Genet. Sel. Evol. 31 (1999) 437–455

Original article

Regression on markers with uncertain allele transmission for QTL mapping in half-sib designs

Haja N. Kadarmideen[a*], Jack C.M. Dekkers[b]

[a] Department of Animal and Poultry Science, University of Guelph, Guelph, ON N1G 2W1, Canada
[b] Department of Animal Science, Iowa State University, Ames, IA 50011-3150, USA

(Received 22 September 1998; accepted 16 August 1999)

Abstract – Recently, regression of phenotype on marker genotypes was described for quantitative trait loci (QTL) mapping in F_2 populations and shown to be equivalent to regression interval mapping (RIM). In this study, regression on markers was extended to half-sib designs with uncertain marker allele transmission, and properties of QTL parameters were examined analytically. In this method, offspring phenotypes are first regressed on the probability of transmission of a given allele from the common parent at flanking marker loci. Resulting regression coefficients can then be interpreted based on an assumed genetic model. With presence of a single QTL in the marker interval, it was shown that expected values of regression coefficients for the flanking markers contained all information about position and effect of the QTL and were independent of the probability of marker allele transmission. Through simulation, it was shown that regression of phenotype on marker allele transmission probabilities is equivalent to RIM under the same assumed genetic model. Regression on marker genotypes is computationally less time consuming than QTL interval mapping, as it eliminates the need to search for the best QTL position across marker intervals. This can form the basis for more efficient methods of analysis with more complex models, including threshold or logistic models for the analysis of categorical traits. © Inra/Elsevier, Paris

genetic marker / QTL mapping / half-sib design

Résumé – Détection de QTLs dans des familles de demi-frères par régression sur des marqueurs avec transmission allélique incertaine. Récemment, la régression des phénotypes sur les génotypes pour les marqueurs a été décrite pour la détection de loci de caractères quantitatifs (QTL) dans des populations F_2. Elle a été montrée équivalente à la détection sur intervalles par régression (RIM). Dans cette étude, la

* Correspondence and reprints: Animal Breeding and Genetics Department, Animal Biology Division, SAC, West Mains Road, Edinburgh EH9 3JG, Scotland, UK
E-mail: h.kadarmideen@ed.sac.ac.uk

régression sur les marqueurs a été étendue aux schémas demi-frères avec transmission incertaine des allèles aux marqueurs et les propriétés des paramètres concernant les QTLs ont été examinées analytiquement. Dans cette méthode, les phénotypes de la descendance ont été d'abord régressés sur la probabilité de transmission d'un allèle donné issu du parent commun à des loci de marqueurs flanquants. Les coefficients de régression résultant peuvent alors être interprétés à partir d'un modèle génétique supposé. En présence d'un seul QTL par intervalle de marqueurs, on a montré que les valeurs espérées des coefficients de régression pour les marqueurs flanquants contenaient toute l'information à propos de la position et de l'effet du QTL, et étaient indépendantes de la probabilité de transmission des allèles aux marqueurs. Par simulation, on a montré que la régression du phénotype sur la probabilité de transmission des allèles aux marqueurs est équivalente au RIM avec le même modèle génétique supposé. La régression sur les génotypes aux marqueurs demande moins de temps de calcul que la détection de QTLs par intervalle, parce qu'éliminant la nécessité de chercher la meilleure position pour le QTL dans les intervalles entre marqueurs. Ceci peut former la base de méthodes plus efficaces avec des modèles plus complexes, incluant les modèles à seuils ou logistiques pour l'analyse des variables discrètes.

marqueur génétique / détection de QTL / schéma demi-frères

1. INTRODUCTION

Identification and mapping of genes affecting quantitative traits, so-called quantitative trait loci or QTL, based on genetic markers has gained much importance in animal and plant genetics in recent years. The main goal behind identifying and mapping QTL is to accelerate genetic progress with the use of information on identified QTL (e.g. [9]). Earlier studies used a single marker approach to detect QTL linked to a marker (e.g. [11]). Lander and Botstein [7] proposed a method to map QTL using two DNA markers that flank a genomic region (so-called interval mapping). Later studies (e.g. [5]) showed that the effect and position of a QTL are confounded in single marker methods and suggested the use of the interval mapping method of Lander and Botstein [7] to overcome this problem. Now, interval mapping of QTL is widely applied in livestock populations based on a variety of statistical methods.

Regression interval mapping (e.g. [3]; henceforth abbreviated to RIM) is based on a genetic model that assumes that a QTL is located in the marker interval. In RIM, phenotypic observations for the quantitative trait are regressed on the probability of offspring inheriting a given QTL allele from a common parent in half-sib designs (e.g. [6, 8, 12]) or from a given parental line in back cross and F_2 designs (e.g. [3]), conditional on a hypothetical position of the QTL in the marker interval. The analysis is repeated for a range of assumed locations of the QTL along the marker interval (grid search). Estimates from the location that gives the minimal residual sum of squares (RSS) are considered to be the best estimates.

Wright and Mowers [14] proposed multiple regression on genetic markers to estimate QTL effect in F_2 designs, which will henceforth be referred to as marker regression mapping (MRM). In contrast to RIM, MRM does not require assumptions about a genetic model in the process of statistical analysis but phenotypic observations are regressed on variables that code which marker allele has been transmitted to offspring, instead of on the probability of the

offspring inheriting a specific QTL allele given QTL position. The resulting estimates of regression coefficients on marker alleles can then be interpreted based on an assumed genetic model. In F_2 designs, Wright and Mowers [14] showed that the sum of partial regression coefficients on flanking markers provides an unbiased estimate of the effect of an additive QTL in the marker interval when interference is complete and when there are no QTL in adjoining marker intervals (isolated QTL). Without complete interference, however, some bias is introduced.

Whittaker et al. [13] showed that the information contained in the regression coefficients on flanking markers in F_2 and back-cross designs is in fact equivalent to that provided by the conventional regression interval mapping of Haley and Knott [3]; with no interference, estimates of QTL position and effect equivalent to those obtained from RIM can be derived as non-linear functions of regression coefficients on flanking markers. Whittaker et al. [13] considered two situations for multiple marker, multiple QTL models: first, isolated QTL, where a marker interval containing a single QTL is flanked by marker intervals devoid of QTL and second, non-isolated QTL, where flanking marker intervals also contain QTL. They showed that, with no interference, expected regression coefficients from a multi-marker multi-QTL model are equivalent to expected regression coefficients from a two-marker single QTL model for markers that flank an isolated QTL. Specifically, Whittaker et al. [13] showed that the partial regression coefficients for markers that flank an isolated QTL depend only on the effects of the QTL in that interval and not on effects at other QTL, as effects of those QTL are accounted for by simultaneous fitting of markers external to the interval. For non-isolated QTL, Whittaker et al. [13] showed that it is impossible to uniquely map two additive QTL in adjoining intervals but that it is possible to map non-isolated QTL if at least one QTL has non-additive effects. The main advantage of MRM for QTL mapping is that estimates are obtained from a single simple linear regression analysis on markers and there is no need for a grid search as in RIM.

Wright and Mowers [14] and Whittaker et al. [13] assumed that transmission of marker alleles from parent to offspring was known with certainty, which is often not the case in half-sib designs. Also, in F_2 or backcrosses between outbred lines, transmission of marker alleles from parental lines may not be known with certainty [4]. In such situations, only a probability statement can be made about marker allele transmission from the parent to progeny. Progenies with incomplete marker information must be included in the statistical analysis to increase the statistical power and reduce bias and standard errors of estimates [12].

The objective of this paper, therefore, was to extend the MRM method of Whittaker et al. [13] to QTL mapping in a half-sib family, with emphasis on uncertain marker allele transmission. Simulation was used to validate methods and to compare MRM to QTL mapping based on RIM.

2. MATERIALS AND METHODS

2.1. The genetic and experimental model

A sire that is heterozygous at two marker loci, 1 and 2, that flank a biallelic QTL is considered. With sire genotype $-M_{11}-Q_1-M_{21}-/-M_{12}-Q_2-M_{22}-$, the QTL is located with recombination rates r_1 and r_2 from marker loci 1 and 2, respectively. Rates r_1 and r_2 are unknown. The recombination rate between marker loci 1 and 2 is θ and is assumed known. The Haldane mapping function [2] is assumed such that $\theta = r_1 + r_2 - 2r_1r_2$.

The sire is randomly mated to n dams, resulting in n offspring. The sire transmits one of four marker haplotypes h_j to its offspring with frequencies $f(h_j)$, where $f(h_j)$ is equal to $(1-\theta)/2$ for marker haplotypes $-M_{11}-M_{21}-$ and $-M_{12}-M_{22}-$, and equal to $\theta/2$ for marker haplotypes $-M_{11}-M_{22}-$ and $-M_{12}-M_{21}-$. Which marker haplotype is transmitted from the sire to progeny cannot always be determined with certainty, but depends on the marker haplotype the progeny received from its dam. The available marker information can, however, be used to compute probabilities of marker allele transmission from the sire to its progeny. The probability of a given paternal marker allele being present in the ith offspring, conditional on the marker information that is available for offspring i (S_i), is denoted as $p(M_{1k}|S_i)$ for marker locus 1 and $p(M_{2\ell}|S_i)$ for marker locus 2. Here, subscripts k ($k = 1, 2$) and ℓ ($\ell = 1, 2$) refer to the paternal marker alleles at marker loci 1 and 2, respectively. The sources of marker information included in S_i could include, besides the known recombination rate between markers, θ, marker genotypes for the flanking markers and possibly other markers on the offspring (g_i), its sire (M_s), its dam (M_d), and other relatives.

2.2. Expected phenotypic value of marker haplotypes

2.2.1. Known marker haplotype transmission

When marker allele transmission from the sire to offspring can be determined unequivocally, the expected value of offspring phenotype given that the offspring received the jth sire marker haplotype can be derived under an assumed genetic model of one QTL in the marker bracket, based on the probability that the paternal marker haplotype carries the Q_1 or Q_2 allele. The expected value of offspring phenotype given marker haplotype h_j is transmitted by the sire can be derived as

$$E(y|h_j) = [w_j - \frac{1}{2}]\alpha \tag{1}$$

Here, $E(y|h_j)$ is the expected value of offspring phenotype given paternal marker haplotype h_j, w_j is the probability that the offspring received the Q_1 allele from the sire conditional on inheritance of paternal marker haplotype h_j, and α is the allele substitution effect at the QTL [1]. Conditional probability w_j can be derived as $w_j = f(Q_1, h_j)/f(h_j)$ where $f(Q_1, h_j)$ is the joint probability of paternal transmission of the Q_1 allele and marker haplotype h_j. Equations for $f(Q_1, h_j)$, $f(h_j)$ and w_j are given in *table I.*

Table I. Joint and conditional probabilities of transmission of allele Q_1 and marker haplotype h_j from sire to progeny for a sire with genotype $M_{11}Q_1M_{21}/M_{12}Q_2M_{22}$.

j	Paternal marker haplotype h_j	Joint probability of Q_1 and h_j transmission $f(Q_1, h_j)$	Frequency of paternal marker haplotype $f(h_j)$	Conditional probability of transmission of Q_1 given marker haplotype h_j w_j
1.	$h_1 = M_{11}M_{21}$	$(1-r_1)(1-r_2)/2$	$(1-\theta)/2$	$(1-r_1)(1-r_2)/(1-\theta)$
2.	$h_2 = M_{11}M_{22}$	$(1-r_1)r_2/2$	$\theta/2$	$(1-r_1)r_2/\theta$
3.	$h_3 = M_{12}M_{21}$	$r_1(1-r_2)/2$	$\theta/2$	$r_1(1-r_2)/\theta$
4.	$h_4 = M_{12}M_{22}$	$r_1r_2/2$	$(1-\theta)/2$	$r_1r_2/(1-\theta)$

j is a numerical code corresponding to sire's marker haplotype h_j.

2.2.2. Unknown marker haplotype transmission

If the paternal marker haplotype transmission is not known with certainty, transmission probabilities can be computed for each paternal marker haplotype based on the marker information that is available for offspring i (S_i). These probabilities, which are denoted as $p(h_j|S_i)$ can then be used to derive the expected value of the ith offspring phenotype, as shown below.

With no interference, $p(h_j|S_i)$ is the product of conditional probabilities for paternal allele transmission at each marker locus:

$$p(h_j|S_i) = p(M_{1k}|S_i) \cdot p(M_{2\ell}|S_i) \tag{2}$$

where k and ℓ are appropriately determined by h_j.

The expected value of the phenotype of offspring i is then obtained as a weighted sum of the expected value of each of the four possible haplotypes, $E(y_i|h_j)$, as:

$$E(y_i|S_i) = \sum_{j=1}^{4} p(h_j|S_i) \cdot E(y_i|h_j) \tag{3}$$

Based on the rules of probability when conditioning on the same source of information S_i, it can be shown that

$$\sum_{j=1}^{4} p(h_j|S_i) = \sum_{1k} \sum_{2\ell} p(M_{1k}|S_i) \cdot p(M_{2\ell}|S_i) = 1$$

Note that probabilities $p(M_{1k}|S_i)$ and $p(M_{2\ell}|S_i)$ are both *dependent* on each others' information (M_{1k} and $M_{2\ell}$) which is included in S_i. Also, note that when probabilities $p(M_{1k}|S_i)$ and $p(M_{2\ell}|S_i)$ are equal to 0 or 1, i.e. when sire marker allele transmission is known, then $E(y_i|S_i) = E(y_i|h_j)$.

2.3. Expected values from regression on flanking markers

Using the expected values for phenotypes of offspring with known and unknown paternal marker haplotype transmission, as derived above, the expected values of coefficients of regression of phenotype on marker allele probabilities can be derived as shown below.

Let $p(M_{11}|S_i) = p_{1i}$ and $p(M_{21}|S_i) = p_{2i}$.

The model for regressing phenotype on marker allele transmission probabilities is

$$y_i = \beta_o + \beta_1 p_{1i} + \beta_2 p_{2i} + e_i \tag{4}$$

where y_i is the phenotype of offspring i, β_o is the overall mean, β_1 is the regression coefficient on marker 1, β_2 is the regression coefficient on marker 2, e_i is the error term for the ith offspring and all other terms are as described earlier.

In matrix notation, the MRM model can be written as $\mathbf{Y} = \mathbf{P}\beta + \mathbf{e}$, where $\mathbf{Y}$ is a vector of observations on n offspring with size $n \times 1$, $\mathbf{P}$ is a matrix of size $n \times 3$, and β is of size 3×1 with $\beta = (\beta_0\ \beta_1\ \beta_2)'$. When phenotypic observations are adjusted for the mean genetic values of parents and for all other systematic environmental effects, the expectation of an observation y_i, with marker information S_i, is equal to $E(y_i|S_i)$, which can be calculated using equation (3). Based on equation (3), the expectation of the vector of adjusted observations $\mathbf{y}$ can be written as a product of two matrices: $E(\mathbf{y}) = \mathbf{Hw}$ where $\mathbf{H}$ is a matrix of haplotype transmission probabilities of size $n \times 4$ and $\mathbf{w}$ is a 4×1 vector with haplotype coefficients w. Based on equation (2), haplotype transmission probabilities, $p(h_j|S_i)$ can be written in terms of $p(M_{11}|S_i) = p_{1i}$ and $p(M_{21}|S_i) = p_{2i}$. Equations for $E(\mathbf{y})$ are:

$$E\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} p_{11}p_{21} & p_{11}(1-p_{21}) & (1-p_{11})p_{21} & (1-p_{11})(1-p_{21}) \\ p_{12}p_{22} & p_{12}(1-p_{22}) & (1-p_{12})p_{22} & (1-p_{12})(1-p_{22}) \\ \vdots & \vdots & \vdots & \vdots \\ p_{1n}p_{2n} & p_{1n}(1-p_{2n}) & (1-p_{1n})p_{2n} & (1-p_{1n})(1-p_{2n}) \end{bmatrix}$$

$$\left(\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} - \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \right) \alpha \tag{5}$$

Matrix $\mathbf{P}$ is given as,

$$\mathbf{P} = \begin{bmatrix} 1 & p_{11} & p_{21} \\ 1 & p_{12} & p_{22} \\ 1 & . & . \\ 1 & . & . \\ 1 & . & . \\ 1 & p_{1n} & p_{2n} \end{bmatrix} \tag{6}$$

Expected values of the regression coefficients can be derived based on

$$E(\widehat{\beta}) = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'E(\mathbf{y}) \tag{7}$$

Derivations for $E(\widehat{\beta})$ in equation (7) are given in Appendix I. The resulting elements in $E(\widehat{\beta})$, after simplification, can be shown to be independent of the paternal marker allele transmission probabilities as

$$E(\widehat{\beta}) = E\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} (1 - w_1) - \frac{1}{2} \\ w_1 - w_3 \\ w_1 - w_2 \end{bmatrix} \alpha = \begin{bmatrix} w_4 - \frac{1}{2} \\ w_2 - w_4 \\ w_3 - w_4 \end{bmatrix} \alpha \tag{8}$$

Substituting formulas from *table I* for w_j in equation (8), it can be shown that the regression coefficients are equal to

$$E(\widehat{\beta}) = E\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \dfrac{r_1 r_2}{(1-\theta)} - \dfrac{1}{2} \\ \dfrac{r_2(1-r_2)(1-2r_1)}{\theta(1-\theta)} \\ \dfrac{r_1(1-r_1)(1-2r_2)}{\theta(1-\theta)} \end{bmatrix} \alpha \tag{9}$$

Equation (9) proves that $E(\widehat{\beta})$ depends only on the coefficients w_j and is *independent* of marker allele transmission probabilities $p(M_{11}|S_i)$ and $p(M_{21}|S_i)$. In other words, $E(\widehat{\beta})$ depends only on contrasts between sire marker alleles M_{11} and M_{12} for locus M_1 and between alleles M_{21} and M_{22} for locus M_2. The expectations of marker regression coefficients are identical to those found by Whittaker et al. [13] for F_2 designs but are shown here to apply also for half-sib family designs and with uncertain marker haplotype transmission. An alternative proof is also given in Appendix II.

2.4. QTL location and its effect

The estimates of the partial regression coefficients $\widehat{\beta}_1$ and $\widehat{\beta}_2$ (equation 9) contain all information to determine the position of a QTL that is flanked by markers M_1 and M_2. The absolute value of $E(\widehat{\beta}_1)$ will be greater than the absolute value of $E(\widehat{\beta}_2)$ if the QTL is located closer to marker M_1, and smaller

if the QTL is located closer to marker M_2. If the QTL is located at the centre of the interval, we would expect $E(\widehat{\beta}_1)$ and $E(\widehat{\beta}_2)$ to be equal. The relative size of the estimates of the regression coefficients β_1 and β_2 leads us to determine the QTL position r_1. As shown by Whittaker et al. [13], estimates of QTL location and QTL effect can be obtained by writing $E(\widehat{\beta}_1)$ and $E(\widehat{\beta}_2)$ as a ratio and solving for r_1, knowing that $r_1 \in (0, 0.5)$.

Following Whittaker et al. [13], the estimate of QTL location ($\widehat{r}_1$) is given as

$$\widehat{r}_1 = \frac{1}{2}^{*}\left[1 - \sqrt{1 - \frac{4\widehat{\beta}_2\theta(1-\theta)}{\widehat{\beta}_2 + \widehat{\beta}_1(1-2\theta)}}\right] \tag{10}$$

Once the QTL location has been estimated, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ can be equated to their expectation, replacing r_1 with $\widehat{r}_1$ and solving for $\widehat{\alpha}$. Following Whittaker et al. [13], $\widehat{\alpha}$ is obtained from

$$\widehat{\alpha}^2 = \frac{\left[[\widehat{\beta}_1 + (1-2\theta)\widehat{\beta}_2][\widehat{\beta}_2 + (1-2\theta)\widehat{\beta}_1]\right]}{(1-2\theta)} \tag{11}$$

Note that a solution to equation (10) only exists if $\widehat{\beta}_1$ and $\widehat{\beta}_2$ have the same sign. If $\widehat{\beta}_1$ and $\widehat{\beta}_2$ have opposite signs, the solution for r_1 is undefined with respect to presence of a single QTL within the marker interval. If $\widehat{\beta}_1$ and $\widehat{\beta}_2$ have the same sign, an estimate of α can be obtained from equation (11) as $\sqrt{\widehat{\alpha}^2}$. If $\widehat{\beta}_1$ and $\widehat{\beta}_2$ have opposite signs, the solution for α is undefined. When a solution for r_1 exists, the sign of α can be determined, based on the signs of $\widehat{\beta}_1$ and $\widehat{\beta}_2$. The sign for α will be negative if $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are both negative and positive if $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are both positive.

2.5. Validation

In the previous section, it was proven analytically that the expectation of the partial regression coefficients are invariable to transmission probabilities. In this section, the analytical proof will be validated by simulation. A single sire family with 100 half-sib progeny was simulated. The recombination rate between QTL and the left marker, r_1, was 0.3 and between flanking markers, θ, was 0.4. Expectations of offspring phenotypes given paternal marker haplotype, $E(y|h_j)$ were then calculated using equation (1). The w_js needed for the computation of $E(y|h_j)$ were obtained from substituting $r_1 = 0.3$, $r_2 = (\theta - r_1)/(1 - 2r_1) = 0.25$ and $\theta = 0.4$ in the formulas for w_j in *table I.* They were: $w_1 = 0.87500$, $w_2 = 0.43750$, $w_3 = 0.56250$ and $w_4 = 0.12500$. To ensure generality, each offspring was randomly assigned a value for the probability that it received alleles M_{11} ($p(M_{11})$) and M_{21} ($p(M_{21})$) from the sire based on random draws from a uniform (0,1) distribution. Based on these probabilities, expectations of offspring phenotypes $E(y_i)$ were simulated using equation (3). Observations were then regressed on sire marker allele probabilities using model [4]. The resulting regression coefficients (from a single replicate) were $\widehat{\beta}_1 = 0.3125$ and

$\widehat{\beta}_2 = 0.4375$, which is identical to results obtained when substituting $r_1 = 0.3$, $r_2 = 0.25$ and $\theta = 0.4$ in the formula for $E(\widehat{\beta}_1)$ and $E(\widehat{\beta}_2)$ in equation (9).

2.6. Comparison of MRM and RIM

2.6.1. Simulation

To compare MRM with RIM for QTL mapping, a single sire family with 500 offspring was simulated. The genome of the sire carried a pair of homologous chromosomes with two biallelic markers with a spacing of 20 cM. A QTL was simulated at 5, 10 or 15 cM from the left marker, which corresponds to recombination rates of 0.04758, 0.09063 and 0.12959 with the left marker. The sire was heterozygous at both marker loci and at the QTL, denoted as $-M_{11} - Q_1 - M_{21} - / - M_{12} - Q_2 - M_{22}-$. Marker-QTL (MQTL) haplotypes produced by this sire were sampled according to their expected frequencies of transmission. Maternal marker haplotypes were sampled based on population frequencies for M_{11} and M_{21}. The marker genotype of each offspring was generated by combining paternal MQTL with the maternal marker haplotype.

Phenotypic values of offspring were generated using the following model

$$y_i = u + q_i + e_i,$$

where y_i is the phenotypic observation on the ith offspring, u is the sire's polygenic effect, q_i is the effect of the paternal QTL allele (Q_1 or Q_2) inherited by offspring i, and e_i is a random residual. Residuals were sampled from $\mathbf{N}[0, \sigma_p^2 - (0.25\sigma_a^2 + 0.5\sigma_{\text{QTL}}^2)]$, where σ_p^2 is the phenotypic variance, σ_a^2 is the polygenic variance and σ_{QTL}^2 is the QTL variance in the dam population, which was based on equal frequencies for the two QTL alleles among dams. A total heritability of 0.25, including the QTL effect, was used. The QTL substitution effect, α, was $0.4\sigma_p$. A total of 1 000 data sets was simulated for each QTL position. Each data set was analysed by MRM and RIM.

2.6.2. Analysis

2.6.2.1. Conditional probabilities for MRM and RIM

For RIM, the conditional probability that the QTL allele (Q_1) which is associated with marker allele M_{11} in the sire was transmitted from the sire to offspring i was computed as shown in Liu and Dekkers [8]. For MRM, computation of conditional probabilities of paternal transmission of alleles M_{11} and M_{21} is given in Appendix III.

2.6.2.2. Parameter estimation: RIM and MRM

For RIM, parameters (QTL location and effect) were estimated with a search for QTL at every cM in the 20 cM marker interval (e.g. [3]). For MRM, parameters were estimated based on the theory described earlier. For MRM, the estimated regression coefficients ($\widehat{\beta}_1$ and $\widehat{\beta}_2$) must have equal signs to obtain estimates of r_1 and α based on equations (10) and (11), respectively. Whittaker

et al. [13] suggested that estimates of regression coefficients with opposite signs could result when i) the data do not support the presence of a single QTL in the marker interval, ii) the data support the presence of two QTL with opposite signs in the interval, and iii) the data suggest that a QTL is located outside the marker bracket. With regard to possibility iii), if the QTL is estimated to be outside marker 1, $\widehat{\beta}_1$ will have a greater absolute value than $\widehat{\beta}_2$. Similarly, if the QTL is estimated to be outside marker 2, $\widehat{\beta}_2$ is expected to have a greater absolute value than $\widehat{\beta}_1$. When data suggest that a QTL is outside the marker bracket, the estimate of r_1 by MRM will be negative or greater than θ or be undefined. In this situation, RIM would show minimum RSS at one of the marker loci because the search with RIM is limited to the marker bracket.

Based on the above and to allow comparison of results from MRM with results from RIM, the QTL was positioned at one of the markers based on the largest absolute value of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ when regression coefficients from MRM had opposite signs: the QTL was located at M_1 if $|\widehat{\beta}_1| > |\widehat{\beta}_2|$ and at M_2 if $|\widehat{\beta}_1| < |\widehat{\beta}_2|$. The estimate of the QTL effect was obtained as $\sqrt{|\widehat{\alpha}^2|}$ based on equation (11). Note that this approach was applied *only* if regression coefficients had opposite signs in a given replicate. Forcing the QTL to lie at one of the markers is analogous to RIM, for which the QTL is located at a marker when the estimate of location falls outside the marker bracket.

2.6.2.3. Test of significance for presence of a QTL

For MRM, a likelihood ratio (LR) test statistic was obtained as for RIM by computing:

$$LR = n\log_e(RSS_{\text{red}}/RSS_{\text{full}})$$

where n is the total number of offspring in the half-sib family, RSS_{red} is the residual sum of squares when fitting only an overall mean and RSS_{full} is the residual sum of squares when the full model was fitted (equation (4)).

For RIM, table values cannot be used for significance testing because the model is fit at multiple positions (e.g. [6]). With regression on markers, only a single model is fit and, hence, table values should apply. For completeness, however, significance threshold values were determined empirically for both MRM and RIM from data generated under the null hypothesis.

3. RESULTS

3.1. QTL location and effect

Empirical means and standard deviations of marker regression coefficients for MRM are given in *table II* for different QTL positions. Equal values for $\widehat{\beta}_1$ and $\widehat{\beta}_2$ were as expected for a QTL that is located in the centre of the marker bracket (10 cM). For other QTL locations (5 and 15 cM), the marker that is closer to the QTL has a greater value for regression coefficient than the other marker.

Table II. Empirical means and standard deviations (in parenthesis) of estimates of marker regression coefficients based on MRM ($\widehat{\beta}_1$ and $\widehat{\beta}_2$) for a QTL located 5, 10 or 15 cM from the left marker and with an effect equal to 0.4 phenotypic standard deviations for a half-sib family with 500 progeny. Results were based on 1 000 replicates.

QTL location	$\widehat{\beta}_1$	$\widehat{\beta}_2$
5 cM	0.30 (0.20)	0.10 (0.21)
10 cM	0.20 (0.20)[a]	0.20 (0.20)[a]
15 cM	0.09 (0.20)	0.31 (0.20)

[a] Values with the same superscript within a combination of parameters are not significantly different: $P < 0.001$.

Empirical means and standard deviations of estimates of QTL position and effect from MRM and RIM are given in *table III*. Estimates of QTL location from MRM and RIM were not significantly different and had a correlation close to unity (0.999) for all situations. Both RIM and MRM resulted in unbiased estimates of r_1 when the QTL was located at the centre of the marker bracket but were significantly biased towards the centre of the marker bracket when the true QTL location was off centre (5 and 15 cM). This bias is as expected, because we are forcing estimates to lie within the interval, in which there is more room for error to the right (or left) of the true location, resulting in the observed bias. For MRM, 38, 33 and 38 % of replicates had estimates of marker regression coefficients with opposite signs when the QTL was located at 5, 10 and 15 cM, respectively. For RIM, the estimate of QTL position was at a marker for 40, 35 and 40 % of replicates, for QTL positions of 5, 10 and 15 cM, respectively. This indicates that MRM and RIM have similar frequencies of locating the QTL within the marker bracket. Estimates of QTL effects did not significantly differ between RIM and MRM and had correlations equal to 0.969, 0.980 and 0.970, for QTL located at 5, 10 and 15 cM, respectively. Estimates of QTL effects were unbiased for both RIM and MRM.

3.2. Significance threshold values and power

Values of the LR test statistic were very similar for RIM and MRM under the alternate hypothesis and had correlations of 0.993, 0.997 and 0.996 for QTL located at 5, 10 and 15 cM, respectively. Two sets of empirical significance threshold values were determined for RIM and MRM for each simulated QTL location: the first set (unrestricted) was derived from 10 000 replicates under the null hypothesis irrespective of existence of a solution for QTL position under MRM. The second set of significance thresholds (restricted) was determined only from replicates for which estimates of QTL position and effect existed under MRM. The purpose of this restriction was to limit analyses to replicates for which the estimates of QTL position was inside the marker interval. To obtain the restricted significance thresholds, 50 000 replicates were run, of which only 9 765, 9 750 and 9 803 had useable solutions for QTL located at 5, 10 and 15 cM, respectively. This is as expected because data sets under the

Table III. Empirical means and standard deviations (in parenthesis) of estimates of QTL location and effect based on RIM and MRM for a QTL located 5, 10 or 15 cM from the left marker and with an effect equal to 0.4 phenotypic standard deviations for a half-sib family with 500 progeny. Results were based on 1 000 replicates.

True QTL location	QTL location		QTL effect	
cM (r_1)	RIM	MRM	RIM	MRM
5	0.05802[a,*]	0.05802[a,*]	0.41[a]	0.40[a]
(0.04758)	(0.05601)	(0.05612)	(0.11)	(0.12)
10	0.08561[a]	0.08564[a]	0.41[a]	0.40[a]
(0.09063)	(0.06000)	(0.05977)	(0.11)	(0.12)
15	0.11535[a,*]	0.11547[a,*]	0.41[a]	0.40[a]
(0.12951)	(0.05348)	(0.05318)	(0.11)	(0.12)

[a] Values with the same superscript within a combination of parameters are not significantly different: $P < 0.001$.
* Estimates significantly different from the true QTL location: $P < 0.001$.

null hypothesis are simulated with no QTL in the marker interval. Significance threshold values for RIM were obtained from the same replicates as used for MRM. Resulting threshold values are given in *table IV*.

Table IV. Empirical significance threshold values for RIM and MRM for a half-sib family with 500 progeny. Results were based on 10 000 replicates irrespective of signs of regression coefficients (Unrestricted) and on 9 765 replicates which resulted in the same signs for regression coefficients (Restricted).

	Unrestricted		Restricted		χ^2_1	χ^2_2
Significance level	RIM	MRM	RIM	MRM		
1 %	7.36	8.84	9.13[a]	9.13[a]	6.63	9.21
5 %	4.66	6.06	6.05[a]	6.05[a]	3.84	5.99
10 %	3.40	4.64	4.64[a]	4.64[a]	2.71	4.61

[a] Values with the same superscript within a combination of parameters are not significantly different: $P < 0.001$.

Restricted threshold values were similar for MRM and RIM (*table IV*). Unrestricted threshold values were similar to restricted threshold values for MRM but smaller than restricted threshold values for RIM. For MRM, unrestricted and restricted threshold values were higher than table values for χ^2_1 [10] but were close to χ^2_2 table values (*table IV*). For RIM, unrestricted significance threshold values were higher than table values for χ^2_1 but lower than table values for χ^2_2. However, restricted significance threshold values for RIM were close to χ^2_2 table values. Correlations of LR test statistics for RIM and MRM under

the null hypothesis were 1.000 when based on the restricted data sets. The empirical power to detect the QTL was also calculated based on the two sets of significance threshold values and are given in *table V*. The power of RIM and MRM was significantly different when based on either unrestricted or restricted significance threshold values, except for the restricted threshold values when the QTL was at the centre of the marker bracket (10 cM).

Table V. Empirical power (%) for RIM and MRM based on 1 000 replicates for unrestricted and restricted significance threshold values.

True QTL location in cM	Significance level	Unrestricted		Restricted	
		RIM	MRM	RIM	MRM
5 cM	1 %	81.0	75.1	71.6	73.0
	5 %	92.7	89.0	87.0	89.0
10 cM	1 %	79.3	70.2	68.1[a]	69.0[a]
	5 %	91.6	87.4	87.0[a]	87.3[a]
15 cM	1 %	79.2	73.3	70.5	72.2
	5 %	92.2	88.0	87.2	88.1

[a] Values with the same superscript within a combination of parameters are not significantly different: $P < 0.001$.

When power was computed only from replicates for which estimates of QTL position existed with MRM (620, 670 and 620 of 1 000 replicates when the QTL was located at 5, 10 and 15 cM, respectively), the power of RIM and MRM was not significantly different for any QTL location.

4. DISCUSSION

In this study, the method of multiple regression of phenotype on marker genotypes for QTL mapping in F_2 populations [13] was extended to a half-sib family design.

In contrast to Wright and Mowers [14] and Whittaker et al. [13], offspring with complete and incomplete marker information on paternal marker allele transmission were included in the analysis. Inclusion of offspring with incomplete marker information in QTL mapping results in higher statistical power and lower standard errors and bias of estimates of QTL location and QTL effect [12].

It was shown that regression coefficients and hence the resulting estimates of QTL parameters did not depend on transmission probabilities. The regression coefficients only depended on contrasts between marker haplotype class means under known marker haplotype transmission. Although, results from this study focused on half-sib designs, uncertainty of marker allele transmission can also apply to F_2 and backcross designs that involve outbred lines and to QTL mapping with markers of limited polymorphism.

Although MRM and RIM are essentially equivalent, the two methods resulted in different test statistics under the null and alternate hypothesis and,

therefore, had different power to detect a QTL (*table V*). These differences were found to be caused by the fact that MRM does not restrict the test for the QTL to within the marker interval. Rather, the test is for a QTL anywhere on the chromosome. Furthermore, MRM does not make assumptions on the genetic model in the process of analysis and any effects that are present in the data, even if they do not conform with a genetic model of one QTL within the marker bracket, are picked up by the regression coefficients. The RIM, on the other hand, assumes a genetic model (usually of one QTL within the marker bracket) and, in the present study, searches for the QTL only within the marker bracket; if data indicated a QTL outside the marker bracket, the QTL was mapped to one of the markers. To compare results from RIM and MRM on an equivalent basis, MRM estimates of location outside the marker interval were forced to be at the nearest marker (*table III*). This was used to illustrate that MRM and RIM are equal when the search is restricted to between the two flanking markers: RIM and MRM had similar LR test statistics under the null and alternate hypotheses (correlation of 1.000, *table V*) and identical power (not shown). An alternate way of comparing these methods would be to also search for the QTL outside the interval by fitting markers outside the marker bracket under study. In this case, MRM and RIM are expected to give identical results. One advantage of RIM over MRM, is that the LR test statistic (or RSS) will be continuous across marker intervals, and can be used to provide a graphical representation of the location of the likelihood which can, therefore, be used as a 'confidence region'.

Empirical thresholds for MRM were similar to standard Chi-square values with two degrees of freedom. Empirical thresholds for MRM were not affected by exclusion of replicates for which a solution for QTL position did not exist (*table IV*). Empirical threshold values for RIM were intermediate to Chi-square values with one and two degrees of freedom when computed from all replicates (unrestricted) but were close to Chi-square values with two degrees of freedom when computed from replicates for which a solution for QTL position existed with MRM (restricted). This raises the question on the number of degrees of freedom that are available for interval versus marker regression mapping methods in relation to the number of parameters that are estimated. Note that for MRM two parameters are estimated (two regression coefficients). Accordingly, significance thresholds were similar to Chi-square table values with two degrees of freedom. For RIM, two parameters are estimated (QTL position and QTL effect) if the QTL is mapped to between the two markers, but only one parameter is estimated if the data suggest the QTL is outside the marker bracket. In the later case, the QTL is mapped to one of the markers. In order to test the existence of such a mixture distribution of the LR test statistics for RIM, 10 000 replicates were generated under the null hypothesis and threshold values were determined based on replicates in which the QTL was mapped outside versus inside the marker bracket (8 216 versus 1 784 replicates, respectively). When the QTL was mapped outside the marker bracket, 1 and 5 % significance threshold values (based on 8 216 replicates) were 7.05 and 4.34, respectively, which were slightly higher than Chi-square table values with one degree of freedom (6.83 and 3.84). When the QTL was mapped inside the marker bracket, 1 and 5 % significance threshold values (based on 1 784 replicates) were 8.58 and 5.93, respectively, which were slightly

less than Chi-square table values with two degrees of freedom (9.21 and 5.99). Therefore, the differences in threshold values between RIM and MRM were due to differences in treatment of QTL fitted outside the marker bracket. As mentioned earlier, RIM and MRM may yield similar results when fitting more markers and searching for a QTL among marker brackets on the chromosome.

When regression is performed on multiple markers, MRM amounts to standard multiple regression, as described by Wright and Mowers [14]. With no interference, only marker brackets which contain a QTL are expected to give non-zero regression coefficients and those that are devoid of QTL are expected to give zero regression coefficients. For multiple QTL located on the same chromosome, results from a two-marker single QTL model is equivalent to a multi-marker multi-QTL model when QTL are isolated, as shown by Whittaker et al. [13]. That is, if a second QTL exists on the same chromosome, its effect on the expected regression coefficients from the two-marker single QTL model, can be removed by fitting a conditional regression on a marker positioned outside the interval but between the interval and the second QTL. The same procedure also applies to RIM [13]. When multiple QTL are located within the same marker interval, no unique and independent estimates of QTL parameters can be obtained with RIM or MRM [13] and possibly with other statistical methods. In such cases, regression coefficients would simply relate to some weighted average of QTL effects and positions for both RIM and MRM.

The MRM studied here was for a single sire family. There are difficulties associated with extension of this method to QTL mapping in a multi-family half-sib design, as studied by, for example, Knott et al. [6] and Liu and Dekkers [8]. In a multi-family analysis with RIM, a nested regression is used with one unique estimate of QTL location but different QTL substitution effects for each sire. Although the MRM method can be extended to multiple families by nesting regression coefficients within family, each family will receive a separate estimate of QTL location and effect. This problem may be overcome by fitting markers as random effects and by expressing estimated variances at markers in terms of a genetic model of one QTL with multiple alleles.

The MRM method described in this study shows that information to map QTL is derived entirely from contrasts between marker-associated effects at flanking markers, regardless of uncertainty of marker allele transmission. However, the uncertainty of marker transmission results in increased standard error for the regression coefficients. This study has provided further insight into properties of the test statistic for RIM. Specifically, results illustrate that the difference between empirical and table threshold values is not due to multiple testing within the marker interval but results from a mixture of fitting one (when the QTL is positioned outside the marker bracket) and two parameters. The computational efficiency of MRM over RIM may be of little importance for least-square analyses because the computational demands of RIM are already limited. The same principle of regression on markers can, however, also be applied to other types of models, for example threshold and other non-linear models, for which computing time is of importance.

In general, the marker regression method can be applied to QTL mapping studies where the RIM is considered to be the method of choice. Because of the simplicity of the MRM method, initial screening of marker data can be performed with this method to identify regions displaying QTL activity before

adopting advanced statistical methods such as maximum likelihood, generalized linear mixed models, non-parametric or Bayesian methods. Once potential QTL regions are identified we can either choose to adopt advanced methods focused on those genomic regions or simply interpret the regression coefficients based on a genetic model.

ACKNOWLEDGEMENTS

This research was supported by Natural Sciences and Engineering Research Council of Canada and by Hatch Act and State of Iowa funds of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, USA (Project No. 3456, Journal Paper No. J-18024). An anonymous reviewer is thanked for providing the alternative proof.

REFERENCES

[1] Falconer D.S., Mackay T.F.C., Introduction to Quantitative Genetics, 4th ed., Longman, Harlow, UK, 1996.

[2] Haldane J.B.S., The combination of linkage values, and the calculation of distances between the loci of linked factors, J. Genet. 8 (1919) 299–309.

[3] 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.

[4] Haley C.S., Knott S.A., Elsen J., Mapping quantitative trait loci in crosses between outbred lines using least squares, Genetics 136 (1994) 1195–1207.

[5] Knott S.A., Haley C.S., Aspects of maximum-likelihood methods for the mapping of quantitative trait loci in line crosses, Genet. Res. 60 (1992) 139–151.

[6] Knott S.A., Elsen J.M., Haley C.S., Methods of multiple-marker mapping of quantitative trait loci in half-sib populations, Theor. Appl. Genet. 93 (1996) 71–80.

[7] Lander E.S., Botstein D., Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, Genetics 121 (1989) 185–199.

[8] Liu Z., Dekkers J.C.M., Least squares interval mapping of quantitative trait loci under the infinitesimal genetic model in outbred populations, Genetics 148 (1998) 495–505.

[9] Meuwissen T.H.E., Goddard M.E., The use of marker haplotypes in animal breeding schemes, Genet. Sel. Evol. 28 (1996) 161–176.

[10] Snedecor G.W., Cochran W.G., Statistical Methods, The Iowa state university press, Ames, USA, 1967.

[11] Soller M., Genizi M.A., The efficiency of experimental designs for detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations, Biometrics 34 (1978) 47–55.

[12] Wang Y., Detection and estimation of associations between genetic markers and quantitative trait loci in segregating populations, Ph.D. thesis, University of Guelph, Guelph, 1996.

[13] 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.

[14] Wright A.J., Mowers R.P., Multiple regression for molecular-marker, quantitative trait data from large F_2 populations, Theor. Appl. Genet. 89 (1994) 305–312.

APPENDIX 1: Derivations of expected values of the regression coefficients for flanking markers

Expected values of the regression coefficients can be derived based on:

$$E(\widehat{\beta}) = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'E(\mathbf{y}) \tag{A1}$$

Since $E(\mathbf{y}) = \mathbf{Hw}$ from equation (5), $E(\widehat{\beta})$ can also be written as

$$E(\widehat{\beta}) = [(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{H}]\mathbf{w} = \mathbf{Cw} \tag{A2}$$

with $\mathbf{C} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{H}$.

Noting that the conditional probability $p(M_{12}|S_i)$, is equal to $(1 - p_{1i})$ and $p(M_{22}|S_i)$, is equal to $(1 - p_{2i})$, and simplifying $\mathbf{C}$ in equation (A2), it can be shown that $\mathbf{C}$ matrix simplifies to

$$\mathbf{C} = \begin{bmatrix} \lambda & -\lambda & -\lambda & (1+\lambda) \\ \gamma & (1-\gamma) & -\gamma & (\gamma-1) \\ \delta & -\delta & (1-\delta) & (\delta-1) \end{bmatrix} \tag{A3}$$

where

$$\begin{aligned}\lambda = [(& \sum_{i=1}^{n} p_{1i}^2 \sum_{i=1}^{n} p_{2i}^2 \sum_{i=1}^{n} p_{1i}p_{2i} - (\sum_{i=1}^{n} p_{1i}p_{2i})^3 + \sum_{i=1}^{n} p_{2i} \sum_{i=1}^{n} p_{1i}p_{2i} \sum_{i=1}^{n} p_{1i}^2 p_{2i} \\ - & \sum_{i=1}^{n} p_{1i} \sum_{i=1}^{n} p_{2i}^2 \sum_{i=1}^{n} p_{1i}^2 p_{2i} + \sum_{i=1}^{n} p_{1i} \sum_{i=1}^{n} p_{1i}p_{2i} \sum_{i=1}^{n} p_{1i}p_{2i}^2 \\ - & \sum_{i=1}^{n} p_{2i} \sum_{i=1}^{n} p_{1i}^2 \sum_{i=1}^{n} p_{1i}p_{2i}^2)/D]\end{aligned}$$

$$\begin{aligned}\gamma = [(& \sum_{i=1}^{n} p_{2i}(\sum_{i=1}^{n} p_{1i}p_{2i})^2 - \sum_{i=1}^{n} p_{1i} \sum_{i1}^{n} p_{2i}^2 \sum_{i=1}^{n} p_{1i}p_{2i} + n \sum_{i=1}^{n} p_{2i}^2 \sum_{i=1}^{n} p_{1i}^2 p_{2i} \\ & -(\sum_{i=1}^{n} p_{2i})^2 \sum_{i=1}^{n} p_{1i}^2 p_{2i} + \sum_{i=1}^{n} p_{1i} \sum_{i=1}^{n} p_{2i} \sum_{i=1}^{n} p_{1i}p_{2i}^2 \\ & -n \sum_{i=1}^{n} p_{1i}p_{2i} \sum_{i=1}^{n} p_{1i}p_{2i}^2)/D]\end{aligned}$$

and

$$\begin{aligned}\delta = [(& \sum_{i=1}^{n} p_{1i}(\sum_{i=1}^{n} p_{1i}p_{2i})^2 - \sum_{i=1}^{n} p_{2i} \sum_{i=1}^{n} p_{1i}^2 \sum_{i=1}^{n} p_{1i}p_{2i} + \sum_{i=1}^{n} p_{1i} \sum_{i=1}^{n} p_{2i} \sum_{i=1}^{n} p_{1i}^2 p_{2i} \\ & - n \sum_{i=1}^{n} p_{1i}p_{2i} \sum_{i=1}^{n} p_{1i}^2 p_{2i} + n \sum_{i=1}^{n} p_{1i}^2 \sum_{i=1}^{n} p_{1i}p_{2i}^2 - (\sum_{i=1}^{n} p_{1i})^2 \sum_{i=1}^{n} p_{1i}p_{2i}^2)/D]\end{aligned}$$

with $$D = n\sum_{i=1}^{n} p_{1i}^2 \sum_{i=1}^{n} p_{2i}^2 - n(\sum_{i=1}^{n} p_{1i}p_{2i})^2$$
$$-(\sum_{i=1}^{n} p_{1i})^2 \sum_{i=1}^{n} p_{2i}^2 + 2\sum_{i=1}^{n} p_{1i} \sum_{i=1}^{n} p_{2i} \sum_{i=1}^{n} p_{1i}p_{2i} - (\sum_{i=1}^{n} p_{2i})^2 \sum_{i=1}^{n} p_{1i}^2$$

Note that elements within rows of **C** in (A3) corresponding to the mean and the two regression coefficients sum to one and zero, respectively.

Based on *table I*, the coefficients for expectations of sire marker haplotypes, w_3 and w_4 are equal to $1-w_2$ and $1-w_1$, respectively. Substituting, $w_3 = 1-w_2$ and $w_4 = 1 - w_1$ in vector **w** given in equation (A2), it can be shown that the resulting equations simplify to $E(\widehat{\beta})$ given in equation (9).

APPENDIX 2: Alternative proof

Let y, g and s be the phenotypic value, genetic value and available marker information, respectively, for an individual, and let $h = (h_l, h_r)$ be the complete marker information at the flanking markers. Then, suppressing the constant term for convenience

$$\begin{aligned} E(y|s) &= E(g|s) = \Sigma_g\, g\, p(g|s) = \Sigma_g \Sigma_h\, g\, p(g|s,h)p(h|s) \\ &= \Sigma_g \Sigma_h\, g\, p(g|h)p(h|s) = \Sigma_h\, E(g|h)p(h|s) \\ &= \Sigma_h(\lambda h_l + \rho h_r)p(h|s) \text{ (from Whittaker et al. [13])} \\ &= \lambda E(h_l|s) + \rho E(h_r|s) \end{aligned}$$

As in Whittaker et al. [13], it follows that λ and ρ are regression coefficients, but now from the regression of phenotype on *expected*, rather than actual, marker genotypes.

APPENDIX 3: Transmission probabilities for paternal marker alleles

The conditional probability for transmission of marker alleles M_{11} and M_{21} from the sire to offspring i, conditional on marker linkage phase L_1 in the sire are denoted by $p(M_{11}|S_i = g_i, M_s, L_1, r_1, \theta)$ and $p(M_{21}|S_i = g_i, M_s, L_1, r_1, \theta)$, respectively. Similarly, for linkage phase L_2, the conditional probability for transmission of marker alleles M_{11} and M_{21} from the sire to offspring i are denoted by $p(M_{11}|S_i = g_i, M_s, L_2, r_1, \theta)$ and $p(M_{21}|S_i = g_i, M_s, L_2, r_1, \theta)$, respectively. The conditional probabilities of marker allele transmission are given below for marker linkage phase L_1. The conditional probability of M_{11} and M_{21} allele transmission from the sire to offspring i across linkage phases, is then computed as $p(M_{11}|S_i) = p(L_1) \cdot p(M_{11}|S_i = g_i, M_s, L_1, r_1, \theta) + p(L_2) \cdot p(M_{11}|S_i = g_i, M_s, L_2, r_1, \theta)$ and $p(M_{21}|S_i) = p(L_1) \cdot p(M_{21}|S_i = g_i, M_s, L_1, r_1, \theta) + p(L_2) \cdot p(M_{21}|S_i = g_i, M_s, L_2, r_1, \theta)$ where $p(L_1)$ and $p(L_2)$ are the probability for linkage phase 1 and 2, respectively.

Transmission probabilities for a sire with linkage phase, $M_{11}M_{21}/M_{12}M_{22}$

Conditional probability of transmission of paternal marker allele M_{11} and M_{21} to offspring with nine possible marker genotypes. Sire marker genotype is $M_{11}M_{21}/M_{12}M_{22}$. Dams have population frequency for four marker haplotype, viz. $M_{11}M_{21} = p_1p_2, M_{11}M_{22} = p_1(1-p_2), M_{12}M_{21} = (1-p_1)p_2$ and $M_{12}M_{22} = (1-p_1)(1-p_2)$ where p_1 and p_2 are population frequencies for marker alleles M_{11} and M_{21}, respectively.

Offspring genotype (g_i)	Frequency of offspring marker genotypes (f_g)	Conditional probability of paternal marker allele transmission	
		$p(M_{11}\|S_i = g_i, M_s, L_1, r_1, \theta)$[1]	$p(M_{21}\|S_i = g_i, M_s, L_1, r_1, \theta)$[2]
$M_{11}M_{21}$ $M_{11}M_{21}$	$(1-\theta)p_1p_2/2$	1	1
$M_{11}M_{21}$ $M_{11}M_{22}$	$p_1(1-p_2-\theta(1-2p_2))/2$	1	$(\theta-1)(p_2-1)/$ $[(1-p_2-\theta(1-2p_2))]$
$M_{11}M_{22}$ $M_{11}M_{22}$	$p_1(1-p_2)/2$	1	0
$M_{11}M_{21}$ $M_{12}M_{21}$	$p_2(1-p_1-\theta(1-2p_1))/2$	$(\theta-1)(p_1-1)/$ $[p_2(1-p_1-\theta(1-2p_1))]$	1
$M_{11}M_{21}$ $M_{12}M_{22}$	$[1-p_2-p_1(1-2p_2)-$ $\theta(1+2p_1+2p_2-4p_1p_2)]/2$	$(p_1-1)(1-p_2-\theta(1-2p_2))/$ $[p_2+p_1(1-2p_2)-1+$ $\theta(1-2p_1-2p_2+4_1p_2)]$	$(p_2-1)(1-p_1-\theta(1-2p_1))/$ $[p_2+p_1(1-2p_2)-1+$ $\theta(1-2p_1-2p_2+4p_1p_2)]$
$M_{11}M_{22}$ $M_{12}M_{22}$	$[(p_2-1)(\theta(2p_1-1)-p_1)]/2$	$\theta(p_1-1)/[p_1(2\theta-1)-\theta]$	0
$M_{12}M_{21}$ $M_{12}M_{21}$	$\theta p_2(1-p_1)/2$	0	1
$M_{12}M_{21}$ $M_{12}M_{22}$	$[(p_1-1)(\theta(2p_2-1)-p_2)]/2$	0	$\theta(p_2-1)/[p_2(2\theta-1)-\theta]$
$M_{12}M_{22}$ $M_{12}M_{22}$	$(1-\theta)(1-p_1)(1-p_2)/2$	0	0

[1] Given marker genotype of the sire (M_s) and its offspring (g_i) the probability for sire marker allele (k) transmission to the ith offspring of a given genotype (g) at marker locus M_1 or M_2 were obtained as $p(M_{11}|S_i = g_i, M_s, L_1, r_1, \theta) = \Sigma_g(f_g|M_{1k})/f_g$, where $\Sigma_g(f_g|M_{1k})$ is the sum of frequency of offspring genotype g to which sire marker allele M_{1k} is transmitted through its marker haplotype and f_g is the frequency of the genotype.

[2] $p(M_{21}|S_i = g_i, M_s, L_1, r_1, \theta)$ was obtained as $\Sigma_g(f_g|M_{2i})/f_g$, similar to $p(M_{11}|S_i = g_i, M_s, L_1, r_1, \theta)$.