Igolkina *et al. BMC Genomics* 2020, **21**(Suppl 8):490
https://doi.org/10.1186/s12864-020-06833-2

BMC Genomics

RESEARCH

Open Access

Multi-trait multi-locus SEM model discriminates SNPs of different effects

Anna A. Igolkina[1*], Georgy Meshcheryakov[1], Maria V. Gretsova[1,2], Sergey V. Nuzhdin[1,3] and Maria G. Samsonova[1*]

Abstract

Background: There is a plethora of methods for genome-wide association studies. However, only a few of them may be classified as multi-trait and multi-locus, i.e. consider the influence of multiple genetic variants to several correlated phenotypes.

Results: We propose a multi-trait multi-locus model which employs structural equation modeling (SEM) to describe complex associations between SNPs and traits - **m**ulti-**t**rait **m**ulti-**l**ocus **SEM** (mtmlSEM). The structure of our model makes it possible to discriminate pleiotropic and single-trait SNPs of direct and indirect effect. We also propose an automatic procedure to construct the model using factor analysis and the maximum likelihood method. For estimating a large number of parameters in the model, we performed Bayesian inference and implemented Gibbs sampling. An important feature of the model is that it correctly copes with non-normally distributed variables, such as some traits and variants.

Conclusions: We applied the model to Vavilov's collection of 404 chickpea (*Cicer arietinum L.)* accessions with 20-fold cross-validation. We analyzed 16 phenotypic traits which we organized into five groups and found around 230 SNPs associated with traits, 60 of which were of pleiotropic effect. The model demonstrated high accuracy in predicting trait values.

Keywords: GWAS, SEM, Multi-trait multi-locus SEM, Bayesian inference, Chickpea

Background

Understanding how genetic variation translates into phenotypic effects is one of the central challenges facing fundamental biology, agriculture, and medicine. Solutions of this problem fall into two main classes: association studies and trait prediction studies. Genome-wide association studies (GWAS) are designed to identify genetic variants associated with a trait. Initially, GWAS was conducted for each trait separately testing SNPs one by one. However, single-locus approaches may lead to biased estimates due to multiple testing correction, and they are not suitable in the common case of genetically correlated traits.

To alleviate the latter challenge, multi-trait models have been proposed [1, 2]. One way to cope with correlated traits is to model the inter-trait covariance as a random effect in linear mixed effects models [3]. Until recently, this model could use only a pair of correlated traits at a time due to the computational intensity [4]. To avoid this complexity, variable reduction techniques were suggested to replace several phenotypic traits with new independent constructs. These constructs play the role of new traits and can be obtained with a standard principal component analysis of traits (PCA), various principal components of heritability (PCH) [5–7] or pseudo-principal components [8]; however, the biological interpretation of these artificial traits is not clear. Moreover, these methods do not distinguish trait-specific and pleiotropic variants. To carry this out, meta-analysis

* Correspondence: igolkinaanna11@gmail.com; msamsonova@spbstu.ru
[1]Peter the Great Saint-Petersburg Polytechnic University, Russian Federation, Polytechnicheskaya, 29, St. Petersburg 195251, Russia
Full list of author information is available at the end of the article

combining several single-trait GWAS of different traits was proposed [9]. It can derive trait-specific variants, but, as correlated traits were not analyzed simultaneously, this method is not multi-trait by definition.

Another challenge in association studies is to develop a powerful multi-locus model. Single-locus models require correction for multiple testing, which dramatically reduces power. To avoid this problem, multi-locus models that consider all markers simultaneously have been proposed. Due to the 'large p (number of SNPs), small n (sample size)' problem, many multi-locus models are based on regularization/penalized techniques: LASSO [10], Elastic Net [11], Bayesian LASSO [12], adaptive mixed LASSO [13]. Other multi-locus methods, which are incorporated in the mrMLM package, involve a two-step algorithm which first selects candidate variants from a single-locus design and then examines them together in a multi-locus manner [14]. Despite their diversity, the multi-locus models are limited in multi-trait cases and seldom pay attention to different types of SNP effects (e.g. pleiotropic, single-trait, direct, indirect).

In contrast to GWAS, the second broad class of studies make genome-wide trait predictions. These studies have gained popularity and enjoy practical application in agriculture, specifically, in estimating individual breeding values and selecting breeding lines [15]. Genomic prediction methods not only search for trait-variant associations but also validate them by demonstrating their predictive ability. Similar to GWAS, these methods are based on various regression models that typically include multiple loci and consider kin relationships between individuals. The latter is usually treated as the random effect, i.e. the multivariate normally distributed variable with zero mean and a covariance matrix proportional to pedigree-based or marker-based kinship [16]. The random effect can be estimated together with marker effects as in BLUP and various GWAS mixed-models [17–19] or before the association analysis as in GRAMMAR [20].

Despite the broad spectrum of multi-trait and multi-locus models in GWAS and trait prediction studies, only a few of them simultaneously incorporate correlated traits and several associated variants [21–25]. In principle, multi-trait and multi-locus models have the potential to reveal complex and important types of associations; for instance, a single variant might have a direct effect on one trait and an indirect impact on the other trait, may act on a single trait or its effect might be pleiotropic affecting several traits. However, none of these traits-variants associations are explicitly embedded into known models. This is why it is tempting to have these relationships described explicitly, as in structural equation models.

Structural equation modeling (SEM) is a multivariate statistical analysis technique first introduced for path analysis by geneticist Sewell Wright [26, 27]. Once predominantly used in genetics, econometric, and sociology, SEM applications have gradually shifted to the field of molecular biology [28]. For example, SEM has been used to explore alterations in gene networks in diseases [29, 30], to provide a quantitative map of relationships between traits and disease [31], and to infer gene regulatory networks involving several hundred genes and eQTLs [32, 33].

SEM models have also been applied in association studies in both multi-trait and multi-locus designs. For example, the GW-SEM method has been developed to test the association of a SNP with multiple phenotypes through a latent construct [34]. In comparison with the existing multi-trait single-locus GWAS software package GEMMA (Zhou and Stephens 2014), GW-SEM provides more accurate estimates of associations; however, GEMMA is almost three times faster than GW-SEM. Another SEM-based model which can be used in association studies has been proposed for multi-trait QTL mapping [35]. This method assumes that phenotypes are causally related forming a core structure without latent constructs, and QTLs play the role of exogenous variable to the structure. This approach allows the model to decompose QTL effects into direct, indirect, and total effects. However, the assumption of causally related traits is limiting because the correlation between traits can additionally be caused by pleiotropy rather than the direct influence of traits on each other. Therefore, the current SEM-based models for genotype-phenotype associations can be improved to address these drawbacks.

Here, we propose a new multi-trait multi-locus SEM-based model – **mtmvSEM** – that considers both correlated traits joined into latent constructs, which can be causally related to each other, and multiple SNPs influencing both traits and latent variables. In contrast to PCA-based approaches, our model does not operate with artificial phenotypes in the form of linear combinations of traits, but rather the phenotypes are regressed on the latent constructs. The proposed configuration of the model distinguishes pleiotropic and single-trait effects of SNPs on latent variables and phenotypes, respectively. Moreover, SNP effects can be differentiated between direct and indirect. This explicit separation of SNP roles may provide a better understanding of genetic mechanisms underlying a trait than other multi-trait multi-locus models.

Our approach faces several challenges. First, in case of a large number of traits and variants, the model potentially belongs to the "large p, small n" class, so that the standard maximum likelihood (ML) method for estimating parameters in SEM models is limited due to the parameter identification criteria. This problem can be solved by applying the Bayesian approach, which uses prior information about model parameters. Bayesian multiple-regression methods are widely used for genomic prediction in agriculture and in GWAS [36] reducing the number of tests, and consequently, increasing robustness and power as compared to standard GWAS analyses [37]. In our model, we performed Bayesian inference and obtained posterior distributions of parameters by Gibbs sampling, a Markov chain Monte Carlo (MCMC) algorithm.

Another challenge in our model is the inclusion of both continuous and ordinal variables given that variants and many phenotypes are measured on ordinal scales. As a result, it is impossible to estimate parameters in SEM models using statistical models relying on the normality assumption. These limitations explain the sparsity of studies conducting SEM analyses in a genome-wide context. In our model, we incorporated techniques to cope with ordinal data – polychoric and polyserial correlations – that provide a correct analysis of genetic variants and traits.

Our model was applied to a dataset of 404 chickpea landraces analyzed recently [38]. Chickpea is the second most widely grown food legume, providing a vital source of nutritional nitrogen for ~ 15% of the world's population. To accelerate chickpea breeding, it is important to identify regions controlling agronomically important traits. However, while performing GWAS, we found that 16 out of 30 phenotypic traits considered were correlated. Therefore, to obtain statistically reliable markers and to understand the causal relationships between traits and variants, the mtmlSEM model developed here was applied to this dataset. We also used the model to predict chickpea phenotypic traits and got sufficiently good results for most of them.

Results

Application of mtmlSEM model to chickpea dataset

To test whether the relations between latent factors in the model are reasonable and to evaluate impacts of different types of SNPs, we compared four types of models (Fig. 1). We denote a model having parameters in the B matrix as *connected* and a model without a B matrix as *zero*. We denote a model without the K matrix as *base* and a model having parameters in the K matrix as *extended*. Four model configurations were considered covering all possible combinations (Fig. 1).

For each of the four models, we assessed its predictive ability with the fixed 20-fold cross-validation. In each of the 20 training sets, we automatically obtained the same set of 5 factors influencing 16 partly correlated phenotypes (Table 1, Additional File 1). The first two factors reflect different types of productivity traits. The third factor reflects joint variation in the color of different plant parts. The fourth can be interpreted as a phenological factor. The fifth reflects joint variation of traits

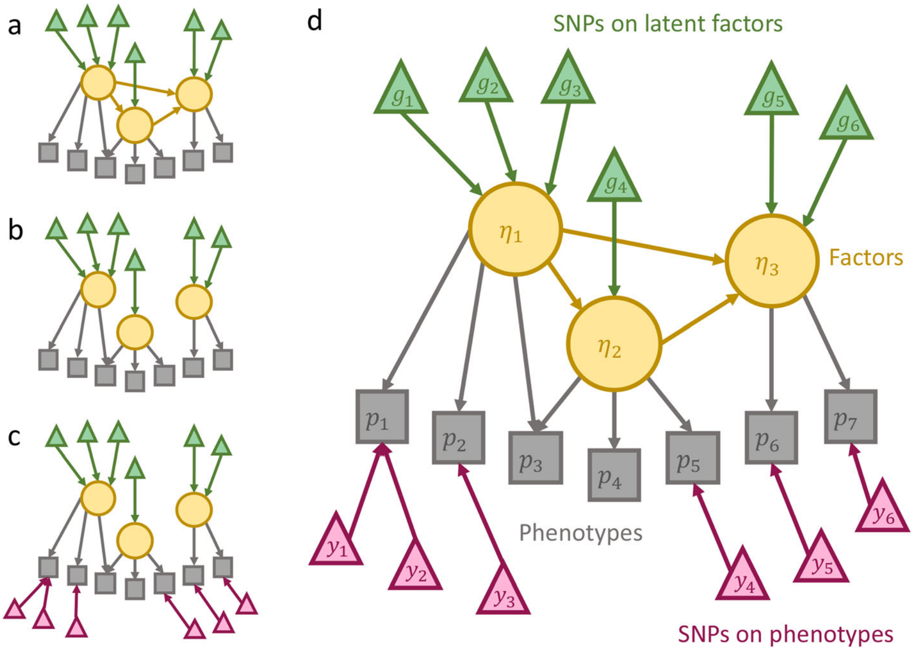

Fig. 1 Examples of the genome-wide multi-trait SEM model. **a** Connected base model; (**b**) Zero base model; (**c**) Zero extended model; (**d**) Connected extended model

Table 1 5 factors influencing 16 partly correlated phenotypes

Factor	Attributed phenotypes	Description
1	NoPodsWeight	Plant weight without pods
	PodsWeight	Pods weigth
	PodsNumber	Number of pods per plant
	SeedsNumber	Number of seeds per plant
	SeedsWeight	Seeds weight per plant
2	PodLength	Pod length
	PodWidth	Pod width
	Seed1000W	Thousand seeds weight
3	FloCol	Flower colour
	StemCol	Stem colour
	FlowStemCol	Peduncle colour
	SeedCol	Seed colour
4	BegFEndF	Days from beginning of flowering to end of flowering
	EndFBegM	Days from end of flowering to beginning of maturation
5	Height	Plant height
	Hlp	Height of lower pod attachment

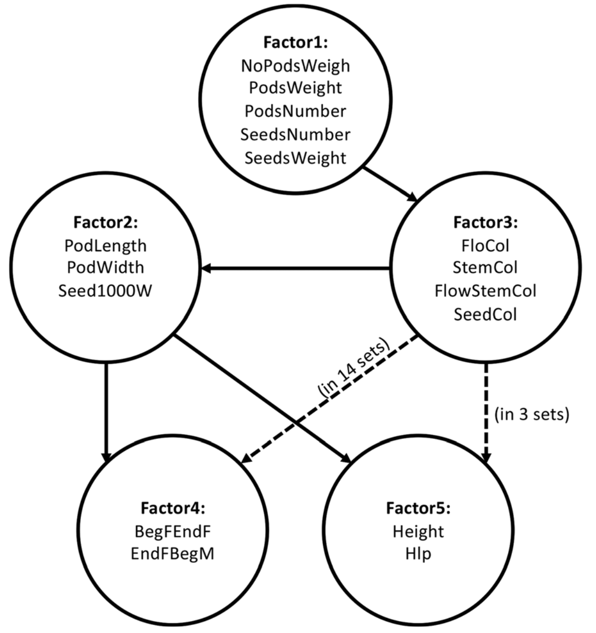

Fig. 2 Latent factors joined to form structural part of connected SEM model. Dashed arrows represent relationships, which were not present is all training sets for directed acyclic graph obtained; Solid lanes represent relationships, which were found in each of 20 training sets

related to plant architecture, in particular, plant height and height of the lover pod attachment.

In the connected model, the latent factors were joined into a directed acyclic graph and this procedure resulted in slightly different structural parts for the 20 training set models. We found that the number of connections between latent variables varied from four to six with four being common to all training sets (Fig. 2). From the statistical viewpoint, relationships between latent variables reflect their common variances that maximize the likelihood of the sample covariance matrix subject to parameters of the model. However, a biological interpretation of the connections may be that the relationships between factors related to productivity and plant color reflect selection on market class: desi chickpeas have a small dark seed, while kabuli have large lightly colored seeds [39]. Relations between productivity and phenology as well as between productivity and plant architecture are also apparent: plant productivity reflects the efficiency of plant metabolism that obviously influences plant architecture and phenology [40].

We first added SNPs influencing the latent factors to obtain both the connected and zero base models. The number of SNPs in the connected base models constructed for 20 training sets varied from 52 to 62; for zero base models, this number was in the range from 36 to 46. The larger number of SNPs in connected models as compared with zero models can be explained by the essential difference between SNPs attributed to these model types. In connected base models, some SNPs are associated with several latent factors and therefore affect a larger number of phenotypic traits than in zero models. Therefore, in connected models, SNPs describe a more complex variance-covariance structure and, as a result, a larger number of SNPs is required to estimate it.

Notably, SNPs influencing latent factors do not explain the variances specific to individual phenotypic traits. To take into account these variances, we built extended models for each training set. The number of SNPs in connected extended models varied from 223 to 256; in zero extended models, this number was in the range from 218 to 242. The significant increase in the number of SNPs in extended models as compared with base models can be explained by the fact that extended models additionally consider around ten SNPs per each of the 16 traits on average.

To obtain parameter estimates for each of the 80 models (4 model types and 20 training sets), we performed five Gibbs sampling chains of length 2000 and checked several diagnostics with tools in the *coda* CRAN package. The Gelman-Rubin diagnostics was higher than 1.05 in only 1% of all parameters. The minimum effective sample size for a parameter was 83 and the mean and median effective sample sizes across all parameters and models were 3193 and 3304, respectively. Based on these diagnostic values, we concluded that there was good convergence of the Gibbs sampling chains and took parameter estimates for testing.

For all model types, the accuracy of trait prediction is good for plant height, some traits related to productivity, and all traits related to plant color (Table 2, Additional File 2). Closer inspection of the table showed that the connected base model outperformed the zero base model for 9 phenotypic traits, the opposite situation was observed for 5 traits, and predictions for the remaining 2 traits were nearly equal. When comparing the connected and zero extended models, the number of times one model outperforms the other is nearly equal (Table 2) and the number of predictions with equal accuracy increases pointing to greater similarity between these models.

Next, we analyzed positions of trait-associated SNPs on the chromosomes in both connected and zero extended model types. For each of these types, we had independently built 20 models due to the fixed 20-fold CV, and, consequently, the sets of SNPs included into the models were different. To evaluate the congruence between chromosomal positions of SNPs from different sets, we applied the sliding window technique (500 kb window size with 100 kb step) and, for each window, we counted the number of models having at least one SNP in it. We applied this technique for five subsets of SNPs separately, such that each subset was associated with a factor and its attributed phenotypes. We visualized the evaluated congruence between 20 models in Fig. 3. We found that the models agree with each other due to the significant amount of windows, where all models have SNPs. We next compared positions of peaks with GWAS-hits obtained by a single-trait, single-locus model for the chickpea dataset [38]. Utilizing the permutation test, we found that positions of the GWAS-hits and the peaks are not independent (p-value < 0.05) indicating that there is some concordance between our models and GWAS analysis. In Fig. 3, some GWAS hits do not have any matches with peaks, because our model does not include correlated SNPs, which naturally occur in GWAS results. Moreover, our model describes essentially more information than single-trait GWAS; therefore, some peaks do not match any GWAS hits.

Table 2 Accuracy of trait prediction for four models (Pearson correlation between actual values and predicted and coefficient of determination). Bold font: connected model outperforms zero model; Italic font - prediction accuracies of connected and zero models are nearly equal

Model type	Connected		Zero		Connected		Zero	
SNPs influence	extended		extended		base		base	
Measure	corr, r	r^2	corr, r	r^2	corr, r	r^2	corr, r	r^2
Seed1000W	*0.75*	*0.56*	*0.75*	*0.56*	*0.75*	*0.56*	*0.75*	*0.57*
FloCol	0.69	0.48	0.71	0.50	**0.65**	**0.42**	**0.64**	**0.42**
FlowStemCol	**0.68**	**0.46**	**0.67**	**0.45**	**0.67**	**0.45**	**0.66**	**0.44**
SeedCol	0.67	0.45	0.68	0.46	0.60	0.36	0.63	0.39
NoPodsWeight	*0.67*	*0.45*	*0.67*	*0.45*	**0.59**	**0.34**	**0.57**	**0.33**
PodLength	0.67	0.44	0.69	0.47	**0.65**	**0.42**	**0.64**	**0.41**
StemCol	0.64	0.42	0.67	0.45	0.65	0.43	0.67	0.45
PodWidth	0.63	0.40	0.66	0.44	0.64	0.40	0.67	0.44
Height	*0.59*	*0.34*	*0.59*	*0.35*	**0.57**	**0.33**	**0.49**	**0.24**
PodsWeight	**0.45**	**0.20**	**0.42**	**0.18**	**0.44**	**0.19**	**0.42**	**0.18**
SeedsWeight	**0.38**	**0.14**	**0.36**	**0.13**	**0.38**	**0.15**	**0.37**	**0.14**
SeedsNumber	**0.36**	**0.13**	**0.32**	**0.10**	*0.15*	*0.02*	*0.15*	*0.02*
EndFBegM	*0.33*	*0.11*	*0.33*	*0.11*	**0.35**	**0.12**	**0.32**	**0.10**
Hlp	0.32	0.10	0.35	0.12	0.31	0.10	0.35	0.12
BegFEndF	**0.30**	**0.09**	**0.28**	**0.08**	0.27	0.07	0.28	0.08
PodsNumber	**0.30**	**0.09**	**0.27**	**0.07**	**0.26**	**0.07**	**0.25**	**0.06**

Discussion

GWAS often relies on data with a number of highly correlated phenotypic traits. Due to these correlations, significant SNPs are frequently associated with several phenotypes, i.e., they are pleiotropic. Until recently, multi-trait multi-locus models could neither distinguish SNP effects between pleiotropic and single-trait ones nor analyze a large number of traits and variants. In a SEM-based model, aggregation of pleiotropic effects into latent constructs makes it possible to distinguish SNP effects and, therefore, shed more light on mechanisms underlying associations. Large numbers of SNPs and traits in the model can lead to a parameter identification problem that, nevertheless, can be solved by applying Bayesian approach for parameter estimation.

Here we developed the mtmlSEM (multi-trait multi-locus SEM) model that estimates and evaluates casual relations between phenotypes and SNPs, reliably discriminates variant effects between single-trait and pleiotropic ones, and has good predictive ability. The developed model is a general one and can be applied to analysis of associations between variants and correlated traits in any dataset. It consists of two main steps. Firstly, the structure of the model is automatically constructed, such that correlated traits are joined into latent factors and explanatory SNPs are introduced to latent factors and phenotypic traits directly. Under this paradigm, one could consider latent factors as aggregating yet unknown biological processes that explain the SNP influence on phenotypes. At the second step, the parameter estimates are obtained

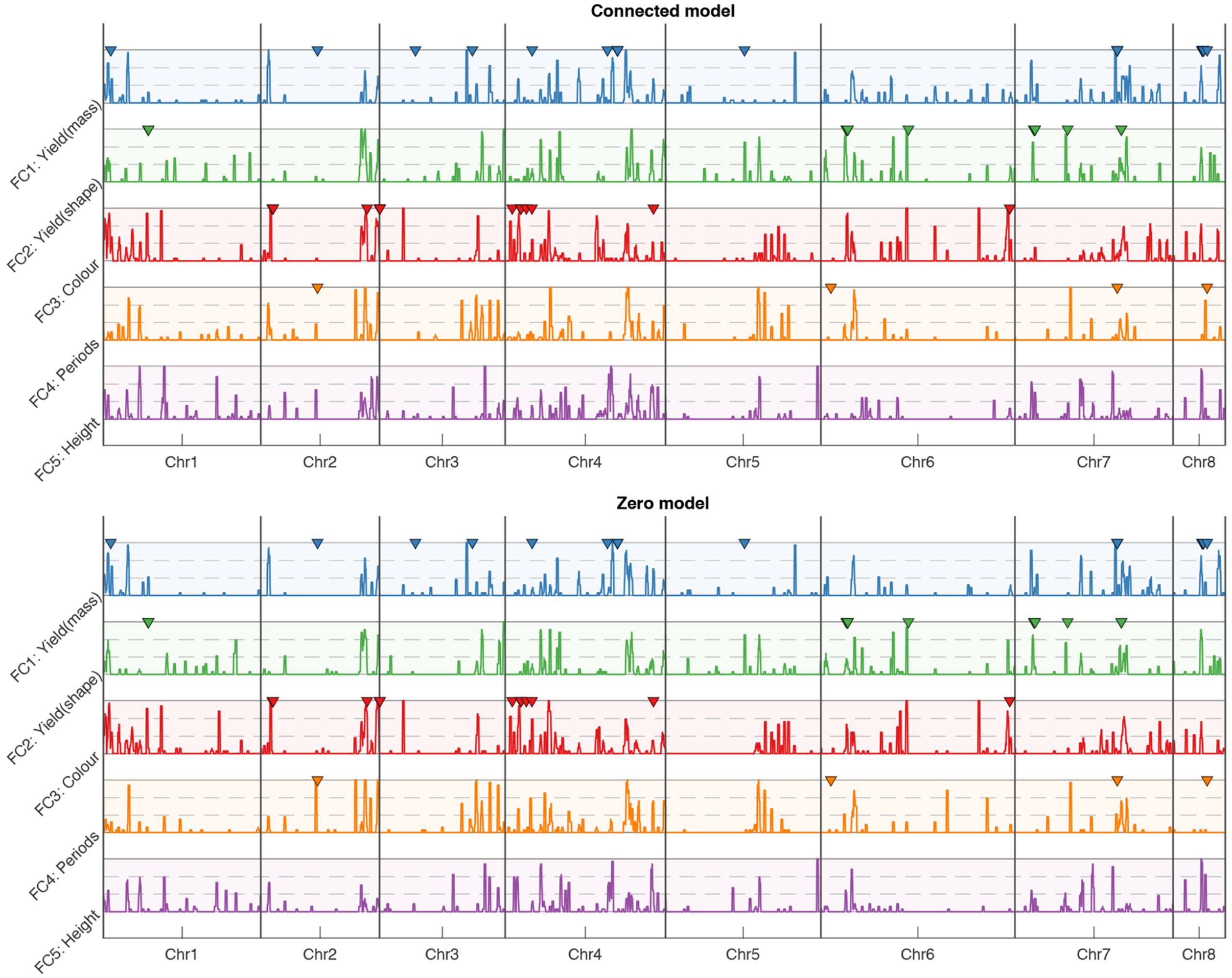

Fig. 3 The sliding-window congruence between models obtained in 20-fold cross validation. The hight of a peak reflects the number of models having at least one SNPs within the window corresponding to the peak

with MCMC (Gibbs sampling) after the Bayesian inference of posterior distributions for parameters.

At the next step, the applicability of the mtmlSEM model was illustrated on a dataset of chickpea accessions. Many phenotypic traits in this dataset are correlated and therefore single-trait GWAS inferences can be biased. We compared four models: *zero* or *connected* means inclusion or not parameters in B, *base* or *extended* means inclusion or not parameters in K. To estimate model accuracy, we applied the 20-fold cross-validation, which led to construction of 20 different models for each model type.

After the accuracy of trait prediction was assessed, it became evident that among base models, connected ones describe the covariance structure of the data more accurately and, therefore, showed better predictive ability than the zero models. Therefore, one may conclude that joining latent factors into a structure was reasonable as all phenotypes are mutually dependent and cannot be considered as isolated blocks of traits.

In the case of extended models, the supplementary SNPs added to phenotypes described the residual variance not covered by the base models, so that the connected and zero extended models were comparable in both total numbers of SNPs and accuracy.

We next tested the utility of the models to predict associations between SNPs and phenotypes. We found that in that base and connected extended models behave similarly supporting their resemblance to one another. The associations revealed with mtmlSEM model and in standard GWAS analysis are consistent and the differences observed arise due to exclusion of correlated SNPs from the mtmlSEM models, and because mtmlSEM models consider individual and pleiotropic effects of SNPs separately. These effects could be singled out by calculating the difference between SNP effects in extended and zero models. However, the pleiotropic SNP effects are central to trait prediction in the models since the addition of SNPs to traits does not result in marked increase of prediction accuracy (see Table 2).

Conclusions

We developed the mtmlSEM model that describes casual relations between between single-trait and pleiotropic SNPs and phenotypic traits. The particular strength of mtmlSEM model developed here is its ability to predict traits from genomic data. Notably, while the chickpea dataset used in this study is relatively small, the accuracy of the predictions for many traits was good and is comparable or even superior to the accuracy of breeding values predictions in genomic selection models. However, the applicability of mtmlSEM models in genomic selection studies requires further investigation.

Methods

Structural equation modeling

First proposed by S. Wright [26] for path analysis, SEM is defined today as a diverse set of tools and approaches covering regression models, path analysis and confirmatory factor analysis. The first SEM model was LISREL, and it has two distinct parts: structural and measurement [41, 42]. The structural part of LISREL reflects the causal relationships between endogenous and exogenous latent variables; the measurement model describes how latent variables influence their manifest variables:

$$\begin{aligned} \eta &= \mathrm{B}\eta + \varepsilon \\ p &= \Lambda\eta + \delta \end{aligned} \tag{1}$$

where η is a vector of n_η latent factors (both exogeneous and endogenous), p is a vector of n_p observed manifest variables, Λ is a matrix of factor loadings, B is a matrix of relationships between latent factors, $\varepsilon \sim N(0, \Theta_\varepsilon)$ and $\delta \sim N(0, \Theta_\delta)$ are random errors, Θ_ε and Θ_δ are diagonal matrices of sizes (n_η, n_η) and (n_p, n_p), respectively.

To adapt this model for genotype-phenotype studies, we considered p as a vector of phenotypes, and η as a vector of latent variables, which describe the shared variance of genetically correlated traits. One possible interpretation of the measurement part of the model in these terms is that latent variables play the role of molecular mechanisms governing the correlation between traits. The structural part describes the interplay between these mechanisms.

To construct the mtmlSEM model, we extended the LISREL model with observed exogenous variables assuming them as SNPs. New exogenous variables influence either latent factors or phenotypes traits1 and mean pleiotropic and single-trait effects, respectively. As a result, latent variables η become only endogenous and the SEM model is transformed as follows:

$$\begin{aligned} \eta &= \mathrm{B}\eta + \Pi g + \varepsilon \\ p &= \Lambda\eta + \mathrm{K}y + \delta \end{aligned} \tag{2}$$

where g and y are variables of SNPs influencing latent factors and phenotypic traits, respectively; Π and K are matrixes of SNP influences on latent factors and phenotypes, respectively. We assumed that each column of both the Π and K matrices can contain only one cell with a parameter such that each SNP can influence only one variable. SNPs in the structural part, g, describe a part of phenotypic variance, which is common for several traits. However, each phenotype has its own variance, which is described by SNPs in the measurement part, y. If the B matrix is not zero, a pleiotropic SNP, which directly influences one latent variable and its related traits, can indirectly affect other latent variables and their traits. Therefore, in mtmlSEM model, SNPs can be subdivided into single-trait, pleiotropic and direct/indirect effects.

The Maximum likelihood method, most often used to estimate parameters in SEM model, assumes that all observed and latent variables are normally distributed. Under this assumption, the sample covariance matrix of observed variables follows the Wishart distribution with the mean equal to the model-implied covariance matrix. In our dataset, some of the phenotypic traits and all SNPs take discrete ordinal values; therefore, the ML approach cannot be applied. To consider ordinal variables as normally distributed, we substituted sample covariances between ordinal variables with polychoric correlations and between ordinal and continuous variables with polyserial correlations (see section Ordinal variables). The ML approach can be applied after this manipulation (see Additional File 3).

Construction of measurement part

We identified latent variables influencing phenotypic traits applying factor analysis (FA). To determine the number of factors, we applied the parallel analysis [43]. Then, we performed FA and attributed a trait to a factor if the absolute value of the factor loading (i.e. standardized regression coefficient) exceeds 0.5. Factors influencing less than two phenotypes and phenotypes not attributed to the factors were filtered out. As a result, we obtained the measurement part of the model (1), which is a set of latent factors that influence the subsets of phenotypic traits:

$$p = \Lambda\eta + \delta \tag{3}$$

where Λ is a sparse matrix. The model does not contain an intercept term because traits are standardized to have mean zero and variance one.

Construction of structural part

In FA, factors are independent and influence all observed variables. By setting some factor loadings to zero, we probably violated the factor independency; therefore, we expect them to be non-independent. To include factor dependency into the model, we allowed factors to be in causal relationships that describe presumable common variance between them:

$$\eta = \mathrm{B}\eta + \varepsilon \tag{4}$$

where B is the coefficient matrix for relationships between latent variables, η. The model does not contain an intercept term because latent variables are assumed to have mean zero. Eq. (4) together with eq. (3) form the traditional LISREL model. To obtain the positions of parameters in the B matrix, we iteratively add them one by one until a stopping criterion is met. At an iteration, we considered each pair of latent factors and examined two possible relationships within the pair: to and back links. For each causal relationship not forming a cycle in the structural part, we estimated the parameters of the corresponding LISREL model by the ML method and checked for statistical significance of all the parameters in both Λ and B matrices (p-value < 0.05). Next, we defined the best relationship between latent factors as having the highest likelihood value and fixed the corresponding position of a new parameter in B. The iterations continued until the log-likelihood value stops decreasing.

SNP selection

Before SNPs were incorporated into the model, we estimated parameters for the constructed LISREL part of the model (Eq. (1)) and fixed all parameter values in B and Λ matrices. This is necessary to do as SNP addition enlarges the number of parameters that makes further ML estimation unstable. Therefore, we added SNPs to the model with fixed B and Λ matrices.

We first automatically introduced SNPs for each latent variable (vector g in Eq. (2)) into the model starting from the exogenous latent variables and breadth-first following the direct acyclic graph (DAG) of the structural part. Then, we performed the same automatic procedure and introduced SNPs for phenotypes (vector y in Eq. (2)).

Selecting a SNP for a variable, whether it is a latent factor or phenotype, consisted of three steps. At the first step, we included SNPs one by one as influencing the variable and perform the ML estimation of model parameters. The sample covariance matrix of all observed variables for both phenotypic traits and SNPs follows the Wishart distribution with the mean equal to model-implied covariance matrix (see Additional File 3). Secondly, based on the ML estimates, we calculate the Wishart density for the sample covariance matrix of phenotypes only taking as the mean parameter of the distribution the model-implied covariance of phenotypes. At the third step, we sort all SNPs according to the calculated densities and put the top SNP into the model fixing the corresponding parameter in Π or K matrix with the ML estimate. This automatic algorithm for selecting SNPs was implemented using the tools of the ***semopy*** [44] Python package.

Ordinal variables

The estimation of parameters in the SEM model is traditionally based on the assumption that all variables, whether they are observed or latent, are normally distributed. However, in the mtmlSEM model, this assumption is inevitably violated because SNPs take only discrete values, for instance, $\{0, 1, 2\}$, in the additive model. Moreover, the ordinal scale is often used for measurements of phenotypic traits.

We considered ordinal data as coming from a hidden continuous normal distribution with a threshold specification [45] and introduced additional latent variables to the model as follows. Let $\tilde{x}$ be a latent normally distributed variable that mimics the ordinal variable x taking values from $\{x_1, x_2, \ldots x_n\}$. Suppose for a given data set the proportions of these values are $\{f_1, f_2, \ldots f_n\}$, respectively. Let thresholds $\{-\infty = t_0, t_1, \ldots t_n = \infty\}$ divide the normal distribution into n parts corresponding to the proportions t_k equal to the standard normal quantile at $\sum_{i=1}^{k} f_i$. Although the exact continuous measurements of $\tilde{x}$ are not available, we consider that if $x = x_k$, then $t_{k-1} < \tilde{x} \leq t_k$ [45]. Thereby, for each SNP and ordinal phenotypic trait, we introduce to the model additional normally distributed latent variables.

Let the vector of phenotypes p be split into two parts: continuous traits, u, modelled as normally distributed, and discrete phenotypes, v, measured on an ordinal scale. For the latter, as well as for g and y variables, we apply the threshold approach described above and introduce vectors of latent variables $\tilde{v}$, $\tilde{g}$ and $\tilde{y}$, respectively. Therefore, the model (2) is transformed to

$$\begin{aligned} \eta &= \mathrm{B}\eta + \Pi\tilde{g} + \varepsilon \\ \begin{pmatrix} u \\ \tilde{v} \end{pmatrix} &= \Lambda\eta + \mathrm{K}\tilde{y} + \delta \end{aligned} \tag{5}$$

Bayesian estimation of model parameters

The ML method is used to estimate parameters of SEM models most of the time. However, if the number of parameters is large, as in our mtmlSEM model, this method is computationally unstable and prone to optimization failure. In contrast to the ML method, the Bayesian approach can cope with this situation taking into account prior information about parameters and maximizing the posterior distribution of parameters and latent variables. We considered values in the Β, Λ, Π, Κ matrices that were fixed during model construction as prior information and performed the Bayesian inference to obtain the posterior distributions for all parameters (denote set of all parameters as $\phi = \{\mathrm{B}, \Lambda, \Pi, \mathrm{K}, \Theta_\varepsilon, \Theta_\delta\}$) and latent variables $(\eta, \tilde{\nu}, \tilde{g}, \tilde{y})$ (see Additional File 4). As a result, we were able to generate posterior distributions of parameters by the Gibbs sampler, a Markov chain Monte Carlo algorithm. We initiated each chain with random values, and, at each iteration of the sampler, we draw

1. datasets for $\tilde{\nu}$, $\tilde{g}$ and $\tilde{y}$ from truncated normal distributions, independently of ϕ;
2. datasets for η from the multivariate normal distribution conditional on ϕ;
3. diagonal values in Θ_ε from the inverse gamma distribution conditional on ϕ;
4. values in rows of the block matrix $[\mathrm{B}, \Pi]$ from multivariate normal distributions conditional on ϕ;
5. diagonal values in Θ_δ from the inverse gamma distribution conditional on ϕ;
6. values in rows of the block matrix $[\Lambda, \mathrm{K}]$ from multivariate normal distributions conditional on ϕ.

To get parameter estimates, we performed Gibbs sampling on 5 chains of length 2000, checked convergence indicators (Gelnman-Rubin diagnostics and the effective sample size), and calculated the parameter estimates.

The chickpea dataset

The chickpea dataset (*Cicer arietinum L.*) consists of 404 accessions from the Vavilov Institute of Plant Genetic

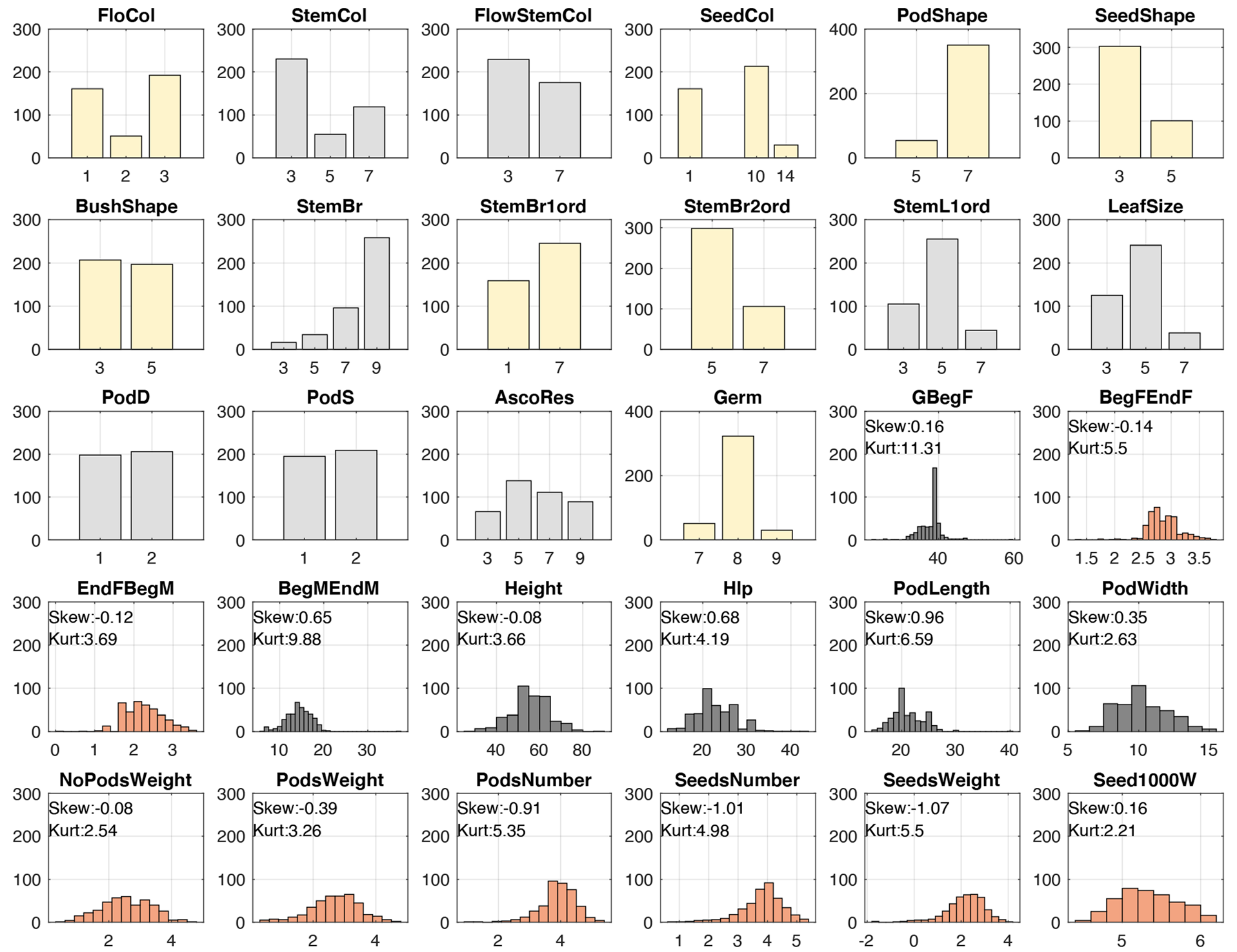

Fig. 4 Distributions of the data after preparation. Grey-coloured traits were not transformed. Yellow-coloured traits are categorial traits that were transformed; orange-coloured traits are non-categorial and were log-transformed

Resources (VIR) seed bank. In 2017, these accessions were phenotyped for 30 phenological, morphological, agronomical, and biological traits. Some of these traits are categorical and others are quantitative. Phenotype abbreviations and units of measurement are in Additional File 2. Genotyping by sequencing (GBS) of chickpea accessions identified 56, 855 segregating single nucleotide polymorphisms (SNPs). These SNPs were further filtered to meet requirements for minor allele frequency (MAF) > 3% and genotype call-rate > 90%. 2579 SNPs in 404 accessions passed all filtering criteria and were retained for further analysis. The phenotype data were further transformed in two ways. Firstly, for some categorial traits, we merged categories to make them more distinct (Additional File 2). Secondly, several quantitative traits were log-transformed to satisfy the assumption of normality (Fig. 4). All quantitative traits were further centered and scaled by calculation of z-score.

Test for predictive ability

The model was validated by 20-fold cross-validation. We randomly partitioned the dataset into 20 training (about 380 samples) and test (20 samples) sets and fixed the splits. For each training set, we independently constructed an mtmlSEM model and obtained parameter estimates after Gibbs sampling on 5 chains taking these parameters to predict values of phenotypic traits in the corresponding test set. The prediction accuracy was estimated by calculating the Pearson correlation between observed and predicted values across all test sets, the coefficient of determination and normalized rooted mean square error (Additional File 5).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-06833-2.

Additional File 1. Absolute values of correlations between phenotypic traits.

Additional File 2. Description of phenotypic trait.

Additional File 3. Maximum Likelihood estimates.

Additional File 4. Bayesian inference and Gibbs sampling.

Additional File 5. Root mean square error.

Abbreviations

SEM: Structural equation modeling; mtmlSEM: Multi-train multi-locus SEM; GWAS: Genome-wide association studies; PCA: Principal component analysis; FA: Factor analysis

Acknowledgements

We would like to thank Katrina Sherbina for the careful proofreading.

About this supplement

This article has been published as part of BMC Genomics Volume 21 Supplement 8, 2020: Selected articles from the 5th International Scientific Conference "Plant genetics, genomics, bioinformatics, and biotechnology" (PlantGen2019): genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-21-supplement-8.

Authors' contributions

All authors read and approved of the final manuscript. Methodology was developed by A.A.I; data analysis and visualization were performed by A.A.I.; software was developed by A.A.I and G.M; Bayesian inference was performed by A.A.I. and M.V.G. Initial writing and draft preparation was done by A.A.I.; review and editing were made by A.A.I., M.G.S., S.V.N, M.V.G. Data is curated by M.G.S. and S.V.N.

Funding

Model development, data analysis, manuscript preparation and writing were supported by the RBFR grant 18–29-13033 to AI, SN and MS. All the research, namely model development and testing, as well as paper writing and publication costs, were covered from the grant funds.

Availability of data and materials

The datasets analyzed and the scripts during the current study are available in the [GitHub] repository, https://github.com/iganna/mtmlSEM.git
All figures were created by A.Igolkina.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Peter the Great Saint-Petersburg Polytechnic University, Russian Federation, Polytechnicheskaya, 29, St. Petersburg 195251, Russia. [2]Centre for Genome Bioinformatics, St. Petersburg State University, St. Petersburg 199034, Russia. [3]Program Molecular & Computational Biology, Dornsife College of Letters Arts and Science, University of Southern California, Los Angeles, CA, USA.

Received: 16 September 2019 Accepted: 16 June 2020
Published: 28 July 2020

References

1. Yang Q, Wang Y. Methods for analyzing multivariate phenotypes in genetic association studies. J Probab Stat. 2012;2012:1–13. https://doi.org/10.1155/2012/652569.
2. Hackinger S, Zeggini E. Statistical methods to detect pleiotropy in human complex traits. Open Biol. 2017;7:170125. https://doi.org/10.1098/rsob.170125.
3. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;38:963–74.
4. Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet. 2012;44:1066–71. https://doi.org/10.1038/ng.2376.
5. Ott J, Rabinowitz D. A principal-components approach based on heritability for combining phenotype information. Hum Hered. 1999;49:106–11. https://doi.org/10.1159/000022854.
6. Wang Y, Fang Y, Jin M. A ridge penalized principal-components approach based on heritability for high-dimensional data. Hum Hered. 2007;64:182–91. https://doi.org/10.1159/000102991.
7. Lange C, van Steen K, Andrew T, Lyon H, DeMeo DL, Raby B, et al. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. Stat Appl Genet Mol Biol. 2004;3:1–27. https://doi.org/10.2202/1544-6115.1067.
8. Gao H, Zhang T, Wu Y, Wu Y, Jiang L, Zhan J, et al. Multiple-trait genome-wide association study based on principal component analysis for residual covariance matrix. Heredity (Edinb). 2014;113:526–32. doi:https://doi.org/10.1038/hdy.2014.57.
9. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 2018;50:229–37. https://doi.org/10.1038/s41588-017-0009-4.
10. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009;25:714–21. https://doi.org/10.1093/bioinformatics/btp041.

11. Cho S, Kim H, Oh S, Kim K, Park T. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. BMC Proc. 2009; 3(Suppl 7):S25. https://doi.org/10.1186/1753-6561-3-s7-s25.
12. Yi N, Xu S. Bayesian LASSO for quantitative trait loci mapping. Genetics. 2008;179:1045–55. https://doi.org/10.1534/genetics.107.085589.
13. Wang D, Eskridge KM, Crossa J. Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. J Agric Biol Environ Stat. 2011;16:170–84. https://doi.org/10.1007/s13253-010-0046-2.
14. Wen Y-J, Zhang H, Ni Y-L, Huang B, Zhang J, Feng J-Y, et al. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. Brief Bioinform. 2018;19:700–12. https://doi.org/10.1093/bib/bbw145.
15. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D. de los Campos G, et al. genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22:961–75. https://doi.org/10.1016/j.tplants.2017.08.011.
16. Goudet J, Kay T, Weir BS. How to estimate kinship. Mol Ecol. 2018;27:4121–35. https://doi.org/10.1111/mec.14833.
17. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet. 2012;44:825–30. https://doi.org/10.1038/ng.2314.
18. Robinson GK. That BLUP is a good thing: the estimation of random effects. Stat Sci. 1991;6:15–32. https://doi.org/10.1214/ss/1177011926.
19. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44:821–4. https://doi.org/10.1038/ng.2310.
20. Aulchenko YS, de Koning D-J, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for Genomewide pedigree-based quantitative trait loci association analysis. Genetics. 2007; 177:577–85. https://doi.org/10.1534/genetics.107.075614.
21. Liu J, Yang C, Shi X, Li C, Huang J, Zhao H, et al. Analyzing association mapping in pedigree-based GWAS using a penalized multitrait mixed model. Genet Epidemiol. 2016;40:382–93. https://doi.org/10.1002/gepi.21975.
22. Zhan X, Zhao N, Plantinga A, Thornton TA, Conneely KN, Epstein MP, et al. Powerful genetic association analysis for common or rare variants with high-dimensional structured traits. Genetics. 2017;206:1779–90. https://doi.org/10.1534/genetics.116.199646.
23. Dutta D, Scott L, Boehnke M, Lee S. Multi-SKAT: general framework to test for rare-variant association with multiple phenotypes. Genet Epidemiol. 2019;43:4–23. https://doi.org/10.1002/gepi.22156.
24. Weighill D, Jones P, Bleker C, Ranjan P, Shah M, Zhao N, et al. Multi-phenotype association decomposition: unraveling complex gene-phenotype relationships. Front Genet. 2019;10. https://doi.org/10.3389/fgene.2019.00417.
25. Lippert C, Casale F, Rakitsch B, Stegle O. LIMIX: genetic analysis of multiple traits. bioRxiv. 2014. http://europepmc.org/article/PPR/ppr7019.
26. Wright S. Correlation and causation. J Agric Res. 1921;20:557–85.
27. Wright S. On the nature of size factors. Genetics. 1918;3:367–74.
28. Igolkina AA, Samsonova MG. SEM: Structural Equation Modeling in Molecular Biology. Biophys (Russian Fed). 2018;63. https://link.springer.com/article/10.1134/S0006350918020100.
29. Igolkina AA, Armoskus C, Newman JRB, Evgrafov OV, McIntyre LM, Nuzhdin SV, et al. Analysis of gene expression variance in schizophrenia using structural equation modeling. Front Mol Neurosci. 2018;11. https://www.frontiersin.org/articles/10.3389/fnmol.2018.00192/full.
30. Pepe D, Grassi M. Investigating perturbed pathway modules from gene expression data via structural equation models. BMC Bioinformatics. 2014;15: 132. https://doi.org/10.1186/1471-2105-15-132.
31. Karns R, Succop P, Zhang G, Sun G, Indugula SR, Havas-Augustin D, et al. Modeling metabolic syndrome through structural equations of metabolic traits, comorbid diseases, and GWAS variants. Obesity. 2013;21:745–54.
32. Liu B, de la Fuente A, Hoeschele I. Gene network inference via structural equation modeling in Genetical genomics experiments. Genetics. 2008;178: 1763–76. https://doi.org/10.1534/genetics.107.080069.
33. Cai X, Bazerque JA, Giannakis GB. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. PLoS Comput Biol. 2013;9.
34. Verhulst B, Maes HH, Neale MC. GW-SEM: a statistical package to conduct genome-wide structural equation modeling. Behav Genet. 2017;47:345–59.
35. Mi X, Eskridge K, Wang D, Baenziger PS, Campbell BT, Gill KS, et al. Regression-based multi-trait QTL mapping using a structural equation model. Stat Appl Genet Mol Biol. 2010;9:38. https://doi.org/10.2202/1544-6115.1552.
36. Fernando RL, Garrick D. Bayesian Methods Applied to GWAS; 2013. p. 237–74. https://doi.org/10.1007/978-1-62703-447-0_10.
37. Yang Y, Basu S, Mirabello L, Spector L, Zhang L. A Bayesian gene-based genome-wide association study analysis of osteosarcoma trio data using a hierarchically structured prior. Cancer Inform. 2018;17:117693511877510. https://doi.org/10.1177/1176935118775103.
38. Sokolkova AB, Chang PL, Carrasquila-Garcia N, Nuzhdina NV, Cook DR, Nuzhdin SV, et al. Signatures of Ecological Adaptation in Genomes of Chickpea Landraces. Biophys (Russian Fed). 2020;65. https://link.springer.com/article/10.1134/S0006350920020244.
39. Purushothaman R, Upadhyaya HD, Gaur PM, Gowda CLL, Krishnamurthy L. Kabuli and desi chickpeas differ in their requirement for reproductive duration. F Crop Res. 2014;163:24–31.
40. Taiz L, Zeiger E. Plant physiology. 5th ed. Sunderland: Sinauer Associates; 2010.
41. Bollen KA. Structural equations with latent variables. Hoboken, NJ: Wiley; 1989. https://doi.org/10.1002/9781118619179.
42. Kline RB. Principles and practice of Structural Equation Modeling (3rd ed.): The Gulford Press; 2011. ISBN 9781462523344.
43. Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965;30:179–85. https://doi.org/10.1007/BF02289447.
44. Igolkina AA, Meshcheryakov G. semopy: A Python Package for Structural Equation Modeling. Struct Equ Model A Multidiscip J. 2020:1–12. https://www.tandfonline.com/doi/abs/10.1080/10705511.2019.1704289?scroll=top&needAccess=true&journalCode=hsem20.
45. Lee S-Y. Structural equation modeling: a Bayesian approach. Wiley: Chichester; 2007. https://doi.org/10.1002/9780470024737.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.