Malaysian Journal of Mathematical Sciences 2(1): 135-148 (2008)

Evaluation of the Performance of Maximum Likelihood and Regression Approach in Quantitative Trait Loci Mapping for Trait in Binary Scale

[1]Farit M. Afendi, [1]Asep Saefuddin, [1]Totong Martono, [2]Muhamad Jusuf
[1]Department of Statistics, IPB
Jl. Meranti Wing 22 Level 4,
Kampus IPB Darmaga, Bogor - Indonesia
[2]Department of Biology, IPB
Jl. Raya Pajajaran Kampus IPB Baranangsiang,
Bogor - Indonesia
E-mail: fmafendi@ipb.ac.id

ABSTRACT

Genes or loci on chromosome underlying a quantitative trait are called Quantitative Trait Loci (QTL). Characterizing genes controlling quantitative trait on their position in chromosome and their effect on trait is through a process called QTL mapping. This research was focusing on the assessment of the performance of Maximum Likelihood (ML) and Regression (REG) approach employed in QTL mapping for binary trait by means of simulation study. The simulation study was conducted by taking into account several factors that may affect the performance of both approaches. The factors are: (1) marker density; (2) QTL effect; (3) sample size; and (4) shape of phenotypic distribution. From simulation study, it was obtained that LB and Piepho method showing similar performance in determining critical value in testing the existence of QTL for binary trait. The simulation study also indicating that both methods could be used in determining critical value in QTL mapping analysis for binary trait. In assessing the performance of ML and REG approach in QTL mapping analysis for binary trait, the two approaches showing comparable performance. As a result, in QTL mapping analysis, ML and REG approach could be used when dealing with binary trait.

Keywords: QTL mapping, binary, maximum likelihood, regression, critical value

INTRODUCTION

Background

Genes or loci on chromosome underlying a quantitative trait are called Quantitative Trait Loci (QTL). Many such traits are both important economically as well as biologically such as milk, meat or crop production. Hence, characterizing genes controlling quantitative trait on their position in chromosome and their effect on trait through a process called QTL mapping are needed. The QTL genotypes are unobserved. In addition, the environment also affects the trait making the characterization of QTL become complex.

The idea in locating QTL is if there is association among trait and DNA markers, then the QTL should located near the DNA markers. The statistical method in utilizing this association has been proposed, which were: (1) single marker (Sax, 1923); (2) interval mapping (Lander and Botstein (1989); (3) composite interval mapping (Jansen and Stam (1994) and Zeng (1994)); and (4) multiple interval mapping (Kao, Zeng, and Teasdale, 1999). However, all these methods assuming that the trait of interest is in continuous scale. On the other hand, many important traits are obtained in categorical scale, such as resistance for certain disease. If the resistance from the disease is obtained as suspect or resistance, then the trait is in binary scale, whether if the resistance scored on ordered scale varying from unaffected to dead then the trait is in ordinal scale. Another trait could also be obtained in nominal scale such as shapes and colors of flowers, fruits, and seeds in plants, as well as coat colors. From a theoretical point of view, QTL mapping method assuming continuous trait could not be applied to categorical trait.

In dealing with binary trait, Xu and Atchley (1996) proposed likelihood based method by assuming there is continuous distribution called liability underlying binary trait by means of threshold model. Similar approach proposed by Hackett and Weller (1995) in dealing with ordinal trait. On the other hand, Hayashi and Awata(2006) proposed likelihood based approach in analyzing trait in nominal scale.

During the development of statistical method in QTL mapping, the likelihood approach becomes the main approach in analyzing data. However, this approach is computationally intensive. In simplifying the computation, in the case of continuous trait, Haley and Knott (1992) proposed regression approach in interval mapping. The idea in their approach is the component of independent variable representing the QTL effect is replacing by their expected value conditional on the two markers flanked the interval. However, the regression approach in the case of categorical scales is not yet developed. Moreover, it is interesting to evaluate the performance of likelihood and regression approach in QTL mapping dealing with categorical trait.

THEORY AND METHODS

Backcross population

In a classical backcross design, the population is generated by a heterozygous F1 backcrossed to one of the homozygous parents (for example, a cross of AaQqBb x AAQQBB)(see Figure 1). The rationale behind the interval mapping can be explained using co-segregation listed in the Table 1

(Liu, 1998). As mentioned above, the QTL genotypes are unobservable, but the probability of QTL genotypes could be obtained using the information from flanking markers genotypes as listed in Table 1.

Trait in Binary Scale

Threshold model and liability

In dealing with binary trait, it is assumed that there is continuous distribution, say U, underlying binary trait, say Y, referred to as liability (Xu and Atchley, 1996). In relation between liability and binary trait (such as resistance to certain disease), it is assumed that there is threshold (γ) in the scale of liability, below which the individual has unaffected phenotype, and above which it is affected (see Figure 2).
The relation can be summarized by:

$$y_i = \begin{cases} 1; \text{if } u_i \geq y \\ 0; \text{if } u_i < y \end{cases} \qquad (1)$$

Maximum likelihood (ML) approach

Using liability model, the one-QTL ML mapping model for a backcross population can be written as:

$$u_i = \mu + bx_i^* + \varepsilon_i, \; i = 1, 2, \ldots, n \qquad (2)$$

where u_i is the liability value for individual *i*, μ is the mean, *b* is the effect of QTL Q, *xi** taking the value of 1(0) for homozygote QQ (heterozygote Qq), denotes the genotypes of Q, ε_i is environmental deviation and is assumed to follow N(0, σ^2). Since the liability is unobserved, the mean μ and variance of ε can be set at any arbitrary value (for simplicity, it is determined that $\mu = 0$ and $\sigma^2 = 1$).

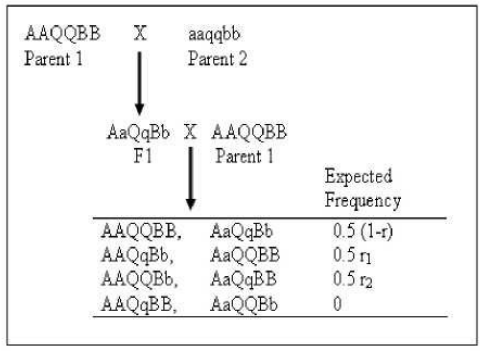

Figure 1: Conventionally defined backcross progeny for a QTL and two flanking markers.

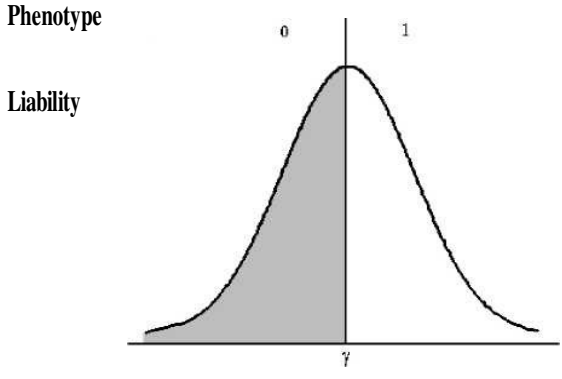

Figure 2. Liability and threshold model for binary trait

TABLE 1: Co-segregation pattern for backcross design in interval mapping

Marker Genotype	Observed Count	Frequency	QTL Genotype		Expected Value g_i
			QQ	Qq	
			Joint frequency		
AABB	N1	0.5(1-r)	0.5(1-r)	0 $0.5r_1$	
AABb	N2	0.5r	$0.5r_1$	$0.5r_2$	
AaBB	N3	0.5r	$0.5r_2$	0.5(1-r)	
AaBb	N4	0.5(1-r)	0		
			Conditional frequency		
AABB	N1	0.5(1-r)	1	0	μ_1
AABb	N2	0.5r	$r_2/r = 1-\rho$	$r_1/r = \rho$	$(1-\rho)$ μ_1+
AaBB	N3	0.5r	$r_1/r = \rho$	$r_2/r = 1-\rho$	$\rho\mu_1+$ $(1-\rho)\mu_2$
AaBb	N4	0.5(1-r)	0	1	μ_2
Mean		0.25	μ_1	μ_2	$0.5(\mu 1+\mu_2)$

Based on the conditional probability of u_i given x_i*, the conditional probability of y_i given x_i*is obtained by:

$$P(y_i \left| x_i^* \right.) = \int_{\gamma}^{\infty} f(u_i \left| x_i^* \right.) d(u_i \left| x_i^* \right.)$$

$$= 1 - \int_{-\infty}^{\gamma} f(u_i \left| x_i^* \right.) d(u_i \left| x_i^* \right.) = 1 - \Phi(\gamma - bx_i^*) = \Phi(bx_i^* - \gamma) \qquad (3)$$

where $\Phi(\xi)$ stands for the standardized cumulative normal distribution function and ξ is the argument. Analysis involving $\Phi(\xi)$ is referred to as probit analysis. However, the probit model is difficult to manipulate because numerical integration is required although the parameters are easy to interpret. So, a logistic model is employed to approximate $\Phi(\xi)$ for estimation purpose and is expressed by:

$$\psi(\xi) = \frac{\exp(\xi)}{1+\exp(\xi)} \tag{4}$$

The relationship between a probit model and a logistic model is $\Phi(\xi) \approx \psi(\mathrm{d}\xi)$, where d = $\pi/\sqrt{3}$. Therefore,

$$P(y_i = 1 \Big| x_i^*) \approx \frac{\exp\left\{d(bx_i^* - \gamma)\right\}}{1+\exp\left\{d(bx_i^* - \gamma)\right\}} \tag{5}$$

Since the QTL genotype x_i^* could be homozygote (1) or heterozygote (0) for an individual, the likelihood is then a mixture distribution with mixing proportions equivalent to the conditional probabilities of QTL genotypes given two flanking markers, q_{i1} and q_{i2} for the QTL genotypes QQ and Qq respectively (see Table 1). For n individuals in the sample, the likelihood function is:

$$L = \prod_{1=1}^{n} \left[\sum_{j=1}^{2} q_{ij} p_{ij}^{y_i} (1 - p_{ij})^{1-y_i} \right].$$

where p_{i1} and p_{i2} denotes the conditional probability of $y_{i=1}$ given the QTL genotypes $x_i^* = 1$ and $x_i^* = 0$, respectively. The log likelihood function is:

$$l = \sum_{i=1}^{n} \log(\sum_{j=1}^{2} q_{ij} p_{ij}^{y_i} (1 - p_{ij})^{1-y_1}). \tag{6}$$

The first partial derivatives are:

$$\frac{\partial l}{\partial b} = \sum_{i=1}^{n} \omega_i (y_i - p_{i1}) \tag{7}$$

and

$$\frac{\partial l}{\partial b} = \sum_{i=1}^{n} [\omega_i (y_i - p_{i1}) + (1 - \omega_i)(y_i - p_{i0})] \tag{8}$$

where

$$\omega_i = \frac{q_{i1} p_{i1}{}^{y_i} (1 - p_{i1})^{1-y_i}}{\sum_{j=1}^{2} q_{ij} p_{ij}{}^{y_i} (1 - p_{ij})^{1-y_i}}. \tag{9}$$

is the posterior probability of $x_i^* = 1$. By treating ω_i as constants, the second partial derivatives are:

$$\frac{\partial^2 l}{\partial b^2} = -\sum_{i=1}^{n} \omega_i p_{i1}(1 - p_{i1}) \tag{10}$$

$$\frac{\partial^2 l}{\partial b \partial \gamma} = -\sum_{i=1}^{n} \omega_i p_{i1}(1 - p_{i1}) \tag{11}$$

$$\frac{\partial^2 l}{\partial \gamma^2} = -\sum_{i=1}^{n} [\omega_i p_{i1}(1 - p_{i1}) + (1 - \omega_i) p_{i0}(1 - p_{i0})] \tag{12}$$

In obtaining the parameter estimates, the EM algorithm could be applied. The idea of EM algorithm is the likelihood solution of complete data is relatively simple compared to incomplete data (Pawitan, 2001).

In QTL mapping, the unobserved QTL genotype x_i^* treated as missing data. The EM steps are as follows:

1. Set up initial values of b and γ
2. Calculate ω_i (E-Step)
3. Given ω_i, solve for b and γ using theNewton-Raphson iteration

(M-Step) as follow. Let g denote the vector of first partial derivatives and H be a matrix of second partial derivatives. If θ(t) is a vector of solutions at the tth step, thesolutions at the (t+1) step is $\theta(t+1) = \theta(t) + H^{-1}g$

4. Update the initial values and go to step 2
5. Repeat steps 2-4 until convergence

Regression (REG) approach

Using liability model, the one-QTL REG mapping model for a backcross population can be written as:

$$u_i = \mu + b\pi_i + \varepsilon_i,\ i = 1,2,...,n \tag{13}$$

where $u_i, \mu + b + \varepsilon_i$ have the same definitions as in model (2), and π_i is the conditional expectation of QTL genotypes given the two flanking markers. The likelihood function is:

$$L = \prod_{i=1}^{n} p_i^{y-1}(1-p_i)^{1-y_i}$$

where p_i denotes the conditional probability of $y_i = 1$ given the π_i. The log likelihood function is:

$$l = \sum_{i=1}^{n} [y_i \log p_i + (1-y_i)(1-p_i)]. \tag{14}$$

The first partial derivatives are:

$$\frac{\partial l}{\partial b} = \sum_{i=1}^{n} \pi_i (y_i - p_i) \tag{15}$$

and

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{n} (y_i - p_i). \tag{16}$$

The second partial derivatives are

$$\frac{\partial^2 l}{\partial b^2} = -\sum_{i=1}^{n} \pi_i^2 p_i (1 - p_i) \tag{17}$$

$$\frac{\partial^2 l}{\partial b \partial \gamma} = -\sum_{i=1}^{n} \pi_i^2 p_i (1 - p_i) \tag{18}$$

and

$$\frac{\partial^2 l}{\partial \gamma^2} = -\sum_{i=1}^{n} p_i (1 - p_i) \tag{19}$$

The procedure in obtaining parameter estimates are as follow. Let g denote the vector of first partial derivatives and H be a matrix of second partial derivatives. If $\theta(t)$ is a vector of solutions at the tth step, the solutions at the $(t+1)$ step is $\theta(t+1) = \theta(t) + H^{-1} g$.

Critical Value

When we conducted QTL analysis at a certain point in the genome, we determine the type I error of the test equal to α. This kind of error is called comparison-wise error rate (CWER). However, in characterizing QTL, the analysis is performed by searching or scanning and conducting test at every point on the genome (genome scan) simultaneously. Then, the type I error of all the tests simultaneously is larger than α (This type I error related with genome scan simultaneously is called family-wise error rate/FWER). Hence, in characterizing QTL by genome scan, we are concerned with controlling FWER.

There are several methods proposed in controlling FWER. Lander and Botstein (1989) proposed a method in controlling FWER by determining critical value which considering the size of the genome and the distribution of the trait. Consider an organism with C chromosomes and genetic length G, measured in Morgans. When no QTLs are present, the probability that the test statistic exceeds a high level T is $\approx (C2Gt)\chi^2(t)$, where t = (2 log 10)T and $x^2(t)$ denotes the cumulative distribution function of the x^2 distribution with 1*df*. In order to make the probability less than α that a false positive occurs

somewhere in the genome, the appropriate LOD threshold is thus $\approx T_\alpha = (2\log 10)\mathrm{t}_\alpha$, where t_α solves the equation $\alpha = (C + 2Gt_\alpha)x^2(t_\alpha)$ (Lander and Botstein,1989).

RESULT AND DISCUSSION

For marker density factor, estimation of the threshold and QTL effect using REG approach was close to the ones obtained using ML approach. The estimation of the QTL position obtained using ML and REG approach also indicating similar result. On the other hand, the empirical statistical power obtained using 5% critical value using LB method showing similar result to the Piepho method. In this simulation, the marker density factor affect the performance of the ML and REG approach on the estimation of the threshold, QTL effect, QTL position, as well as the statistical power. Here, as the marker denser, the estimation of the threshold, QTL effect, and QTL position tends to be close to the true value. Moreover, the statistical power of the ML and REG approach was higher for the denser marker than for the less dense ones.

The performance of REG approach was comparable to ML approach in the investigation of the shape of phenotypic distribution (Table 3). As for marker density factor, LB and Piepho method also has similar performance in determining critical value in testing the existence of QTL. In this simulation, it was obtained that the skewed phenotypic distribution has effect in lowering the statistical power for both statistical approaches, especially for REG approach.

The investigation on effect of sample size factor on the performance of statistical approach yield result that REG approach also has similar performance with ML approach (Table 4).The performance of LB method and Piepho method in determining critical value are also comparable. In addition, the sample size factor affecting the performance of the ML and REG approach on the estimation of threshold and QTL effect, the QTL position as well as the statistical power in detecting QTL. Here, the QTL has higher power to be detected and the threshold, QTL effect, and QTL position were more precisely estimated for larger sample size than for smaller sample size.

In the evaluation of the effect of QTL effect, the REG approach showing similar performance with ML approach on the estimation of the threshold, QTL effect, QTL position, as well as statistical power in detecting QTL (Table 5). On the other hand, the LB and Piepho method showing similar result in detecting QTL as in the evaluation of the other factors. In addition, the QTL effect factor affecting the performance of the ML and REG approach on the statistical power in detecting QTL as well as the QTL position estimated. Here, QTL with larger effect tends to have a higher power to be detected and the QTL position was more precisely estimated than for smaller one.

TABLE 2: Comparison of the performance of ML and REG approach for various marker densities (d) for binary trait

d (cM)	Parameter		Estimation with likelihood approach					Estimation with regression approach				
	Name	True Value	Mean	STD[a]	Power (%)[b]		Position[c]	Mean	STD[a]	Power (%)[b]		Position[c]
					LB	Piepho				LB	Piepho	
20	Threshold	0.3334	0.3098	0.2697	46.0	40.5	33.43	0.3343	0.2857	46.0	40.0	33.36
	QTL Effect	0.6667	0.6148	0.4842			(20.46)	0.6081	0.4909			(20.19)
10	Threshold	0.3334	0.3115	0.2785	48.5	50.0	30.66)	0.3058	0.2651	48.0	49.5	30.65
	QTL Effect	0.6667	0.6175	0.4802			(23.15	0.6149	0.4811			(23.13)
5	Threshold	0.3334	0.3374	0.2486	52.5	54.5	26.42	0.3164	0.2380	52.0	54.0	26.08
	QTL Effect	0.6667	0.6499	0.4259			(19.56)	0.6556	0.4089			(19.17)

[a]STD stands for standard deviation of the estimated parameters obtained from 200 replicated simulations.

[b]Empirical statistical power was calculated as the proportion of the simulated samples among 200 replicates with the highest test statistical value along the genome greater than the critical value obtained using LB and Piepho method at 5% significant value.

[c]The true QTL position is at 25 cM of the simulated chromosome. The standard deviations of the estimated QTL positions (obtained from 200 replicates) are given in parentheses.

TABLE 3: Comparison of the performance of ML and REG approach for various shapes of phenotypic distribution for binary trait

Phenotypic distribution	Parameter		Estimation with likelihood approach					Estimation with regression approach				
	Name	True Value	Mean	STD	Power (%)		Positio	Mean	STD	Power (%)		Position
					LB	Piepho				LB	Piepho	
Uniform distribution (1:1)	Threshold	0.3334	0.3254	0.1648	85.5	85.0	26.64	0.3361	0.1416	84.5	86.0	26.49
	QTL Effect	0.6667	0.6277	0.2741			(15.92)	0.6740	0.2214			(11.17)
Skewed distribution (7:3)	Threshold	0.8578	0.8273	0.1715	80.0	78.5	26.35	0.8149	0.1155	75.0	77.5	26.88
	QTL Effect	0.6667	0.6827	0.2699			(11.57)	0.6767	0.1920			(11.05)

see the legends in Table 2

TABLE 4: Comparison of the performance of ML and REG approach for various sample sizes (n) for binary trait

Sample size	Parameter		Estimation with likelihood approach					Estimation with regression approach				
	Name	True Value	Mean	STD	Power (%)		Position	Mean	STD	Power (%)		Position
					LB	Piepho				LB	Piepho	
100	Threshold	0.3334	0.3062	0.2445	46.5	48.0	27.86	0.2975	0.2530	47.0	47.5	27.84
	QTL Effect	0.6667	0.6040	0.4482			(22.32)	0.6036	0.4481			(22.31)
200	Threshold	0.3334	0.3353	0.1258	86.0	86.0	26.27	0.3222	0.1177	86.0	86.0	26.24
	QTL Effect	0.6667	0.6574	0.1695				0.6573	0.1689			
300	Threshold	0.3334	0.3215	0.0933	94.5	95.5	25.28	0.3289	0.1021	94.5	95.5	25.28
	QTL Effect	0.6667	0.6507	0.1488			(5.56)	0.6500	0.1486			(5.57)
500	Threshold	0.3334	0.3162	0.0768	100	100	25.13	0.3167	0.0769	100	100	25.13
	QTL Effect	0.6667	0.6330	0.1138			(1.90)	0.6329	0.1139			(1.95)

see the legends in Table 2

TABLE 5: Comparison of the performance of ML and REG approach under various levels of QTL effect for binary trait

Heritability (h^2)	Parameter		Estimation with likelihood approach					Estimation with regression approach				
	Name	True Value	Mean	STD	Power (%)		Position	Mean	STD	Power (%)		Position
					LB	Piepho				LB	Piepho	
0.05	Threshold	0.2294	0.2046	0.1814	38.0	40.0	29.70	0.029.70	0.1805	38.0	39.5	29.72
	QTL Effect	0.4588	0.4032	0.3223			(23.68)	0.4030	0.3218			(23.70
0.10	Threshold	0.3334	0.3207	0.1353	77.0	78.5	26.47	0.3140	0.1305	77.0	78.5	26.50
	QTL Effect	0.6667	0.6346	0.2047			(9.86)	0.6346	0.2047			(9.86
0.20	Threshold	0.5000	0.4776	0.1322	100	100	24.87	0.24.87	0.1307	100	100	24.87
	QTL Effect	1.0000	0.9480	0.1827			(2.27)	0.9474	0.1818			(2.31
0.40	Threshold	0.8165	0.7813	0.1632	100	1000	24.75	0.7809	0.1498	100	100	24.76
	QTL Effect	1.6330	1.5629	0.2436			(1.38)	1.5633	0.2439			(1.61

see the legends in Table 2

CONCLUSION

From simulation study, it was obtained that LB and Piepho method showing similar performance in determining critical value in testing the existence of QTL for binary trait. The simulation study also indicating that both methods could be used in determining critical value in QTL mapping analysis for binary trait. In assessing the performance of ML and REG approach in QTL mapping analysis for binary trait, the two approaches showing comparable performance. Consequently, when dealing with binary trait, QTL mapping analysis could be performed by ML or REG approach.

REFERENCES

Hackett, C. A., and Weller, J. L.1995. Genetic mapping of quantitative trait loci for traits with ordinal distributions, *Biometrics* **51**: 1252-1263.

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

Hayashi, T. and Awata, T. 2006. Interval mapping for loci affecting unordered categorical traits, *Heredity* **96**:185-194.

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

Kao, C. H., Zeng, Z. B., and Teasdale, R. D. 1999. Multiple interval mapping for quantitative trait loci, *Genetics* **152:** 1203-1216.

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

Liu, B. H. 1998.*Statistical Genomics*, CRC Press.

Martinez, O. and Curnow, R. N. 1992. Estimating the Locations and the Sizes of the Effects of Quantitative Trait Loci Using Flanking Markers, *Theor. Appl. Genet.* **85**:480-488.

Pawitan, Y. 2001. *In All Likelihood Statistical Modelling and Inference Using Likelihood*, Clarendon Press Oxford.

Piepho, H.P. 2001. A Quick Method for Computing Approximate Thresholds for Quantitative Trait Loci Detection, *Genetics,* **157**:425-432.

Sax, K. 1923. The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris, Genetics,* **8**:552-560.

Xu, S. 1995. A comment on the Simple Regression Method for Interval Mapping, *Genetics,* **141**:1657-1659.

Xu, S., and Atchley, W.R. 1996. Mapping quantitative trait loci for complex binary disease using line crosses, *Genetics,***143**:1417-1424.

Xu, S. 1998. Further Investigation on the Regression Method of Mapping Quantitative Trait Loci, *Heredity,* **80**: 364-373.

Zeng, Z. B. 1994. Precision mapping of quantitative trait loci, *Genetics,* **136**:1457-1468.