BMC Genetics

BioMed Central

Methodology article

Mapping quantitative trait loci in line cross with repeat records

Runqing Yang*[1] and Ming Fang[2]

Address: [1]School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai, 201101, P.R. China and [2]Life Science College, Heilongjiang August First Land Reclamation University, Daqing, 163319, P.R. China

Email: Runqing Yang* - runqingyang@sjtu.edu.cn; Ming Fang - Fangming618@126.com

* Corresponding author

Abstract

Background: Phenotypes with repeat records from one individual or multiple individuals were often encountered in practices of mapping QTL in linecross. The current genetic mapping method for a trait with repeat records is adopted by simply replacing the phenotype by the average value of the repeat records. This simple treatment has not sufficiently utilized the information from the replication and ignored the impacts of the permanent environmental effects on the accuracy of the estimated QTL.

Results: We propose to map QTL by using the repeatability model to directly analyze the repeat records rather than simply analyze the mean phenotype, improving the efficiency of QTL detecting because of adequately utilizing the information from data and allowing for the permanent environmental effects. A maximum likelihood method implemented via the expectation-maximization (EM) algorithm is applied to perform the parameter estimation of the repeatability model. The superiority of the mapping method based on the repeatability model over simple analysis using the mean phenotype was demonstrated by a series of simulations.

Conclusion: Our results suggest that the proposed method can serve as a powerful alternative to existing methods. By mean of the repeatability model, utilizing the repeat records on individual may improve the efficiency of QTL detecting in line cross.

Background

Replication is the fundamental of the experimental design, the important advantages of which are that it allows for an estimate of experimental error and increases the reliability of information obtained at each experimental point [1,2]. Replication denotes sampling or measuring multiple times under the same experimental condition (within one treatment), where the experimental unit may be either one individual or multiple individuals with the identical genetic background.

Often plants or animals are observed more than once for a particular trait. For examples, fleece weight of sheep in different years, blood pressure and pulse of a human over time, litter size of sows over time, antler size of deer in different seasons, racing results of horses from several races, exam scores of students during university and so on. These records observed belong to replicate ones if they are not influenced by the measuring environments, such as the years, seasons, parities, races.

In classical quantitative genetics, a trait with repeat records is generally analysed by means of the repeatability

model [3,4], in which, there is an additional permanent environmental effect besides an individual's additive genetic value for a trait. The permanent environmental effect as a measure of the differences among experimental units, is a non-genetic effect common to all observations on the same individual [5]. Such environmental effects are usually accounted for in the model to ensure accurate prediction of breeding values [4]. However, the repeatability model has not been paid adequate attention to mapping QTL by using data with repeat records.

The current genetic mapping method for a trait with repeat records is adopted by simply replacing the phenotype by the average value of the repeat records [6,7]. This simple treatment has not sufficiently utilized the information from the replication and ignored the impacts of the permanent environmental effects on the accuracy of the estimated QTL, although it enables to improve the power of detecting QTL with a certain extent.

In this study, we apply the repeatability model to mapping quantitative trait loci with repeat records and demonstrate the higher efficiency of this model by the simulations.

Theory and methods
Mapping QTL based on the mean phenotype
Take a simple F_2 population of size n derived from two homozygous lines as an example. There are the three possible genotypes denoted by Q_1Q_1, Q_1Q_2, and Q_2Q_2, respectively, at a quantitative trait locus Q. The phenotypic value of an individual i is usually described by the following linear model,

$$y_i = \mu + z_i a + w_i d + e_i, \tag{1}$$

Where μ is the population mean, a and d are additive and dominant effects of the QTL, e_i is the residual error with a $N(0, \sigma^2)$ distribution, and

$$z_i = \begin{cases} +1 & for\ Q_1Q_1 \\ 0 & for\ Q_1Q_2 \\ -1 & for\ Q_2Q_2 \end{cases} \quad and \quad w_i = \begin{cases} -1 & for\ Q_1Q_1 \\ +1 & for\ Q_1Q_2 \\ -1 & for\ Q_2Q_2 \end{cases}.$$

If m_i records are repeatedly sampled from each individual and the phenotypic value of an individual i is measured by the average of m_i records, the model is modified as

$$\bar{y}_i = \mu + z_i a + w_i d + \bar{e}_i, \tag{2}$$

where

$$\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}, \bar{e}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} e_{ij},$$

and the variable with additional subscript j indicates the corresponding variable for the jth record of the ith F_2 individual. The residual error now follows a $N(0, \sigma^2/m_i)$ distribution, given that $e_{ij} \sim N(0, \sigma^2)$.

Let

$$f(\bar{y}_i \mid \theta, z_i, w_i) = \sqrt{\frac{m_i}{2\pi\sigma}} \exp\left[-\frac{m_i}{2\sigma^2} (\bar{y}_i - \mu - z_i a - w_i d)^2 \right] \tag{3}$$

be the conditional density of $\bar{y}_i$, where $\theta = [\mu\ a\ d\ \sigma^2]^T$ are the parameters; the log likelihood function defined under the missing variables z_i and w_i is

$$L(\theta) = \sum_{i=1}^{n} \ln\left\{ E[f(\bar{y}_i \mid \theta, z_i, w_i)] \right\} \tag{4}$$

The expectation-maximization (EM) algorithm [8] can be used to obtain the MLE, as shown below,

$$\begin{bmatrix} \sum_{i=1}^{n} m_i & \sum_{i=1}^{n} m_i E(z_i) & \sum_{i=1}^{n} m_i E(w_i) \\ \sum_{i=1}^{n} m_i E(z_i) & \sum_{i=1}^{n} m_i E(z_i^2) & \sum_{i=1}^{n} m_i E(z_i w_i) \\ \sum_{i=1}^{n} m_i E(w_i) & \sum_{i=1}^{n} m_i E(z_i w_i) & \sum_{i=1}^{n} m_i E(w_i^2) \end{bmatrix} \begin{bmatrix} \mu \\ a \\ d \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} m_i \bar{y}_i \\ \sum_{i=1}^{n} m_i E(z_i) \bar{y}_i \\ \sum_{i=1}^{n} m_i E(w_i) \bar{y}_i \end{bmatrix} \tag{5}$$

and

$$\sigma^2 = \left[\sum_{i=1}^{n} m_i \right]^{-1} \sum_{i=1}^{n} m_i E\left[(\bar{y}_i - \mu - z_i a - w_i d)^2 \right] \tag{6}$$

The expectation shown in Equation 6 can be further expressed as

$$E[\bar{y}_i - \mu - z_i a - w_i d)^2]$$
$$= (\bar{y}_i - \mu)^2 + a^2 E(z_i^2) + d^2 E(w_i^2) - 2(\bar{y}_i - \mu)[aE(z_i) + dE(w_i)] + 2ad E(z_i w_i)$$

Define the posterior probabilities of the three QTL genotypes for jth individual as

$$p_{ik}^* = \frac{p_{ik} f(\bar{y}_i \mid \theta, z_i, w_i)}{\sum_{k=1}^{3} p_{ik} f(\bar{y}_i \mid \theta, z_i, w_i)} \quad (k = 1, 2, 3), \tag{7}$$

where p_{ik} are the conditional probabilities inferred by marker information, then

$$E(z_i) = p_{i1}^* - p_{i3}^*, E(z_i^2) = p_{i1}^* + p_{i3}^*,$$
$$E(w_i) = p_{i2}^* - p_{i1}^* - p_{i3}^*, E(w_i^2) = 1 \text{ and } E(z_i w_i) = p_{i3}^* - p_{i1}^*.$$
(8)

Because p_{ik}^* is a function of the unknown parameters, iterations are required for EM algorithm. The iterations are described as

Step 0: Set up initials for $\theta^{(0)}$.

Step 1: Calculate the posterior probabilities p_{ik}^* with equation (7).

Step 2: Substituting (8) into equation (5), estimate

$$\mu^{(1)} = \left[\sum_{i=1}^n m_i\right]^{-1}\left[\sum_{i=1}^n m_i \bar{y}_i - a^{(0)}\sum_{i=1}^n m_i(p_{i1}^* - p_{i3}^*) - d^{(0)}\sum_{i=1}^n m_i(p_{i2}^* - p_{i1}^* - p_{i3}^*)\right]$$
$$a^{(1)} = \left[\sum_{i=1}^n m_i(p_{i1}^* + p_{i3}^*)\right]^{-1}\left[\sum_{i=1}^n m_i(p_{i1}^* - p_{i3}^*)\bar{y}_i - \mu^{(1)}\sum_{i=1}^n m_i(p_{i1}^* - p_{i3}^*) - d^{(0)}\sum_{i=1}^n m_i(p_{i3}^* - p_{i1}^*)\right]$$
$$d^{(1)} = \left[\sum_{i=1}^n m_i\right]^{-1}\left[\sum_{i=1}^n m_i(p_{i2}^* - p_{i1}^* - p_{i3}^*)\bar{y}_i - \mu^{(1)}\sum_{i=1}^n m_i(p_{i2}^* - p_{i1}^* - p_{i3}^*) - a^{(1)}\sum_{i=1}^n m_i(p_{i3}^* - p_{i1}^*)\right]$$
(9)

Step 3: Substituting (8) into equation (6), estimate

$$\sigma^{2(1)} = \left[\sum_{i=1}^n m_i\right]^{-1}\left\{\sum_{i=1}^n m_i\left[p_{i1}^*\left(\bar{y}_i - \mu^{(1)} - a^{(1)} + d^{(1)}\right)^2 + p_{i2}^*\left(\bar{y}_i - \mu^{(1)} - d^{(1)}\right)^2 + p_{i3}^*\left(\bar{y}_i - \mu^{(1)} + a^{(1)} + d^{(1)}\right)^2\right]\right\}$$

Step 4: Go to step 1, which complete one round of iteration.

Mapping QTL based on the repeatability model

Partitioning residual error e_i in model (1) into an individual-specific permanent environmental effect ζ_i and random environmental effect ε_{ij}, the jth phenotypic value of an individual i is represented as

$$y_{ij} = \mu + z_i a + w_i d + \zeta_i + \varepsilon_{ij} \qquad (10)$$

This is a mixed effects model, also called repeatability model, with a and d being treated as the fixed effects and p_i as the random effect. i.i.d. $N(0, \sigma_\zeta^2)$ distribution and ε_{ij} i.i.d. $N(0, \sigma_\varepsilon^2)$ distribution.

We use an $m_i \times 1$ vector $y_i = [y_{i1} y_{i2} \dots y_{im}]^T$, for $n = 1, 2, \dots, n$ to denote the array of phenotypic values for the ith individual and define $\varphi_i = [1\ 1\ \dots\ 1]^T$ as a vector of dimension m_i. In matrix notation, model (9) can be written as

$$y_i = \varphi_i\mu + z_i\varphi_i a + w_i\varphi_i d + \varphi_i\zeta_i + \varepsilon_i \qquad (11)$$

where $\varepsilon_i = [\varepsilon_{i1}\ \varepsilon_{i2}\ \dots\ \varepsilon_{im}]^T$ is an $m_i \times 1$ vector for the random environmental effects which follows $N(0, I_i, \sigma_\varepsilon^2)$ with I_i being an $(m_i \times 1) \times (m_i \times 1)$ identity matrix. The conditional expectation of model (11) given the fixed effects is

$$E(y_i) = M_i = \varphi_i\mu + z_i\varphi_i a + w_i\varphi_i d \qquad (12)$$

and the variance-covariance matrix is

$$Var(y_i) = V_i = \phi_i\phi_i^T\sigma_\zeta^2 + I_i\sigma_\varepsilon^2 \qquad (13)$$

which applies to all $i = 1, 2, \dots, n$.

The conditional density of y_i based on M_i and V_i is

$$f(y_i \mid \theta, z_i, w_i) = \frac{1}{\sqrt{2\pi}|V_i|^{1/2}}\exp\left[-\frac{1}{2}(y_i - M_i)^T V_i^{-1}(y_i - M_i)\right]$$
(14)

where $\theta = [\mu\ a\ d\ \sigma_p^2\ \sigma_\varepsilon^2]$. Corresponding log-likelihood function defined is

$$L(\theta) = \sum_{i=1}^n \ln\{E[f(y_i \mid \theta, z_i, w_i)]\} \qquad (15)$$

With derivative for μ, a and d, we can obtain

$$\begin{bmatrix} \sum_{i=1}^n \phi_i^T V_i^{-1}\phi_i & \sum_{i=1}^n E(z_i)\phi_i^T V_i^{-1}\phi_i & \sum_{i=1}^n E(w_i)\phi_i^T V_i^{-1}\phi_i \\ \sum_{i=1}^n E(z_i)\phi_i^T V_i^{-1}\phi_i & \sum_{i=1}^n E(z_i^2)\phi_i^T V_i^{-1}\phi_i & \sum_{i=1}^n E(z_i w_i)\phi_i^T V_i^{-1}\phi_i \\ \sum_{i=1}^n E(w_i)\phi_i^T V_i^{-1}\phi_i & \sum_{i=1}^n E(z_i w_i)\phi_i^T V_i^{-1}\phi_i & \sum_{i=1}^n E(w_i^2)\phi_i^T V_i^{-1}\phi_i \end{bmatrix}\begin{bmatrix} \mu \\ a \\ d \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n \phi_i^T V_i^{-1} y_i \\ \sum_{i=1}^n E(z_i)\phi_i^T V_i^{-1} y_i \\ \sum_{i=1}^n E(w_i)\phi_i^T V_i^{-1} y_i \end{bmatrix},$$
(16)

but the explicit equations for σ_ζ^2 and σ_ε^2 can not be derived in the same way. Instead of above likelihood function, we construct the following likelihood function by using joint conditional density of y_i,

$$L(\theta) = \sum_{i=1}^n \ln\{E[f(y_i \mid \theta_1, z_i, w_i)g(\zeta_i \mid \sigma_\zeta^2)]\} \qquad (17)$$

Where $\theta_1 = [\mu\ a\ d\ \zeta_i\ \sigma_\varepsilon^2]$

$$f(y_i \mid \theta_1, z_i, w_i) = \frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\exp\left[-\frac{1}{2\sigma_\varepsilon^2}(y_i - M_i - \phi_i\zeta_i)^T(y_i - M_i - \phi_i\zeta_i)\right]$$

$$g(\zeta_i \mid \sigma_\zeta^2) = \frac{1}{\sqrt{2\pi}\sigma_\zeta}\exp\left(-\frac{\zeta_i^2}{2\sigma_\zeta^2}\right)$$

With derivative for θ_1, we obtain

$$
\begin{bmatrix}
\sum_{i=1}^{n} m_i & \sum_{i=1}^{n} m_i E(z_i) & \sum_{i=1}^{n} m_i E(w_i) & m_1 & \cdots & m_n \\
\sum_{i=1}^{n} m_i E(z_i) & \sum_{i=1}^{n} m_i E(z_i^2) & \sum_{i=1}^{n} m_i E(z_i w_i) & m_1 E(z_1) & \cdots & m_n E(z_i) \\
\sum_{i=1}^{n} m_i E(w_i) & \sum_{i=1}^{n} m_i E(z_i w_i) & \sum_{i=1}^{n} m_i E(w_i^2) & m_1 E(w_1) & \cdots & m_n E(w_i) \\
m_1 & m_1 E(z_1) & m_1 E(w_1) & m_1 + \frac{\sigma_\varepsilon^2}{\sigma_\zeta^2} & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
m_n & m_n E(z_i) & m_n E(w_i) & 0 & \cdots & m_n + \frac{\sigma_\varepsilon^2}{\sigma_\zeta^2}
\end{bmatrix}
\begin{bmatrix} \mu \\ a \\ d \\ \zeta_1 \\ \cdots \\ \zeta_n \end{bmatrix}
=
\begin{bmatrix}
\sum_{i=1}^{n} m_i \bar{y}_i \\
\sum_{i=1}^{n} m_i E(z_i) \bar{y}_i \\
\sum_{i=1}^{n} m_i E(w_i) \bar{y}_i \\
m_1 \bar{y}_1 \\
\cdots \\
m_n \bar{y}_n
\end{bmatrix}
$$

$$(18)$$

and

$$
\sigma_\varepsilon^2 = \frac{1}{\sum_{i=1}^{n} m_i} \sum_{i=1}^{n} E[(y_i - \phi_i \mu - z_i \phi_i a - w_i \phi_i d - \phi_i \zeta_i)^T (\tilde{y}_i - \phi_i \mu - z_i \phi_i a - w_i \phi_i d - \phi_i \zeta_i)]
$$

$$(19)$$

Where

$$
E[(y_i - \phi_i \mu - z_i \phi_i a - w_i \phi_i d - \phi_i \zeta_i)^T (y_i - \phi_i \mu - z_i \phi_i a - w_i \phi_i d - \phi_i \zeta_i)]
$$
$$
= (y_i - \phi_i \mu)^T (y_i - \phi_i \mu)^2 + m_i a^2 E(z_i^2) + m_i d^2 E(w_i^2) - 2(y_i - \phi_i \mu)^T [\phi_i a E(z_i) + \phi_i d E(w_i)] + 2 m_i a d E(z_i w_i) + m_i E(\zeta_i^2)
$$

$$(20)$$

so, we can simply utilize existing mixed model EM algorithm to find the MLE of parameters [9]. Followings are the EM steps for the mixed model analysis.

Step 0: Initialize all parameters with values in their legal domain, denoted by $\theta^{(0)}$.

Step 1: Compute the posterior probabilities of the three genotypes for each individual

$$
p_{ik}^* = \frac{p_{ik} f(y_i \mid \boldsymbol{q_1}, z_i, w_i)}{\sum_{k=1}^{3} p_{ik} f(y_i \mid \boldsymbol{q_1}, z_i, w_i)} \quad \text{for } k = 1, 2, 3
$$

$$(21)$$

Step 2: Compute all the expectations involved in the following maximization steps (same with the equation (8)).

Step 3: Find the posterior distribution of the random effect p_i from equation (18). This posterior distribution turns out to be a mixture of three normal distributions with a mean

$$
E(\zeta_i) = \sigma_p^2 \phi V^{-1} \left[y_i - \phi \mu - (p_{i1}^* - p_{i3}^*) \phi a - (p_{i2}^* - p_{i1}^* - p_{i3}^*) \phi d \right]
$$

$$(22)$$

and a variance

$$
Var(\zeta_i) = \sigma_\zeta^2 - \phi_i V^{-1} \phi_i^T \sigma_\zeta^4 \tag{23}
$$

Step 4: Update the population mean, additive effect and dominance effect by equation (16). The resulting equa-

tions are equivalent to equations (9) replacing m_i with $\phi^T V_i^{-1} \phi$.

Step 5: Update the covariance matrix of the random effect

$$
\sigma_\zeta^{2(1)} = \frac{1}{n} \sum_{i=1}^{n} E(\zeta_i^2) = \frac{1}{n} \sum_{i=1}^{n} \left[Var(\zeta_i) + E^2(\zeta_i) \right]
$$

$$(24)$$

Step 6: Update the residual variance by equation (19)

$$
\sigma^2 = \left[\sum_{i=1}^{n} m_i \right]^{-1} \left\{ \sum_{i=1}^{n} m_i \left[p_{i1}^* (y_i - \mu - a + d - \phi_i \zeta_i)^T (y_i - \mu - a + d - \phi_i \zeta_i) + \right. \right.
$$
$$
+ p_{i2}^* (\bar{y}_i - \mu - d - \phi_i \zeta_i)^T (\bar{y}_i - \mu - d - \phi_i \zeta_i)
$$
$$
\left. \left. + p_{i3}^* (\bar{y}_i - \mu + a + d - \phi_i \zeta_i)^T (\bar{y}_i - \mu + a + d - \phi_i \zeta_i) \right] \right\}
$$

Step 7: Repeat from step 1 to step 6 until a certain convergence criterion is reached.

MLE of parameters in both model (2) and (10) are iteratively solved at specific location on chromosomes using EM algorithm and the QTL position and effects are determined by means of likelihood ratio statistics in chromosome or genome scanning.

Simulation studies

A series of simulation experiments were used to compare the efficiency and behaviour of two mapping methods based on the repeatability model with simple analysis using the mean phenotype for a trait with repeat records. We simulated a single chromosome of 100 cM long with 11 evenly spaced codominant markers for an F2 population with sample size $n = 100$ and a single QTL was put at position 25 cM (between markers 3 and 4). Under the null model, the QTL was assigned a value of zero for both the additive and dominance effects. The empirical critical values of likelihood ratio statistics for testing the presence of the QTL were obtained by simulating 1000 replicates. Under the alternative model, nonzero and equal additive and dominance effects were simulated. The simulations were replicated 100 times. Empirical power was calculated by counting the number of runs in which test statistics were greater than the critical values.

Factor considered include the QTL size, measured as the proportion of the phenotypic variance explained by the QTL (also called the QTL heritability), the number of replicates and $\sigma_\zeta^2 : \sigma_\varepsilon^2$ i.e the variance ratio of permanent environmental effect to random environmental effect. The QTL size was set at three levels: $a = d = 0.265, 0.577, 0.943$ correspond to the three levels of $h^2 = 0.05, 0.10, 0.20$

Table 1: Effects of the number of replications on the mapping analysis based on the repeatability model

								Estimate			
a and d	h^2	Replicate	Power	Position	a	d	h^2	σ_p^2	σ_ε^2	LOD
0.4189	0.05	3	37	28.65(1.946)	0.615(0.061)	0.594(0.028)	0.116(0.007)	2.183(0.069)	2.544(0.046)	15.31(0.810)
		5	46	27.63(1.506)	0.626(0.291)	0.545(0.226)	0.108(0.041)	2.224(0.050)	2.555(0.018)	15.39(0.493)
		10	63	26.55(0.923)	0.698(0.224)	0.661(0.178)	0.134(0.042)	2.375(0.037)	2.548(0.014)	18.85(0.535)
		15	86	26.71(1.441)	0.496(0.028)	0.544(0.190)	0.093(0.003)	2.294(0.035)	2.532(0.010)	14.85(0.386)
0.6086	0.10	3	80	25.98(0.635)	0.724(0.036)	0.670(0.025)	0.142(0.006)	2.239(0.049)	2.605(0.030)	18.18(0.635)
		5	83	26.55(0.922)	0.699(0.022)	0.661(0.018)	0.134(0.042)	2.375(0.034)	2.548(0.014)	18.85(0.535)
		10	87	26.43(0.640)	0.650(0.024)	0.668(0.017)	0.130(0.004)	2.310(0.029)	2.547(0.010)	20.09(0.538)
		15	93	25.78(0.678)	0.637(0.232)	0.632(0.138)	0.119(0.037)	2.412(0.028)	2.537(0.078)	18.77(0.515)
0.9129	0.20	3	99	24.04(0.507)	0.937(0.038)	0.960(0.023)	0.223(0.007)	2.322(0.051)	2.643(0.028)	29.25(0.928)
		5	99	24.89(0.319)	0.881(0.027)	0.922(0.015)	0.207(0.005)	2.374(0.024)	2.636(0.013)	29.12(0.640)
		10	100	24.90(0.330)	0.949(0.025)	0.917(0.157)	0.213(0.005)	2.397(0.027)	2.573(0.010)	32.86(0.655)
		15	100	25.16(0.359)	0.914(0.241)	0.884(0.014)	0.120(0.005)	2.473(0.030)	2.569(0.008)	31.26(0.651)

h^2 is the proportion of phenotypic variance explained by the QTL. The variance ratio of permanent environmental effect to random environmental effect is fixed as 1:1. Standard deviations are in parentheses.

respectively. The number of replicates was examined at five levels: m = 1, 3, 5, 10, 15, and $\sigma_\zeta^2 : \sigma_\varepsilon^2$ = 1:4, 2:3, 2.5:2.5, 3:2, 4:1, remaining $\sigma_\zeta^2 + \sigma_\varepsilon^2$ = 5.0.

The jth phenotypic value of individual i was simulated by using the repeatability model:

$$\gamma_{ij} = \mu + z_i a + w_i d + \xi_i \sigma_\zeta + \eta_{ij} \sigma_\varepsilon \qquad (25)$$

Where both ξ_i and η_{ij} are the random numbers from standard normal distribution.

The results of all simulations consistently show that under the same experimental condition, (1) using the repeatability model can significantly increase the statistical power

of QTL detecting compared with simple analysis using the mean phenotype, (2) the position and effects of QTL, especially the proportion of phenotypic variance contributed by QTL were more accuracy estimated by using the repeatability model than using the genetic mapping model without permanent environmental effects to analyze mean phenotype. The superiority of the repeatability model over the simple analysis using the mean phenotype performs in evidence under the condition of the low QTL heritability.

The effects of number of replications on the efficiency and behaviour of the two methods were investigated only at variance ratio of permanent environmental effect to random environmental effect of 1:1. The results of simulations were listed in Table 1 and 2, respectively, by different

Table 2: Effects of the number of replications on the simple analysis using the mean phenotype

								Estimate		
a and d	h^2	Replicate	Power	Position	a	d	h^2	σ^2	LOD
0.4189	0.05	1	21	38.36(5.188)	0.673(0.101)	0.505(0.097)	0.096(0.004)	4.266(0.139)	17.00(0.812)
		3	34	26.65(0.888)	0.559(0.317)	0.560(0.207)	0.197(0.007)	2.189(0.034)	15.39(0.493)
		5	42	29.50(1.026)	0.653(0.294)	0.541(0.296)	0.178(0.059)	2.743(0.512)	17.34(0.529)
		10	56	27.04(0.954)	0.719(0.241)	0.694(0.180)	0.218(0.061)	2.911(0.337)	20.92(0.559)
		15	81	26.16(1.521)	0.496(0.029)	0.560(0.021)	0.173(0.005)	2.469(0.038)	16.87(0.416)
0.6086	0.10	1	57	23.89(1.774)	0.767(0.050)	0.777(0.040)	0.120(0.039)	4.785(0.082)	17.22(0.606)
		3	78	25.39(0.660)	0.661(0.024)	0.639(0.016)	0.234(0.063)	2.256(0.027)	23.31(0.531)
		5	81	27.04(0.954)	0.719(0.241)	0.694(0.018)	0.219(0.061)	2.911(0.034)	20.92(0.559)
		10	84	26.23(0.602)	0.667(0.245)	0.683(0.167)	0.223(0.065)	2.600(0.279)	21.65(0.564)
		15	87	25.79(0.672)	0.652(0.233)	0.647(0.147)	0.211(0.060)	2.586(0.026)	20.77(0.529)
0.9129	0.20	1	97	25.21(0.563)	1.003(0.043)	0.970(0.030)	0.208(0.005)	4.800(0.082)	23.44(0.725)
		3	100	25.10(0.302)	0.909(0.015)	0.916(0.015)	0.357(0.007)	2.311(0.025)	38.04(0.773)
		5	99	25.00(0.305)	0.886(0.027)	0.930(0.016)	0.306(0.007)	2.974(0.033)	30.93(0.653)
		10	100	25.08(0.307)	0.952(0.026)	0.932(0.016)	0.335(0.007)	2.689(0.027)	34.94(0.673)
		15	99	25.07(0.288)	0.929(0.025)	0.914(0.015)	0.330(0.007)	2.659(0.026)	33.65(0.678)

h^2 is the proportion of phenotypic variance explained by the QTL. The variance ratio of permanent environmental effect to random environmental effect is fixed as 1:1. Standard deviations are in parentheses.

mapping method. Notices that the simulated results at *m* = 1 (no replication) only correspond to the mapping method based on the mean phenotype for no solution by using the repeatability model. As expected, the statistical power of QTL detecting with replication is higher than no replication, based on either the mean phenotype or the repeatability model. The estimation of QTL parameters show a general tendency to improve as the number of replications increases.

We have also investigated the impact of the variance ratio of permanent environmental effect to random environmental effect on differences in mapping performance between the two methods. The results of simulations fixing five replications were listed in Table 3. The difference in variance between permanent environmental effect and random environmental effect is greater under fixing total variance of random effects, the superiority of the mapping method based on the repeatability model over the mean phenotype is clearer in the statistical power of QTL detecting. The possible reasons are that either the large variance of random environmental effect made reliability of the individual's mean phenotype value low or the variance of residual error in model (2) increases with the variance of permanent environmental effect increased.

Discussion

For a trait with repeat records, we proposed use of the repeatability model to map QTL, which distinguishes from simple analysis using the mean phenotype not only in the data analyzed but essentially in the model adopted. Simple analysis using the mean phenotype was based on regular genetic model for mapping QTL in linecross, which excluded the permanent environmental effects. The excluded permanent environmental effects were deposited to the residual error, decreasing the accuracy of estimation for QTL parameters, which was strictly proved in the relevant books to statistic models [e.g., [10,11]]. Of course, the loss of data information has also influenced the performance of mapping QTL based on the mean phenotype.

Replication required either the experimental conditions must be the same when multiple records were observed only from one individual or the genetic backgrounds must be the identical for each individual while those records were from multiple individuals. If the former was not satisfied, then such "repeat" records observed became longitudinal data, such as test-day records of milk production and body weight in cattle, were genetically analysed using the random regression model which is essentially the repeatability model nested submodels of time [12-14]. Besides cloned individuals and progencies from each plant in RIL, the later was hard to be satisfied. For example, there were incompletely same genetic backgrounds among individuals within a family and F3 progenies from one F2 individual. To improve the efficiency of detecting QTL using such data, the genetic backgrounds should be

Table 3: Comparisons of the mapping analysis based on the repeatability model with the simple analysis using the mean phenotype under the conditions of different the variance ratios of permanent environmental effects to random environmental effects

h^2	$\sigma_p^2 : \sigma_\varepsilon^2$	Method	Power	Position	a	d	h^2	σ_p^2	σ_ε^2 or σ^2	LOD
0.05	1: 4	Repeat	72	25.59(0.919)	0.471(0.021)	0.494(0.013)	0.076(0.003)	0.889(0.022)	4.080(0.025)	17.27(0.484)
		Mean	60	25.52(0.870)	0.491(0.024)	0.507(0.015)	0.199(0.006)		1.711(0.234)	19.68(0.525)
	2: 3	Repeat	44	27.59(1.662)	0.544(0.032)	0.540(0.022)	0.098(0.006)	1.776(0.039)	3.023(0.024)	15.79(0.495)
		Mean	42	26.47(1.415)	0.549(0.030)	0.527(0.027)	0.177(0.006)		2.418(0.039)	17.39(0.500)
	3: 2	Repeat	38	24.97(1.441)	0.607(0.034)	0.576(0.023)	0.110(0.004)	2.745(0.060)	2.048(0.018)	14.54(0.389)
		Mean	37	25.76(1.456)	0.601(0.034)	0.585(0.026)	0.162(0.006)		3.158(0.054)	16.16(0.452)
	4: 1	Repeat	33	30.57(2.141)	0.668(0.050)	0.558(0.035)	0.128(0.063)	3.604(0.067)	1.007(0.010)	14.04(0.437)
		Mean	26	30.62(2.321)	0.717(0.051)	0.598(0.043)	0.171(0.007)		3.694(0.071)	16.74(0.570)
0.10	1: 4	Repeat	97	25.01(0.408)	0.643(0.019)	0.622(0.013)	0.115(0.036)	0.917(0.023)	4.042(0.019)	25.05(0.586)
		Mean	94	24.93(0.411)	0.648(0.020)	0.628(0.013)	0.267(0.007)		1.765(0.023)	26.56(0.600)
	2: 3	Repeat	86	26.32(0.815)	0.667(0.022)	0.635(0.014)	0.122(0.033)	1.890(0.030)	3.085(0.015)	19.28(0.440)
		Mean	84	26.34(0.827)	0.669(0.023)	0.643(0.014)	0.215(0.054)		2.518(0.029)	20.80(0.459)
	3: 2	Repeat	83	25.53(0.612)	0.655(0.028)	0.679(0.017)	0.137(0.004)	2.718(0.035)	2.079(0.013)	17.76(0.422)
		Mean	83	25.73(0.750)	0.659(0.029)	0.689(0.018)	0.199(0.006)		3.145(0.033)	19.37(0.451)
	4: 1	Repeat	64	25.14(1.043)	0.703(0.029)	0.686(0.018)	0.143(0.004)	3.812(0.051)	1.007(0.007)	16.27(0.430)
		Mean	61	25.44(0.997)	0.725(0.032)	0.751(0.022)	0.192(0.006)		3.898(0.051)	18.32(0.437)

The "Estimate" header spans the Power, Position, a, d, h^2, σ_p^2, σ_ε^2 or σ^2, and LOD columns.

h^2 is the proportion of phenotypic variance explained by the QTL. There are 100 individuals each with five records. Standard deviations are in parentheses.

at least taken into account in the analysis [7], furthermore, the repeatability model may be a good choice for directly analyzing such "repeat" records.

Although we demonstrate the statistical method of QTL mapping using a F_2 population as an example, other more simple or complex designs, such as backcross population and full-sib family can also be extended. Assuming only one QTL in the model considered here is to conveniently investigate efficiency of presented method based on various estimates. If a trait is controlled by multiple loci, the composite interval mapping [15,16] or Bayesian mapping [e.g., [17,18]] will be proposed for mapping those QTLs by incorporating marker-cofactors outside the scanning interval or all the QTLs into the model (9).

Authors' contributions

RQY coordinated the study, developed the foundational principle of the method and wrote the computing program and the paper. FM was responsible for the simulation experiment and carried out the analysis of results.

References

1. Fisher RA: **The design of experiments.** 9th edition. New York, Hafner Publishing Company; 1971.
2. Steel RGD, Torrie JH: **Principles and procedures of statistics: a biometrical approach.** 2nd edition. Tokyo, McGraw-Hill Kogakusha; 1980.
3. Henderson CR: **Applications of Linear Models in Animal Breeding.** Guelph ON Univ of Guelph; 1984.
4. Mrode RA: **Linear Models for the Prediction of Animal Breeding Values.** UK, CAB International; 1996.
5. Falconer DS: **Introduction to Quantitative Genetics.** London,Oliver & Boyd; 1960.
6. Zhang TY, Yuan J, Yu W, Guo Z, Kohel RJ: **Molecular tagging of a major QTL for fiber strong in upland cotton and its marker-assisted selection.** *Theor Appl Genet* 2003, **106**:262-268.
7. Zhang Y, Xu S: **Mapping Quantitative Trait Loci in F2 Incorporating Phenotypes of F3 Progeny.** *Genetics* 2004, **166**:1981-1993.
8. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via EM algorithm.** *J R Stat Soc Ser B* 1977, **39**:1-38.
9. Henderson CR: **Recent developments in variance and covariance estimation.** *J Anim Sci* 1986, **63**:208-216.
10. Zar JH: **Biostatistical Analysis.** 3rd edition. Prentice Hall; 1996.
11. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W: *Applied Linear Statistical Models* 4th edition. RD Irwin, Homewood, IL; 1996.
12. Henderson CR: **Analysis of covariance in the mixed model: Higher level, no homogenous, and random regressions.** *Biometrics* 1982, **38**:623-640.
13. Schaeffer LR: **Application of random regression model in animal breeding.** *Livest Prod Sci* 2004, **86**:35-45.
14. Macgregor S, Knott SA, White I, Visscher PM: **Quantitative trait locus analysis of longitudinal quantitative trait data in complex pedigrees.** *Genetics* 2005, **171**:1365-1376.
15. Jansen RC: **Controlling the type I and type II errors in mapping quantitative trait loci.** *Genetics* 1994, **138**:871-881.
16. Zeng ZB: **Precision mapping of quantitative trait loci.** *Genetics* 1994, **136**:1457-1468.
17. Satagopan JM, Yandell BS, Newton MA, Osborn TC: **A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo.** *Genetics* 1996, **144**:805-816.
18. Yi N, Xu S: **Bayesian mapping of quantitative trait loci under complicated mating designs.** *Genetics* 2001, **157**:1759-1771.