Statistical Applications in Genetics and Molecular Biology

Volume 9, *Issue* 1 2010 *Article* 38

Regression-Based Multi-Trait QTL Mapping Using a Structural Equation Model

Xiaojuan Mi* Kent Eskridge† Dong Wang‡
P. Stephen Baenziger** B. Todd Campbell†† Kulvinder S. Gill‡‡
Ismail Dweikat§ James Bovaird¶

*University of Nebraska–Lincoln, xjmixu@yahoo.com
†University of Nebraska–Lincoln, keskridg@unlserve.unl.edu
‡University of Nebraska–Lincoln, dwang3@unlnotes.unl.edu
**University of Nebraska–Lincoln, pbaenziger1@unl.edu
††USDA-ARS Coastal Plains Soil, Water, and Plant Research Center, todd.campbell@ars.usda.gov
‡‡Washington State University, ksgill@mail.wsu.edu
§University of Nebraska–Lincoln, idweikat2@unl.edu
¶University of Nebraska–Lincoln, jbovaird2@unl.edu

Regression-Based Multi-Trait QTL Mapping Using a Structural Equation Model*

Xiaojuan Mi, Kent Eskridge, Dong Wang, P. Stephen Baenziger, B. Todd Campbell, Kulvinder S. Gill, Ismail Dweikat, and James Bovaird

Abstract

Quantitative trait loci (QTL) mapping often results in data on a number of traits that have well-established causal relationships. Many multi-trait QTL mapping methods that account for the correlation among multiple traits have been developed to improve the statistical power and the precision of QTL parameter estimation. However, none of these methods are capable of incorporating the causal structure among the traits. Consequently, genetic functions of the QTL may not be fully understood. Structural equation modeling (SEM) allows researchers to explicitly characterize the causal structure among the variables and to decompose effects into direct, indirect, and total effects. In this paper, we developed a multi-trait SEM method of QTL mapping that takes into account the causal relationships among traits related to grain yield. Performance of the proposed method is evaluated by simulation study and applied to data from a wheat experiment. Compared with single trait analysis and the multi-trait least-squares analysis, our multi-trait SEM improves statistical power of QTL detection and provides important insight into how QTLs regulate traits by investigating the direct, indirect, and total QTL effects. The approach also helps build biological models that more realistically reflect the complex relationships among QTL and traits and is more precise and efficient in QTL mapping than single trait analysis.

KEYWORDS: QTL mapping, multiple traits, structural equation model, least squares

*We thank the anonymous reviewers and associate editor for their comments and suggestions which contributed to great improvement to this paper.

Introduction

In QTL studies, it is common to collect data on a number of traits where the causal relationships among these traits are well-established. In wheat genetics for example, yield components develop sequentially with later-developing components under the control of earlier-developing ones. Gain yield (GYLD) and its components such as, 1000-kernel weight (TKWT), spikes per square meter (SPSM), and kernels per spike (KPS) have well-established causal relationships (Dofing and Knight, 1992) (Figure 1). A QTL may affect SPSM, KPS, and TKWT, which biologically may act as intermediate variables and ultimately affect GYLD. The common procedure has been to capture the total QTL effects without investigating the distinction between direct and indirect effects. However, these effects can help answer important questions that are not addressed by examining the total effect alone. For instance, a pleiotropic QTL can have a positive direct effect on grain yield, but a negative effect on a yield component. Without knowing the full pathway of the causal relationship, a breeder might select against the QTL thinking it only affects the yield component detrimentally, not knowing it is actually beneficial on the important trait of grain yield. Thus the total effect can provide a misleading impression. To understand the genetic effects of a QTL thoroughly, it is necessary to understand not only the total QTL effect, but also the direct and indirect effects of a QTL through other traits by taking advantage of causal relationships among traits. Such a strategy of QTL mapping can provide additional insight into how QTLs regulate traits directly and indirectly through other traits. It should also improve the power to detect the QTL and the precision of the location estimate.

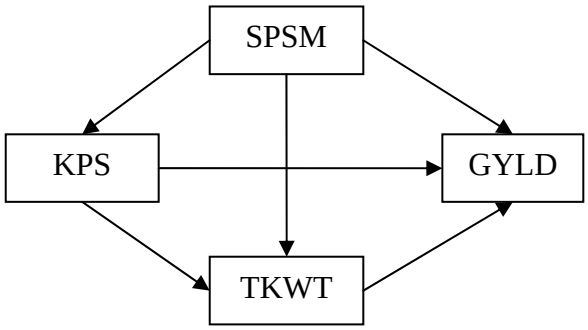

Figure 1: The path diagram of the causal relationship among grain yield and yield components

Although it is common to collect data observations on multiple causally related or genetically correlated traits frequently, QTLs are mapped for each trait

separately using single trait analyses (Lander and Botstein, 1989; Haley and Knott, 1992; Jansen and Stam, 1994; Zeng, 1994). Alternatively, several multiple trait QTL analysis (joint analysis) methods have been developed that take into account the correlation among multiple traits. These methods have been shown to improve statistical power for QTL detection and precision of parameter estimates compared to single trait analysis. Among the most effective approaches are multi-trait maximum-likelihood (ML) (Jiang and Zeng, 1995 and Korol et al., 1995, 1998), multi-trait least squares (LS) (Korol et al., 1995, 1998; Knott and Haley, 2000; Hackett et al., 2001), principal component analysis (PCA) (Weller et al., 1996; Mangin et al., 1998; Calinski et al., 2000), and discriminant analysis (DA) (Gilbert and Le Rol, 2003). Multi-trait ML, implemented with the expectation/conditional maximization (ECM) algorithm, extracts maximum information from the data, but might be very difficult to be implemented in complex data structures because of computational difficulties. Multi-trait LS, which regresses the quantitative trait value on the conditional expected genotypic value, produces results very similar to ML and simplifies computation (Haley and Knott, 1992). The PCA method transforms multiple traits into canonical variables so that single trait analyses can be carried out for each canonical variable. Similarly the DA method is based on the linear combination of the traits, specific to each tested position and analyzed by a univariate method. However, the approaches of PCA and DA may cause spurious linkages and difficulties in the biological interpretation of study results (Mähler et al., 2002; Gilbert and Le Rol, 2003). In addition, none of the above methods take advantage of causal structure among the traits. Multi-trait QTL mapping should provide additional insight into the genetic functions of QTL when causal structure is considered.

SEM is a generalization of simultaneous equation procedures originating from path analysis (Wright, 1921) and initially popularized in Econometrics and genetics. It is a useful method for estimating and evaluating simultaneous causal relationships among variables which allows variables to be both dependents and predictors. It is best explained by considering a path diagram. In particular, SEM allows researchers to decompose the effects of one variable on another into direct, indirect, and total effects. The direct effect is the path coefficient between an independent variable and the dependent variable that are not causally explained by any other intermediary variable. The indirect effects of a variable are mediated by at least one other intervening variable. The indirect effects are calculated by multiplying the path coefficients for each path of the associated variable to the dependent variable. The total effect is the sum of direct and all indirect effects. By explicitly accounting for the complex multi-component causal structure among traits, SEM can provide better understanding of multiple trait QTL analysis.

Recently SEM has been applied to functionally related traits in genetic research with the goal of characterizing genetic architecture precisely and

DOI: 10.2202/1544-6115.1552

intuitively (Nadeau et al., 2003; Gianola and Sorensen, 2004; Li et al., 2006; Neto et al., 2008; Zhu and Zhang, 2009). However, their approaches were limited to testing and quantifying the relationships among identified QTLs and phenotypes without QTL detection. In addition, Wu and others recently developed a series of joint statistical models based on nonlinear power functions for detecting QTLs that are responsible for allometric scaling laws and testing the hypotheses about the genetic control of allometry (Wu et al., 2002; Ma et al., 2003; Li et al., 2007). However, their methods may not separate the direct and indirect QTL effects, and were different from our approach.

In this paper, we developed a multi-trait SEM method of QTL mapping using a population of recombinant inbred lines (RILs), which are usually derived from a cross between two inbred parents followed by self-pollination and single seed descent to reach homozygosity. The proposed model is compared with multi-trait LS composite interval mapping and single-trait LS composite interval mapping in terms of the statistical power of QTL detection and the precision of parameter estimation. The comparison and performance of the proposed method is evaluated by simulation and applied to agronomic trait data collected on a population of wheat chromosome 3A recombinant inbred chromosome substitution lines (RICLs).

Materials and Methods

Statistical Method

Let y_1, y_2 …, and y_p be the phenotypic values of p causally related traits from an RIL individual. The SEM in matrix form is:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} 0 & \beta_{12} & \cdots & \beta_{1p} \\ 0 & 0 & \cdots & \beta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} x_{QTL} + \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1q} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pq} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{bmatrix} \quad (1)$$

where y_k is the phenotypic value for trait k; β_{kl} is the regression coefficient of trait l on trait k; α_k is the additive effect of the putative QTL on trait k; x_{QTL} is a indicator variable taking values of 1 for one homozygous parent type QQ and -1 for other homozygous parent type qq; γ_{kj} is the regression coefficient of cofactor marker j on trait k, assuming q markers are selected as cofactor markers to control the variation from these QTLs; x_j is the genotype of the jth cofactor marker, which takes values of 1 and -1 for marker genotype MM and mm respectively; and e_k, the residual effect on trait k, is assumed to be multivariate normal distributed with means zero and covariance matrix

$$\boldsymbol{\Psi} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^2 \end{pmatrix}$$

In practice we observe the marker genotypes and trait values but not the putative QTL genotypes x_{QTL}. However, it can be replaced by its conditional expectation given flanking marker genotypes (Haley and Knott, 1992, Xu, 1998). Suppose the flanking markers are *M* and *N*. Then, there are four types of marker genotypes, *MMNN*, *MM Nn*, *MmNN*, and *MmNn*. We denote p_1 and p_2 as conditional probabilities for QTL genotypes *QQ* and *qq* given the four marker genotypes. The mixture model (1) can be approximated by

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} 0 & \beta_{12} & \cdots & \beta_{1p} \\ 0 & 0 & \cdots & \beta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} E(x_{QTL}|MN) + \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1q} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pq} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{bmatrix} \quad (2)$$

where $E(x_{QTL} \mid MN) = (+1)p_1 + (-1)p_2 = p_1 - p_2$. Model (2) is more compactly written as

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta} \quad \text{where } \boldsymbol{\zeta} \sim MVN(\mathbf{0}, \boldsymbol{\Psi}) \text{ and } \mathbf{x} \sim MVN(\mathbf{0}, \boldsymbol{\Phi})$$

And the reduced model

$$\mathbf{y} = (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}) \qquad (3)$$

where **y** is a *p* x 1 vector of y_k, **B** is the *p* x *p* coefficient matrix (contains βs) that describes the causal relationships among the *p* traits, where $(\mathbf{I} - \mathbf{B})^{-1}$ exists; $\boldsymbol{\Gamma}$ is the *p* x (*q*+1) coefficient matrix (contains α_ks and γ_{kj}s) that describes causal relationship between endogenous variables (traits) and exogenous variables (QTL and cofactor markers); **x** is a (*q*+1) x 1 vector of exogenous variables, which include $E(x_{QTL} \mid MN)$ and *q* cofactor markers used for background control, and is assumed to be multivariate normally distributed with a mean vector of zeros and a covariance matrix $\boldsymbol{\Phi}$; and $\boldsymbol{\zeta}$, a *p* x 1 vector of errors, is assumed to be multivariate normally distributed with a mean vector of zeros and a diagonal covariance matrix $\boldsymbol{\Psi}$. Elements in $\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\Psi}$ are parameters to be estimated.

Maximum Likelihood (ML): In SEM, the statistical tests are based on the assumption of a multivariate normal distribution for the observed variables. A

DOI: 10.2202/1544-6115.1552

commonly used fitting function is the likelihood function. Let $\boldsymbol{\theta}$ be all the unknown parameters. Given model (3), the model implied covariance matrix of observed variables **y** and **x** were derived as

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{\theta}) &= \begin{pmatrix} E(\mathbf{yy}`) & E(\mathbf{yx}`) \\ E(\mathbf{xy}`) & E(\mathbf{xx}`) \end{pmatrix} \\ &= \begin{pmatrix} (\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}`+\boldsymbol{\Psi})\left[(\mathbf{I}-\mathbf{B})^{-1}\right]` & (\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{\Gamma}`\left[(\mathbf{I}-\mathbf{B})^{-1}\right]` & \boldsymbol{\Phi} \end{pmatrix}\end{aligned}$$

If we combine **y** and **x** into a single (p + q + 1) x 1 vector **z**, then its probability density is

$$f(\mathbf{z};\Sigma(\boldsymbol{\theta})) = (2\pi)^{-(p+q+1)/2} \mid \Sigma(\boldsymbol{\theta}) \mid^{-1/2} \exp[(-\tfrac{1}{2})\mathbf{z}`\Sigma(\boldsymbol{\theta})^{-1}\mathbf{z}] \qquad (4)$$

where p is the number of traits, and q +1 is the number of cofactor markers including one for the QTL. Since the N observations are assumed to be sampled independently, the marginal likelihood is the product of the contributions from all observations,

$$L(\boldsymbol{\theta}) = (2\pi)^{-N(p+q+1)/2} \mid \Sigma(\boldsymbol{\theta}) \mid^{-N/2} \exp[-\frac{1}{2}\sum_{i=1}^{N}\mathbf{z}_i`\Sigma^{-1}(\boldsymbol{\theta})\mathbf{z}_i]$$

The log of the likelihood function is

$$\begin{aligned}\log L(\boldsymbol{\theta}) &= \frac{-N(p+q+1)}{2}\log(2\pi) - \frac{N}{2}\log\mid\Sigma(\boldsymbol{\theta})\mid -\frac{1}{2}\sum_{i=1}^{N}\mathbf{z}_i`\Sigma^{-1}(\boldsymbol{\theta})\mathbf{z}_i \\ &= \text{constant} - \frac{N}{2}\log\mid\Sigma(\boldsymbol{\theta})\mid -\frac{1}{2}\sum_{i=1}^{N} tr[\mathbf{z}_i`\Sigma^{-1}(\boldsymbol{\theta})\mathbf{z}_i] \\ &= \text{constant} - \frac{N}{2}\log\mid\Sigma(\boldsymbol{\theta})\mid -\frac{N}{2}\sum_{i=1}^{N} tr[N^{-1}\mathbf{z}_i\mathbf{z}_i`\Sigma^{-1}(\boldsymbol{\theta})] \\ &= \text{constant} - \frac{N}{2}\{\log\mid\Sigma(\boldsymbol{\theta})\mid + tr[\mathbf{S}*\Sigma^{-1}(\boldsymbol{\theta})]\}\end{aligned} \qquad (5)$$

where **S*** is the maximum likelihood estimator of the sample covariance matrix $\mathbf{S} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{YY}} & \boldsymbol{\Sigma}_{\mathbf{YX}} \\ \boldsymbol{\Sigma}_{\mathbf{YX}}` & \boldsymbol{\Sigma}_{\mathbf{XX}} \end{pmatrix}$. These two matrices are essentially equal in large samples. The 'constant' term in equation (5) has no impact on estimating $\boldsymbol{\theta}$. For a given sample, **S** and (p+q+1) are constant. The unknown parameters are estimated by maximizing (5), which is equivalent to minimizing the function (6)

$$F_{ML} = \log|\Sigma(\boldsymbol{\theta})| + tr[\mathbf{S}\Sigma^{-1}(\boldsymbol{\theta})] - \log|\mathbf{S}| - (p+q+1) \qquad (6)$$

Parameter estimation: The fitting function that is minimized is equation (6). The unknown parameters are estimated so that the model implied covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is close to the sample covariance matrix $\mathbf{S}$. F_{ML} is zero when $\Sigma(\hat{\boldsymbol{\theta}}) = \mathbf{S}$. Numerical solutions are typically used to estimate the parameters since the first-order partial derivatives are nonlinear in the parameter and explicit solutions for the parameters usually are not accessible. Three steps are involved.

Step 1. Select the initial or starting values $\boldsymbol{\theta}^{(0)}$.
Step 2. Move from one step to the next step, in general $\hat{\boldsymbol{\theta}}^{(i+1)}$ is determinated by

$$\hat{\boldsymbol{\theta}}^{(i+1)} = \hat{\boldsymbol{\theta}}^{(i)} - \left[\frac{\partial^2 F_{ML}}{\partial\hat{\boldsymbol{\theta}}\partial\hat{\boldsymbol{\theta}}`}\right]^{-1}\left[\frac{\partial F_{ML}}{\partial\hat{\boldsymbol{\theta}}}\right]$$

Step 3. Stop the iteration when the differences in the fitting function from one iteration to the next differ by less than very small value. The final estimates will be used for the calculation of the maximum likelihood value for hypothesis testing.

Note that the indirect and total QTL effects are functions of the path coefficients $\boldsymbol{\theta}$. They are calculated based on the final values of $\boldsymbol{\theta}$. The indirect QTL effect for a particular indirect path from the QTL to the trait is calculated by multiplying all the coefficients in the path. The total indirect QTL effect on the trait is the sum of all the indirect effects from all indirect paths. The total QTL effect on the trait is the sum of direct and indirect QTL effects.

Although ML is based on the assumption of multivariate normality, its estimation procedures are robust to moderate violation of this assumption (Joreskog and Sorbom, 1989; Bollen, 1989).

Hypothesis tests: For QTL mapping, we are most interested in the existence of a QTL. The hypothesis test can be formulated as:

H_0: $\alpha_1 = \alpha_2, \ldots, = \alpha_p = 0$ (restricted model, *i.e.*, the putative QTL does not exist)
$H_{A:}$ at least one of them is not zero (unrestricted model, *i.e.*, the putative QTL exists)

DOI: 10.2202/1544-6115.1552

The Likelihood Ratio (LR) statistic is

$$LR = -2[\log L(\hat{\boldsymbol{\theta}}_r) - \log L(\hat{\boldsymbol{\theta}}_u)]$$

where $\hat{\boldsymbol{\theta}}_r$ is the ML estimator under the restricted model, and $\hat{\boldsymbol{\theta}}_u$ is the ML estimator under the unrestricted model. The LR is approximately chi-square distributed with *p* (number of traits) degrees of freedom when the restricted model is true. In the SEM framework, the LR statistic is calculated as the difference in the usual chi-square estimators for the restricted and unrestricted model (Bollen, 1989), with the difference in model degrees of freedom as the degrees of freedom for the LR statistic. The LR test compares the fit of restricted model to the fit of the unrestricted model. A significant test indicates that the model with QTL effects fits significantly better than the model without QTL effects. Because the test is performed for a number of intervals, the distribution of the maximum LR statistic is very complicated. Therefore, it is difficult to determine an exact significance critical value. Zeng (1994) suggested that the error rate of the test per interval, α, can be approximated by using the Bonferroni correction where $\chi^2_{\alpha/M,\, m+1}$ is used to approximate the critical value of the test; *M* is the number of intervals involved in the test; *m* is the number of traits and 1 is the position of the putative QTL. Alternatively, the permutation test can be applied to our proposed model to empirically estimate the genome-wise critical value for a given data set (Churchill and Doerge, 1994).

Note that, the multi-trait SEM and the multi-trait LS models have similar structures, but are used to test different hypotheses of QTL effects, and estimate different parameters. With the multi-trait SEM, the existence of direct QTL effects is tested, and the total, indirect and direct QTL effects are estimated. While with the multi-trait LS approach, the existence of the total QTL effects are tested, and only the estimates of the total QTL effects are provided.

Simulation

We investigated the multi-trait SEM method using data simulated for 100 replicates of 150 lines of a RILs population. On a single chromosome segment of length 100 cM, 11 evenly spaced markers were simulated. Three additive QTLs were placed at 22, 42, and 78cM to affect three traits, which are causally related as in equation (7).

The phenotypic values for each individual are determined by equation (7), the sum of QTL effects (where x_{QTL} s take values of 1 and -1 for genotype *QQ* and *qq* respectively) plus the random residual effects sampled from a multivariate normal distribution with mean zero and covariance matrix (8).

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{pmatrix} 0 & 0.5 & 0.25 \\ 0 & 0 & -0.5 \\ 0 & 0 & 0 \end{pmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} -0.125 \\ 0.5 \\ 0.25 \end{bmatrix} x_{QTL1} + \begin{bmatrix} 0.25 \\ -0.5 \\ -0.125 \end{bmatrix} x_{QTL2} + \begin{bmatrix} 0.5 \\ 0.5 \\ 0.25 \end{bmatrix} x_{QTL3} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \quad (7)$$

$$\boldsymbol{\Psi} = \begin{pmatrix} 1.6 & 0 & 0 \\ 0 & 1.8 & 0 \\ 0 & 0 & 2.5 \end{pmatrix} \quad (8)$$

Prior to the analysis, a test of multivariate normality was performed for all observed variables. The results showed that the Mardia and Henze-Zirkler multivariate tests rejected the multivariate normality but the multivariate plot indicated approximate normality, suggesting a minor violation of multivariate normality. The parameter estimates from multi-trait SEM are obtained by minimizing the ML fitting function in this study. The proposed method was implemented using PROC TCALIS in SAS Version 9.2 software, to account for the causal relationships among multiple traits.

A single QTL was tested sequentially at each 1-cM point along the chromosome. Since markers are evenly distributed and widely separated, all except flanking markers are fitted in the model to control the genetic background (Jiang and Zeng, 1995). Multi-trait SEM, multi-trait LS, and single-trait LS were applied to test for the presence of a QTL at 1cM segments of a chromosome. Means and standard deviations of all parameter estimates were calculated from 100 replications. The statistical power was determined by the proportion of the number of runs with the test statistic values greater than a critical value, over 100 replicates. We used $\chi^2_{0.005/4}$=14.86 (approximated by the Bonferroni correction (Jiang and Zeng, 1995)) as the critical value for the multi-trait SEM and the multi-trait LS, and $\chi^2_{0.005/2}$=10.60 for the single-trait LS. The overall power was calculated as the proportion of times the QTL was detected for at least one of the three traits.

RICL Wheat Experiment

The proposed method was applied to data from a wheat experiment with a population of 98 3A RICLs derived from a cross between 'Cheyenne' (CNN) and Cheyenne with a 'Wichita' 3A chromosome substitution (CNN(WI3A)) and thus, the lines differed only for chromosome 3A. This population was evaluated in multi-environment field trials from 1999-2001 to identify QTL and QTL-environment interactions for grain yield and other agronomic traits in seven environments. Details of the experiment and results of the data analysis performed

DOI: 10.2202/1544-6115.1552

by univariate QTL detection techniques have been described by Campbell et al. (2003). In the current study, we focus on GYLD and the yield component traits TKWT, SPSM, and KPS. We constructed a genetic map of chromosome 3A using 14 molecular markers covering 120.8 cM of the chromosome with an average marker interval of 8.5 cM. The causal relationships among grain yield and these yield component traits are described in Figure 1 (Dofing and Knight, 1992; Dhungana et al., 2007).

Grain yield and yield component trait data were analyzed using the multi-trait SEM method and single-trait LS analysis. Prior to the analysis, analysis of variance (ANOVA) linear model was fit for each trait to remove the main effects of environments and blocks. Residuals of the four traits were used as observed trait values after removing effects of environment and block within environment. For each trait, single marker analysis was performed to select cofactors. A stepwise selection procedure with significance of $P < 0.1$ was used. The maximum number of cofactors was five. Multi-trait SEM and single-trait LS were applied to test for the presence of a QTL at 1cM segments of the chromosome with a window size of 10 cM. Five-hundred permutations of the data were analyzed for multi-trait SEM and single-trait LS to establish significance threshold values for declaring significant QTL effects in a genome at α=0.05 (Churchill and Doerge, 1994).

Results

Simulation

Table 1 shows the observed statistical power of QTL detection over 100 replicates by three different mapping methods. The power of multi-trait SEM is calculated as the percentage of detecting a direct QTL effect on at least one of three traits, while the power of multi-trait LS is obtained as the percentage of detecting a total QTL effect on at least on one of three traits over 100 replicates. The power of multi-trait SEM is higher than that of the multi-trait LS for all three QTLs. This is because the pleiotropic QTL1 has a larger positive direct effect on Y_2 but a negative indirect effect, which in turn reduces the total QTL1 effect on Y_2. Similarly the total effect of QTL2 on Y_1 and Y_2 and QTL3 on Y_2 decreased due to their opposite indirect effects. QTLs with relatively smaller total effects may not be detected by multi-trait LS method. Generally, the QTL detection power for the two multi-trait analysis methods was much higher than that of the single-trait

analysis which tests existence of a total QTL effect. However, the power of detection of QTL3 on trait Y_1 and overall power on QTL3 in single-trait analysis are higher than both multi-trait analyses. This is because both indirect and direct QTL_3 effects on Y_1 are in the same direction, resulting an increased total QTL3 effect on Y_1. Therefore multi-trait QTL analysis would be most effective when the direct and indirect effects of a QTL are in opposite directions. If the direct and indirect QTL effects are in the same direction, the power of the multi-trait analysis may be less than the overall power of the single-trait analysis. In such a situation, the total QTL effect should be larger than either of the direct or indirect effects tested with the multi-trait SEM approach.

Table 2 shows the estimates (and standard deviations) of QTL effects and positions resulting from the three different mapping methods. All estimates are relatively unbiased with high precision except the QTL position estimates from the single-trait LS method, which display markedly higher standard deviation. In general, the precision and accuracy of estimating QTL positions and effects by multi-trait SEM and multi-trait LS are much greater than single-trait LS. However, as mentioned previously, the multi-trait SEM is favored over the multi-trait LS and single-trait LS analyses because direct and indirect QTL effects can be detected.

Table 1: Observed statistical powers (%) of QTL detection of multi-trait SEM, multi-trait LS, and single-trait LS methods obtained from 100 replicates in the simulation study

	Multi-Trait	Multi-Trait	Single-Trait analysis			
QTL	SEM	LS	Y1	Y2	Y3	Overall
1	64	57	6	29	10	40
2	77	69	3	43	1	45
3	96	94	97	33	17	97

DOI: 10.2202/1544-6115.1552

Table 2: Parameters and estimates of QTL positions and effects in the simulation

Methods	QTL	Trait	Position (CM)	Putative QTL Effect		
				Total	Direct	Indirect
Parameters	1	Y_1	22	0.125	-0.125	0.25
		Y_2		0.375	0.5	-0.125
		Y_3		0.25	0.25	0
	2	Y_1	42	0	0.250	-0.250
		Y_2		-0.435	-0.50	0.065
		Y_3		-0.125	-0.125	0
	3	Y_1	78	0.75	0.50	0.250
		Y_2		0.375	0.50	-0.125
		Y_3		0.25	0.25	0
Multi-trait SEM	1	Y_1	21.57 (2.83)	0.117 (0.200)	-0.128 (0.174)	0.245 (0.101)
		Y_2		0.356 (0.204)	0.464 (0.174)	-0.107 (0.107)
		Y_3		0.210 (0.205)	0.210 (0.205)	0
	2	Y_1	41.95 (2.58)	-0.034 (0.223)	0.231 (0.174)	-0.265 (0.114)
		Y_2		-0.450 (0.218)	-0.511 (0.192)	0.061 (0.096)
		Y_3		-0.119 (0.187)	-0.119 (0.187)	0
	3	Y_1	78.03 (2.16)	0.750 (0.174)	0.479 (0.176)	0.271 (0.097)
		Y_2		0.393 (0.175)	0.527 (0.166)	-0.133 (0.111)
		Y_3		0.257 (0.205)	0.257 (0.205)	0
Multi-trait LS	1	Y_1	21.56 (3.12)	0.110 (0.217)		
		Y_2		0.353 (0.203)		
		Y_3		0.239 (0.204)		
	2	Y_1	41.92 (2.63)	-0.019 (0.216)		
		Y_2		-0.448 (0.231)		
		Y_3		-0.112 (0.205)		
	3	Y_1	78.46 (2.42)	0.746 (0.185)		
		Y_2		0.384 (0.190)		
		Y_3		0.256 (0.217)		
Single-trait	1	Y_1	19.99 (6.75)	0.105 (0.247)		
		Y_2	41.80 (5.86)	0.335 (0.229)		
		Y_3	78.38 (2.82)	0.223 (0.246)		
	2	Y_1	20.60 (4.56)	-0.015 (0.210)		
		Y_2	41.81 (4.09)	-0.428 (0.227)		
		Y_3	78.38 (4.71)	-0.100 (0.202)		
	3	Y_1	20.80 (6.04)	0.749 (0.175)		
		Y_2	42.44 (6.43)	0.383 (0.180)		
		Y_3	78.18 (6.48)	0.259 (0.209)		

*Estimates are means over 100 replicates with standard deviation in parentheses, by multi-trait SEM, multi-trait LS and single-trait LS.

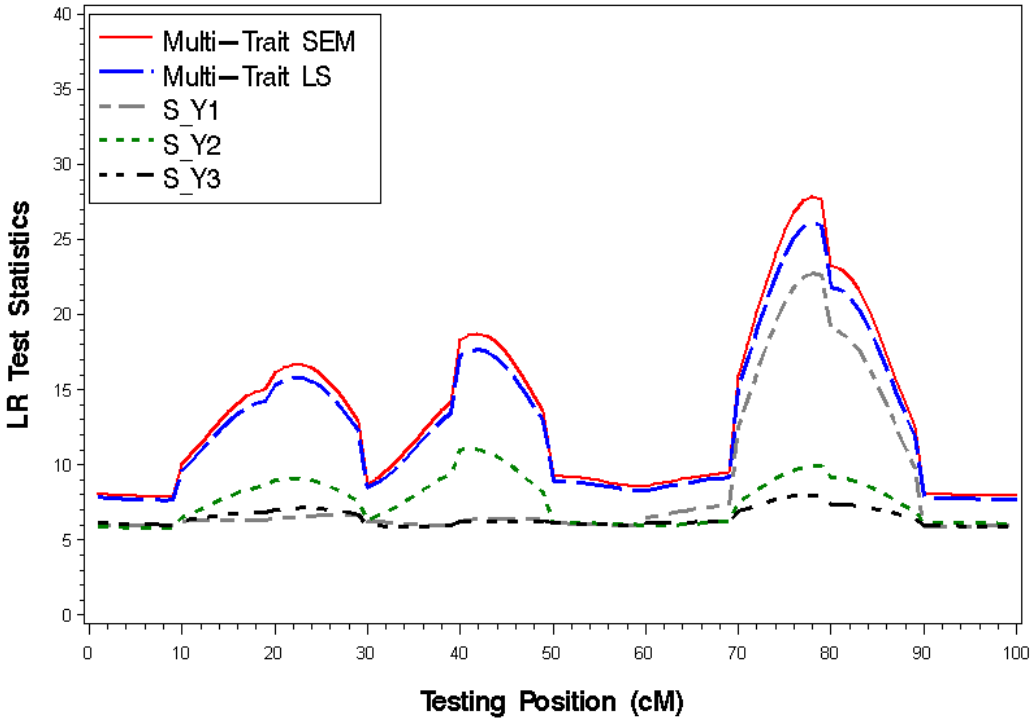

Figure 2: Results of QTL mapping on three traits from a simulated RILs population using multi-trait SEM, multi-trait LS and single-trait LS.

The LR test statistic profiles were plotted against the chromosome position (Figure 2) to compare the three methods of QTL detection. The positions where the LR test statistics exceed critical value indicate the possible QTL locations. Three QTLs are identified: one between 16 and 28 cM, one between 38 and 48 cM and one between 70 and 88 cM. The QTL profiles of the two multi-trait methods (multi-trait SEM and multi-trait LS) were very close. The QTL profiles of the single-trait LS method are very low without clear peak except for QTL3 on Y_1. This indicates the single-trait LS method has a lower chance of detecting the QTL effects when the total QTL effect is reduced due to compensating effects among traits.

DOI: 10.2202/1544-6115.1552

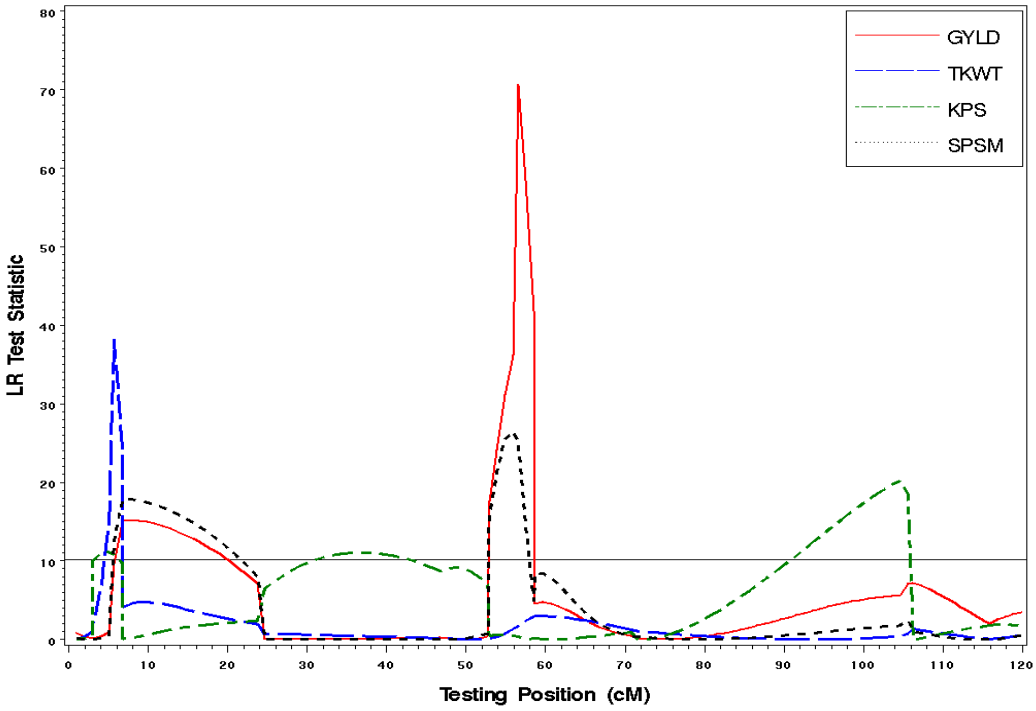

Figure 3: LR test statistic QTL profiles for chromosome 3A using single-trait LS QTL analysis for grain yield, thousand kernel weight, kernels per spike, and spikes per square meter.

RICL Wheat Experimental Data Analysis

Likelihood ratio plots of the single-trait LS analysis (Figure 3) indicated that there were three regions containing QTLs. Region 1 (4-18cM) contained QTLs associated with GYLD, KPSM, KPS, and TKWT, region 2 (53-58 cM) contained QTLs associated with GYLD and SPSM, and region 3 (96-105 cM) contained a single QTL associated with KPS. Interestingly, the largest effect QTL was found in region 2 and had a large effect on GYLD (LR= 70.56). These QTL positions corresponded to those found by Campbell et al. (2003).

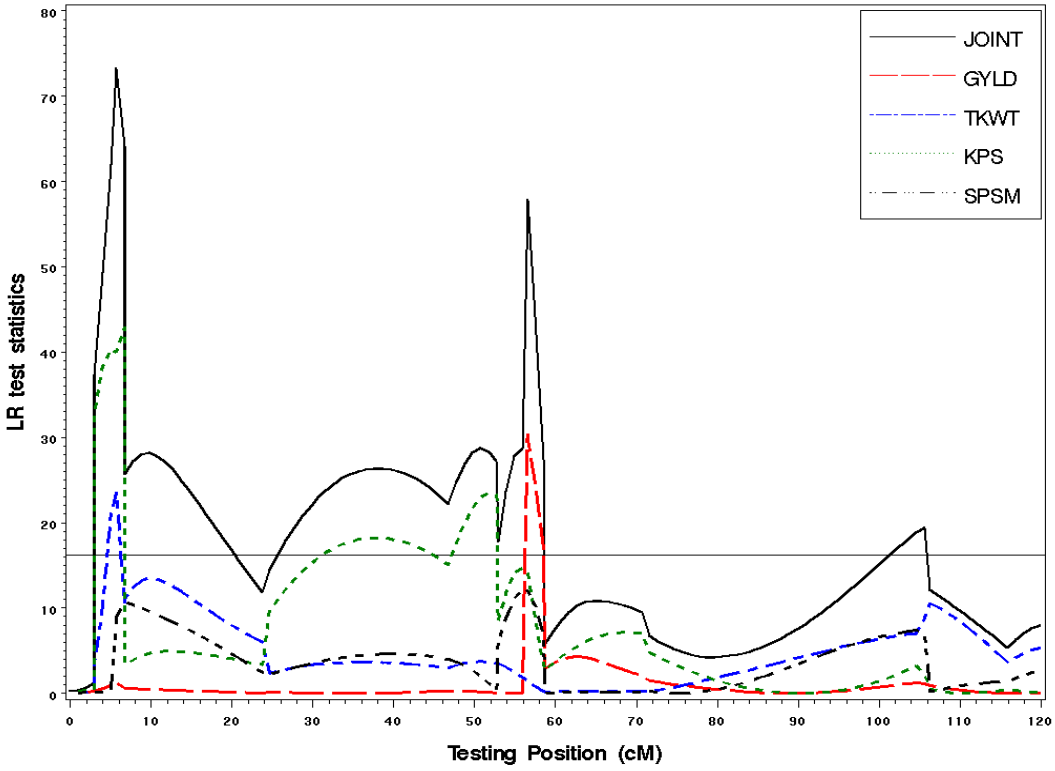

Figure 4: Map of direct QTL effects for chromosome 3A using multi-trait SEM method. Profiles of LR test statistic: solid line corresponds to the joint LR statistic, dashed lines corresponds to the contribution of the traits GYLD, TKWT, KPS and SPSM. JOINT=at least one trait is directly affected by the QTL

Table 3: Contribution to LR for individual QTLs and traits for Chromosome 3A

	Trait				
QTL at	GYLD	TKWT	KPS	SPSM	JOINT
5.677 cM	1.26	23.54	39.96	8.92	73.26
50.727 cM	0.17	3.71	22.96	1.96	28.70
56.498 cM	30.39	1.47	14.18	12.06	57.86
105.592 cM	1.16	8.75	2.25	7.27	19.39

Results of the multi-trait SEM method provided joint LR statistics and the contribution of individual traits to the joint LR simultaneously. Figure 4 summarizes results of composite interval mapping on chromosome 3A of the

DOI: 10.2202/1544-6115.1552

direct QTL effects using the multi-trait SEM method. At least four major regions containing QTLs are indicated by the analyses. Region one is between 3 and 20 cM with a clear peak located near *Xbarc12* (LR statistic of 73.26). The largest contribution to this peak is attributed to KPS (LR statistic of 39.96) and TKWT (LR statistic of 23.54) (Table 3). The second region is between 25 and 52 cM with a clear peak at 50.72 cM close in proximity to *XksuA6*. The largest contribution to this peak is attributed to the direct QTL effect on KPS. This region was not identified with the single-trait LS method (Figure 3) or in the previous QTL analysis by Campbell et al. (2003). The third region is located between 53 and 58 cM, and the highest LR is observed adjacent to *Xbarc67* (56.49 cM, LR statistic of 56.86). The largest contribution to this peak comes from the direct QTL effect on GYLD (LR statistic of 30.39) (Table 3). A weak QTL in the fourth region (101 and 105 cM) was detected adjacent to *Xbcd141* that was associated with SPSM and TKWT.

Figures 5 and 6 show the standardized path coefficients with multi-trait SEM located in regions one and three at *Xbarc12* and *Xbarc67*, respectively. The path coefficients are used to calculate the indirect and total QTL effects. For example, the indirect QTL effects on TKWT at *Xbarc12* are calculated by multiplying the path coefficients for each path of associated trait QTL to TKWT (QTL->SPSM->KPS->TKWT is 0.0585*(-0.6233)*(-0.1605)=0.00585; QTL->SPSM->TKWT is 0.0585*(-0.4689)=-0.02743; QTL->KPS->TKWT is 0.0966*(-0.1605)=-0.0155). Hence, the total indirect effect of QTL on TKWT is the sum of all the indirect effects of associated trait QTL to TKWT (0.00585+(-0.02743)+(-0.00155)=-0.0371). The total QTL effect is the sum of direct and indirect QTL effects on TKWT (-0.0868+(-0.0371)=-0.1239). Table 4 shows the standardized direct, indirect and total QTL effects on each trait at *Xbarc12* and *Xbarc67* with multi-trait SEM. At *Xbarc12*, the QTL has a large positive direct effect on KPS ($p < 0.001$) and a negative indirect effect ($p<0.01$) resulting in a smaller absolute total effect ($p<0.01$) and thus a higher peak than with the single-trait LS analysis, which only captures total QTL effects. The direct and indirect QTL effects on TKWT are in the same direction resulting in an increased total effect (Table 4) and a lower LR peak than the single-trait approach (Figure 3 and 4). Similarly, at marker *Xbarc67*, the direct and indirect QTL effects on GYLD are in the same direction leading to a large GYLD total effect QTL ($p<0.001$) and thus a lower LR peak for GYLD than the single-trait LS analysis (Figures 3 and 4). These results show that our multi-trait SEM QTL analysis method provides additional information on how the QTLs on chromosome 3A affect agronomic performance directly and indirectly, which was not possible with any previously proposed methods.

Table 4: Estimates of the putative QTL effects using multi-trait SEM in two major QTL regions in the chromosome 3A

Trait	QTL at Position (cM)	Putative QTL Effect		
		Total	Direct	Indirect
GYLD	Region 1 5.6771 cM	0.0624 (0.0169)***	0.011 (0.0098)	0.0514 (0.0169)**
TKWT		-0.1239 (0.0194)***	-0.0868 (0.0178)***	-0.0371 (0.0078)***
KPS		0.0601 (0.0195)**	0.0966 (0.0153)***	-0.0364 (0.0122)**
SPSM		0.0585 (0.0195)**	0.0585 (0.0195)**	0.0000
GYLD	Region 3 56.4983 cM	0.1733 (0.0230)***	0.0532 (0.0008)***	0.1201 (0.0230)***
TKWT		-0.0302 (0.0213)	0.0217 (0.0179)	-0.0519 (0.0115)***
KPS		0.02267 (0.0293)	0.0861 (0.0229)***	-0.0634 (0.0183)***
SPSM		0.1011 (0.0290)***	0.1011 (0.0290)***	0.0000

Values in parentheses are respective standard deviation values, ***P<0.001; **P<0.01; *P<0.05

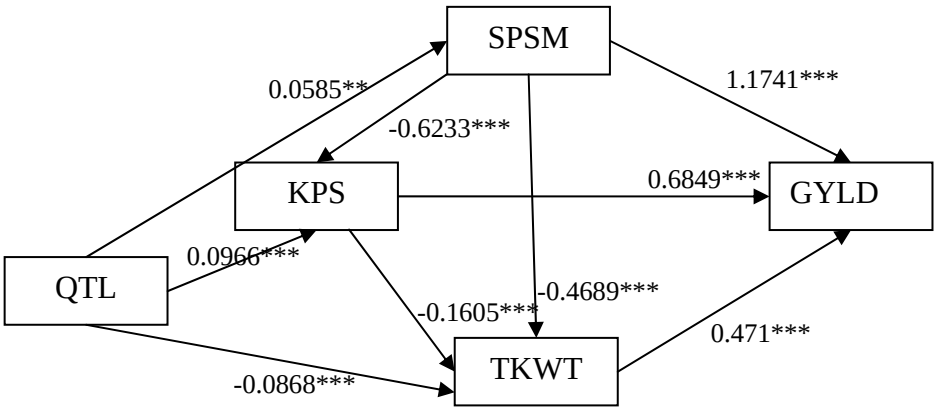

Figure 5: Path estimates of multi-trait SEM for chromosome 3A at position 5.6771 cM (*Xbarc12*). Single arrows indicate causal relationships. Numbers by the arrow lines represent the estimated standardized coefficients with significance level: ****P*<0.001; ***P*<0.01; **P*<0.05.

DOI: 10.2202/1544-6115.1552

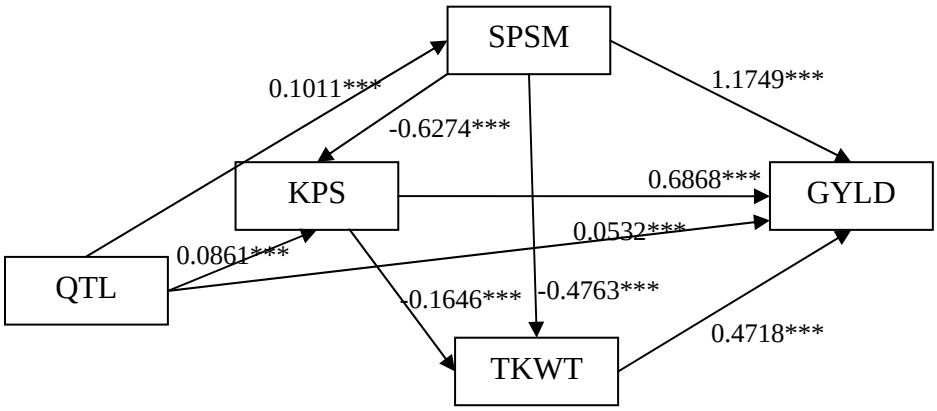

Figure 6: Path estimates of multi-trait SEM for chromosome 3A at position 56.4983 cM (*Xbarc67*). Single arrows indicate causal relationships. Numbers by the arrow lines represent the estimated standardized coefficients with significance level: ***P<0.001; **P<0.01; *P<0.05.

Discussion

We have presented a multi-trait SEM method for QTL mapping, extending the work of Knott and Haley (2000), which takes into account the causal relationships among multiple traits. The performance of the method was illustrated using simulated data. The primary advantage of the multi-trait SEM over the multi-trait LS and the single-trait LS is that it improves the power of QTL detection, which is consistent with previous findings (Zhu and Zhang, 2009). It also allows one to investigate the direct, indirect, and total QTL effects of yield and yield component traits. This ultimately allows for important insight into how QTLs interact and regulate correlated traits. Knowledge of the direct and indirect QTL effects can be very important for plant breeders interested in (1) breaking unfavorable indirect QTL effects; (2) obtaining more precise and efficient estimates and tests in QTL mapping, and (3) using statistical methods that can be simply performed using commonly available statistical software such as SAS.

Using our method, we are able to detect QTLs for SPSM in both region one and three which have not been reported in Campbell et al. (2003) where univariate QTL detection techniques were used. Furthermore, they detected a minor QTL for GYLD in region one, while the corresponding QTL that we detected had a greater effect on GYLD. In addition, some claim there is really no QTL per se for GYLD and every GYLD QTL must work through yield components. However, our research reported here clearly shows how the GYLD

QTL may be due to component traits, which might be investigated using candidate genes from other species using a comparative genomics approach.

A prerequisite of the proposed method is prior knowledge of the causal relationships among the multiple traits, since SEM is generally used as a confirmatory rather than exploratory procedure. Theoretical insight and judgment by the researcher is very important in building a model. One can obtain some basic background about the key structure of the model either from knowledge of the related field or from preliminary data analysis. Misspecification could potentially bias estimates of the parameters (Bollen, 1989) and possibly result in the multi-trait LS outperforming the multi-trait SEM approach. The researcher needs to be sure that the model is at least approximately correct, and the parameters are interpretable. In practice, one can try to correct misspecification by building a different model such as adding a new path, removing a path, or reversing the causal direction of a path to the initial model. The final model is obtained based on modification indices (Sorbom, 1989). However, caution is needed. The credibility of any causal hypothesis must be judged by biological interpretation and not solely on statistical evidence.

The model considered in this paper was illustrated using a RICL population to provide a general idea of the nature of QTLs affecting the traits. However, the general approach can be easily applied to different population structures (such as F2 and backcross) and genetic models by setting up the corresponding conditional QTL genotype probability. In addition, it is possible to test pleiotropic effects against closely linked QTL and QTL-environment interactions at a given genomic position where the presence of a QTL is indicated by joint mapping. Here, we assumed that the residual errors of the traits were independent which we believe is a reasonable assumption since we removed the effects of environment by using, for the y values in equation (2), the residuals from a main effects linear model with environment and block within environment as the main effects. However, even after removing environment main effects, the traits may still be correlated and the trait error terms can be modeled as being correlated with each other. Such a specification indicates that the traits associated with those error terms share common variation that is not explained by predictor relations in the model such as the genotypes interacting with environmental factors. We have focused on the linear relationships between traits. However, there are situations where the nonlinearity may be more appropriate such as with allometric scaling relationships between size and rate in the biological processes. One may incorporate such behavior in the model to improve the performance. In addition, the proposed model here did not account for the genotype-environment or genotype-block interactions. Methods incorporating these innovations could result in increased statistical power of QTL detection, precision in estimation of QTL effects and position, and an improved understanding of how QTL interact

DOI: 10.2202/1544-6115.1552

with environmental factors. Furthermore, researchers may collect data of different types for a sample set (e.g., both binary and continuous traits). Methods that are capable of dealing with a mixture of continuous and binary traits could be valuable in a variety of situations. Although the proposed multi-trait SEM approach may not always be appropriate for every QTL mapping application, it does provide an attractive complementary method to understand complicated biological pathways and systems using available molecular marker and phenotypic trait data.

References

Bollen, K. A. *Structural equations with latent variables*. Wiley, New York (1989).

Calinski, T., Kaczmarek, Z., Krajewski, P., Frova, C. & Sari-Gorla, M. A multivariate approach to the problem of QTL localization. *Heredity* **84**, 303-310 (2000).

Campbell, B. T., Baenziger, P. S., Gill, K. S., Eskridge, K. M., Budak, H., Erayman, M., & Yen, Y. Identification of QTLs and environmental interactions associated with agronomic traits on chromosome 3A of wheat. *Crop Sci*. **43**, 1493–1505 (2003).

Churchill, G. A. & Doerge, R. W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-971 (1994).

Dhungana, P., Eskridge, K. M., Baenziger, P. S., Campbell, B. T., Gill, K. S. & Dweikat, I. Analysis of genotype-by-environment interaction in wheat using a structural equation model and chromosome substitution lines. *Crop Sci*. **47**, 477-484 (2007).

Dofing, S. M. & Knight. C. W. Alternative model for path analysis of small-grain yield. *Crop Sci.* **32**, 487–489 (1992).

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

Gilbert, H. & Le Roy, P. Comparison of three multitrait methods for QTL detection. *Genet. Sel. Evol*. **35**, 281–304 (2003).

Hackett, C. A., Meyer, R. C. & Thomas, W. T. B. Multi-trait QTL mapping in barley using multivariate regression. *Genetic Research* **77,** 95-106 (2001).

Haley, C. S. & Knott, S. A. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69,** 315-324 (1992).

Jansen, R. C. & Stam, P. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447– 1455 (1994).

Jiang, C. J. & Zeng, Z. B. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**, 1111-1127 (1995).

Joreskog, K. G. & Sorbom, D. *LISREL 7: A Guide to the program and Applications, 2nd edition.* SPSS Inc., Chicago (1989).

Knott, S. A. & Haley, C. S. Multitrait least squares for quantitative trait loci detection. *Genetics* **156**, 899-911 (2000).

Korol, A. B., Ronin, Y. I. & Kirzhner, V. M. Interval mapping of quantitative trait loci employing correlated trait complexes. *Genetics* **140**, 1137-1147 (1995).

Korol, A. B., Ronin, Y. I., Nevo, E. & Hayes, P. M. Multi-interval mapping of correlated trait complexes. *Heredity* **80**, 273-284 (1998).

Lander, E. S. & Botstein, D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185-199 (1989).

Li, H., Huang, Z., Gai, J., Wu, S., Zeng, Y., Li, Q. & Wu, R. A Conceptual Framework for Mapping Quantitative Trait Loci Regulating Ontogenetic Allometry. *PLoS ONE* **2(11)**, e1245 (2007).

Li, R., Tsaih, S. W., Shockley, K., Stylianou, I. M., Wergedahl, J. et al. Structural model analysis of multiple quantitative traits. *PLoS Genet.* **2**, e114 (2006).

Ma, C-X., Casella, G., Littell, R. C., Khuri, A. I. & Wu, R. Exponential mapping of quantitative traits governing allometric relationships in organisms. *J Math Biol* **47,** 313–324 (2003).

DOI: 10.2202/1544-6115.1552

Mähler, M., Most, C., Schmidtke, S., Sundberg, J. P., Li, R., Hedrich, H. J. & Churchill, G. A. Genetics of colitis susceptibility in IL-10-deficient mice: backcross versus F2 results contrasted by principal component analysis. *Genomics* **80**, 274-282 (2002).

Mangin, B., Thoquet, P. & Grimsley, N. Pleiotropic QTL analysis. *Biometrics* **54**, 88–99 (1998).

Nadeau, J. H., Burrage, L. C., Restivo, J., Pao, Y. H., Churchill, G., et al. Pleiotropy, homeostasis, and functional networks based on assays of cardiovascular traits in genetically randomized populations. *Genome Res* **13**, 2082–2091 (2003).

Neto, E. C., Ferrara, C. T., Attie, A. D., & Yandell, B. S. Inferring causal phenotype networks from segregating populations. *Genetics* **179**, 1089–1100 (2008).

Sorbom, D. Model modification. *Psychometrika* **54**, 371-384 (1989).

Weller, J. I., Wiggans, G. R., Van Raden, P. M. & Ron, M. Application of a canonical transformation to detection of quanti-tative trait loci with the aid of genetic markers in a multi-trait experiment. *Theor. Appl. Genet*. **92**, 998–1002 (1996).

Wright, S. Correlation and causation. *J. Agricultural Research* **20**, 557-585 (1921).

Wu, R., Ma, C-X., Littell, R. C. & Casella, G. A statistical model for the genetic origin of allometric scaling laws in biology. *J Theor Biol* **219**, 121–135 (2002).

Xu, S. Further investigation on the regression method of mapping quantitative trait loci. *Heredity*, **80**, 364-73 (1998).

Zeng, ZB. Precision mapping of quantitative trait loci. *Genetics* **136**, 1457-1468 (1994).

Zhu, WS. & Zhang, HP. Why Do We Test Multiple Traits in Genetic Association Studies? *J Korean Stat Soc* **38(1),** 1–10 (2009).