RESEARCH ARTICLE

Mapping quantitative trait loci for binary trait in the $F_{2:3}$ design

CHENGSONG ZHU[1], YUAN-MING ZHANG[1]* and ZHIGANG GUO[2]

[1]*Section on Statistical Genomics, State Key Laboratory of Crop Genetics and Germplasm Enhancement/National Center for Soybean Improvement, College of Agriculture, Nanjing Agricultural University, Nanjing 210095, People's Republic of China*

[2]*Department of pathology, Kansas State University, Manhattan, Kansas 66506, USA*

Abstract

In the analysis of inheritance of quantitative traits with low heritability, an $F_{2:3}$ design that genotypes plants in F_2 and phenotypes plants in $F_{2:3}$ progeny is often used in plant genetics. Although statistical approaches for mapping quantitative trait loci (QTL) in the $F_{2:3}$ design have been well developed, those for binary traits of biological interest and economic importance are seldom addressed. In this study, an attempt was made to map binary trait loci (BTL) in the $F_{2:3}$ design. The fundamental idea was: the F_2 plants were genotyped, all phenotypic values of each $F_{2:3}$ progeny were measured for binary trait, and these binary trait values and the marker genotype informations were used to detect BTL under the penetrance and liability models. The proposed method was verified by a series of Monte–Carlo simulation experiments. These results showed that maximum likelihood approaches under the penetrance and liability models provide accurate estimates for the effects and the locations of BTL with high statistical power, even under of low heritability. Moreover, the penetrance model is as efficient as the liability model, and the $F_{2:3}$ design is more efficient than classical F_2 design, even though only a single progeny is collected from each $F_{2:3}$ family. With the maximum likelihood approaches under the penetrance and the liability models developed in this study, we can map binary traits as we can do for quantitative trait in the $F_{2:3}$ design.

[Zhu C. S., Zhang Y. M. and Guo Z. 2008 Mapping quantitative trait loci for binary trait in the $F_{2:3}$ design *J. Genet.* **87**, 201–207]

Introduction

In plants and some model animals, quantitative trait loci (QTL) mapping is commonly performed using F_2 or backcross populations derived from the cross between two inbred lines. Typically, QTL mapping statistics assumes that each F_2 individual is genotyped for the markers and phenotyped for the trait. However, the power in the detection of QTL for a trait with low heritability is relatively low. To increase the power, an $F_{2:3}$ design, in which F_2 plants are genotyped and $F_{2:3}$ progeny are phenotyped is developed. In such a design, progeny do not have to be typed for markers, thus leading to substantial cost saving. Acutally, one can arbitrarily increase the number of generations from 3 to y, leading to an $F_{2:y}$ design. It is even possible to genotype plants in generation x and phenotype plants in generation y to conduct QTL mapping. This design is called an $F_{x:y}$ design for $y \geq x$ (Fisch *et al.* 1996; Jiang and Zeng 1997; Chapman *et al.* 2003; Kao 2006). In addition, some traits, e.g. endosperm traits, can be measured only in tissues controlled by genotypes of progeny (Xu *et al.* 2003; Kao 2004; Cui and Wu 2005). Therefore, it is important to focus on the $F_{2:3}$ design.

Many characters of biological interest and economic importance vary in a dichotomous form, i.e., presence or absence, but are not inherited in a simple Mendelian fashion. These traits are called complex binary traits and are presumably controlled by a number of genetic and environmental factors, thus belonging to the category of quantitative traits (Falconer and Mackay 1996). From a theoretical point of view, however, standard QTL mapping for continuous traits cannot be applied to discrete trait mapping, and genetic analyses may be more challenging for dichotomous traits than for continuous traits, because the former requires modelling the link between the observable phenotype and the corresponding latent variable. McIntyre *et al.* (2001) developed a proba-

*For correspondence. E-mail: soyzhang@njau.edu.cn.

Keywords. binary trait; $F_{2:3}$ design; liability model; maximum likelihood; Monte–Carlo simulation; penetrance model.

bility model particularly suitable for binary trait mapping by handling the probabilities of disease (penetrance) as parameters of interest. So the disease penetrance difference among the alternative genotypes may be estimated and tested. Under the liability model, Xu *et al.* (2003) proposed a new Expectation–Maximization (EM) algorithm by treating both the unobserved genotypes and the disease liability as missing values. However, there are seldom studies on mapping binary trait loci (BTL) in the specific design like the $F_{2:3}$ design.

Taking full advantage of the mixture distribution for $F_{2:3}$ families of heterozygous F_2 plants, Zhang and Xu (2004) and Zhu *et al.* (2007) showed that the $F_{2:3}$ design can be used significantly to increase the power of QTL detection relative to the classic F_2 design, even if only a single F_3 progeny is collected from each $F_{2:3}$ family. However, the sum of phenotypic values of $F_{2:3}$ progeny derived from each F_2 plant were analysed in Zhu *et al.* (2007). If all phenotypic values of each $F_{2:3}$ family are measured, large sample results in a high resolution of the genetic analysis. If all the phenotypic values are replaced by the corresponding averages, some information will be lost. Therefore, it is necessary to use all phenotypic values in the $F_{2:3}$ design to map BTL under both the penetrance and the liability models. The new approaches developed in this article are compared with those for analysing a continuously distributed quantitative trait in the F_2 and $F_{2:3}$ designs, and verified by a series of Monte–Carlo simulation studies.

Theory and methods

Genetic model of $F_{2:3}$

We consider cosegregation of a gene at the autosomal locus that affects a dichotomous trait. The phenotype of the trait is assumed to be distributed as a binary variable. The phenotype of the *j*th individual within the *i*th family is modelled by

$$y_{ij} = \begin{cases} 1 & \text{if } z_{ij} \geq t \\ 0 & \text{if } z_{ij} < t \end{cases}, \tag{1}$$

where *t* is the threshold for the underlying liability z_{ij} of the trait which is formulated as,

$$z_{ij} = g_{ij} + \varepsilon_{ij}, \tag{2}$$

in which g_{ij} is the genotypic value of the individual at the trait locus, and ε_{ij} a normally distributed residual variable with mean zero and standard deviation 1.0, which accounts both for polygenes that are linked to the markers and for environmental variation. Three genotypes at this locus, say *AA*, *Aa*, and *aa*, assumed to have genotypic values $a - d/2$, $d/2$ and $-a - d/2$, respectively, with *a* and *d* being additive and dominant effects. Thus, the genetic effects of the trait locus are measured in units of the residual deviations of the liability.

The liability of each genotype will follow a truncated distribution with a cumulative probability indicated by

$$f_k = \Pr(z_{ij} \geq t | G_{ij} = k, \theta) = \Phi[(2-k)a + (-1)^k d/2 - t], \tag{3}$$

where $\theta = \{a, d, t, \delta\}$, δ is the position of detected BTL, G_{ij} is the genotype of the *j*th individual within the *i*th family; $k = 1, 2, 3$ referring, correspondingly, to the genotypes *AA*, *Aa*, and *aa* at the trait locus, and f_k is referred to as the penetrance of the *k*th genotype of the locus. $\Phi(\bullet)$ denotes the standardized normal probability function.

Maximum likelihood estimation

The log-likelihood of the observed phenotypes of the trait can be written as,

$$L(Y, \Omega) = \sum_{i=1}^{m} \sum_{j=1}^{n} 1\text{n} \left[\sum_{k=1}^{3} h_{ik} f_k^{y_{ij}} (1 - f_k)^{(1-y_{ij})} \right], \tag{4}$$

where h_{ik} is the conditional probability of the *k*th QTL genotype of the $F_{2:3}$ progeny from the *i*th $F_{2:3}$ family given flanking marker genotypes of F_2 plant (table 1), *m* is the number of plants in F_2, *n* is the number of $F_{2:3}$ progeny for each F_2 plant, and y_{ij} is defined in equation 1.

The likelihood may be analysed under two models: (i) the penetrance model, which involves unknown parameters $\Omega = (f_1, f_2, f_3)$ with f_k $(k = 1, 2, 3)$ being the penetrance of the *k*th genotype at the trait loci; (ii) the liability model in which the unknown parameters are $\Omega = (t, a, d)$.

The maximum likelihood estimates (MLEs) of the unknown parameters in equation 4 can be calculated by the use of the EM algorithm (Dempster *et al.* 1977). Implementation of the EM algorithm in the present context involves the iteration of the following two steps:

E-step: The parameters ($a^{(0)}, d^{(0)}$ and $t^{(0)}$) are initialized with zero, so $f_k = 0.5$, $k = 1, 2, 3$. Calculating the posterior probabilities w_{ijk} using the parameter estimates at the *s*th iteration for both penetrance and liability models yields:

$$w_{ijk} = h_{ik} f_k^{y_{ij}} (1 - f_k)^{(1-y_{ij})} \Big/ \sum_{0=1}^{3} h_{i0} f_0^{y_{ij}} (1 - f_0)^{(1-y_{ij})}. \tag{5}$$

M-step: For the penetrance model, the updated estimates of the three penetrance parameters can be directly obtained as

$$f_k^{(s+1)} = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ijk} y_{ij} \Big/ \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ijk}. \tag{6}$$

Once the f_1, f_2 and f_3 are obtained, the genetic parameters, $a^{(1)}$ and $d^{(1)}$, can be estimated from equation 3 using the numerical algorithm (Press *et al.* 2001) conditional on the fact that the threshold t_0 is known.

Table 1. Conditional probabilities of QTL genotypes of $F_{2:3}$ progeny given marker genotypes of F_2 plant.

Marker genotype for F_2 plant	Conditional probabilities of QTL genotypes or $F_{2:3}$ progeny		
	$P\ (AA\backslash I_M)$	$P\ (Aa\backslash I_M)$	$P\ (aa\backslash I_M)$
$M_1M_1M_2M_2$	$\frac{(1-r_1)(1-r_2)(2-2r_1-2r_2+3r_1r_2)}{2(1-r)^2}$	$\frac{r_1r_2(1-r_1)(1-r_2)}{(1-r)^2}$	$\frac{r_1r_2(1-r_1-r_2+3r_1r_2)}{2(1-r)^2}$
$M_1M_1M_2m_2$	$\frac{(1-r_1)[4r_2(1-r_1)(1-r_2)+r_1(1-2r_2+2r_2^2)]}{4r(1-r)}$	$\frac{r_1(1-r_1)(1-2r_2+2r_2^2)}{2r(1-r)}$	$\frac{r_1[4r_1r_2(1-r_2)+(1-r_1)(1-2r_2+2r_2^2)]}{4r(1-r)}$
$M_1M_1m_2m_2$	$\frac{r_2(1-r_1)[2r_2(1-r_1)+r_1(1-r_2)]}{2r^2}$	$\frac{r_1r_2(1-r_1)(1-r_2)}{r^2}$	$\frac{r_1(1-r_2)[2r_1(1-r_2)+r_2(1-r_1)]}{2r^2}$
$M_1m_1M_2M_2$	$\frac{(1-r_2)[4r_1(1-r_1)(1-r_2)+r_2(1-2r_1+2r_1^2)]}{4r(1-r)}$	$\frac{(1-2r_1+2r_1^2)r_2(1-r_2)}{2r(1-r)}$	$\frac{r_2[4r_1(1-r_1)r_2+(1-r_2)(1-2r_1+2r_1^2)]}{4r(1-r)}$
$M_1m_1M_2m_2$	$\frac{4r_1r_2(1-r_1)(1-r_2)+(1-2r+2r^2)}{4(1-2r+2r^2)}$	$\frac{(1-2r_1+2r_1^2)(1-2r_2+2r_2^2)}{2(1-2r+2r^2)}$	$\frac{4r_1r_2(1-r_1)(1-r_2)+(1-2r+2r^2)}{4(1-2r+2r^2)}$
$M_1m_1m_2m_2$	$\frac{r_2[4r_1(1-r_1)r_2+(1-r_2)(1-2r_1+2r_1^2)]}{4r(1-r)}$	$\frac{(1-2r_1+2r_1^2)r_2(1-r_2)}{2r(1-r)}$	$\frac{(1-r_2)[4r_1(1-r_1)(1-r_2)+r_2(1-2r_1+2r_1^2)]}{4r(1-r)}$
$m_1m_1M_2M_2$	$\frac{r_1(1-r_2)[2r_1(1-r_2)+r_2(1-r_1)]}{2r^2}$	$\frac{r_1r_2(1-r_1)(1-r_2)}{r^2}$	$\frac{r_2(1-r_1)[2r_2(1-r_1)+r_1(1-r_2)]}{2r^2}$
$m_1m_1M_2m_2$	$\frac{r_1[4r_1r_2(1-r_2)+(1-r_1)(1-2r_2+2r_2^2)]}{4r(1-r)}$	$\frac{r_1(1-r_1)(1-2r_2+2r_2^2)}{2r(1-r)}$	$\frac{(1-r_1)[4r_2(1-r_1)(1-r_2)+r_1(1-2r_2+2r_2^2)]}{4r(1-r)}$
$m_1m_1m_2m_2$	$\frac{r_1r_2(1-r_1-r_2+3r_1r_2)}{2(1-r)^2}$	$\frac{r_1r_2(1-r_1)(1-r_2)}{(1-r)^2}$	$\frac{(1-r_1)(1-r_2)(2-2r_1-2r_2+3r_1r_2)}{2(1-r)^2}$

r is the recombination fraction between the two markers, and r_1 and r_2 are the recombination fractions of the QTL with the two markers.

Let u_0 be the proportion of the affected individuals in the sample. Thompson (1972) suggested the use of u_0 to calculate the MLE of the threshold $t^{(1)}$. In the present context, the following equation,

$$u_0 = \sum_{k=1}^{3} \frac{\sum_{i=1}^{m} h_{ik}}{m} \frac{1}{\sqrt{2\pi}} \int_{t^{(1)}}^{\infty} \exp\left\{ -\frac{[z-(2-k)a^{(1)}-(-1)^k d^{(1)}/2]^2}{2} \right\} dz. \tag{7}$$

is searched numerically for the MLE of $t^{(1)}$ on the basis of the above estimates of the other model parameters.

As stated above, a numerical algorithm must be used to estimate the genetic parameters ($a^{(1)}$ and $d^{(1)}$). For the liability model, EM algorithm is an alternative approach for estimating the model parameters $\Omega = (t, a, d)$. Therefore, it is necessary to propose the algorithm in this study. Provided that we treat both the unobserved genotype and the disease liability as missing values (Xu *et al.* 2005), the EM procedure is as follows.

E-step: The parameters ($a^{(0)}$, $d^{(0)}$ and $t^{(0)}$) are initialized with zero, so $f_k = 0.5$, $k = 1, 2, 3$. The posterior probability of kth QTL genotype of the jth F_3 progeny from the ith $F_{2:3}$ family given flanking marker genotypes of F_2 plant can also be calculated based on equation 5.

The expectation involving disease liability z is:

$$\begin{aligned} E(z_{ij}) &= \sum_{k=1}^{3} w_{ijk} E(z_{ijk}, y_{ij}, \Omega) \\ &= \sum_{k=1}^{3} w_{ijk} \Bigg\{ (2-k)a^{(0)} + (-1)^k d^{(0)}/2 - t^{(0)} + \\ &\quad \frac{(2y_{ij}-1)\phi[(2-k)a^{(0)} + (-1)^k d^{(0)}2 - t^{(0)}]}{\Phi[(1-2y_{ij})((2-k)a^{(0)} + (-1)^k d^{(0)}2 - t^{(0)})]} \Bigg\} \end{aligned} \tag{8}$$

where $\Phi(\bullet)$ is the standardized normal density function.

M-step: For the liability model, the updated estimates of the three parameters can be directly obtained as,

$$\alpha_h = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ijh} E(z_{ij}) \Big/ \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ijh} \quad \text{for } h = 1, 2, 3. \tag{9}$$

We can then easily obtain:

$$a^{(1)} = \frac{1}{2}(\alpha_1 - \alpha_3) \quad d^{(1)} = \alpha_2 - \frac{1}{2}(\alpha_1 + \alpha_3). \tag{10}$$

The threshold t can be updated as follows:

$$\begin{aligned} t^{(1)} &= \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{3} w_{ijk} \\ &\left\{ t^{(0)} + \frac{(1-2y_{ij})\phi[t^{(0)} - (2-k)a^{(1)} - (-1)^k d^{(1)}/2]}{\Phi[(1-2y_{ij})(t^{(0)} - (2-k)a^{(1)} - (-1)^k d^{(1)}/2)]} \right\} \end{aligned} \tag{11}$$

The E-step and M-step are iterated until convergence.

We can now test the null hypothesis that there is no QTL for the particular location δ. The null hypothesis is formulated as H_0: $a = d = 0.0$, which can be tested using the likelihood-ratio (LR) test statistic,

$$\text{LR} = -2\{1\text{n}[L(0,0,\delta)] - 1\text{n}[L(a,d,\delta)]\} \tag{12}$$

The MLE for the position of the QTL can be obtained by examining the likelihood-ratio profile along the chromosome as is commonly done in interval mapping (Lander and Botstein 1989).

Simulation studies

The purpose of the simulation studies is (i) to illustrate the validation of the theoretical analyses under both the penetrance and the liability models, and (ii) to compare the power between the binary trait and the continuous trait models in the F_2 and $F_{2:3}$ designs.

Eleven equally spaced markers were simulated on a single-chromosome segment of length 100 cM. A single BTL was located at position 25 cM. In all simulations, the environmental error variance on the individual plant was set at $\sigma^2 = 1$. The conditional probabilities of BTL genotypes, given the marker information, were calculated on the basis of the multipoint method (Jiang and Zeng 1997). The broad heritability involved in the simulation experiments are all expressed on the individual F_2 plant basis. One complication from the multiple-generation problem is that different generations usually have different genetic variances and different definitions of heritability. Here, we simply defined the genetic variance and the environmental residual variance on a single plant basis, thus eliminate this complication.

Each simulation run consisted of 100 replicates. For each QTL simulated, the samples for which LOD (logarithm of odds) exceeded the threshold value were counted. The ratio of the number of detected QTL to the total number of replicates (100) represents the empirical power of the method. It should be noted that the threshold value of the test statistic used to declare statistical significance at the 5% experiment-wise type I error rate was obtained from analyses of 1000 additional samples simulated under the null model (zero heritability).

To demonstrate the first objective of the simulation experiments, under the penetrance model, we first simulated 200 $F_{2:3}$ families each with five plants. A single QTL was simulated with heritability of 0.01, 0.05 and 0.10, respectively. The maximum likelihood method was used to estimate the parameters in the penetrance model. Table 2 illustrates the means and the corresponding standard deviations of the MLEs of the model parameters, as well as the empirical powers of the LR test under the penetrance model. It shows that the parameters are well estimated by their corresponding MLEs, except for the location of the BTL when the heritability was set to 0.01. Table 2 also provides the empirical powers, on the basis of 100 simulations, for detecting BTL. There is a trend toward increase in the power as the heritability of the trait increases. Comparison among the three populations shows that the test tends to be more efficient when the trait genes are at either low or high frequencies. This may reflect the fact that the contrast of the differences among the penetrances of three genotypes is enhanced when the genes are at low or high frequencies.

Under the liability model, we then simulated 200 $F_{2:3}$ families each with five plants. A single QTL was also simulated with heritability of 0.01, 0.05 and 0.10, respectively. These methods are evaluated under two levels of threshold: 0 and 0.5. Table 3 gives the means and standard deviations of the MLEs of the parameters used in the liability model, as well as the empirical powers of the LR test. It can be seen from the table that the parameters are well predicted by their corresponding MLEs. Table 3 also shows the empirical powers on the basis of 100 simulations. Again, the observed trend is consistent with what was expected, i.e. the method performs well as the heritability increases. Comparison between tables 2 and 3 shows that the penetrance model is as efficient as the liability model.

To demonstrate the second objective of the simulation experiments, we compare the statistical power for analysing the continuous quantitative traits to those for detecting BTL in the present study in the F_2 and $F_{2:3}$ designs. We simulated 200 $F_{2:3}$ families each with only one plant. A single QTL was simulated with heritability of 0.05 and 0.10 respectively. Analysing a continuously distributed quantitative trait, we assumed that the phenotypic value of individual was known. The environmental variance in place of the threshold was estimated via EM algorithm. The corresponding results are summarized in table 4,

Table 2. Means of the maximum likelihood estimates of model parameters and their corresponding standard deviations, along with empirical power for mapping BTL under the penetrance model.

				Estimates					
f_1	f_2	f_3	h^2	Power (%)	Position (cM)	$\hat{\mathrm{f}}_1$	$\hat{\mathrm{f}}_2$	$\hat{\mathrm{f}}_3$	LOD
0.5231	0.5213	0.4309	0.01	29	34.94 (25.54)	0.5288 (0.0566)	0.5167 (0.0819)	0.4393 (0.0415)	2.36 (1.18))
0.5527	0.5527	0.3455	0.05	99	25.65 (7.99)	0.5565 (0.1496)	0.5632 (0.0684)	0.3476 (0.0341)	7.04 (2.44)
0.5763	0.5763	0.2818	0.10	100	24.11 (4.14)	0.5762 (0.2215)	0.5766 (0.0634)	0.2839 (0.0304)	13.30 (3.09)

There are 200 F_2 plants each having five F_3 progeny. Standard deviations are in parentheses.

Table 3. Effects of BTL heritability on mapping BTL under the liability model.

				Estimates					
a	d	t	h^2	Power (%)	Position (cM)	$\hat{a}$	$\hat{d}$	$\hat{t}$	LOD
0.116	0.116	0	0.01	28	37.67 (27.89)	0.1273 (0.0812)	0.0.0946 (0.2721)	0.0004 (0.0623)	2.26 (1.26)
		0.5	0.01	30	0.35.29 (28.77)	0.1211 (0.0954)	0.0939 (0.2891)	0.5001 (0.0791)	2.61 (1.11)
0.265	0.265	0	0.05	98	25.86 (8.74)	0.2629 (0.0583)	0.2714 (0.2399)	0.0009 (0.0716)	7.24 (2.39)
		0.5	0.05	96	24.13 (9.2381)	0.2600 (0.0691)	0.2609 (0.2046)	0.4937 (0.0582)	6.80 (2.66)
0.385	0.385	0	0.10	100	25.88 (4.1194)	0.3833 (0.0611)	0.3859 (0.2068)	−0.0041 (0.0671)	12.36 (3.47)
		0.5	0.10	100	24.04 (3.8971)	0.3824 (0.0773)	0.3922 (0.1641)	0.5006 (0.0594)	13.86 (3.11)

There are 200 F_2 plants each having five F_3 progeny. Standard deviations are in parentheses. $\hat{t}$ is the estimate of the threshold.

Table 4. Comparisons of results for mapping BTL and QTL in the F_2 and $F_{2:3}$ design.

					Estimates					
A	d	h^2	Method		Power (%)	Position (cM)	$\hat{a}$	$\hat{d}$	$\hat{t}^*$	LOD
0.265	0.265	0.05	F_2	BTL	25	37.19 (26.66)	0.2663 (0.1891)	0.2532 (0.5478)	0.0194 (0.1239)	1.97 (1.00)
				QTL	48	35.13 (23.05)	0.2700 (0.1467)	0.0.2633 (0.2099)	0.9558 (0.0991)	3.26 (1.52)
			$F_{2:3}$	BTL	35	33.89 (27.01)	0.2603 (0.1943)	0.2477 (0.3188)	0.0021 (0.0001)	2.85 (1.13)
				QTL	64	31.21 (21.01)	0.2594 (0.1171)	0.2626 (0.1920)	1.0002 (0.0903)	3.31 (1.47)
0.385	0.385	0.10	F_2	BTL	69	25.42 (18.68)	0.0.3820 (0.1684)	0.3893 (0.6013)	0.0107 (0.1558)	3.76 (2.39)
				QTL	90	24.97 (10.97)	0.3886 (0.1408)	0.0.3867 (0.1926)	1.0106 (0.1044)	5.61 (2.61)
			$F_{2:3}$	BTL	73	24.22 (15.33)	0.3859 (0.1329)	0.3882 (0.2182)	0.0003 (0.0002)	4.06 (1.47)
				QTL	93	25.59 (9.75)	0.3823 (0.1009)	0.3845 (0.1110)	0.9835 (0.0955)	5.70 (2.42)

There are 200 F_2 plants each having only one F_3 progeny. Standard deviations are in parentheses. *indicate either the estimates of threshold under binary data model or the environmental variance under the normal data model.

showing that the methods modelling the continuous trait phenotype are usually more powerfull than the methods analysing the binary phenotype. Comparison of the empirical powers for detecting QTL between the F_2 and $F_{2:3}$ designs shows that the $F_{2:3}$ design is more efficient than the classical F_2 design under the binary trait and continuous trait models, even if only a single progeny is collected from each $F_{2:3}$ family.

Discussion

Following the concept of the mixture distribution of the $F_{2:3}$ progeny derived from the heterozygous F_2 plants in Zhang and Xu (2004), we extended the statistical method of analysing the continuous trait to map BTL in the $F_{2:3}$ design. Although the $F_{x:y}$ design has been widely used for QTL mapping, there are few reports on analysing discrete traits. Genetic analyses with binary traits may be theoretically more challenging than with continuous traits, because the former requires modelling the link between the observable phenotype and the corresponding latent variable. As for the specified $F_{2:3}$ designs, the method presented here differs from those of Zhang and Xu (2004); Zhu *et al.* (2007). The difference in this study is that instead of summing/averaging trait scores for all F_3 plants in a family (Zhang and Xu 2004; Zhu *et al.* 2007), all the individual F_3 data are used. Obviously, the new method provides an approach for analysing all the individual F_3 data. From a theoretical point of view, our analysis is similar to the method of Xu *et al.* (2003) in spirit, but we conducted the BTL mapping, whereas, they mapped the endosperm trait loci. Although we develop the approach using the $F_{2:3}$ design as an example, it is possible to extend this approach to the $F_{x:y}$ design and some similar designs, e.g. grand-daughter design (Weller *et al.* 1990; Bovenhuis and Weller 1994; Mackinnon and Weller 1995; Ron *et al.* 2001), NC design III (Cockerham and Zeng 1996), and TTC design (Zhu *et al.* 2007), with minor modifications.

Many complex traits show a binary phenotypic distribution. Mapping loci of such traits requires methods that specifically take into account these phenotypic distributions. McIntyre *et al.* (2001) proposed a probability model in which the probabilities of disease (penetrance) are parameters of interest. However, the genetic parameters that most plant or animal breeders are interested were not estimated. Moreover, the penetrance model cannot estimate another important parameter, i.e. threshold. Under the threshold model of binary disease, Xu *et al.* (2003) developed an EM-implemented ML method by treating both the unobserved genotype and the disease liability as missing values in a four-way-cross mouse family. Yet, the $F_{2:3}$ design differs from the four-way-cross design in that the $F_{2:3}$ progeny are not genotyped. Again, the threshold was not estimated in Xu *et al.* (2003). Herein, the proportion of the affected individuals in the sample is used to estimate the threshold (Thompson 1972; Luo and Wu 2001) on the condition that both dichotomous forms, i.e. presence or absence, are observed. Once the threshold is empirically estimated, the population average in equations 2 and 3 (Zhu *et al.* 2007) may be estimated. In this study, the mean was set at zero. The estimate of the mean in Zhu *et al.* (2007) confirmed the hypothesis. Provided that only disease phenotype (case-only) is observed, for example, truncated traits, the threshold cannot be estimated (Luo *et al.* 2005) using the suggestion of Thompson (1972).

A commonly used model is the logistic model for binary traits. However, here we adopted the penetrance and the liability models. Our previous results from simulation experiments show that the estimates for logistic analysis and the liability model are close to the true values simulated for both methods, and the statistical power of the two methods is also comparable, and both follow the expected trend (Xu *et al.* 2005). Under the single BTL model, we proposed to combine the penetrance model with the liability model to simultaneously estimate the penetrance of each genotype and the genetic parameters of interest, as well as the threshold value that links the quantitative trait with binary trait. In actual BTL-mapping experiments, the number of BTL is most likely greater than one and is usually unknown. In that case, the single locus model may be used to scan for multiple BTL, like the original interval mapping procedure of Lander and Botstein (1989). In the future, it is possible to develop multiple BTL model in the $F_{2:3}$ design like Li *et al.* (2006). Moreover, it is possible to broaden the new method to include ordinal traits like Yi *et al.* (2004).

Acknowledgements

This work was supported in part by 973 programme (2006CB101708), the National Natural Science Foundation of China (No. 30470998), Specialized Research Fund for the Doctoral Program of Higher Education (20060307008), 863 programme (2006AA10Z1E5), NCET (NCET-05-0489), the 111 project (B08025), and the Talented Foundation of Nanjing Agricultural University to YMZ and, China and Jiangsu province Postdoctoral Science Foundation to CZ.

References

Bovenhuis H. and Weller J. I. 1994 Mapping and analysis of dairy cattle quantitative traits loci by maximum likelihood methodology using milk protein genes as genetic markers. *Genetics* **136**, 267–280.

Chapman A., Pantalone V. R., Ustun A., Allen F. L. and Landau-Ellisetal D. 2003 Quantitative trait loci for agronomic and seed quality traits in an F_2 and $F_{4:6}$ soybean population. *Euphytica* **129**, 387–393.

Cockerham C. C. and Zeng Z. B. 1996 Deign III with marker loci. *Genetics* **143**, 1437–1456.

Cui Y. H. and Wu R. L. 2005 Mapping genome-genome epistasis: a high-dimensional model. *Bioinformatics* **21**, 2447–2455.

Dempster A. P., Laird N. M. and Rubin D. B. 1977 Maximum likelihood from incomplete data via EM algorithm. *J. Royal Stat. Soc. B* **39**, 1–38.

Falconer D. S. and Mackay T. F. C. 1996 Introduction to Quantitative Genetics. Longman, London.

Fisch R. D., Ragot M. and Gay G. 1996 A generalization of the mixture model in the mapping of quantitative trait loci for progeny from a bi-parental cross of inbred lines. *Genetics* **143**, 571–577.

Jiang C. J. and Zeng Z. B. 1997 Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* **101**, 47–56.

Kao C. H. 2004 Multiple interval mapping for quantitative trait loci controlling endosperm traits. *Genetics* **167**, 1987–2002.

Kao C. H. 2006 Mapping Quantitative trait loci using the experimental designs of recombinant inbred populations. *Genetics* **174**, 1373–1386.

Lander E. and Botstein D. 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Li J., Wang S. C. and Zeng Z.-B. 2006 Multiple-interval mapping for ordinal traits. *Genetics* **173**, 1649–1663.

Luo L., Zhang Y. M. and Xu S. 2005 A quantitative genetics model for viability selection. *Heredity* **94**, 347–355.

Luo Z. W. and Wu C. I. 2001 Modeling linkage disequilibrium between a polymorphic marker locus and a locus affecting complex dichotomous traits in natural populations. *Genetics* **158**, 1785–1800.

McIntyre L. M., Coffman C. J. and Doerge R. W. 2001 Detection and localization of a single binary trait locus in experimental populations. *Genet. Res.* **78**, 79–92.

Mackinnon M. J. and Weller J. I. 1995 Methodology and accuracy of estimation of quantitative trait loci parameters in a half-sib design using maximum likelihood. *Genetics* **141**, 755–770.

Press W. H., Teukolsky S. A., Vetterling W. T. and Flannery B. P. 2001 Numerical Recipes. *In C++: The Art of Scientific Computing*, 2nd version. Cambridge University Press, New York.

Ron M. D., Kliger E. F., Seroussi E. and Rzra E. 2001 Multiple quantitative trait locus analysis of bovine chromosome 6 in the Israeli Holstein population by a daughter design. *Genetics* **159**, 727–735.

Thompson R. 1972 The maximum likelihood approach to the estimate of liability. *Ann. Hum. Genet.* **36**, 221–231.

Weller J. I., Kashi Y. and Soller M. 1990 Power of "daughter" and "granddaughter" designs for genetic mapping of quantitative traits in dairy cattle using genetic markers. *J. Dairy Sci.* **73**, 2525–2537.

Xu C. W., He X. H. and Xu S. Z. 2003 Mapping quantitative loci underlying triploid endosperm traits. *Heredity* **90**, 228–235.

Xu C. W., Zhang Y. M. and Xu S. Z. 2005 An EM algorithm for mapping quantitative resistance loci. *Heredity* **94**, 119–128.

Xu S. Z., Yi N. J., Burke D., Galecki A. and Miller R. 2003 An EM algorithm for mapping binary disease loci: application to *fibrosarcoma* in a four-way-cross mouse family. *Genet. Res.* **82**, 127–138.

Yi N. J., Xu S., George V., and Allison D. B. 2004 Mapping multiple quantitative trait loci for complex ordinal traits. *Behavior Genetics* **34**, 3–15

Zhang Y. M. and Xu S. Z. 2004 Mapping quantitative trait loci in F_2 incorporating phenotypes of F_3 progeny. *Genetics* **166**, 1981–1993.

Zhu C. S., Huang J. and Zhang Y. M. 2007 Mapping binary trait loci in the $F_{2:3}$ design. *J. Heredity* **98**, 337–344.

Zhu C. S. and Zhang R. M. 2007 Efficiency of triple test cross for detecting epistasis with marker information. *Heredity* **98**, 401–410.

Received 4 November 2007, in revised form 6 April 2008; accepted 15 April 2008

Published on the Web: 21 November 2008