Genet. Sel. Evol 32 (2000) 357–381

Original article

Detection and parameter estimation for quantitative trait loci using regression models and multiple markers

Yang DA[a,*], Paul M. VANRADEN[b], Lawrence B. SCHOOK[c]

[a] Department of Animal Science, University of Minnesota, Saint Paul, MN 55108, USA
[b] Animal Improvement Programs Laboratory, ARS-USDA, Beltsville, Maryland 20705, USA
[c] Department of Veterinary Pathobiology, University of Minnesota, Saint Paul, MN 55108, USA

(Received 12 July 1999; accepted 12 April 2000)

Abstract – A strategy of multi-step minimal conditional regression analysis has been developed to determine the existence of statistical testing and parameter estimation for a quantitative trait locus (QTL) that are unaffected by linked QTLs. The estimation of marker-QTL recombination frequency needs to consider only three cases: 1) the chromosome has only one QTL, 2) one side of the target QTL has one or more QTLs, and 3) either side of the target QTL has one or more QTLs. Analytical formula was derived to estimate marker-QTL recombination frequency for each of the three cases. The formula involves two flanking markers for case 1), two flanking markers plus a conditional marker for case 2), and two flanking markers plus two conditional markers for case 3). Each QTL variance and effect, and the total QTL variance were also estimated using analytical formulae. Simulation data show that the formulae for estimating marker-QTL recombination frequency could be a useful statistical tool for fine QTL mapping. With 1 000 observations, a QTL could be mapped to a narrow chromosome region of 1.5 cM if no linked QTL is present, and to a 2.8 cM chromosome region if either side of the target QTL has at least one linked QTL.

multiple markers / regression analysis / quantitative trait loci / QTL detection / QTL parameters

Résumé – Détection d'un QTL et estimation de son effet par analyses de régression sur plusieurs marqueurs. On a développé une stratégie basée sur l'analyse de régression en plusieurs étapes et à partir d'un nombre minimum de marqueurs pour détecter un QTL et évaluer son effet individuel sur le caractère indépendamment de l'existence d'autres QTL liés. Trois cas sont à considérer pour estimer la fréquence de recombinaison entre le marqueur et le QTL : 1) Il y a un seul QTL, 2) Il existe au moins un autre QTL sur un des côtés du QTL recherché, 3) Il existe au moins un

* Correspondence and reprints
E-mail: yda@tc.umn.edu

QTL sur chacun des deux côtés du QTL recherché. La fréquence de recombinaison a été estimée analytiquement dans les trois cas. La formule obtenue utilise l'information sur les deux marqueurs flanquants dans le cas 1), sur les deux marqueurs flanquants et sur un marqueur plus éloigné dans le cas 2), sur les deux marqueurs flanquants et sur deux marqueurs plus éloignés dans le cas 3). Pour chaque QTL ainsi détecté, on a aussi développé analytiquement une estimation de son effet et de sa variance, et, pour l'ensemble des QTL ainsi validés, de leur contribution totale à la variance génétique. On a montré par simulation que les formules pour la fréquence de recombinaison pouvaient être utiles pour la cartographie fine de QTL. Ainsi, 1 000 observations permettaient de placer un QTL dans un intervalle de seulement 1,5 cM s'il n'était pas lié à un autre QTL, et de 2,8 cM s'il était lié à un autre QTL à sa droite ou à sa gauche.

marqueurs multiples / analyse de régression / QTL / détection de QTL / paramètres du QTL

1. INTRODUCTION

The mapping of a quantitative trait locus (QTL) using genetic markers includes two central issues, detection of the QTL and estimation of the QTL location and effect. Most current methods for QTL detection and estimation are based on a likelihood analysis [13,16,17,19,20,29] or a regression analysis [9,10,22,30]. These likelihood and the regression analyses require numerical maximization of likelihood functions and yield similar results [9], but the regression analysis is computationally more efficient and generally robust for QTL detection [9,10,28]. For a quantitative trait affected by a single QTL, most of these methods could detect the QTL with good accuracy. However, fine QTL mapping and mapping linked QTLs remain to be challenging tasks in QTL analysis [6,21,31]. Several methods are available for mapping linked QTLs, but unsolved problems exist. Whittaker *et al.* [26] developed analytical formulae based on regression analysis to estimate the marker-QTL recombination frequency and the QTL effect for an F2 population without the influence of linked QTLs under the assumption of an "isolated" QTL as defined by Martínez and Curnow [22]. With these formulae, numerical maximization is no longer necessary such that statistical analysis becomes more efficient computationally. This computational efficiency is appealing for complex QTL mapping issues such as multiple traits and categorical data. In addition, parameter estimation based on regression coefficients is robust against violations in the underlying distribution assumptions. However, these formulae in fact do not apply when two or more linked QTLs exist on the same chromosome even if these QTLs are isolated, because necessary conditional analysis to separate linked QTLs was not applied. The use of a pair of flanking markers is the main idea of interval mapping [19] but this method may yield wrong QTL locations when linked QTLs are present [22]. A multi-marker analysis [15] and composite interval mapping based on multi-marker analysis [29] have been proposed to improve the precision of QTL mapping. However, these methods do not solve the problem of wrong QTL locations of interval mapping if linked QTLs are not appropriately separated by markers, because these methods do not have

a mechanism to distinguish between a QTL independent of linked QTLs and a QTL correlated with linked QTLs. Consequently, mis-identification of QTLs and seriously biased parameter estimation may occur (as shown in this study). Without a mechanism allowing to make such a distinction, it is also difficult to conclude whether the QTL location was identified correctly even if strong statistical evidence exists to support the existence of the QTL. A multi-marker analysis using all available markers have practical problems such as reduced statistical power [29]. Although multi-marker analysis is implemented in some software packages, this analysis does not yet seem to be widely used [25]. One factor affecting the widespread usage of multi-marker analysis in animal populations is the joint marker informativeness. Fitting multiple markers may result in reduced sample size due to the reduced number of informative offspring for multiple markers, because the joint marker informativeness decreases as the number of markers increases [5]. Jansen [14] suggested selecting only statistically significant markers using a standard regression analysis, but this selection may still require fitting a large number of markers before a small number of markers is selected, and may result in incorrect QTL locations and biased estimates due to improperly separated QTLs.

The purpose of this study was to develop a new approach to map each individual QTL to a specific chromosome region and to obtain independent testing and estimation of QTL parameters (QTL location and effect) for each QTL without the influence of linked QTLs. This approach uses a multi-step regression analysis for detecting each QTL. Once a QTL is identified, estimation of the exact QTL location, the size of the QTL effect, and the total QTL variance is accomplished using analytical formulae. Both the QTL detection and parameter estimation use a minimal number of markers per analysis while ensuring the target QTL is unaffected by linked QTLs.

2. METHOD

2.1. General assumptions

Two designs for QTL detection were considered, a one-way backcross design and an F2 design resulting from matings between the F1 offspring. Parental lines were assumed to have homozygous marker and QTL genotypes. Under this assumption, offspring of the one-way backcross had two marker genotypes and two QTL genotypes, the F2 offspring had three marker and three QTL genotypes, and alternative marker and QTL alleles in both designs had equal allele frequencies in the offspring (see appendix). For the F2 design, homozygous marker genotypes were used to estimate the additive effect. Heterozygous F2 marker genotypes were not used for estimating the additive effect because they contained redundant information about the additive effect contained in homozygous marker genotypes. When dominant effect is absent, heterozygous and homozygous marker averages contain completely redundant information, because the heterozygous marker average can be expressed by the averages of the homozygous markers [4]. When dominance effect is present, the inference about additive effects using both homozygous and heterozygous markers is

affected by dominance effects. Chiasma interference was assumed to be absent so that the relationship between recombination frequencies of three adjacent loci in the order *A-B-C* is given by $\theta_{AC} = \theta_{AB} + \theta_{BC} - 2\theta_{AB}\theta_{BC}$ [8], where θ = recombination frequency. The QTL genotypic difference is the difference between the homozygous QTL genotype and the heterozygous genotype for the one-way backcross and is the difference between two homozygous QTL genotypes for the F2 design (Appendix). The one-way backcross design does not have information to separate additive and dominance effects whereas the F2 design does have such information [4]. Application of the results obtained for the one-way backcross and F2 designs to a segregating population such as the granddaughter design [27] will be discussed.

2.2. A strategy of multi-step analysis

The main goal of this strategy was to obtain independent testing and parameter estimation for each QTL without the influence of linked QTLs. A minimal number of markers was used per analysis to achieve this goal, because this is more practical for animal populations and yields simple analytical formulae for QTL parameter estimation. The first step was to evaluate whether the chromosome contained one QTL or more than one QTL, and whether a marker interval containing a QTL was a "continuous" or "discrete" interval based on the statistical significance of the partial regression coefficient of each marker. A continuous interval contains a single QTL flanked by two markers that share a common marker with another marker interval containing at least one QTL. A discrete interval contains a single QTL flanked by two markers but does not share a common marker with another marker interval containing at least one QTL. For examples, in *A-Q1-B-Q2-C*, *A-Q1-B* and *B-Q2-C* are both continuous intervals because these two intervals share a common marker *B*; in *A-Q1-B-C-Q2-D*, *A-Q1-B* and *C-Q2-D* are both discrete intervals because these two intervals do not share a common marker, where *Q1* and *Q2* are QTLs, and *A*, *B*, *C*, and *D* are markers. For a continuous interval, independent testing for the QTL effect is possible but independent parameter estimation for each QTL is impossible using flanking markers because the number of unknowns is more than the number of equations for the unknowns. For example, in *A-Q1-B-Q2-C*, the significance testing for marker *A* conditional on marker *B* offers an independent test for the existence of *Q1*. However, it is impossible to estimate the QTL effect of *Q1* or the recombination frequency between *Q1* and marker *A* or *B* without the influence of *Q2*. With appropriate conditioning on genetic markers, inference on QTL within a discrete interval can be made independent of QTLs existing outside this interval [15,23,29,30]. Based on this result, both independent testing and parameter estimation are available for a discrete interval.

This section presents a method that distinguishes between one QTL and more than one QTL and identifies continuous and discrete intervals if linked QTLs exist using a multi-step minimal conditional analysis. This analysis is a multi-step regression analysis because cne conditional analysis is conducted for each marker; it is a minimal conditional analysis because only one to three

markers are involved per analysis for QTL detection and parameter estimation. The statistical model for the multi-step analysis can be described by

$$y = \sum_{i=1}^{t} \beta_i + x_j b_j + \sum_{k=1}^{c} x_k b_k + \varepsilon \tag{1}$$

where y = the phenotypic observation of an individual, β_i = fixed effect i such as the general mean of the phenotypic value or herd effect, t = number of fixed effects, x_j = independent variable of flanking marker j taking the value of 1 or -1, b_j = partial regression coefficient of y on flanking marker j, x_k and b_k are the independent variable and partial regression coefficient for conditional marker k, c = number of conditional markers c = (1 or 2), and ε = phenotypic residual value. Polygenic effects on other chromosomes can be modeled by including random additive effects in model (1) with usual assumptions as in a mixed model [11].

The significance of the partial regression coefficient of each flanking marker was used to identify the presence of a QTL on either side of the marker. For example, given the chromosome interval *A-B-C*, if the partial regression coefficient of marker *A* given conditional marker *B* is significant, then a QTL may be present on either side of marker *A*. However, this partial coefficient alone does not have information about which side of marker *A* may contain a QTL. If the partial regression coefficient of marker *B* given conditional markers *A* and *C* is insignificant, then neither side of marker *B* contains a QTL and the significant effect of marker *A* must be due to a QTL to the left of marker *A*. Note that a conditional marker is used to separate linked QTLs so that linked QTLs do not affect the target QTL for QTL testing and parameter estimation, but the conditional marker itself is not used to test the presence of a particular QTL or to estimate the QTL parameters. With this type of multi-step conditional analysis for each marker, a method can be developed to distinguish between one QTL and more than one QTL, and between continuous and discrete intervals. This method can be summarized as follows.

1. If $b_{A.B} = 0$, $b_{B.AC} > 0$, $b_{C\,BD} > 0$ and $b_{D.C} = 0$, then only one QTL exists in the chromosome region of *A-D* and the marker-QTL order is *A-B-Q-C-D*.
2. If $b_{A.B} = 0$, $b_{B.AC} > 0$, $b_{C\,BD} > 0$, $b_{D.CE} > 0$, and $b_{E\,D} = 0$, then two QTLs exist in two continuous intervals, *i.e.*, *A-B-Q*1-*C-Q*2-*D-E*.
3. If $b_{A.B} = 0$, $b_{B.AC} > 0$, $b_{C\,BD} > 0$, $b_{D.CE} = 0$, $b_{E.DF} > 0$, $b_{F.EG} > 0$, and $b_{G.F} = 0$, then two QTLs are present in two discrete intervals, *i.e.*, *A-B-Q*1-*C-D-E-Q*2-*F-G*.

In the above algorithms, $b_{A.B}$ = partial regression coefficient of marker *A* conditional on marker *B*, $b_{B\,AC}$ = partial regression coefficient of marker *B* conditional on markers *A* and *C*, "> 0" indicates a significant marker effect, and "$= 0$" indicates an insignificant marker effect from the conditional analysis. With sufficient marker coverage, these algorithms should be able to define a discrete interval for each QTL. If the marker coverage is insufficient, these algorithms could identify chromosome locations to place more markers to obtain discrete intervals. For example, algorithm 2 indicates that adding a marker to each side of marker *C* could define two discrete intervals similar to those

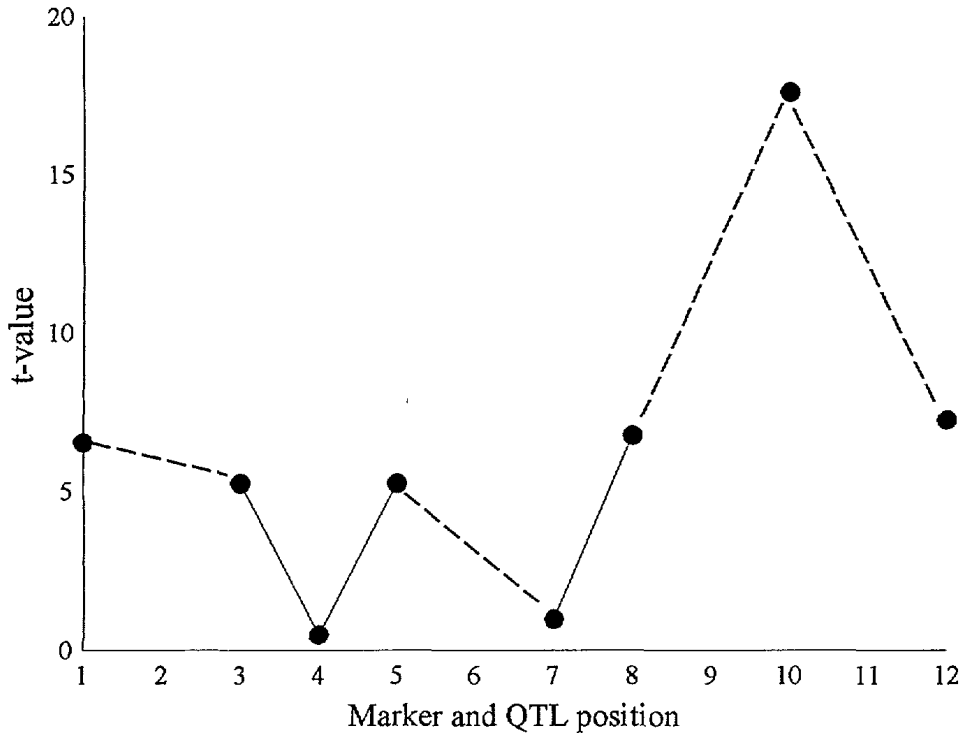

Figure 1. Multi-step analysis for simulated QTL genotypic values consist of four QTLs. The locus spanned by a dotted line is a QTL. Multi-step analysis correctly identified all four QTLs, two in discrete intervals (loci 2 and 6) and two in continuous intervals (loci 9 and 11). Estimates of marker-QTL recombination frequencies and the QTL effects involving the two QTLs in discrete intervals are nearly unbiased. The discovery of two continuous intervals suggests that adding a marker on either side of locus 10 may result in independent parameter estimation for the two QTLs (loci 9 and 11) in the continuous intervals.

defined by algorithm 3. In the case of three neighboring statistically significant markers, such as loci 7, 8 and 10 in Figure 1, it may not be clear whether each of the two adjacent intervals contains a QTL. This potential ambiguity can be clarified by removing the identified QTLs to one side of a flanking marker and then test the significance of the flanking marker. This removal can be achieved by subtracting $(1 - 2\theta_{Mq})\alpha$ from the right-hand-side of the normal equations for model (1), where α is the effect of the identified QTL, and θ_{Mq} is the recombination between marker M and the identified QTL. If this marker is still significant, then the other side of this marker also has a QTL. The valley point at marker 4 was a result of this test, indicating no QTL in the interval between loci 3 and 4 (Fig. 1).

The multi-step analysis yields simple mathematical formulae for independent estimation of QTL parameters for each QTL. To derive formulae for QTL parameter estimation, only three cases need to be considered: 1) the target QTL is the only QTL on the chromosome, 2) one side of the target QTL

within a discrete interval has one or more QTLs, and 3) either side of the target QTL within a discrete interval has one or more QTLs. For convenience of presentation and discussion, we use "side interval" to refer to case 2), and "middle interval" to refer to case 3).

2.3. QTL parameters for a single QTL

The derivation of formulae for QTL parameters starts with a mixed linear model for a single QTL fitted with a single marker. Assuming a single QTL on each chromosome, this mixed model can be described as

$$q_{ij} = \mu + m_j + r_{ij} \tag{2}$$

where q_{ij} = quantitative value of QTL genotype Q_i $(i = 1, 2)$, μ = fixed effect of common mean, m_j = effect of marker j $(j = 1, 2)$, and r_{ij} = recombination residual of the QTL value not explained by the common mean and the marker effect. Treating the marker effect (m_j) as random, the first and second moments of model (2) are $\mathrm{E}(q_{ij}) = \mu$ and $\mathrm{var}(q_{ij}) = \mathrm{var}(m_j) + \mathrm{var}(r_i) = \sigma_q^2 = \sigma_m^2 + \sigma_r^2$, where σ_q^2 = variance of the QTL value, σ_m^2 = variance of marker effects, and σ_r^2 = variance of recombination residuals. For both the backcross design and the F2 design using homozygous markers only, the variances of QTL value, the marker effect and recombination residual can be expressed as:

$$\sigma_q^2 = \frac{1}{4}\alpha^2 \tag{3}$$

$$\sigma_m^2 = \frac{1}{4}(1 - 2\theta)^2\alpha^2 \tag{4}$$

$$\sigma_r^2 = \theta(1 - \theta)\alpha^2 \tag{5}$$

where α = the difference between two alternative QTL genotypes for the backcross design or the F2 design, and θ = recombination frequency between the marker and the QTL. Note that recombination residual variance is non-zero only when the marker-QTL linkage is incomplete and that the marker variance equals the QTL variance when the marker-QTL linkage is complete. It is important to note that the definitions for α under the two designs are different in terms of additive effects (see Appendix). From equations (3–5), the marker-QTL recombination frequency and the QTL effect can be expressed as

$$\theta = \frac{1}{2}\left[1 - \sqrt{\sigma_m^2/(\sigma_m^2 + \sigma_r^2)}\right] \tag{6}$$

$$|\alpha| = 2\sigma_q. \tag{7}$$

Equation (6) has the advantage that estimates of θ are guaranteed to be within the parameter space if an estimate of each variance component is nonnegative, *i.e.*, $0 \leq \theta \leq 1/2$ because $0 \leq \sigma_m^2/(\sigma_m^2 + \sigma_r^2) \leq 1$. Equation (7) shows that the QTL effect is simply twice the standard deviation of the QTL values. The phenotypic correspondence of model (2) can be denoted by

$$y_{ij} = \mu + m_i + r_{ij} + e_{ij} = \mu + m_i + \varepsilon_{ij} \tag{8}$$

where y_{ij} = the phenotypic value of individual j with marker genotype i, e_{ij} = random residual and $\varepsilon_{ij} = r_{ij} + e_{ij}$ with $\sigma^2_\varepsilon = \sigma^2_r + \sigma^2_e$. In equation (6), σ^2_m is available from model (8) as a variance component of the marker effect but σ^2_r is unavailable because σ^2_r and σ^2_e cannot be separated by a single-marker model due to confounding between r_{ij} and e_{ij}. When two flanking markers are available, σ^2_r can be avoided by using $\sigma^2_q = \sigma^2_m + \sigma^2_r$ in equation (6) or estimated as shown by the two equations below equation (14), either using two separate analyses of single-marker models, or a joint two-marker analysis. This study used two separate single-marker analyses according to our strategy to use a minimal number of markers per analysis. The statistical models for the two single-marker analyses to estimate σ^2_r can be described as $y_{ij} = \mu + m_{Ai} + (r_{ij} + e_{ij})$ and $y_{ik} = \mu + m_{Bk} + (r_{ik} + e_{ik})$. Let the marker-QTL order be A-q-B, where A and B are flanking markers and q is the QTL, θ_{Aq} = recombination frequency between marker A and the QTL, θ_{Bq} = recombination frequency between marker B and the QTL, and θ_{AB} = recombination frequency between flanking markers A and B (assumed known), and let the estimates of marker variances for these two models be denoted by σ^2_A and σ^2_B respectively. Then, noting $\sigma_A\sigma_B = 1/4(1-2\theta_{AB})\alpha^2$, $\sigma^2_q = \sigma_A\sigma_B/(1-2\theta_{AB})$, and substituting σ^2_q into equation (6), θ_{Aq} can be expressed as

$$\theta_{Aq} = \frac{1}{2}\left[1 - \sqrt{(1-2\theta_{AB})\sigma_A/\sigma_B}\right]. \tag{9}$$

When the QTL is located exactly in the middle of markers A and B, *i.e.* $\theta_{Aq} = \theta_{Bq}$ and $\sigma_{Aq}/\sigma_{Bq} = 1$, equation (9) is reduced to the estimation of θ_{Aq} and θ_{Bq} based on $\theta_{AB} = \theta_{Aq} + \theta_{Bq} - 2\theta_{Aq}\theta_{Bq}$. Therefore, the ratio σ_A/σ_B can be considered as an adjustment for unequal marker-QTL distances between the QTL and the two flanking markers to the estimation obtained by assuming equal marker-QTL distance. When an estimate of θ_{AB} is available, θ_{Bq} can be estimated using the relationship between recombination frequencies involving the flanking markers and the QTL under the assumption of no interference, *i.e.*,

$$\theta_{Bq} = (\theta_{BC} - \theta_{Aq})/(1 - 2\theta_{Aq}). \tag{10}$$

Equivalently, σ_{Aq} and σ_{Bq} in equation (9) can be replaced by the ratio of the absolute values of the regression coefficients each from a single marker analysis, *i.e.*,

$$\theta_{Aq} = \frac{1}{2}\left[1 - \sqrt{(1-2\theta_{AB})|b_A/b_B|}\right]. \tag{11}$$

A proof for equation (11) is given in the Appendix. The absolute sign for the regression coefficients is to account for the possibility that the two regression coefficients have different signs due to a repulsion phase of the flanking markers with respect to QTL effects. During our data simulation, we noted that random data errors could cause regression coefficients of flanking markers to have opposite signs. The chance for this problem to occur decreases as the sample size and marker-QTL distance increase. The regression coefficients in equation (11) can be obtained in three ways, (i) from normal equations based

variances and covariances [30], (ii) from model (8) as the difference between the two marker effects, or (iii) from model (1). For purposes of estimating marker-QTL recombination frequency and significance testing, these three methods yield identical results in terms of estimates and statistical significance. However, in terms of estimating the QTL variances (equations 3–5) and the size of the QTL effect (equation 7), the result from method (iii) is different from those of methods (i) and (ii), because the partial regression coefficient from method (iii) is half of that from methods (i) and (ii). In equation (11) and in formulae derived below for parameter estimation, regression coefficients are assumed to be from method (iii), because this method uses a marker contrast that is independent of dominance effect [3] and is convenient to deal with the problem of non-informative offspring in segregating populations.

Once the marker-QTL recombination frequency for each flanking marker is available, two separate estimates of QTL variance are available based on regression coefficients of flanking markers, *i.e.*,

$$\sigma^2_{qA} = b^2_A/(1-\theta_{Aq})^2 \tag{12}$$

$$\sigma^2_{qB} = b^2_B/(1-\theta_{Bq})^2. \tag{13}$$

A weighted average of the two estimates of QTL variance can then be devised. While the best weighting method is a subject of further study, it is reasonable to use a weight that is a function of the sample size for each marker and an inverse function of the marker-QTL recombination frequency, *i.e.*, $w_A = n_A/\theta_{Aq}$, $w_B = n_B/\theta_{Bq}$, where n_A and n_B are the sample sizes for markers A and B. Then, the weighted average of the QTL variance is

$$\sigma^2_q = \left(w_A\sigma^2_{qA} + w_B\sigma^2_{qB}\right)/(w_A + w_B). \tag{14}$$

The recombination residual variance for each marker can then be obtained based on equations (3), (5), and (14), *i.e.*, $\sigma^2_{rA} = 4\theta_{Aq}(1-\theta_{Aq})\sigma^2_q$, $\sigma^2_{rB} = 4\theta_B(1-\theta_B)\sigma^2_q$. Estimate for the size of the QTL effect is obtained by substituting equation (14) into equation (7).

When linked QTLs are present, analytical formulae for QTL parameters are available only for discrete intervals. For continuous intervals, analytical formulae for QTL parameters are possible only after additional markers are used to divide these intervals into discrete intervals. Two sets of formulae are required to estimate QTL parameters for discrete intervals: one set for side intervals, and one set for middle intervals.

2.4. QTL parameters for a side interval

For a side interval, $Q1$-A-B-$Q2$-C is used as an example to derive formulae of QTL parameters, where $Q1$ and $Q2$ are two QTLs, and A, B and C are markers. Let $q = q_{1i} + q_{2j}$ = the QTL genotypic value of an individual, where q_{1i} = the value of genotype i ($i = 1, 2$) of $Q1$ and q_{2j} = the value of genotype j ($j = 1, 2$) of $Q2$, and let $b_{qB.A}$ and $b_{qC.B}$ be the partial regression coefficients

of markers A and C respectively, then

$$b_{qB\,A} = \frac{1}{2}(1 - 2\theta_{B2})\alpha_2 \tag{15}$$

$$b_{qC.B} = \frac{(1 - 2\theta_{B2})\theta_{BC}(1 - \theta_{BC})}{2(1 - 2\theta_{BC})\theta_{B2}(1 - \theta_{B2})}\alpha_2. \tag{16}$$

Replacing the QTL genotypic value (q) by the phenotypic value (y) in the regression coefficients of equations (15–16) and letting $w = |b_{yB.A}/b_{yC\,B}|$, the recombination frequency between marker B and $Q2(\theta_{B2})$ can be expressed as

$$\theta_{B2} = \frac{1}{2}\left[1 - \sqrt{1 - \frac{4\theta_{BC}(1 - \theta_{BC})}{(1 - 2\theta_{BC})w + 4\theta_{BC}(1 - \theta_{BC})}}\right]. \tag{17}$$

From equations (3) and (15–16), estimates of the QTL variance of $Q2$ based on markers B or C can be obtained as

$$\sigma^2_{qB} = \left[\frac{b_{yB\,A}}{1 - 2\theta_{B2}}\right]^2 \tag{18}$$

$$\sigma^2_{qC} = \left[\frac{(1 - 2\theta_{BC})\theta_{B2}(1 - \theta_{B2})}{(1 - 2\theta_{B2})\theta_{BC}(1 - \theta_{BC})}b_{yC\,B}\right]^2. \tag{19}$$

Then, the weighted average of the QTL variance for $Q2$ is obtained by substituting equations (18–19) into equation (14), and the size of the QTL effect is obtained by substituting equation (14) into equation (7).

2.5. QTL parameters for a middle interval

For a middle interval, the marker-QTL order $Q1$-A-B-$Q2$-C-D-$Q3$ is used as an example and $Q2$ is assumed to be the target QTL for parameter estimation, where $Q1$, $Q2$ and $Q3$ are QTLs and A, B, C, and D are markers. Let $q = q_{1i} + q_{2j} + q_{3k}$ = the QTL genotypic value of an individual, where q_{1i} = the value of genotype i ($i = 1, 2$) of $Q1$, q_{2j} = the value of genotype j ($j = 1, 2$) of $Q2$, q_{3k} = the value of genotype k ($k = 1, 2$) of $Q3$. Based on a general expression for the partial regression coefficient of a marker conditional on two flanking markers in [30], the partial regression coefficients of markers B and C that involves only $Q2$ can be expressed as

$$b_{qB\,AC} = \frac{\theta_{BC}(1 - \theta_{BC}) - \theta_{B2}(1 - \theta_{B2})}{2(1 - 2\theta_{B2})\theta_{BC}(1 - \theta_{BC})}\alpha_2 \tag{20}$$

$$b_{qC.BD} = \frac{\theta_{B2}(1 - \theta_{B2})(1 - 2\theta_{BC})}{2(1 - 2\theta_{B2})\theta_{BC}(1 - \theta_{BC})}\alpha_2. \tag{21}$$

Replacing the QTL genotypic value (q) by the phenotypic value (y) in the regression coefficients of equations (20–21) and letting $w = |b_{yB\,AC}/b_{yC.BD}|$,

the recombination frequency between each flanking marker and the QTL can be expressed as

$$\theta_{B2} = \frac{1}{2}\left[1 - \sqrt{1 - \frac{4\theta_{BC}(1-\theta_{BC})}{1+(1-2\theta_{BC})w}}\right]. \tag{22}$$

From equations (3) and (22), estimates of the QTL variance of $Q1$ based on partial regression coefficient of markers A or B can be obtained as

$$\sigma^2_{qB} = \left[\frac{(1-2\theta_{B2})\theta_{BC}(1-\theta_{BC})}{\theta_{BC}(1-\theta_{BC}) - \theta_{B2}(1-\theta_{B2})} b_{yB\ AC}\right]^2 \tag{23}$$

$$\sigma^2_{qC} = \left[\frac{(1-2\theta_{B2})\theta_{BC}(1-\theta_{BC})}{\theta_{B2}(1-\theta_{B2})(1-2\theta_{BC})} b_{yC.BD}\right]^2. \tag{24}$$

2.6. Parameter space

Of the three formulae for estimating marker-QTL recombination frequency, two formulae (equations 17 and 22) guarantee the estimates to be within the parameter space of $0 \leq \theta \leq 1/2$, but one (equation 11) does not offer such a guarantee. Estimates from equation (11) are within the parameter space of $0 \leq \theta \leq 1/2$ only when $|b_A/b_B| \leq 1/(1-2\theta_{AB})$. To guarantee estimates to be within the parameter space, equation (11) can be replaced by either equation (17) or equation (22), at the expense of potentially larger variations in estimates, as discussed in Section 3.

2.7. Multiple alleles

Formulae presented above apply to multiple marker alleles if each QTL has only two alleles and a nested model with marker nested within family is used, such as the nested models used in Ashwell *et al.* [1] and Heyen *et al.* [12]. When multiple QTL alleles are present, these formulae apply when each family is analyzed separately, as is done in Ashwell *et al.* [1] and Heyen *et al.* [12]. Across-family analysis assuming multiple QTL alleles remains to be studied.

2.8. Total QTL variance

Once independent estimates of the variance for each QTL are obtained, the total QTL variance of detected QTLs is available as a sum of QTL variances across the genome, *i.e.*,

$$\sigma^2_q = \sigma^2_{q1} + \cdots + \sigma^2_{qn}.$$

This total QTL variance would provide a complete description of additive inheritance and an estimate of heritability of the quantitative trait using information about the genes of the quantitative trait rather than the traditional approach

of resemblance between relatives if all QTLs are identified by genetic markers. For a given set of markers, however, this total QTL variance does not include variances of QTLs unlinked with the available genetic markers. To account for variances of unidentified QTLs, fitting a random polygenic effect in model (1) should help.

2.9. Application to segregating population

The analysis of a segregating population such as the granddaughter design [27] is similar to the backcross design or the F2 design using homozygous marker genotypes, except that the QTL effect (α) in the mathematical formulae is now defined as the difference of the parental QTL alleles. For the granddaughter design, the QTL effect is the difference between two alternative sire QTL alleles. Therefore, results obtained for the backcross and F2 designs are in principle applicable to segregating populations. The only problem that requires additional statistical treatment is non-informative offspring. Genotypes of non-informative offspring cannot unequivocally determine the marker allele transmission from the parents to the offspring [5] and result in loss of information for QTL analysis. For microsatellite markers in dairy cattle, non-informative offspring account for about 25% of the sons [1], resulting in about 25% sample size reduction if these sons are not used. Dentine and Cowan [7] proposed a regression model that makes use of non-informative offspring by linking the quantitative trait with the grandparental marker alleles using population allele frequencies. We propose an alternative method to predict parental allele transmission for non-informative offspring using linked markers for QTL analysis. This method is described below for the granddaughter design. Let y_k = observation of offspring k, μ = general mean of the quantitative trait, b = the marker regression coefficient = the difference between two marker allele effects, and e_k = the random phenotypic residual of the quantitative trait. Then, the statistical model can be described as

$$y_k = \mu + xb + e_k \tag{25}$$

where $x = 1$ or -1 for informative offspring, or $x = p - q$ or $q - p$ for non-informative offspring, and where p = the transmitting probability that the offspring inherits marker allele i from the sire, q = the transmitting probability that the offspring inherits marker allele j from the sire. The assignment of the x value is dependent on the sire marker allele or haplotype of linked marker(s) the offspring inherits, as shown by the calculation of the transmitting probabilities (p and q) in Table I, where the non-informative marker is denoted by A, and two linked markers are denoted by B and C. Two alleles of two adjacent markers are assumed as nonrecombinant type if they have the same subscript (1 or 2) and are recombinant type if they have different subscripts (1 and 2). Chiasma interference is assumed absent. The calculations of p and q for the order A-B-C (non-informative marker not flanked by two markers) shows that a marker separated from the non-informative marker does not contribute information to the inference about the allele transmission of the non-informative marker. Therefore, the use of more than one marker on one side of the non-informative marker is unnecessary. The calculations of p and q

Table I. Calculation of allele transmitting probabilities for non-informative marker [a] using information of linked markers.

Locus order	Marker alleles B and C	Prob$\{A_1/BC\}$	Marker alleles B and C	Prob$\{A_2/BC\}$
A-B-C	B_1C_1	$p = 1 - \theta_{AB}$	B_2C_2	$q = \theta_{AB}$
	B_1C_2	$p = 1 - \theta_{AB}$	B_2C_1	$q = \theta_{AB}$
B-A-C	B_1C_1	$p = \frac{(1-\theta_{AB})(1-\theta_{AC})}{d}$ [b]	B_2C_2	$q = \frac{\theta_{AB}\theta_{AC}}{d}$
	B_1C_2	$p = \frac{(1-\theta_{AB})\theta_{AC}}{\theta_{BC}}$	B_2C_1	$q = \frac{\theta_{AB}(1-\theta_{AC})}{\theta_{BC}}$

[a] Marker A in this table is the non-informative marker.
[b] $d = (1 - \theta_{AB})(1 - \theta_{AC}) + \theta_{AB}\theta_{AC}$.

for the order B-A-C shows that flanking markers contribute more information than a single adjacent marker. Therefore, whenever possible, flanking markers should be used to calculate the probability of allele transmission of the non-informative marker. Model (25) can be readily extended to include more than one marker. Once the partial regression coefficient of each marker is obtained, formulae for QTL parameter estimation can be applied. The benefit of including non-informative offspring using information of linked markers can be measured by the relative efficiency of the two alternative models. Following Dentine and Cowan [7], the relative efficiency of the regression model using full data to the regression model using informative offspring only will be defined as the ratio of two variances. Let v_1 = variance of marker regression coefficient using full data, and v = variance of marker regression coefficient using informative offspring only. Then the relative efficiency (R) of using full data *versus* using informative offspring only is $R = v/v_1$. Using full data is more beneficial if $R > 1$, and has no benefit if $R = 1$. The variance of the regression coefficient using informative offspring only is $v = \sigma^2/n$, where n = number of informative offspring. Similarly, it can be shown that the variance of the regression coefficient using full data is

$$v_1 = \sigma^2 / \left[n + n_0(p - q)^2\right] \tag{26}$$

where n_0 = number of non-informative offspring. Therefore, the relative efficiency of using full data *versus* using informative offspring only is

$$R = 1 + (n_0/n)(p - q)^2. \tag{27}$$

Equation (26) has an important implication to the calculation of degrees of freedom for testing QTL effect using a statistical test that requires the calculation of degrees of freedom, such as the t-test and the F-test. To calculate the degrees of freedom for the residual sum of squares, the total number of observations is required. Equation (26) implies that the total number of observations for using full data is $n + n_0(p - q)^2$, not $n + n_0$. The use of total number of offspring

$(n + n_0)$ as the degrees of freedom could greatly exaggerate the efficiency of including non-informative offspring in the statistical analysis. Equation (27) shows that including non-informative offspring is always beneficial unless the linked markers are far from the non-informative marker ($p \approx q$). The benefit increases as the number of informative offspring and the distance between the non-informative marker and linked markers decrease. The maximum information increase is the frequency of non-informative offspring measured by n_0/n. This maximum is reached when the non-informative marker and flanking markers are completely linked so that $p - q = 1$.

3. SIMULATION RESULTS AND DISCUSSION

Seven sets of simulation data were used to test the method of the multi-step analysis for QTL detection and the formulae for QTL parameter estimation developed in this study.

3.1. Data simulation

Data set 1 was used to test the method of the multi-step analysis for QTL detection. This set had 1 558 observations that yielded equal recombination frequencies between adjacent markers for 12 marker and QTL genotypes on the same chromosome. Loci 2, 6, 9 and 11 were assumed to be the four QTLs, and the rest of the loci were markers. The recombination frequency between each pair of adjacent loci was 0.1502. Each QTL had a variance of 0.25. The QTL genotypic value was assumed to be the sum of the genotypic values of the four QTLs. The four QTLs were assumed to be in coupling linkage phase, *i.e.*, each parent had the following QTL genotype Q_1-Q_2-Q_3-Q_4/q_1-q_2-q_3-q_4, where the quantitative value for each QTL allele was 2 for Q_1, Q_2, Q_3, and Q_4, and was 1 for q_1, q_2, q_3, and q_4. To study the true patterns of the multi-step analysis and other methods to be compared without the influence of random errors, no random errors were added to the QTL genotypic values. Data set 2 was used to evaluate the performance of the formulae for parameter estimation for medium recombination frequencies ($0.10 \sim 0.20$). This set had 30 samples with 500 observations in each sample. Data sets 3 through 7 were used to test the formulae for narrow marker intervals ($1 \sim 2$ cM). These five sets had sample sizes of $1\,000 \sim 5\,000$ in the increment of 1 000, and each of the five sets had 30 samples. These relatively large sample sizes were used for two reasons, to ensure that a sufficient number of recombinants were present, and to test whether the accuracy for parameter estimation improved as the sample size increased. Data sets 2 through 7 had the true order $Q1$-A-B-$Q2$-C-D-$Q3$, where $Q1, Q2$, and $Q3$ were QTLs with genotypic values q_1, q_2, and q_3 respectively, and A, B, C, and D were genetic markers. For data set 2, recombination frequencies were $Q1$-(0.2)-A-(0.1)-B-(0.2)-$Q2$-(0.1)-C-(0.2)-D-(0.1)-$Q3$, where the number in each ( ) is the recombination frequency between adjacent loci. For data sets $3 \sim 7$, a recombination frequency of 0.01 was assumed for each pair of adjacent loci. The genotypic data for the marker loci and QTLs were generated in such a way that the assumed true parameters used to generate the data were reversely

obtainable from the simulated data if no rounding of the genotypic probability occurred. With multiple loci, the actual parameters in the simulated data may not have been exactly the same as the assumed true parameters due to the rounding of each genotypic probability. In the simulated data set, the realized parameters were either exactly the same as or only slightly different from the true parameters. In the statistical analysis, the realized parameters were used as the true parameters. For each sample, random residuals following a N (0,1) distribution were generated using SAS [24]. For data sets 2 through 7, the target QTL was $Q2$, *i.e.*, recombination frequencies between $Q2$ and its flanking markers B and C were to be estimated. The total QTL genotypic value (q) is $q = q_2$ for a single QTL, $q = q_1 + q_2$ for a side interval, and $q = q_1 + q_2 + q_3$ for a middle interval. Each QTL had a variance of 0.25. Each phenotypic value was a sum of the QTL genotypic value and a random residual, *i.e.*, $y = q + e$, with phenotypic variance of $\sigma_y^2 = \sigma_q^2 + \sigma_e^2$. The QTL heritability ($= \sigma_q^2/\sigma_y^2$) of each QTL was 0.2 for the case of one QTL, 0.167 for side interval, and 0.143 for middle interval.

3.2. QTL detection

Figure 1 shows that the multi-step analysis correctly identified all four QTLs according to algorithms 1–3, two in discrete intervals (loci 2 and 6) and two in continuous intervals (loci 9 and 11). Estimates of marker-QTL recombination frequencies and QTL effects for the two QTLs in discrete intervals were nearly unbiased. For the two QTLs in continuous intervals, independent parameter estimation is impossible. However, with the information that either side of locus 10 may have a QTL, adding a marker to either side of locus 10 may result in independent parameter estimation for each of the two QTLs, and hence significantly improve the accuracy of QTL detection and parameter estimation. Interval mapping [19] and composite interval mapping [29] were also applied to the data in Figure 1, using the ZMAPQTL and EQTL programs in the computer package of QTL Cartographer Version 1.12 [2]. Interval mapping identified seven QTLs, as shown by the seven peaks in Figure 2, whereas only four QTLs exist. None of the parameter estimates was close to the true parameters. This indicates that interval mapping is an inappropriate analysis when linked QTLs are present. Composite interval mapping identified three QTLs, as shown by the three peaks in Figure 2. The two QTLs in discrete intervals (loci 2 and 6) were correctly identified and the parameter estimates for these two QTLs were nearly unbiased. However, the two QTLs in continuous intervals (loci 9 and 11) were mistakenly considered as one QTL by composite interval mapping, with the marker-QTL order of loc10-(0.2201)-Q-(0.0791)-loc12, where the number in ( ) is the recombination frequency between adjacent loci. However, the true order in fact is Q-(0.1502)-loc10-(0.1502)-Q-(0.1502)-loc12. The estimate of QTL effect for the wrong QTL was 1.488, whereas the true value was 1.0 for each of the two true QTLs. These wrong results can be a serious problem for gene cloning because gene cloning requires highly accurate location estimation. Moreover, it is difficult to determine which of the three QTLs identified by composite interval mapping in Figure 2 is a correct result without defining discrete and continuous intervals. Therefore, a mechanism to distinguish between discrete and continuous intervals is important for accurate

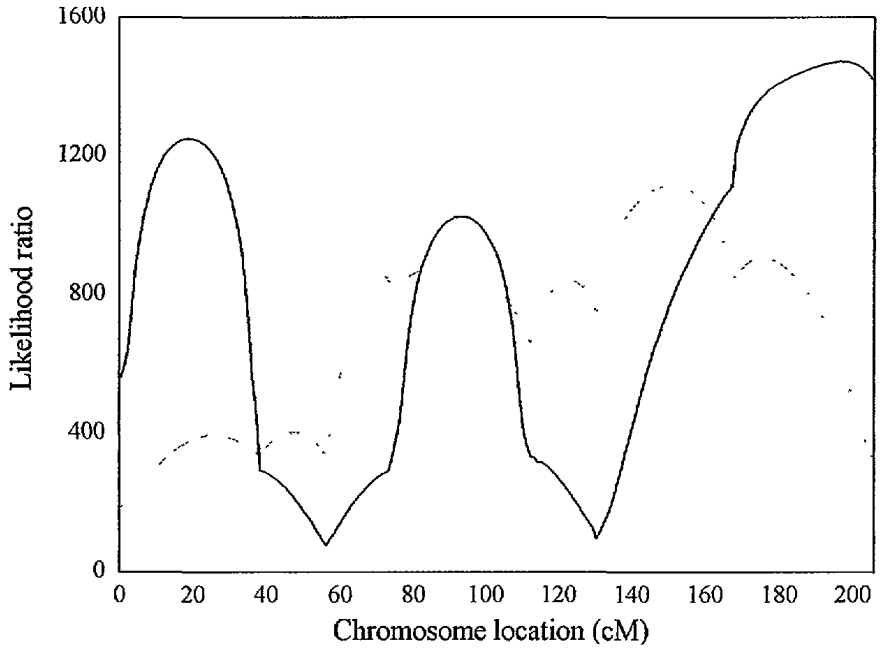

Figure 2. Interval mapping and composite interval mapping for simulated QTL genotypic values consist of four QTLs The true locations of the four QTLs are shown in Figure 1. Interval mapping (dotted line) identified seven QTLs and yielded incorrect parameter estimates. Composite interval mapping (solid line) identified three QTLs. The two QTLs in discrete intervals were correctly identified, whereas the two QTLs in continuous intervals were mis-identified as one QTL.

QTL detection and parameter estimation when linked QTLs are present. The QTL Cartographer [2] also offers an option to analyze a subset of markers per analysis. With this option, the number of markers per analysis can be reduced. However, this option does not provide a mechanism to distinguish between discrete and continuous intervals.

3.3. Estimation of recombination frequencies

The mean of the estimates from 30 simulations for each parameter of both medium and narrow recombination frequencies (Tabs. II and III) show that most averages of the estimates for marker-QTL recombination frequencies were close to the true parameters for the given sample sizes. Bias in estimates measured by the difference between the mean of estimates and the true parameter was mostly a few percentages of the true parameter. Therefore, bias of the estimates from all formulae appeared to be negligible even for narrow recombination frequencies. However, in terms of variation of the estimates, different formulae showed different performances.

For medium recombination frequency ($\theta = 0.10$ or 0.20), the standard deviation (SD) of the estimates were in the range of $0.022 \sim 0.039$ (Tab. II), the 95% confidence interval of the estimates for marker-QTL recombination frequencies was in the range of mean $\pm 0.044 \sim 0.078$ (obtained by multiplying

Table II. Estimates of medium marker-QTL recombination frequency [a].

Formula for	θ_L [b] (true $\theta = 0.204$) Mean $\pm$ SD [e]	CV [c] (%)	θ_R [d] (true $\theta = 0.10$) Mean $\pm$ SD	CV (%)
One QTL	0.208 ± 0.022	10 58	0.087 ± 0.030	34.48
Side interval	$0.207 \pm 0\ 025$	12 08	0.087 ± 0.035	40.22
Middle interval	0.213 ± 0.028	13.15	0.079 ± 0.039	49.36

[a] Based on 30 samples of simulated data sets each with 500 observations.
[b] Recombination frequency between the QTL and the left flanking marker and the QTL.
[c] CV = coefficient of variation = SD/Mean.
[d] Recombination frequency between the QTL and the right flanking marker and the QTL
[e] SD = standard deviation.

Table III. Estimation of narrow marker-QTL recombination frequency for simulated data [a].

Sample size (true θ_L [b]) (true θ_R [c])	Formula	Estimate of θ_L			Estimate of θ_R		
		Mean	SD [d]	CV [e] (%)	Mean	SD	CV (%)
1 000	One QTL	0 0103	0.0038	37.4	0.0099	0.0038	38.7
(0 01[b])	Side interval	0 0105	0.0053	51.2	0.0096	0.0055	56 8
(0.01[c])	Middle interval	0.0097	0.0090	92.9	0.0103	0.0091	87 9
2 000	One QTL	0.0102	0.0038	37.2	0.0100	0.0038	37.8
(0 01[b])	Side interval	0 0094	0.0036	38.1	0.0108	0.0045	33 1
(0.01[c])	Middle interval	0.0093	0.0045	48.3	0.0109	0.0044	40.7

[a] For each sample size, 30 samples of simulated data sets were generated to obtain the average, standard deviation, and coefficient of variation of estimates of marker-QTL recombination frequency between each marker and the QTL.
[b] Recombination frequency between the QTL and the marker to the left of the QTL.
[c] Recombination frequency between the QTL and the marker to the right of the QTL.
[d] SD = standard deviation.
[e] CV = coefficient of variation = SD/Mean

the SD by 1.96), and the coefficient of variation (CV) was in the range of 10.58–49.46%. For the same sample size and true parameter, the formulae with more conditional markers yielded larger variances of the estimates than formulae with fewer conditional marker(s). For example, for the data in Table II, conditioning on two markers had a CV of 13.15% whereas the formula without conditioning had a CV of 10.58% for estimates of the true parameter $\theta_L = 0.204$. A possible interpretation of these increased variation is that the relative heritability of a QTL decreases as the number of QTLs increases, because the total phenotypic

Table IV. Estimates of marker-QTL recombination frequency from different formulae[a].

Formula for	θ_L[b] (true $\theta = 0.202$) Mean ± SD[d]	θ_R[c] (true $\theta = 0.10$) Mean ± SD
One QTL	0.200 ± 0.0161	0.102 ± 0.0209
Side interval	0.200 ± 0.0174	0.102 ± 0.0226
Middle interval	0.199 ± 0.0181	0.103 ± 0.0233

[a] Based on 30 samples of simulated data sets each with 1 000 observations. Only one QTL is assumed for all three sets of formulae.
[b] Recombination frequency between the QTL and the left flanking marker and the QTL.
[c] Recombination frequency between the QTL and the right flanking marker and the QTL.
[d] SD = standard deviation.

variance increases as the number of genes increases. To exclude this possibility, the three sets of formulae were applied to the same set of data with only one QTL, 1 000 observations, and a total of 30 simulations repeats. Results of these simulations confirm that conditioning on more markers does result in larger variations of parameter estimates (Tab. IV) and hence decreases the performance of the formulation. Therefore, conditional markers should be used only when necessary, such as when obtaining independent estimates of QTL parameters. Variations of the estimates for different true parameters ($\theta_L \approx 0.20$ and $\theta_R \approx 0.10$) show that estimating a narrower recombination frequency is more difficult than estimating a larger recombination frequency (Tabs. II and III). The CV's of estimates for $\theta_L \approx 0.20$ were in the range of 3.1–9.2% whereas the CV's of estimates for $\theta_R \approx 0.10$ were in the range of 8.4–20.9%.

For narrow recombination frequency ($\theta = 0.01$) with sample sizes of 1 000 observations, SD ranged from 0.0038 to 0.0091 and CV ranged from 37.4% to 92.9% (Tab. III). The width of the 95% confidence interval of the estimates for $\theta = 0.01$ was in the range of 1.5 cM for a single QTL to 2.8 cM for a middle interval (Fig. 3). The calculation of this range of 95% confidence intervals was based on mean ± 1.96 (SD) and the approximation that 1% recombination frequency was roughly equal to 1 cM when the recombination frequency was narrow. These results indicate that 1 000 observations could map a QTL to a narrow chromosome region of 1.5 cM if no linked QTLs are present, or about 3 cM if either side of the target QTL has a linked QTL. As sample size increases, accuracy in estimating narrow recombination frequencies can be further improved. As shown in Figure 3, confidence intervals become narrower as sample sizes increase. However, for sample sizes beyond 2 000, the decrease in the width of the confidence intervals becomes slower (Fig. 3). This suggests that using sample sizes beyond 2 000 may be unnecessary for the purpose of fine QTL mapping. Overall, the simulation results in Figure 3 indicate that obtaining reliable estimates for narrow marker-QTL distance is possible and the limits of fine QTL mapping are mainly in the resources available such as sample size and marker coverage of the genome.

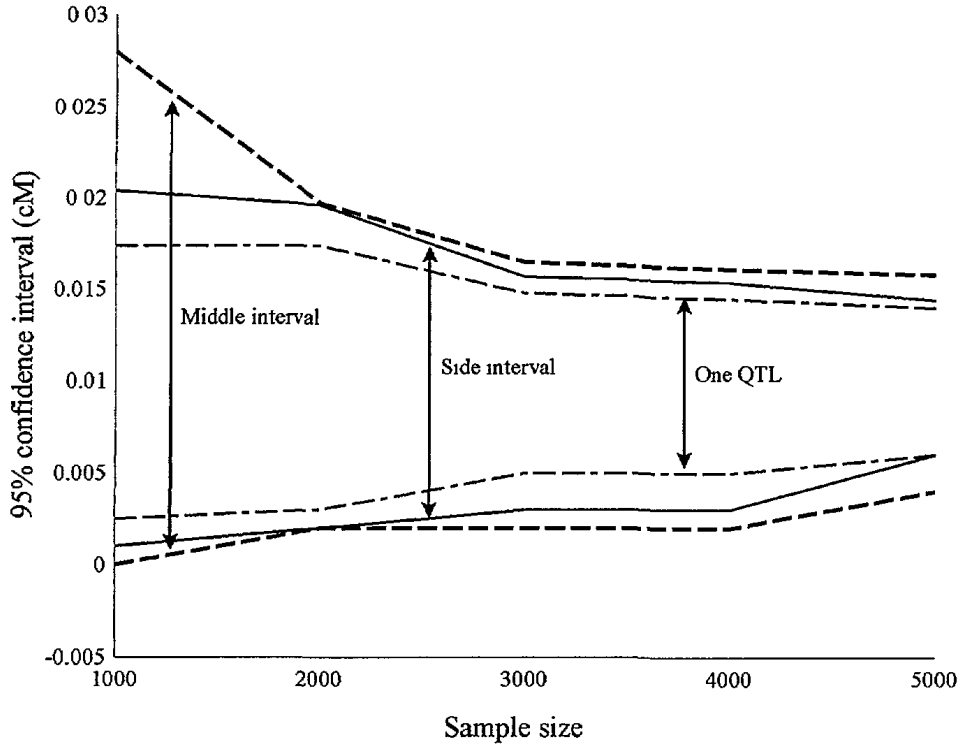

Figure 3. 95% confidence intervals of a 1% marker-QTL recombination frequency.

3.4. Discussion

The understanding of a quantitative inheritance requires the understanding of each gene associated with the quantitative trait. The strategy of the multi-step analysis and formulae for parameter estimation in this study provide an approach to obtain independent testing and parameter estimation for each QTL and to obtain a complete description of a quantitative inheritance. The multi-step analysis is also an alternative multi-marker analysis that uses a minimum number of markers (2 or 3 markers) per analysis and provides a mechanism to distinguish between discrete and continuous intervals. Because of the minimal number of markers used per analysis, this approach is more practical than methods that use all available markers simultaneously for populations where gene fixation in parental lines is unavailable. The formulae for parameter estimation in this study are mathematically simple, because no numerical maximization is required for both QTL detection and parameter estimation. This is appealing for complex QTL mapping problems such as multiple traits and categorical data. Further studies are needed to evaluate the false positive and negative detections. As the approach in this research is based on regression analysis, the statistical power is expected to be similar to exiting methods for QTL detection based on regression analysis. A limitation of the approach in this study is that the QTL detection and estimation for the F2 design are based on additive effects only. However, this limitation could be removed by developing new formulae.

Another limitation is the number of markers available, because the number of markers required per chromosome for independent QTL testing and parameter estimation increases as the number of QTLs on the chromosome increases. However, this limitation is disappearing rapidly because the number of markers available is rapidly increasing in most species. The formulae to estimate the variance of recombination residuals provide an alternative approach to account for existing QTLs in testing for a new QTL by subtracting the variance of recombination residuals of each QTL from the phenotypic residual variance. This approach removes the influence of existing QTLs without the need of fitting multiple markers simultaneously in the statistical model. Research is needed to establish guidelines for the significance threshold to declare the presence of multiple QTL using this approach. Simulation studies for various sample sizes and data structure should yield such guidelines. The current formulae for parameter estimation do not provide estimates of confidence intervals. One way to estimate the confidence intervals of parameter estimates would be to use simulations. Once estimates of the true parameters are available, those estimates can be used in place of true parameters to simulate marker and QTL genotypes for the given design and sample sizes. The simulation study leading to Figure 3 is an example to obtain confidence intervals. Other methods for obtaining confidence intervals could also be considered, such as deriving asymptotic variances of the parameter estimates.

ACKNOWLEDGEMENTS

We would like to thank Dr. I. Hoeschele and an anonymous reviewer for helpful comments and suggestions. The senior author would like to thank Dr. C.J. Basten for communications on the use of QTL Cartographer.

REFERENCES

[1] Ashwell M.S., Da Y., VanRaden P.M., Rexroad C.E. Jr., Miller R.H., Detection of putative loci affecting conformational type traits in an elite population of United States Holsteins using microsatellite markers, J. Dairy Sci. 81 (1998) 1120–1125.

[2] Basten C.J., Weir B.S., Zeng Z.-B., QTL Cartographer: A Reference Manual and Tutorial for QTL mapping, Department of Statistics, North Carolina State University, Raleigh, NC, 1997.

[3] Da Y., Detection of dominance and epistasis effects through marker contrasts, in: HGM'99 Program and Abstract Book, The Human Genome Meeting, March 1999, Southbank, Australia, pp. 27–30.

[4] Da Y., VanRaden P.M., Li N., Weller J.I., Schook L.B., Beattie C.W., Designs of resource families for mapping quantitative trait loci using genetic markers in domestic animals, J. Agric. Biotechnol. (1999) (Suppl.) 15–28.

[5] Da Y., Lewin H.A., Linkage information content and efficiency of full-sib and half-sib designs for gene mapping, Theor. Appl. Genet. 90 (1995) 699–706.

[6] Darvasi A., Vinreb A., Minke V., Weller J I., Soller M., Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map, Genetics 134 (1993) 943–951.

[7] Dentine M.R., Cowan C.M., An analytical model for the estimation of chromosome substitution effects in the offspring of individuals heterozygous at a segregating marker locus, Theor. Appl. Genet. 79 (1990) 775–780.
[8] Haldane J B.S., The combination of linkage values, and the calculation of distance between the loci of linked factors, J. Genet. 8 (1919) 299–309.
[9] Haley C.S., Knott S.A., A simple regression method for mapping quantitative trait loci in line crosses using flanking markers, Heredity 69 (1992) 315–324
[10] Haley C.S., Knott S.A., Elsen J.-M., Mapping quantitative trait loci in crosses between outbred lines using least squares, Genetics 136 (1994) 1195–1207.
[11] Henderson C.R , Theoretical basis and computational methods for a number of different animal models, J. Dairy Sci. 71 (1988) (Suppl. 2) 1–26.
[12] Heyen D.W., Weller J.I , Ron M., Band M., Beever J E , Feldmesser E., Da Y., Wiggans G.R., VanRaden P.M., Lewin H.A , A Genome Scan for Quantitative Trait Loci Influencing Milk Production and Health Traits in Dairy Cattle, Physiol. Genomics 1 (1999) 165–175.
[13] Hoeschele I., Uimari P., Grignola F.E., Zhang Q., Gage K.M., Advances in statistical methods to map quantitative trait loci in outbred populations, Genetics 147 (1997) 1445–1457.
[14] Jansen R.C., Controlling the type I and type II errors in mapping quantitative trait loci, Genetics 138 (1994) 871–881.
[15] Jansen R.C., Interval mapping of multiple quantitative trait loci, Genetics 135 (1993) 205–211
[16] Jansen R C., Johnson D.L., van Arendonk J A M., A mixture model approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families, Genetics 148 (1998) 391–399.
[17] Knott S.A., Haley C.S., Maximum likelihood mapping of quantitative trait loci in half-sib families, Genetics 132 (1992) 1211–1222.
[18] Kruglyak L., Lander E.S., High-resolution genetic mapping of complex traits, Am. J. Hum. Genet. 56 (1995) 1212–1223.
[19] Lander E S., Botstein D., Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, Genetics 121 (1989) 185–199
[20] Mackinnon M.J , Weller J.I., Methodology and accuracy of estimation of quantitative trait parameters in half-sib designs using maximum likelihood, Genetics 141 (1995) 755–770.
[21] Manly K.F., Olson J.M., Overview of QTL mapping software and introduction to map manager QT, Mamm. Genome 10 (1999) 327–334.
[22] Martínez O., Curnow R.N., Estimating the location and sizes of effects of quantitative trait loci using flanking markers, Theor. Appl. Genet. 85 (1992) 480–488.
[23] Rodolphe F., Lefort M., A multi-marker model for detection of chromosomal segments displaying QTL activity, Genetics 134 (1993) 1277–1288.
[24] SAS® User's Guide, SAS® Institute Inc. Cary, North Carolina, 1990.
[25] Sillanpää Mikko J., Elja Arjas, Bayesian mapping of multiple quantitative trait loci from incomplete inbred cross data, Genetics 148 (1997) 1373–1388.
[26] Whittaker J.C., Thompson R., Visscher P.M., On the mapping of QTL by regression of phenotype on marker-type, Heredity 77 (1996) 23–32.
[27] Weller J.I., Kashi Y., Soller M., Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle, J. Dairy Sci. 73 (1990) 2525–2537.
[28] Xu S., Further investigation on the regression method of mapping quantitative trait loci, Heredity 80 (1998) 364–373.

[29] Zeng Z.-B., Precision mapping of quantitative trait loci, Genetics 136 (1994) 1457–1468.
[30] Zeng Z.-B., Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci, Proc. Natl Acad. Sci., USA 90 (1993) 10972–10976.
[31] Zhang Q., Boichard D., Hoeschele I., Ernst C., Eggen A., Murkve B., Pfister-Genskow M., Witte L.A., Grignola F., Uimari P., Thaller G., Bishop M.D., Mapping quantitative trait loci for milk production and health of dairy cattle in a large outbred pedigree, Genetics 149 (1998) 1957–1973.

APPENDIX

Proof of equations (4) and (11)

Definitions for QTL effect under the one-way backcross design and the F2 design

For either the backcross design or the F2 design, two parental lines with homozygous marker and QTL genotypes are assumed. Assuming that the QTL and marker each has two alleles, then the two marker and QTL genotypes in the parental lines can be denoted by $MMQQ$ for line 1 and $mmqq$ for line 2. Also assume the one-way backcross is formed by $MQ/mq \times mmqq$ and the F2 design is formed by the mating $MQ/mq \times MQ/mq$, where "/" signifies the linkage phase, *i.e.*, M and Q are on one chromosome and m and q are on the other chromosome in the F1 generation. Then, the backcross offspring have two QTL genotypes, Qq and qq, and the F2 offspring have three QTL genotypes, QQ, Qq, and qq. Let q_{11} = the genotypic value of QQ, q_{12} = the genotypic value of Qq, and q_{22} = the genotypic value of qq. Then the QTL effect is defined as

$$\alpha = q_{12} - q_{22} \quad \text{for the one-way backcross of } MQ/mq \times mmqq \tag{A.1}$$

$$= q_{11} - q_{22} \quad \text{for the F2 design of } MQ/mq \times MQ/mq. \tag{A.2}$$

The rationale for defining α as the difference between the two alternative QTL genotypes is to make the formulae in the text applicable to both designs. Although α is used to denote the QTL effect in both designs, it is important to note the difference shown by (A.1) and (A.2). Another important difference in the α's under the two designs is the interpretation in terms of additive and dominance effect. To show the relationship between the α and additive and dominance effects, the following models can be used:

$$q_{11} = \mu + 2a_1 \quad \text{for } QQ \text{ genotype} \tag{A.3}$$

$$q_{22} = \mu + 2a_2 \quad \text{for } qq \text{ genotype} \tag{A.4}$$

$$q_{12} = \mu + (a_1 + a_2) + d \quad \text{for } Qq \text{ genotype} \tag{A.5}$$

where μ = the overall average of the QTL values, a_i = additive effect of QTL allele i ($i = 1$ or 2), and d = dominance effect of the QTL. From equations (A.1–A.5),

$$\alpha = (a_1 - a_2) + d \quad \text{for the one-way backcross design} \tag{A.6}$$

$$= 2(a_1 - a_2) \quad \text{for the F2 design.} \tag{A.7}$$

An important fact of equations (A.6) and (A.7) is that the QTL effect under the F2 design is twice as large as that under the one-way backcross design if dominance effect is absent. This fact not only is important for correct interpretation of the QTL effect under these two designs, but also is the reason why the F2 design requires a sample size about half as large as that required by the backcross design for detecting additive effects (Da *et al.*, 1999).

Derivation of equation (4) for the one-way backcross design

Equations (3) and (5) can be found in Zeng [30]. Therefore, a proof for equation (4) is given in this section and the next. The frequencies of marker and QTL genotypes in the offspring of the one-way backcross of $MQ/mq \times mmqq$ are given in the following table.

F1 gamete	Gametic frequency	Genotype of backcross offspring	QTL value
MQ	$\frac{1}{2}(1-\theta)$	$MmQq$	q_{12}
Mq	$\frac{1}{2}\theta$	$Mmqq$	q_{22}
mQ	$\frac{1}{2}\theta$	$mmQq$	q_{12}
mq	$\frac{1}{2}(1-\theta)$	$mmqq$	q_{22}

Let m_1 = the average QTL value of individuals with Mm marker genotype, and m_2 = the average QTL value of individuals with mm marker genotype. Then,

$$m_1 = (1-\theta)q_{12} + \theta q_{22}$$
$$m_2 = (1-\theta)q_{22} + \theta q_{12}$$

where θ = recombination frequency between the marker and the QTL. Since Mm and mm have an equal frequency under the one-way backcross, the overall average of the QTL genotypic value is $\mu = (m_1 + m_2)/2$. Then, the variance of marker genotypic averages is given by

$$\begin{aligned}\sigma_m^2 &= \frac{1}{2}(m_1-\mu)^2 + \frac{1}{2}(m_2-\mu)^2 = \frac{1}{2}[(m_1-m_2)/2]^2 + [-(m_1-m_2)/2]^2 \\ &= \frac{1}{4}(m_1-m_2)^2 = \frac{1}{4}(1-2\theta)^2(q_{12}-q_{22})^2 \\ &= \frac{1}{4}(1-2\theta)^2\alpha^2. \end{aligned} \tag{A.8}$$

Derivation of equation (4) for the F2 design

For the F2 design, the marker-QTL genotypes and their frequencies in the F2 generation are given in the following table.

	F1 gamete and frequency			
F1 gamete and frequency	MQ $\frac{1}{2}(1-\theta)$	mQ $\frac{1}{2}\theta$	Mq $\frac{1}{2}\theta$	mq $\frac{1}{2}(1-\theta)$
MQ $\frac{1}{2}(1-\theta)$	$MMQQ$ $\frac{1}{4}(1-\theta)^2$	$MmQQ$ $\frac{1}{4}\theta(1-\theta)$	$MMQq$ $\frac{1}{4}\theta(1-\theta)$	$MmQq$ $\frac{1}{4}(1-\theta)^2$
mQ $\frac{1}{2}\theta$	$MmQQ$ $\frac{1}{4}\theta(1-\theta)$	$mmQQ$ $\frac{1}{4}\theta^2$	$MmQq$ $\frac{1}{4}\theta^2$	$mmQq$ $\frac{1}{4}\theta(1-\theta)$
Mq $\frac{1}{2}\theta$	$MMQq$ $\frac{1}{4}\theta(1-\theta)$	$MmQq$ $\frac{1}{4}\theta^2$	$MMqq$ $\frac{1}{4}\theta^2$	$Mmqq$ $\frac{1}{4}\theta(1-\theta)$
mq $\frac{1}{2}(1-\theta)$	$MmQq$ $\frac{1}{4}(1-\theta)^2$	$mmQq$ $\frac{1}{4}\theta(1-\theta)$	$Mmqq$ $\frac{1}{4}\theta(1-\theta)$	$mmqq$ $\frac{1}{4}(1-\theta)^2$

Let m_1 = the average QTL value of individuals with MM marker genotype, and m_2 = the average QTL value of individuals with mm marker genotype. Then, from the above table and model (2),

$$\begin{aligned} m_1 &= \left\{\frac{(1-\theta)^2}{4}q_{11} + 2\left[\frac{\theta(1-\theta)}{4}\right]q_{12} + \frac{\theta^2}{4}q_{22}\right\}\Big/\frac{1}{4} \\ &= (1-\theta)^2 q_{11} + 2\theta(1-\theta)q_{12} + \theta^2 q_{22}, \\ m_2 &= \left\{\frac{\theta^2}{4}q_{11} + \left[\frac{2\theta(1-\theta)}{4}\right]q_{12} + \frac{(1-\theta)^2}{4}q_{22}\right\}\Big/\frac{1}{4} \\ &= \theta^2 q_{11} + 2\theta(1-\theta)q_{12} + (1-\theta)^2 q_{22}. \end{aligned}$$

Since MM and mm have an equal frequency, the overall average of the QTL genotypic values is $\mu = (m_1 + m_2)/2$. Then, the variance of marker genotypic averages is given by

$$\begin{aligned} \sigma_m^2 &= \frac{1}{2}(m_1-\mu)^2 + \frac{1}{2}(m_2-\mu)^2 = \frac{1}{4}(m_1-m_2)^2 \\ &= \frac{1}{4}(1-2\theta)^2(q_{11}-q_{22})^2 = \frac{1}{4}(1-2\theta)^2\alpha^2. \end{aligned} \tag{A.9}$$

Although equations (A.8) and (A.9) have exactly the same form, it is important to note the difference in definitions for α by equations (A.1), (A.2), (A.6) and (A.7). Correct distinction of the definitions for α is necessary for the correct interpretation of the QTL effect and variance estimated from equations (7), (12–14), (18–19) and (23–24).

Proof of equation (11)

For model (1) fitted with flanking marker A, the regression coefficient of marker A is $b_A = \frac{1}{2}(1 - 2\theta_{Aq})\alpha$. Similarly, for model (1) fitted with flanking marker B, the regression coefficient of marker B is $b_B = 1/2(1-2\theta_{Bq})\alpha$. Hence, $w = |b_A/b_B| = (1 - 2\theta_{Aq})/(1 - 2\theta_{Bq}) = (1 - 2\theta_{Aq})^2/(1 - 2\theta_{AB})$, because $(1 - 2\theta_{Aq})(1 - 2\theta_{Bq}) = (1 - 2\theta_{AB})$ under the assumption that chiasma interference is absent. Solving $w = (1-2\theta_{Aq})^2/(1-2\theta_{AB})$ for θ_{Aq} and requiring the positive root yield equation (11).