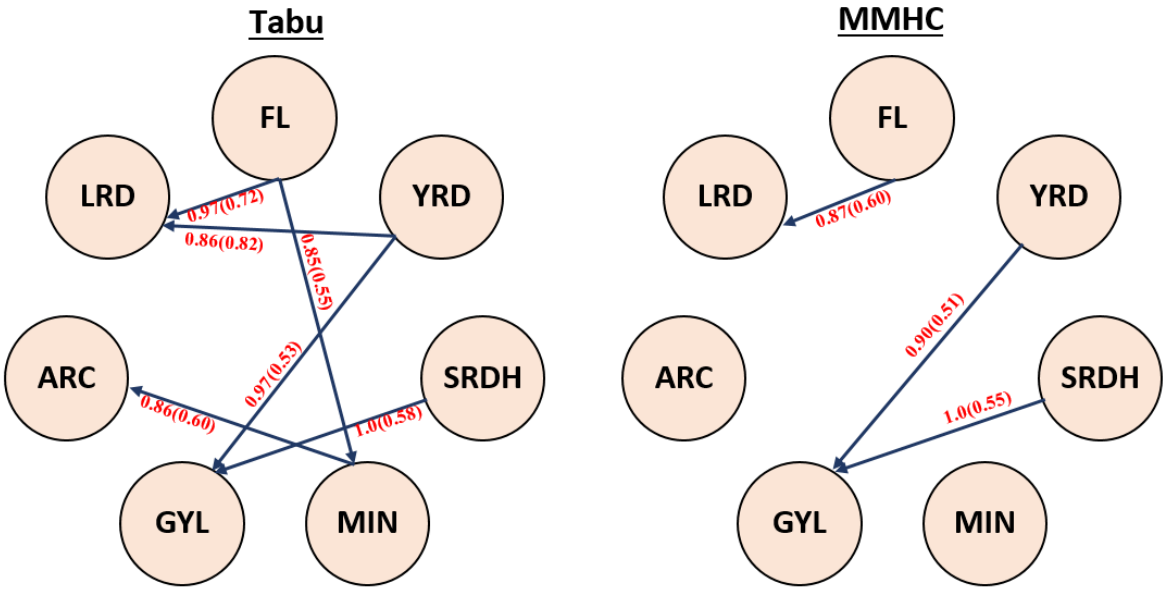

Figure 5: Bayesian networks learned from Tabu search (Tabu) and Max-Min Hill-Climbing (MMHC). Structure learning test was performed with 5,000 bootstrap samples. Labels of the edges refer to the strength and direction (parenthesis) which measure the confidence of the directed edge. The strength indicates the frequency of the edge is present and the direction measures the frequency of the direction conditioned on the presence of edge. GYL: grain yield related traits, ARC: architecture related trait, FL: flag and leaf related traits, MIN: mineral-related traits, YRD: yellow rust related traits, SRDK: stem rust related traits at Kastamonu, SRDH: stem rust related traits at Haymana, LRD: leaf rust related traits.