Genet. Res., Camb. (2001), **77**, *pp.* 199–207. *With 3 figures.* Printed in the United Kingdom © 2001 Cambridge University Press

Using the mixed model for interval mapping of quantitative trait loci in outbred line crosses

Y. NAGAMINE[1]* AND C. S. HALEY[2]

[1] *Department of Animal Production, Tohoku National Agricultural Experiment Station, Morioka 020-0198, Japan*
[2] *Division of Genetics and Biometry, Roslin Institute, Roslin, Midlothian EH25 9PS, UK*

(*Received 12 June 2000 and in revised form 7 September and 25 October 2000*)

Summary

Interval mapping by simple regression is a powerful method for the detection of quantitative trait loci (QTLs) in line crosses such as F_2 populations. Due to the ease of computation of the regression approach, relatively complex models with multiple fixed effects, interactions between QTLs or between QTLs and fixed effects can easily be accommodated. However, polygenic effects, which are not targeted in QTL analysis, cannot be treated as random effects in a least squares analysis. In a cross between true inbred lines this is of no consequence, as the polygenic effect contributes just to the residual variance. In a cross between outbred lines, however, if a trait has high polygenic heritability, the additive polygenic effect has a large influence on variation in the population. Here we extend the fixed model for the regression interval mapping method to a mixed model using an animal model. This makes it possible to use not only the observations from progeny (e.g. F_2), but also those from the parents (F_1) to evaluate QTLs and polygenic effects. We show how the animal model using parental observations can be applied to an outbred cross and so increase the power and accuracy of QTL analysis. Three estimation methods, i.e. regression and an animal model either with or without parental observations, are applied to simulated data. The animal model using parental observations is shown to have advantages in estimating QTL position and additive genotypic value, especially when the polygenic heritability is large and the number of progeny per parent is small.

1. Introduction

Interval mapping was originally developed for the detection of quantitative trait loci (QTLs) in data from a cross between inbred lines. Such a cross is an ideal structure for the detection of QTLs, because all individuals in the F_1 are genetically identical and they show complete linkage disequilibrium for genes differing between lines. Haley & Knott (1992) and Haley *et al.* (1994) introduced a simple regression method for QTL detection by interval mapping in populations derived by crossing inbred or outbred lines, respectively. Haley *et al.* (1994) assumed that the crossed outbred lines were fixed or nearly fixed with respect to alternative alleles (QQ or qq) at a QTL and applied simple regression of the observed trait values onto the predicted genotypic values. One advantage of the regression method is its computational rapidity, which permits substantial flexibility in the models fitted. Fixed effects influencing traits in F_2 animals are easily taken into account in the model. Including fixed effects, such as sex, parity and farm, in a model can increase the power of detection of QTLs and improve the accuracy of the estimation of their effects.

In QTL analysis regions with moderate to large effect on quantitative traits are targeted. However, it is likely that quantitative traits are also influenced by genes scattered through the genome of relatively small effect: so-called polygenes (Falconer & Mackay, 1996, p. 102). These polygenes have individually too small an effect to be picked up by QTL analysis. In a cross between two inbred lines these effects may go undetected and act to inflate the residual variance. Alternatively, the joint effect of several or many such loci may be detected as a QTL (Visscher & Haley, 1996). In a cross between two outbred lines the

* Corresponding author. Present address: Division of Genetics and Biometry, Roslin Institute, Roslin, Midlothian EH25 9PS, UK. Tel: +44 (0)131 527 4358. Fax: +44 (0)131 440 0434. e-mail: Yoshi.Nagamine@bbsrc.ac.uk

situation is more complex. Where two very different outbred lines are crossed, QTLs of major effect may be assumed homozygous (QQ or qq) in the outbred lines and heterozygous (Qq) in the F_1 generation. Under this assumption, if the trait has a heritability (i.e. ratio of additive genetic variance to phenotypic variance) that is greater than zero within the outbred lines, this genetic variance is caused by polygenes. Genetic variance in the F_2 generation is caused both by the effects of these polygenes and by QTLs that contribute to the line difference. When a trait has a high heritability in both outbred (grandparental) lines, a significant part of the phenotypic variance in the F_2 will be caused by polygenes. If we ignore the polygenic effects, they will add to the error term in the statistical model, potentially reducing the power for the detection of QTLs.

In this study, we extend the fixed model for the regression analysis into a mixed model using Henderson's mixed model equations (Henderson, 1984, p. 335). In the context of an animal model it is possible to use the observations not only from progeny (F_2) but also from parents (F_1) to evaluate major gene and polygene effects. We show how the animal model using parental observations can be applied to the analysis of the F_2 population and increase the accuracy of QTL analysis. We also introduce a restricted maximum likelihood (REML) estimator to obtain the variance ratio for the mixed model equations.

2. Methods

We follow the notation adopted by Knott *et al.* (1998). The model applied assumes a QTL (Q) lying between two co-dominant flanking markers (A and B) and is developed for mapping the F_2 generation, although analysis of a backcross would follow very similar procedures. An F_2 is derived from a cross between two outbred (grandparental) lines, which have different alleles for the three loci. The genotypes of the two grandparental lines are $A_1A_1Q_1Q_1B_1B_1$ and $A_2A_2Q_2Q_2B_2B_2$. The parents (F_1) have only one genotype: $A_1Q_2Q_1Q_2B_1B_2$. Three genotypes at a QTL are possible in the F_2. The effects of Q_1Q_1, Q_1Q_2 and Q_2Q_2 are denoted $+a$, $+d$ and $-a$, respectively. The phenotypic value y_i (the observation) of a progeny (F_2) can be written as a regression model in terms of an additive and a dominance contribution at a QTL:

$$y_i = \mu + c_{ai}\,a + c_{di}\,d + e_i,$$

where μ is the mean, e_i is the error, including an additive polygenic effect and residual effect, and c_{ai} and c_{di} are coefficients for additive (a) and dominance (d) genotypic effect of individual animal i, respectively. Further details of these terms are given by Haley & Knott (1992).

(i) *Animal model*

To consider an additive polygenic effect, we use a mixed model:

$$y_i = \mu + c_{ai}\,a + c_{di}\,d + u_i + e_i,$$

where, $u_i \sim N(0, \sigma_u^2)$ is the additive polygenic effect (random effect) and $e_i \sim N(0, \sigma_e^2)$ is the residual error. Since the additive polygenic effect of each individual animal is being considered, this is an animal model. We assume there are n animals, then

$$\mathbf{y} = \mathbf{L}\mu + \mathbf{w}_1\,a + \mathbf{w}_2\,d + \mathbf{Z}\mathbf{u} + \mathbf{e}.$$

Here, $\mathbf{y}$ is $n \times 1$ vector of the observation y_i. $\mathbf{w}_1$ and $\mathbf{w}_2$ are $n \times 1$ vectors of c_{ai} and c_{di}, respectively. $\mathbf{Z}$ is the design matrix ($n \times n$) for the additive polygenic effect and $\mathbf{u}$ is $n \times 1$ vector of the additive polygenic effects of n animals. $\mathbf{e}$ is $n \times 1$ vector for the residual effect. $\mathbf{L}$ is a vector for μ. If all animals have observations, then $\mathbf{L}$ is $n \times 1$ vector with all elements 1.

In the mixed model equations,

$$\begin{pmatrix} \mathbf{L}'\mathbf{L} & \mathbf{L}'\mathbf{w}_1 & \mathbf{L}'\mathbf{w}_2 & \mathbf{L}' \\ \mathbf{w}_1'\mathbf{L} & \mathbf{w}_1'\mathbf{w}_1 & \mathbf{w}_1'\mathbf{w}_2 & \mathbf{w}_1'\mathbf{Z} \\ \mathbf{w}_2'\mathbf{L} & \mathbf{w}_2'\mathbf{w}_1 & \mathbf{w}_2'\mathbf{w}_2 & \mathbf{w}_2'\mathbf{Z} \\ \mathbf{L} & \mathbf{Z}'\mathbf{w}_1 & \mathbf{Z}'\mathbf{w}_2 & \mathbf{Z}'\mathbf{Z} + \mathbf{k}\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \mu \\ a \\ d \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{L}'\mathbf{y} \\ \mathbf{w}_1'\mathbf{y} \\ \mathbf{w}_2'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix}.$$

Here, $\mathbf{A}^{-1}$ is the inverse of relationship matrix among animals. $k\,(=\sigma_e^2/\sigma_u^2)$ is the ratio of the variance of the additive polygenic effects (σ_u^2) and the error variance (σ_e^2). In this model, we use observations from all animals through the relationship matrix. Observations from F_1 parents as well as from F_2 progeny can be used to estimate the effect and position of QTLs.

(ii) *Example data set*

We assume that a QTL (Q) is lying between two co-dominant flanking markers (A and B). We have an F_2 generation from a cross between two outbred lines (1 and 2) which carry different alleles for the three loci. The example data set, comprising two sires (animals 1 and 5) and two dams (animals 2 and 6) from the F_1 generation, is shown in Table 1. Animals 3 and 4 (in the F_2 generation) are the progeny of animals 1 and 2, and animals 7 and 8 (in the F_2 generation) are progeny of animals 5 and 6. Since sires and dams are from the F_1 generation, they have the same genotype, i.e. $A_1A_2Q_1Q_2B_1B_2$. Marker genotypes of the progeny are also shown in Table 1. The recombination fraction between the two flanking markers A and B is designated as r. The recombination fraction between

https://doi.org/10.1017/S0016672301004931 Published online by Cambridge University Press

Table 1. *Example data set for an animal model*

Animal	Observation	Sire	Dam	Marker genotype	Expectation in terms of a
1 (F_1)	20	–	–	$A_1A_2B_1B_2$	0
2 (F_1)	0	–	–	$A_1A_2B_1B_2$	0
3 (F_2)	10	1	2	$A_1A_1B_1B_1$	0·9803
4 (F_2)	2·5	1	2	$A_1A_2B_2B_2$	−0·4902
5 (F_1)	0	–	–	$A_1A_2B_1B_2$	0
6 (F_1)	−10	–	–	$A_1A_2B_1B_2$	0
7 (F_2)	0	5	6	$A_1A_1B_1B_2$	0·4902
8 (F_2)	−7·5	5	6	$A_2B_2A_2B_2$	−0·9803

A and Q is r_A and that between Q and B is r_B. Haley & Knott (1992) gave the coefficient values, c_{ai} and c_{di}, for all possible genotypes. For example, the coefficient c_{ai} of marker genotype $A_1A_1B_1B_1$ is $[(1-r_A)^2(1-r_B)^2 - r_A^2 r_B^2]/(1-r)^2$. When the putative QTL is midway between two markers 20 cM apart (i.e. $r = 0{\cdot}1648$, $r_A = r_B = 0{\cdot}0906$) then

$$[(1-r_A)^2(1-r_B)^2 - r_A^2 r_B^2]/(1-r)^2 = 0{\cdot}9803.$$

We use the same assumptions as Haley & Knott (1992) used, where μ and d are set to zero for simplicity in demonstrating the analysis. The F-statistic is calculated for the simulated position of the putative QTL, 10 cM from both the A and B loci. In practice, we move the putative QTL through the interval to find the position that gives the highest F value. We use an animal model:

$$\mathbf{y} = \mathbf{w}a + \mathbf{Z}\mathbf{u} + \mathbf{e}.$$

Here, $\mathbf{y}$ is 8×1 vector of observations from 8 animals. Term a is a scalar for additive genotypic effect of a QTL, $\mathbf{w}$ is the 8×1 vector of c_{ai} and $\mathbf{u}$ is the 8×1 vector of additive polygenic effects of animals. $\mathbf{Z}$ is the design matrix (8×8) for the additive polygenic effects of animals. In this example, $\mathbf{Z}$ is a diagonal matrix with all elements 1.

In the mixed model equations:

$$\begin{pmatrix} \mathbf{w}'\mathbf{w} & \mathbf{w}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{w} & \mathbf{Z}'\mathbf{Z} + k\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} a \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{w}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix}.$$

The vector of u and a can be estimated as

$$\begin{pmatrix} a \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{w}'\mathbf{w} & \mathbf{w}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{w} & \mathbf{Z}'\mathbf{Z} + k\mathbf{A}^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{w}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix} = \begin{pmatrix} C_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{w}'\mathbf{y} \\ \mathbf{Z}'\mathrm{y} \end{pmatrix}.$$

$\mathbf{A}^{-1}$ is the inverse of the relationship matrix among 8 animals. k ($= \sigma_e^2/\sigma_u^2$) is the ratio of the additive polygenic variance (σ_u^2) and the error variance (σ_e^2). Assume the variance ratio k is 1, i.e. the heritability is 0·5 ($= \sigma_e^2/(\sigma_u^2 + \sigma_e^2)$). We use observations from all individuals through the relationship matrix.

These values are given as:

$$\mathbf{Z}'\mathbf{Z} = \mathrm{diag}(1,1,1,1,1,1,1,1)$$
$$\mathbf{w}'\mathbf{Z} = (0 \quad 0 \quad 0{\cdot}9803 \quad -0{\cdot}4902 \quad 0 \quad 0 \quad 0{\cdot}4902 \quad -0{\cdot}9803)$$
$$\mathbf{w}'\mathbf{w} = 0{\cdot}9803^2 + (-0{\cdot}4902)^2 + 0{\cdot}4902^2 + (-0{\cdot}9803)^2 = 2{\cdot}403$$
$$\mathbf{Z}'\mathbf{y} = (20 \quad 0 \quad 10 \quad 2{\cdot}5 \quad 0 \quad -10 \quad 0 \quad -7{\cdot}5)'$$
$$\mathbf{w}'\mathbf{y} = 0{\cdot}9803 \times 10 + (-0{\cdot}4902) \times 2{\cdot}5 + (-0{\cdot}9803) \times (-7{\cdot}5) = 15{\cdot}930.$$

The following values would be estimated:

$$\mathbf{u}' = (10{\cdot}000 \quad 0{\cdot}000 \quad 5{\cdot}000 \quad 5{\cdot}000 \quad 0{\cdot}000 \quad -5{\cdot}000 \quad -2{\cdot}500 \quad -2{\cdot}500)$$
$$a = 5{\cdot}100$$
$$\sigma_e^2 = (\mathbf{y}'\mathbf{y} - \mathbf{u}'\mathbf{y} - a\,\mathbf{w}'\mathbf{y})/(n-1) = (662{\cdot}500 - 331{\cdot}253 - 5{\cdot}100 \times 15{\cdot}930)/(8-1) = 35{\cdot}714$$
$$F = a \times a/(C_{11} \times \sigma_e^2) = 5{\cdot}100 \times 5{\cdot}100/(0{\cdot}6403 \times 35{\cdot}714) = 1{\cdot}138.$$

Element C_{11}, 0·6403, from the inverted matrix corresponds to $\mathbf{w}'\mathbf{w}$, 2·403, in the mixed model equations.

(iii) *Simulations*

The genome of each individual consisted of a pair of chromosomes 100 cM in length, carrying marker loci at the ends and at 20 cM intervals (i.e. six markers in total). All markers were fully informative. A single QTL in the centre of the chromosome was used. The phenotypic variance was set at 100 and the additive genotypic effect, a, of the QTL was 5. The dominance genotypic effect, d, was set at zero. Three additive polygenic variances (σ_u^2) of 20, 40 and 60 (equivalent to three heritabilities ($= \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$) of 0·2, 0·4 and 0·6) were used. Two types of population structures were simulated. In both structures the individuals to be grandparents were chosen randomly from outbred lines and it was assumed that there was no relationship among them. Dams were nested within sires. Structure 1 had 10 sires and 20 dams in the F_1 generation. Two dams were mated per sire and every dam had 10 progeny in the F_2 generation. Structure 2 had 10 sires and 200 dams in the F_1 generation. Twenty dams

https://doi.org/10.1017/S0016672301004931 Published online by Cambridge University Press

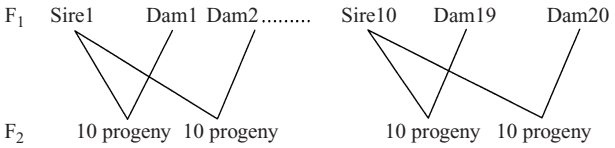

Fig. 1. Simulated population structure 1. Ten sires and 20 dams are in the F_1 generation and dams are nested within sire. Two dams are mated per sire and every dam has 10 progeny in the F_2 generation.

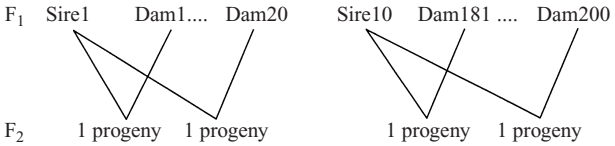

Fig. 2. Simulated population structure 2. Ten sires and 200 dams are in the F_1 generation and dams are nested within sire. Twenty dams are mated per sire and every dam has only one progeny in the F_2 generation.

were mated per sire and every dam had only one progeny. In total, there were 200 progeny in the F_2 generation in both population structures (Figs 1, 2).

(iv) *Statistical methods*

(1) Regression (RG): Simple regression using c_{ai} and a was applied and only observations (phenotypic values) of progeny (F_2) were used.

$$y_i = \mu + c_{ai}\, a + e_i,$$

where, $e_i \sim N(0, \sigma_e^2)$ is the error comprising the additive polygenic effect and the residual effect.

(2) Mixed model with no parental observations (MXN): Mixed model equations were applied. Regression of a and additive polygenic effects (u) were considered. Observations on progeny (F_2) were used but no observations on parents (F_1) were used.

$$y_i = \mu + c_{ai}\, a + u_i + e_i,$$

where $u_i \sim N(0, \sigma_u^2)$ is the additive polygenic effect (random effect) and $e_i \sim N(0, \sigma_e^2)$ is the error term comprising the residual effect. Since no parental observations were used, the solutions from MXN were the same as from a sire and dam model.

(3) Mixed model with parental observations (MXP): Mixed model equations were applied and the same statistical model as (2) was used. Observations on both parents (F_1) and progeny (F_2) were used.

In practice, we used the reduced-animal model to reduce computational cost. Mixed model equations for the animal model demand equations for all individuals, that is, parents and progeny in our case. Quaas & Pollak (1980) developed a method which allows the equations to be set up only for parents and the predicted values for the progeny to be obtained by back-solving from the predicted values of parents. When we used the reduced-animal model, the size of the mixed model equations that needed to be inverted was reduced from 231 (30 parents, 200 progeny and 1 covariate) to 31 (30 parents and 1 covariate) in structure 1 and from 411 (210 parents, 200 progeny and 1 covariate) to 211 (210 parents and 1 covariate) in structure 2.

(v) *Null hypothesis*

Data were generated with no QTL but with markers at 20 cM spacing. Under structure 1, 1000 replicates, and under structure 2, 500 replicates, with all combinations of heritabilities, were generated and analysed.

(vi) *Estimation of variance ratio k by REML*

The mixed model equations require the variance ratio $k(=\sigma_e^2/\sigma_u^2)$ to be known. When the heritability (h^2) is known, k is easily calculated:

$$h^2 = \sigma_u^2/(\sigma_u^2+\sigma_e^2), \quad k = (1-h^2)/h^2 = \sigma_e^2/\sigma_u^2.$$

If the grandparental lines are ordinary domestic breeds, it is not difficult to find a reliable estimate of the heritability of most traits from the literature. However, if one uses an uncommon breed or animal as grandparent, for example wild boar in a study of pigs, or a trait that is not commonly recorded, it might be difficult to find a reliable estimate of the heritability. In this case we must find the ratio, k, from our own data set. Thompson (1977) developed the iterative minimum variance quadratic unbiased estimator (MIVQUE) as the restricted maximum likelihood (REML). Sorensen & Kennedy (1986) derived REML, iterative MIVQUE, under the reduced-animal model. We applied their method to estimate the variances. Iteration procedures for REML using the reduced-animal model are as follows. The length of chromosome is assumed to be 100 cM.

Step 1. Estimation of the following values every 1 cM from 0 cM to 100 cM using an initial ratio, k

- Additive polygenic effect (u) of parents and progeny (u of progeny is obtained by back-solution from parents' u)
- Additive genotypic effect of QT gene ($=a$)
- F-statistic

Step 2. Decision on QTL position

- The QTL position is estimated as the position that gives the highest F-statistic
- Estimated values (a and u) at this QTL position are used for the next step

Step 3. Estimation of variances by MIVQUE

- Estimate additive polygenic variance (σ_u^2) and error variance (σ_e^2) by MIVQUE for the reduced-animal model
- Calculate new variance ratio $k(=\sigma_e^2/\sigma_u^2)$
- Go back to step 1 with a new k until the estimators have converged

https://doi.org/10.1017/S0016672301004931 Published online by Cambridge University Press

When variances, σ_u^2 and σ_e^2, do not change, variances and the ratio k are assumed to have converged and these are the REML estimators.

(vii) *Wrong variance ratio k and QTL position*

When the wrong variance ratio k is given, this could lead to a biased position of the QTL being estimated. To investigate this problem, we gave incorrect variance ratios as the initial values and iterated until convergence. The true k was 1·5 (i.e. heritability 0·4) and the wrong initial ratios k, 4·0 (heritability 0·2) and 0·666 (heritability 0·6), were given. Twenty replicates were generated. When the difference between ratio k_{n-1} from the $(n-1)$th iteration and k_n from nth iteration was less than 0·001 % of k_n, the procedure was judged to have converged.

3. Results

(i) *Comparison of the three methods*

This comparison was carried out under the assumption that the heritability $(=\sigma_u^2/(\sigma_u^2+\sigma_e^2))$ was known. When the heritabilities were 0·2, 0·4 and 0·6, the variance ratios k were 4·0, 1·5 and 0·666, respectively. The simulated value of the QTL (cMA) position was 50 and estimated values from RG, MXN and MXP based upon 200 replicates were close to the simulated value in both population structures (Tables 2, 3). When the heritability was low, the difference in the empirical standard deviation of cMA across replicates from the three methods remained small in both population structures. With higher polygenic heritabilities, the empirical standard deviation of cMA was smaller for MXN and MXP, because the error variance was reduced and this increased the accuracy of prediction. This was especially the case in structure 2, with only one progeny per dam, where the standard deviations of cMA from MXP were markedly reduced with increasing heritability.

The simulated value of the additive genotypic value, a, of the QTL was 5 and estimated values from RG, MXN and MXP were all close to 5. The empirical standard deviation of a from MXN and MXP reduced with increasing heritability. The standard deviation of a and cMA from RG did not reduce with increasing

Table 2. *Comparison of estimates from the three methods in simulated population structure 1*

h^2		cMA	SD	a	SD	F-statistic	SD
0·2	RG	48·48	10·61	5·05	1·28	15·10	7·07
	MXN	49·28	10·09	5·09	1·26	16·03	7·40
	MXP	49·13	9·89	5·08	1·26	16·19	7·49
0·4	RG	48·08	11·33	4·98	1·39	14·99	7·25
	MXN	49·11	9·68	5·05	1·23	17·26	7·86
	MXP	49·39	8·91	5·05	1·22	17·51	7·97
0·6	RG	48·80	9·39	4·92	1·51	14·95	7·47
	MXN	49·78	8·41	5·03	1·17	18·93	8·41
	MXP	49·72	8·16	5·01	1·17	19·29	8·54

Ten sires, 20 dams and 200 progeny are used. The simulated values of cMA and a are 50 cM and 5, respectively. RG is the regression method. MXN and MXP are the mixed models without and with parental observations, respectively.

Table 3. *Comparison of estimates from the three methods in simulated population structure 2*

h^2		cMA	SD	a	SD	F-statistic	SD
0·2	RG	49·40	11·41	5·09	1·60	15·94	9·33
	MXN	49·74	10·87	5·05	1·60	16·18	9·55
	MXP	49·58	10·76	5·07	1·62	16·72	9·69
0·4	RG	48·44	13·39	5·02	1·57	15·58	8·96
	MXN	49·88	10·71	4·95	1·57	16·25	9·48
	MXP	49·08	10·47	4·98	1·58	17·65	10·03
0·6	RG	48·02	13·00	4·96	1·55	15·17	8·46
	MXN	48·70	11·65	4·86	1·51	16·35	9·32
	MXP	49·72	7·68	4·94	1·53	19·23	10·44

Ten sires, 200 dams and 200 progeny are used. The simulated values of cMA and a are 50 cM and 5, respectively. RG is the regression method. MXN and MXP are the mixed models without and with parental observations, respectively.

https://doi.org/10.1017/S0016672301004931 Published online by Cambridge University Press

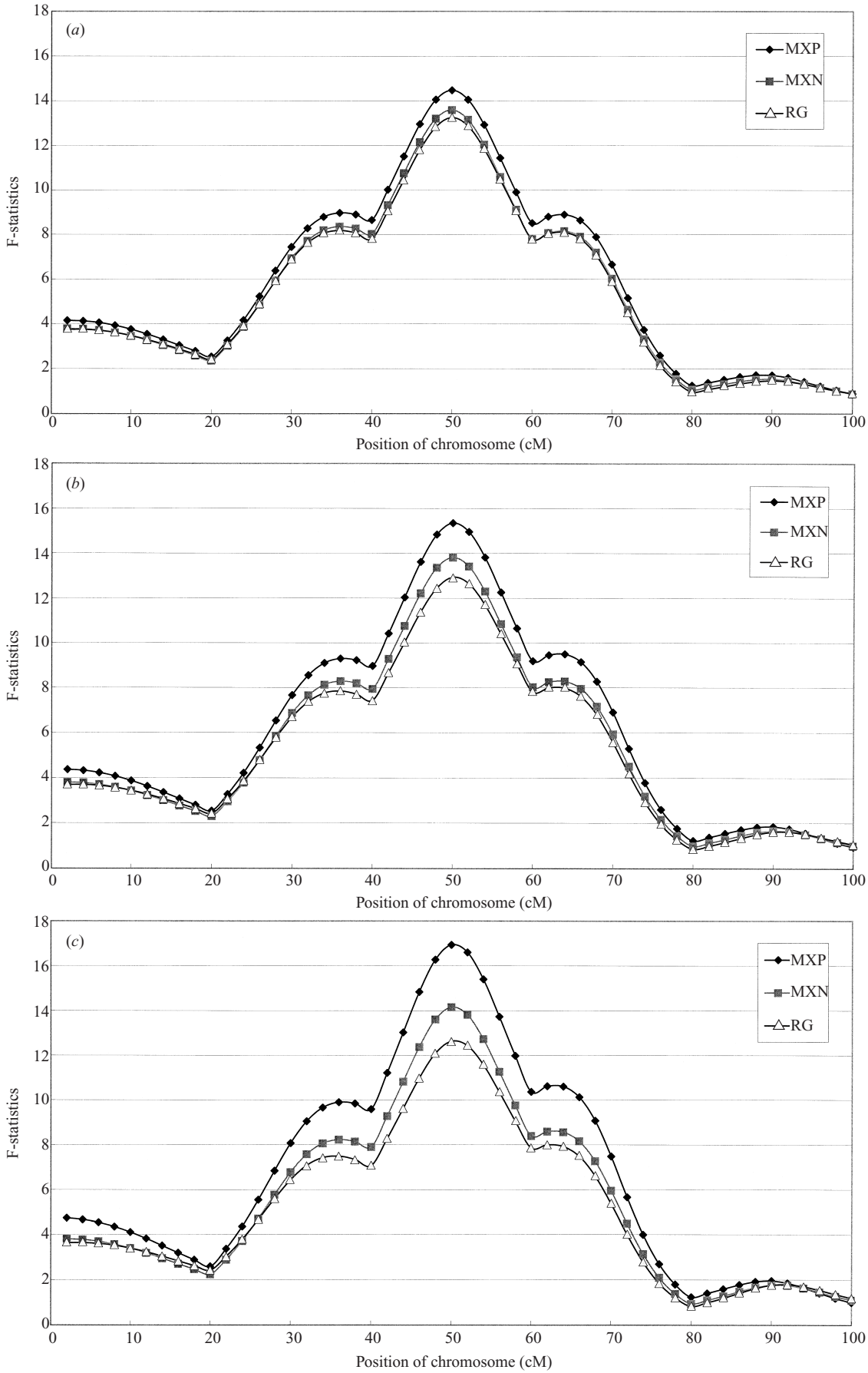

Fig. 3. *F*-statistics curves from the three methods with heritabilities (*a*) 0.2, (*b*) 0.4 and (*c*) 0.6.

https://doi.org/10.1017/S0016672301004931 Published online by Cambridge University Press

Table 4. *Distribution of* F-*statistics under the null hypothesis*

Population structure	h^2	Method	Highest F	Empirical 5% threshold
Structure 1	0·2	RG	3·28 (2·23)	8·00
		MXN	3·29 (2·22)	7·78
		MXP	3·27 (2·21)	7·96
	0·4	RG	3·29 (2·24)	7·83
		MXN	3·29 (2·24)	7·73
		MXP	3·26 (2·23)	7·77
	0·6	RG	3·29 (2·26)	7·80
		MXN	3·27 (2·28)	7·88
		MXP	3·24 (2·27)	7·79
Structure 2	0·2	RG	3·21 (2·21)	6·91
		MXN	3·22 (2·18)	7·15
		MXP	3·19 (2·15)	7·02
	0·4	RG	3·25 (2·24)	7·05
		MXN	3·25 (2·23)	7·17
		MXP	3·22 (2·19)	7·18
	0·6	RG	3·29 (2·31)	7·45
		MXN	3·29 (2·29)	7·10
		MXP	3·25 (2·27)	7·58

The highest F-statistics represents the mean over 1000 replicates for structure 1 and 500 replicates for structure 2. Standard deviation is given in parentheses. The empirical 5% threshold is calculated as the mean of the 50th and 51st highest F-statistics for structure 1 and the mean of the 25th and 26th highest F-statistics for structure 2.

heritability. The polygenic effect was not considered by the RG method. Having a higher heritability increased the difference between families and this affected the estimate of RG.

F-statistics from MXP are higher than those from MXN or RG for both population structures. This is particularly so when the heritability is high; thus for a heritability 0·6, the F-statistic from MXP is larger than those from the other methods. Average F-statistics over 20 replicates from structure 2 have been plotted every 2 cM along the chromosome (Fig. 3). It is obvious that F-statistics from MXP are increasing with higher heritability. The peak F-statistics from MXP are much higher than those from MXN or RG. Since MXP takes into account the polygenic effect by means of parent and progeny observations, the higher heritability reduces the error variance and increases the power to detect QTLs. On the contrary, increased heritability reduces F-statistics from RG slightly in both population structures.

(ii) *Null hypothesis*

Results of analysis of data generated with no QTL under structure 1 and structure 2 are shown in Table 4. Means of highest F-statistics from RG, MXN and MXP at each heritability from the two population structures did not differ. The approximate empirical 5% thresholds were calculated as the mean of 50th and 51st highest F-statistic over replicates under structure 1 and the mean of the 25th and 26th highest F-statistics over replicates under structure 2. These values did not differ among RG, MXN and MXP at each heritability from structure 1 and structure 2.

(iii) *Variance ratio k by REML and QTL position*

When an initial value of variance ratio k is given, we can solve the mixed model equations and estimate the position of QTLs, values of a and the additive polygenic effect (u). These solutions are used to estimate the new ratio k. These procedures are used iteratively until k is converged. Table 5 is an example of these procedures using structure 1. The simulated additive polygenic variance and error variance were 40·0 and 60·0, respectively, and the variance ratio k and heritability were 1·5 ($= 60{\cdot}0/40{\cdot}0$) and 0·4 ($= 40{\cdot}0/(40{\cdot}0+60{\cdot}0)$), respectively. At the first iteration, the wrong variance ratio 9 ($h^2 = 0{\cdot}1$) was given as the initial k. After estimating cMA, a and u, a new

Table 5. *Example of iterative procedures by REML*

Iteration number	k	σ_u^2	σ_e^2	cMA	a	F-statistic
1	9·0000	39·5175	69·7149	48	5·725	15·733
2	1·7642	39·9905	67·0417	50	5·384	17·032
3	1·6764	40·0083	66·8676	50	5·372	17·172
4	1·6713	40·0094	66·8568	50	5·371	17·181
5	1·6710	40·0094	66·8563	50	5·371	17·182
6	1·6710	40·0094	66·8562	50	5·371	17·182
…	…	…	…	…	…	…
10	1·6710	40·0094	66·8562	50	5·371	17·182

Initial ratio $k = 9{\cdot}0$ ($= \sigma_e^2/\sigma_u^2$), i.e. heritability ($= \sigma_u^2/\sigma_e^2+\sigma_u^2$) 0·1, in structure 1 is used. Variances and $k = 1{\cdot}6710$ ($= 66{\cdot}8562/40{\cdot}0094$) are converged at the sixth iteration.

https://doi.org/10.1017/S0016672301004931 Published online by Cambridge University Press

variance ratio k 1·7642 (= 69·7149/39·5175) was obtained. All values – cMA, a, σ_u^2 and σ_e^2 – remained unchanged at iteration 6, indicating that the estimates had converged.

We generated 20 replicates with the wrong initial variance ratios k at 1st iteration and iterated until the estimates converged. The simulated variance ratio k was 1·5 ($h^2 = 0{\cdot}4$) and the given wrong initial ratios were 4·0 ($h^2 = 0{\cdot}2$) and 0·666 ($h^2 = 0{\cdot}6$). The maximum number of iterations required for convergence was 7 in both cases. They converged at the ratio k of 1·72 ($h^2 = 0{\cdot}37$) as the average. The converged value of cMA was 49·05. The values of cMA were 49·15 and 48·95 with initial ratio 4·0 and 0·666, respectively. In both cases, 1 cMA was the maximum difference between initial and converged values. The simulated value of a was 5·0 and the estimates converged at 4·98. The mean differences of a from the converged values are 0·16 with the initial ratio $k = 4{\cdot}0$ and 0·07 with the initial ratio $k = 0{\cdot}666$.

4. Discussion

The results presented here show that the inclusion of the additive polygenic effect in a mixed model can improve the estimate of QTL effects from data derived by crossing outbred lines. Since MXP (the mixed model with parental observations) and MXN (the mixed model without parental observations) treat the polygenic effect as a random effect in the model, the accuracy of QTL position estimates increases with higher heritabilities. Each dam was 10 progeny in structure 1 and each dam has only one progeny in structure 2. Thus structure 1 is appropriate for swine populations and structure 2 for cattle or sheep populations. When the heritability is high, using observations on parents significantly increases the accuracy of QTL position estimation (Table 3) and increases the peak F-statistics (Fig. 3). The solution from MXN for structure 1 would be the same as from a sire and dam model because the relationships at the F_1 level among F_2 animals were used, but no parental (i.e. F_1) observations were taken into consideration. When the number of progeny is large, the solutions from the sire and dam model are as accurate as the solutions from the animal model, because accuracy in a sire and dam model depends mainly on the number of progeny per parent (Wilmink & Dommerholt 1985; Meyer 1989). On the other hand, when the number of progeny per parent is small and the heritability is high, using phenotypic observations on parents has a relatively large influence on accuracy. It can be seen in structure 2, which is based on one progeny per dam, that when the heritability is high, MXP gives much better results than MXN (Table 3).

We should note that the mixed model approach assumes that the polygenic and environmental variances are the same in both F_1 and F_2 populations. Since we consider only the additive polygenic effect and assume that QTLs of major effect are fixed within lines, the polygenic variance in both F_1 and F_2 may reasonably be assumed to be the same. The assumption of the constant environmental variance should also be reasonable in many cases, especially where these populations are not random samples from the field, but are controlled populations from a designed experiment. We have only used the observations and pedigree information from two generations, F_1 and F_2, in the mixed model equations. Phenotypic data from grandparental lines were not used because if the two grandparental lines had different heritabilities, this would introduce the problem of variance heterogeneity.

An alternative approach to the estimation of a random effect in a mixed model framework would be to retain the least squares approach but to include a fixed effect representing family, e.g. full or half-sib family (Knott *et al.*, 1998). This has the effect of removing between-family variation that is partly genetic in origin and hence should have some of the beneficial consequences of the MXN approach. However, it may be costly in terms of lost degrees of freedom when family units are small. Some further work to compare the merits of the MXN approach with a least squares approach with families as fixed effects would be merited.

The highest F-statistics and empirical 5% threshold values under the null hypothesis are the same for the three methods (Table 4). Taking into account these results under the null hypothesis, the higher peak F-statistics from MXP in the presence of a QTL (see Fig. 3) lead to higher power for detection of QTLs. When the variance ratio, k, started with a wrong value – e.g. 9 ($h^2 = 0{\cdot}1$), which was far from the real ratio 1·5 ($h^2 = 0{\cdot}4$) – the estimated QTL position moved 2 cM from 48 to 50 cM (Table 5). The value of a changed from 5·725 to 5·371 from the first to the final iteration. When the wrong initial ratios of k were 4·0 ($h^2 = 0{\cdot}2$) and 0·666 ($h^2 = 0{\cdot}6$), the maximum difference from the converged value in cMA was 1 cM. The maximum difference in value of a is 0·48 from the initial ratio 4·0 and 0·18 from the initial ratio 0·666. These results indicate that the animal model approach has sufficient robustness to estimate QTL position and to estimate additive genotypic values without requiring an exact value for the variance ratio, k. Even information on whether the heritability is low, medium or high in the grandparental lines might be sufficient in practice.

The disadvantages of applying a mixed model are: (1) mixed model equations demand prior information of variance ratio k, (2) the F-statistic is an approximate test, not an exact test, of the null hypothesis regarding a, and (3) when the number of parents is large, computational costs are high. To solve the first

https://doi.org/10.1017/S0016672301004931 Published online by Cambridge University Press

problem, we showed that the prior ratio k, even if it was not the exact value, did not cause a serious bias in estimation of genotypic values and QTL position estimates. We also showed that the variance components could be estimated by REML using iterative MIVQUE.

The second problem relates to the testing of fixed effects in a mixed model. If one gives the exact variance ratio in mixed model equations, a test using F-statistics can be an exact test of estimators in a mixed model (Henderson, 1984, p. 89). Otherwise, it will be an approximate test. In practice this will not be a major problem in any event if the significance threshold is set by a Monte Carlo approach such as simulation or permutation. Setting the threshold in this way is required because of the difficulty in setting a threshold for a chromosomal or whole genome scan. In this case, the problem becomes one of the computation time required to perform sufficient replicates of simulation or permutation to provide a reasonable estimate of threshold.

The third problem of high computational costs is partly due to the size of matrix to be inverted. We can use a reduced-animal model, which does not demand equations for progeny, but equations for parents are required. Structure 2, which has 10 sires, 200 dams and 200 progeny, can be applied to animals such as cattle or sheep. The size of the matrix is reduced to 211, comprising 210 parent and 1 covariate, in the reduced-animal model. This size of the matrix does not present a large problem for the single QTL model but could cause problems with respect to computational time for replicated simulations in a multiple QTL model, where the analysis of multiple QTLs demands a simultaneous search of two or more dimensions to find the best QTL positions. In the past, many methods have been investigated in the study of animal models to get around the problem of large field data sets (Quaas & Pollak, 1980; Schaeffer & Kennedy, 1986). Methods to reduce computation cost should be investigated in QTL analysis for large data sets with more complicated analytical models in future.

Only fully informative markers were used in our simulated data. Haley *et al*. (1994) extended the regression method to use all markers including those that were not completely informative. The same approach could be used to extend our method.

We used the additive polygenic effect as our only random effect. Since our approach shows generally how to extend the model with random effects, other random genetic and environmental effects, such as a maternal genetic effect and a permanent environmental effect, could also be included in animal models. Where appropriate, use of these models may also be valuable to increase power and accuracy in QTL analysis.

Y. N. acknowledges support from the Science and Technology Agency (Japan). C. S. H. is grateful for support from the Biotechnology and Biological Science and Research Council (UK).

References

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th edn. New York: Longman Scientific and Technical.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

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

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. Guelph: University of Guelph Press.

Knott, S. A., Marklund, L., Haley, C. S., Andersson, K., Davies, W., Ellegren, H., Fredholm, M., Hansson, I., Hoyheim, B., Lundstrom, K., Moller, M. & Andersson, L. (1998). Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics* **149**, 1069–1080.

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

Meyer, K. (1989). Approximate accuracy of genetic evaluation under an Animal Model. *Livestock Production Science* **21**, 87–100.

Quaas, R. L. & Pollak, E. J. (1980). Mixed model methodology for farm and ranch beef cattle testing programs. *Journal of Animal Science* **51**, 1277–1287.

Schaeffer, L. R. & Kennedy, B. W. (1986). Computing strategies for solving mixed model equations. *Journal of Dairy Science* **69**, 575–579.

Sorensen, D. A. & Kennedy, B. W. (1986). Analysis of selection experiments using mixed model methodology. *Journal of Animal Science* **63**, 245–258.

Thompson, R. (1977). The estimation of heritability with unbalanced data. II. Data available on more than two generations. *Biometrics* **33**, 497–504.

Visscher, P. M. & Haley, C. S. (1996). Detection of quantitative trait loci in line crosses under infinitesimal genetic models. *Theoretical and Applied Genetics* **93**, 691–702.

Wilmink, J. B. M. & Dommerholt, J. (1985). Approximate reliability of best linear unbiased prediction in models with and without relationship. *Journal of Dairy Science* **68**, 946–952.