MULTIPLE MARKER MAPPING OF QUANTITATIVE TRAIT LOCI IN HALF-SIB POPULATIONS

S.A. Knott, J-M. Elsen and C.S. Haley
ICAPB, University of Edinburgh, West Mains Rd., Edinburgh, EH9 3JT, U.K.
INRA Station d'Amélioration Génétique des Animaux, Castanet-Tolosan, France
Roslin Institute (Edinburgh), Roslin, Midlothian, EH25 9PS, U.K.

SUMMARY

A least squares method using multiple markers for the detection of quantitative trait loci is presented for the analysis of data from half-sib populations. A simulation study illustrates the benefits in power of this method over one using markers one at a time. The multiple marker method provides reasonable estimates of location for the QTL with smaller standard deviation than those obtained using the markers one at a time. The method provides a rapid and relatively powerful way of locating genomic regions containing QTLs for more detailed analysis.

INTRODUCTION

Maps of the major livestock species based on molecular genetic markers are being developed rapidly (e.g. Archibald *et al.*, 1994). These maps provide the tools to begin the dissection of the genetic variation underlying quantitative traits of importance to animal breeders and the location of some of the genes responsible for this variation - so called quantitative trait loci or QTLs (Geldermann, 1975).

A number of authors have considered designs for the analysis of data from outbred populations (e.g. Neimann-Sørensen and Robertson, 1961; Soller and Genizi, 1978; Geldermann *et al.*, 1985; Weller *et al.*, 1990). The drawback of these methods is that they use information from a single marker at a time. No marker will have a heterozygosity of unity, so for any given marker some sires will be homozygous and thus uninformative. This wastes information and has a potentially greater problem of introducing bias into the estimated location of QTLs. Furthermore, the least squares methods proposed cannot separately estimate the position and effect of any detected QTLs. Maximum likelihood (ML) methods (Weller, 1986; Knott and Haley, 1992a) can potentially estimate both effects, but estimates are generally poor using only a single marker (Weller, 1986; Knott and Haley, 1992a and b) and location is relative to the marker (i.e. could be either side of it).

Interval mapping (Lander and Botstein, 1989; Haley and Knott, 1994), where a region of the chromosome between a pair of adjacent markers is explored, has been found to be more powerful than the use of single markers for the analysis of populations derived from a cross between inbred lines and to provide more accurate estimates of the position and effect of a QTL. The application of interval mapping approaches to data from outbred populations is not straightforward and can be computationally demanding. Furthermore, because markers are not completely heterozygous, the information content varies from interval to interval depending on the markers flanking that interval. This presents the potential problem that there may be a bias towards locating a QTL in the most informative interval, rather than the correct one (Knott and Haley, 1992a; Haley *et al.*, 1994). This problem can be overcome by the simultaneous use of all markers in a linkage group. In this paper we demonstrate the use of all markers in a linkage group for the analysis of data from half-sib populations.

METHODS

The methods used assume that the trait data has been collected from the half-sib progeny of a number of unrelated sires ('Daughter' or 'Granddaughter' designs discussed by Weller *et al.*, 1990). Sires are assumed to be randomly mated to unrelated dams and each dam has only a single progeny. Marker data is available on the sires and their offspring and may or may not be available for the dams. We assume that the order of markers in a linkage group and the distance between them is known. The basic philosophy is similar to that in interval mapping (Lander and Botstein, 1989) which we have previously applied to the analysis of data from crosses between inbred and outbred lines (Haley and Knott, 1992; Haley *et al.*, 1994). For given positions (e.g. 1 cM intervals) through a linkage group the

genotype at a putative QTL is calculated conditional on the genotypes of flanking markers. In the half-sib design, where little or no useful information is available from the dam, it is only of value to calculate these probabilities for the sire gamete. The analysis proceeds by firstly considering the marker alleles inherited from the sire by each progeny, secondly reconstructing the gametes of the sire and thirdly using this information in a regression analysis.

Each marker in each sire-family is considered in turn, markers for which a sire is homozygous are uninformative and omitted from consideration. For markers which are heterozygous in the sire it may be possible to determine which allele a progeny has inherited (if they possess only one of the two possible sire alleles). Dam genotype information will increase the probability that the sire allele inherited by a progeny can be determined unequivocally.

Once marker inheritance has been determined, the most likely construction of the two sire gametes is calculated. This is done by considering in turn each pair of adjacent markers for which the sire is heterozygous. Progeny in which the allele inherited from the sire can be determined at both loci are ascertained and the linkage phase is taken as that which minimises the number of recombination events in the sire. If both phases are equally likely, one is selected at random. This is repeated for each pair of adjacent heterozygous markers to reconstruct the two sire gametes.

The probabilities for each progeny inheriting the two sire gametes are calculated for fixed positions through the linkage group conditional upon their marker genotypes. Given the sire gametes these probabilities are calculated based upon the genotypes at the two nearest markers flanking the chosen position that are informative in that progeny. For any position, the markers used to calculate these conditional probabilities will vary from sire to sire and from progeny to progeny within sire. For some individuals a chosen position may be outside the last informative marker in the linkage group and so the conditional probability will depend only on the last informative marker. If all markers in a linkage group are uninformative in an individual the conditional probabilities would be 0.5 for both sire gametes at all positions in the linkage group.

For a given position the conditional probabilities of the offspring inheriting the first gamete of the sire provide an independent variable on which the trait score can be regressed. For a single sire this would provide an estimate of the substitution effect (Falconer, 1989) for a heterozygous QTL at that position. In the simultaneous analysis of data from several sires the regression must be nested within sires. This is because not all sires will be heterozygous for any QTL and, for those that are, the linkage phase between the QTL and the sire gamete which has been designated as the first will vary from sire to sire. The between gamete within sire regression term, with degrees of freedom equal to the number of sires, is compared to the residual mean square, providing an F ratio test for the presence of a QTL. The position maximising this statistic is considered to be the most likely location for any QTL.

The use of multiple markers is compared to the single marker analysis described by Weller *et al.* (1990). For each marker in turn a test is provided by a comparison of the between marker allele within sire mean square to the residual mean square. Chance variation in the markers that are heterozygous in the sire and those informative in the offspring will cause the degrees of freedom of both the numerator and denominator of this test to vary from marker to marker.

SIMULATION STUDY

In order to compare the above methods, data were simulated for 20 sires each with 100 half-sib progeny. Each individual was composed of a 100 cM chromosome with markers at either 10 cM, 20 cM or 50 cM intervals. Markers had either 2 or 4 alleles at equal frequency segregating in the population. A phenotype was simulated with a heritability of 0.2. For each situation 10000 simulations and analyses were carried out in order to obtain suitable significance thresholds. In one set of replicates dam genotype information was utilised in the other it was ignored. Differences in degrees of freedom between replicates prevents the direct comparison of the F ratios. Hence the probability of the observed F ratio was determined and the most significant position selected for each replicate chromosome. Over

the 10000 replicates the significance level was determined which would give a whole chromosome type I error of 5% and 1% (i.e. the level which 5% or 1% of replicates, respectively, would be expected to exceed by chance somewhere on the chromosome if no QTL were segregating). To investigate the power of the proposed method 100 replicates were simulated with the genome and population described above with the addition of a QTL at position 25 cM (for the 10 and 50 cM spaced markers) or 30 cM (for the 20 cM spaced markers). The effect of the QTL was additive with one residual standard deviation (within QTL genotype) between the homozygotes.

RESULTS

We have chosen to present results for the 0.01 significance threshold as this more clearly illustrates the difference between the methods. The simulated empirical 0.01 thresholds are given in Table 1. For the single marker analyses these are close to those that would be obtained on the assumption that the tests at individual loci were independent (i.e. the overall 0.01 significance threshold ≈ 0.01/(number of tests)). Despite a larger number of tests being performed in the multiple marker approach, the significance thresholds tended to be higher, particularly when markers were closer together. This presumably occurred because individual tests at adjacent positions were highly correlated, even tests at the positions of adjacent markers being more highly correlated than they were in the individual marker analysis because of the use of information from multiple markers.

Table 1. Empirical 0.01 significance thresholds from 10000 replicate simulations of each situation.

	Marker	10 cM interval		20 cM interval		50 cM interval	
Method	alleles	Dams	No dams	Dams	No dams	Dams	No dams
Single marker	2	0.00081	0.00090	0.00182	0.00163	0.00378	0.00378
	4	0.00111	0.00081	0.00156	0.00173	0.00396	0.00362
Multiple marker	2	0.00164	0.00172	0.00302	0.00274	0.00410	0.00386
	4	0.00120	0.00111	0.00145	0.00191	0.00277	0.00278

The percentage of analyses of data in which a QTL was simulated which were significant at the empirical 0.01 threshold are shown in Table 2. It can be seen that the use of multiple markers increases the power in all situations (this was also true using the 0.05 threshold except for one instance where the power was 98% with both methods). The increase in power from use of multiple markers was greatest when the markers were closer together and the power was not already high. When there were only three markers 50 cM apart, there is presumably less chance to replace information lost at a marker with that from an adjacent one. Note also that if the power is not already high, there is a useful increase in power going from 20 cM to 10 cM marker spacing, which is not the case for data from an inbred line cross (Darvasi *et al.*, 1993; Haley and Knott, 1994). The use of dam genotype information increases the power as expected, and generally has greater effect with the less informative markers.

Table 2. Percentage of replicates significant at the empirical 0.01 threshold.

	Marker	10 cM interval		20 cM interval		50 cM interval	
Method	alleles	Dams	No dams	Dams	No dams	Dams	No dams
Single marker	2	80	57	67	41	34	21
	4	92	91	95	86	76	56
Multiple marker	2	95	89	93	74	37	25
	4	99	98	97	96	85	74

Mean estimates of position and of the empirical standard deviation of the position estimate are shown in Table 3. Estimates from the single marker analyses have been converted to a cM position for comparative purposes. The mean estimates of position were all reasonable except for the markers with 2 alleles in the map with 50 cM intervals. The standard deviation of the position estimate was always less from the multiple marker analysis, only half the value of that from the equivalent single marker

analysis in extreme cases. This is unsurprising as the QTL was always simulated at the midpoint between two markers and so even in the best possible case for the single marker analysis the most likely estimate of position will be one or other of these. The standard deviation was also decreased by the use of dam genotype information.

Table 3. Mean estimates of the position of the QTL and their empirical standard deviation (in parentheses) from the 100 replicate simulations of each situation. The simulated positions were 25 cM, 30 cM and 25 cM for the 10, 20 and 50 cM intervals respectively.

Method	Marker alleles	10 cM interval		20 cM interval		50 cM interval	
		Dams	No dams	Dams	No dams	Dams	No dams
Single marker	2	25.5 (15.7)	30.2 (20.3)	28.8 (15.4)	34.4 (20.3)	30.5 (32.5)	36.5 (35.4)
	4	26.0 (9.6)	26.3 (11.9)	31.0 (11.8)	30.6 (15.2)	24.0 (26.1)	27.5 (27.9)
Multiple marker	2	24.6 (7.3)	25.7 (13.4)	30.8 (13.8)	34.8 (18.1)	31.9 (25.1)	39.0 (32.6)
	4	25.4 (4.8)	26.2 (7.9)	29.8 (8.1)	29.7 (9.2)	23.2 (14.3)	25.1 (18.9)

DISCUSSION

A least squares approach to gene mapping can provide a relatively fast and simple method for the detection and location of QTLs. The use of multiple markers can increase the power and improve the parameter estimates. The incorporation of dam genotype information is beneficial, but obtaining these data would involve typing almost twice as many animals and, hence, the gains might not be economically justified.

Although estimates of the effect of the QTL are not obtainable directly using this approach, its speed and simplicity allow rapid scanning of the genome. With an accurate estimate of the location of the QTL and the potential of high power when informative markers are used, this method could be followed by the use of a more computationally demanding one, such as ML, on a restricted region of the genome enabling the additional parameters (including the effect and allele frequency of the QTL) to be estimated. Note that the least square approach used here may be less affected by departures from normality than ML and also requires no assumptions about the QTL allele frequencies in the sires and dams or Hardy-Weinberg equilibrium which may be needed to make ML computationally tractable. Thus it may be more robust, and hence preferable to ML, for selected livestock populations.

ACKNOWLEDGEMENTS

We are grateful for support from the BBSRC and MAFF and for AFRC/INRA fellowships to CSH and SAK. This work is associated with PiGMaP and supported by the EC BRIDGE programme.

REFERENCES

ARCHIBALD, A.L., BURT, D.W. and WILLIAMS, J. (1994) These proceedings.
DARVASI, A. WEINREB, A., MINKE, V., WELLER, J.I. and SOLLER, M. (1993) *Genetics* **134**: 943-951.
FALCONER, D.S. (1989) *Introduction to quantitative genetics. (3rd. ed.))* Longman, UK.
GELDERMANN, H. (1975) *Theor. Appl. Genet.* **46**: 319-330.
GELDERMANN, H.,, PIEPER, U. and ROTH, B. (1985) *Theor. Appl. Genet.* **46**: 319-330.
HALEY, C.S. and KNOTT, S.A. (1992) *Heredity* **69**: 315-324.
HALEY, C.S. and KNOTT, S.A. (1994) These proceedings.
HALEY, C.S., KNOTT, S.A. and ELSEN, J.M. (1994)*Genetics* **136**: (in press)
KNOTT, S.A. and HALEY, C.S. (1992a) *Genetics* **132**: 1211-1222.
KNOTT, S.A. and HALEY, C.S. (1992b) *Genet. Res.* **60**: 139-151.
LANDER, E.S. and BOTSTEIN, D. (1989) *Genetics* **121**: 185-199.
NIEMANN-SØRENSON, A. and ROBERTSON, A. (1961) *Acta Agric. Scand.* **11**: 163-196.
SOLLER, M. and GENIZI, A. (1978) *Biometrics* **34**: 47-55.
WELLER, J.I., KASHI, Y. and SOLLER, M. (1990) *J. Dairy Sci.* **73**: 2525-2537.
WELLER, J.I. (1986) *Biometrics* **42**: 627-640.