Genet. Res., Camb. (2002), **79**, *pp.* 75–83. *With* 1 *figure.* © 2002 Cambridge University Press
DOI: 10.1017/S0016672301005432 Printed in the United Kingdom

Comparison of methods used for recovering the line origin of alleles in a cross between outbred lines

Y. S. AULCHENKO[1,2], F. TEUSCHER[2], H. H. SWALVE[2] AND V. GUIARD[2*]
[1] *Institute of Cytology and Genetics SB RAS, 630090 Novosibirsk, Russia*
[2] *Research Institute for the Biology of Farm Animals, 18196 Dummerstorf, Germany*

(*Received 10 April 2001 and in revised form 2 August 2001*)

Summary

Here, we introduce the idea of probabilities of line origins for alleles in general pedigrees as found in crosses between outbred lines. We also present software for calculating these probabilities. The proposed algorithm is based on the linear regression method of Haley, Knott and Elsen (1994) combined with the Markov chain Monte Carlo (MCMC) method for estimating quantitative trait locus coefficients used as regressors. We compared the relative precision of our method and the original method as proposed by Haley *et al.* (1994). The scenarios studied varied in the allelic distribution of marker alleles in parental lines and in the frequency of missing marker genotypes. We found that the MCMC method achieves a higher accuracy in all scenarios considered. The benefits of using MCMC approximation are substantial if the frequency of missing marker data is high or the number of marker alleles is low and the allelic frequency distribution is similar in both parental lines.

1. Introduction

The problem of detecting genes responsible for differences between populations is an important issue in modern theoretical and applied genetics. The populations of natural and domesticated species are frequently divergent in phenotypes and provide a valuable source of genetic variation. One of the tools used to reveal the genetic basis of intrapopulation differences is an intercross experiment, in which the individuals from the populations under study are crossed and intercrossed. The animals are genotyped for marker loci and phenotyped, then the data are analysed by statistical methods.

Intercross experiments are widely used in model organisms. However, statistical methods used for linkage analysis in crosses of inbred model organisms are inapplicable to the analysis of data arising from a cross between outbred lines (breeds of livestock, laboratory and industrial stocks, etc.) because the parental populations used are not homozygous. Outbred lines typically exhibit a number of DNA, biochemical and quantitative trait locus (QTL) polymorphisms (Tagliaro *et al.*, 1993; Aulchenko *et al.*, 1998; Aggrey *et al.*, 1999; Riquet *et al.*, 1999) and can even be divergent in karyotype (Yosida, 1982; Rogatcheva *et al.*, 1998).

For some species, such as mice, the production of inbred strains prior to the establishment of a gene-mapping study is possible, although it is a relatively slow and expensive option. For slow breeding plants and animals, especially for those suffering from inbreeding depression, this option is not realistic. Therefore, suitable methods have been developed for linkage analysis of crosses between outbred lines (Haley *et al.*, 1994; Knott *et al.*, 1998; Perez-Enciso & Varona, 2000).

The power of gene detection is crucially dependent on two factors: the sample size and the efficiency of the method applied. The methods and software currently developed for multipoint linkage analysis in large pedigrees often use approximate methods to extract information on the inheritance of chromosomal segments in the pedigrees and perform linkage analysis in the framework of an assumed genetic model. The power of the methods depends crucially on the efficiency of approximation accepted and the validity of assumptions made.

[*] Corresponding author. Research Institute for the Biology of Farm Animals, Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany. Fax: +49 38208 68 602. e-mail: guiard@fbn-dummerstorf.de

One of the most popular approaches (Andersson *et al*., 1994; Knott *et al*., 1998; Brockmann *et al*., 1998; de Koning *et al*., 1999; Jeon *et al*., 1999; de Koning *et al*., 2000) for mapping QTLs in crosses between outbred lines is the least squares approach proposed by Haley *et al*. (1994). The attractive features of this approach are its relative simplicity, high speed and flexibility. Furthermore, suitable software is available. Currently, this software is restricted to the analysis of three-generation intercross pedigrees, which are common in livestock. In this type of pedigree, all animals are assumed to be genotyped, whereas only the animals from the F_2 generation are phenotyped.

The least squares approach is based on the line-cross concept, which assumes that original populations are homozygous for different QTL alleles (or, at least, that the QTL allele frequency distributions are notably different), but can have marker alleles in common. The key idea underlying the method is to calculate QTL coefficients $Pr_\psi(Q_k Q_m|M_{Pi,i})$ for each individual i at each putative QTL position ψ. Here, $M_{Pi,i}$ denotes the vector of marker phenotypes of the individual i and its direct ancestors (that is two F_1 parents and four P_0 grandparents); $Q_k Q_m$ with $km \in \{11,12,21,22\}$ denotes the genotype of the QTL, where allele Q_k came from population k via maternal meiosis and allele Q_m came from population m via paternal meiosis (Haley *et al*., 1994). The regressive model describing expectation of the value of the trait of individual i is,

$$y_i \approx \mu_{F2} + [Pr_\psi(Q_1Q_1|M_{Pi,i}) - Pr_\psi(Q_2Q_2|M_{Pi,i})]\, a + [Pr_\psi(Q_1Q_2|M_{Pi,i}) + Pr_\psi(Q_2Q_1|M_{Pi,i})]\, d$$

where μ_{F2} is the general mean of the F_2 and a and d are additive and dominance effects of the QTL, respectively. For each putative QTL position, ψ, the test of linkage is the F-ratio test between a model in which a and d are not constrained and a model in which they are constrained to be zero.

As can readily be seen, this model assumes that alleles from different breeds affect the value of quantitative trait differently. Although this assumption might be violated (which we will not discuss here), our objective is to evaluate the efficiency of Haley's original method to achieve information on line origin of genomic segments.

We should emphasize that the original method of Haley *et al*. calculates the line origin probabilities of QTL alleles of an animal using the information from its direct ancestors only. In many cases, the total pedigree information on the line origin of alleles is concentrated in this ancestor information only (e.g. if the original lines are very polymorphic or even fixed for alternative alleles at marker loci and the frequency of missing marker genotypes is low). However, if parental populations share marker alleles in common and/or there are missing marker genotypes, then not only direct ancestors but also other types of relative might provide important information on the line origin of alleles of the individual. In many situations, the benefits from the rapidity and simplicity of Haley's method should outweigh the benefits of slower, more complex methods that make use of complete pedigree information (Haley *et al*., 1994). However, there are conditions under which the amount of information ignored by the method becomes large.

Ideally, one would try to develop a fast method that uses all available information to detect a QTL. This development might be based on, for example, likelihood techniques developed in human statistical genetics. Two major groups of exact methods are used for the analysis of pedigrees with data on multiple markers. The first uses different modifications of the Elston–Stewart algorithm (Elston & Stewart, 1971; Cannings *et al*., 1978) and the second uses the Lander–Green algorithm (Lander & Green, 1987; Kruglyak *et al*., 1996). However, the use of these algorithms for multipoint analysis is restricted by the limitations of the computing time required. The former method does not allow analyses of more than about eight marker loci simultaneously and, in the case of a pedigree with loops (as is common in livestock breeding), the number of markers allowed is even smaller. The latter method restricts the size of the pedigree under analysis (typically < 20 meioses of interest), whereas much larger pedigrees are common in livestock breeding. From this, it is clear that implementations of these exact multipoint methods are not feasible in the context of QTL mapping in livestock and hence approximate methods might be of interest.

Markov chain Monte Carlo (MCMC) methods within different paradigms are increasingly used for the analysis of large pedigrees. The implementations include Bayesian (Heath, 1997; Daw *et al*., 1999) and frequentist (Guo & Thompson, 1992; Guo & Thompson, 1994) segregation and linkage analysis, and non-parametric linkage analysis (Sobel & Lange, 1996; Thompson & Heath, 1999; Thompson, 2000). The stochastic techniques are increasingly used for the calculation of multipoint probabilities of shared alleles that are identity-by-descent (IBD) between pairs of relatives. Several software packages are available that implement these methods (e.g. LOKI, SIMWALK, SOLAR). The main advantage of MCMC methods is their potential to use all information that is available to calculate IBD-sharing probabilities. Theoretically, the results from MCMC should converge to exact ones if the number of iterations approaches infinity. However, the computational demand for the algorithm to converge can be substantial.

The definitions of IBD-sharing probabilities and QTL coefficients are somewhat similar. For two populations, the QTL coefficients can be redefined as

https://doi.org/10.1017/S0016672301005432 Published online by Cambridge University Press

the probabilities of an individual sharing alleles (IBD) with its ancestors from one population, the other population or both populations. Therefore, similar computational techniques can be used to calculate both IBD-sharing probabilities and QTL coefficients. Although there is currently no program package that calculates QTL coefficients using an MCMC algorithm, such a package could be developed on the basis of existing software for calculation of IBD-sharing probabilities.

In this article, we propose the idea of and present the software for calculating the probabilities of line origins for alleles in general pedigrees, as found in crosses between individuals from two populations. The algorithm proposed is based on the MCMC method as implemented in the LOKI software (Heath, 2000) and uses all information available. We compare the precision of our method with the precision achieved by the approach of Haley *et al.* (1994) for a range of scenarios that are tractable with the latter method (i.e. three-generation pedigrees with marker genotypes measured in all generations) and the software available for this method. The precision was estimated by two criteria: the efficiency with which it predicts true QTL genotypes; and the curves of the mean F-ratio linkage statistic, which reflects the power to detect QTLs.

2. Material and methods

(i) *Simulations*

For a comparison of the accuracy of QTL coefficients calculated by the original method of Haley *et al.* (1994), henceforth denoted as OH, and MCMC, we simulated several data sets using the following pedigree structure: in the parental generation, it was assumed that all sires come from one population and dams from another. The pedigree structure consisted of three subfamilies. Each subfamily was obtained by crossing one sire with five dams. Each mating was assumed to result in two F_1 dams and one F_1 sire. Within the F_1 generation, each of three sires were mated to all dams of a different subfamily, each mating producing eight F_2 offspring. Thus, the F_2 generation consisted of 240 outbred animals.

One chromosome of length 60 cM was studied. Four marker loci were assumed to be located at positions 0 cM, 20 cM, 40 cM and 60 cM of the chromosome. It was assumed that, in the parental populations, the alternative alleles of the QTL were fixed. The additive effect of the QTL was set to $a = 1$ and the dominance effect set to 0. The environmental variance was set to 4·5. Under these conditions, the QTL explained 10 % of the total variance in F_2. The QTL was assumed to be located at the 50 cM position of the chromosome.

The simulated sets varied in the distribution of marker alleles in the parental populations and the frequency of missing marker genotypes. For each of the marker loci, we assumed two, five or ten marker alleles segregating at equal frequency in both parental populations (i.e. frequencies of 0·5, 0·2 and 0·1, respectively). Missing marker genotypes were distributed uniformly in the pedigree, with the frequency set to 0, 0·05 or 0·2. For each of the scenarios, 100 data sets were simulated that differed in allelic distribution and frequency of missed genotypes. The computer package MGA-SIMULATE v.0.06 (Aulchenko, 2000) was used for simulations. For each data set, QTL coefficients were calculated using OH and MCMC approximation.

(ii) *Methods*

The QTL coefficients were calculated in two ways: by the COEFF program (Haley *et al.*, 1994), which uses the OH approximation, and by the use of LOKI 2.3 software package (Heath, 2000), which applies the MCMC algorithm. The LOKI program, originally created for reversible jump MCMC oligogenic linkage and segregation analysis, also allows estimation of the proportion of alleles sharing IBD. We modified the part of the source code of LOKI 2.3 that is responsible for calculating this proportion to allow MCMC approximation of the QTL coefficients.

LOKI 2.3 uses the following algorithm. Two unique alleles are assigned to each founder at the genome points of interest. Thus, the total number of founder alleles in a pedigree is twice the number of founders. The genotypes of non-founder animals in a pedigree are simulated based on this assignment and conditional on marker data. LOKI, in our modified version, outputs the frequencies of possible genotypes, composed from founder alleles rather than IBD-sharing probabilities. Using this, it is possible to reconstruct the sharing frequency of alleles between F_2 individuals and founder individuals. Additional programs were written to calculate QTL coefficients based on this output and information on population origin of animals (and, consequently, founder alleles) in the parental generation.

For each of scenarios studied, the precisions of the OH and MCMC approaches to approximate QTL coefficients were estimated in two ways. First, a measure similar to the Euclidian measure of genetic distance used in population genetics (Weir, 1990): the normalized distance between the predicted and true probability distributions of the QTL genotype was calculated for each F_2 animal using the formula $D = \{[(p^{\mathrm{T}}_{qq} - p_{qq})^2 + (p^{\mathrm{T}}_{Qq} - p_{Qq})^2 + (p^{\mathrm{T}}_{QQ} - p_{QQ})^2]/2\}^{0\cdot5}$, where ($p^{\mathrm{T}}_{qq}$, p^{T}_{Qq} and p^{T}_{QQ}) is the true probability distribution of the QTL genotypes and (p_{qq}, p_{Qq}, p_{QQ}) is the estimated

https://doi.org/10.1017/S0016672301005432 Published online by Cambridge University Press

probability of different genotypes calculated at the QTL position by OH or MCMC approximation. The true distribution is either (1, 0, 0), (0, 1, 0) or (0, 0, 1) and thus the zero distance corresponds to the prediction of the true genotype with probability 1. The maximal distance 1 corresponds to a situation in which a false genotype is predicted with probability 1. For each simulated data set, the average distance over all animals in F_2 was calculated. The mean of this value, calculated over all simulated data sets, was used as a characteristic of the accuracy of a method to predict the QTL genotypes.

The second method was to use the curves of the average *F* ratio to reflect the power. *F*-ratio statistics were calculated using the ANAL program (S. Knott & C. Haley).

If marker information is not available (or, equivalently, the QTL is unlinked) then the estimated probability distribution of the QTL genotypes would be (0·25, 0·5, 0·25), whereas the true distribution would be (1, 0, 0), (0, 1, 0) or (0, 0, 1) with probabilities 0·25, 0·5 and 0·25, respectively. In this situation, the expected value of measure (*I*) would be 0·547. By contrast, *I* would be 0·183 if the locus of interest is located halfway between two fully informative markers spaced 20 cM apart. In our simulations, we therefore expect *I* to be between 0·183 and 0·547. If *I* is close to the latter value, this indicates a very poor reconstruction of the true QTL genotypes.

Prior to the simulations outlined above, we addressed the issue of convergence. For this purpose, five data sets were generated under each of the nine scenarios. For each of these data sets the 'reference' *F* ratio was calculated at the true QTL location after generating QTL coefficients using 10,000 dememorization and 40,000 effective MCMC iterations. For scenarios considering ten or five alleles, the *F* ratio converged to the reference $F \pm 5\%$ when 100 dememorization and 1000 effective MCMC iterations were used. However, for scenarios involving two alleles, the same result was obtained only after using 1000 dememorization and 4000 effective iterations. As a conclusion from these results, we decided to use 1100 MCMC iterations when studying scenarios with ten or five alleles and 5000 iterations for scenarios with two alleles.

3. Results

We studied the dependence of the accuracy of the approximations on the distribution of allelic frequencies in parental lines and on the frequency of missing marker genotypes. Table 1 shows the mean distance between estimated and true probability distributions of QTL genotypes for different scenarios varying in number of marker alleles and frequency of missing marker genotypes. As a first case, we studied three scenarios that varied in the number of marker alleles. It was assumed that there are no missing marker genotypes.

From these data, it follows that, for both methods, the distance increases with decreasing number of marker alleles in the parental lines. For the scenario with ten and five marker alleles, the average distances are very close to the distance of 0·18, as expected when using completely informative markers. When the number of marker alleles is two, distances arising from both MCMC and OH are far from 0·18. It is easy to see that, in general, the MCMC approximation gives better results than the OH approximation: the mean distance is smaller when QTL coefficients are calculated using MCMC. The use of MCMC is advantageous for a low number of alleles: when there are ten alleles, the difference between the mean distances of MCMC-based and OH-based prediction is only 0·002, whereas, when the number of alleles is two, it increases substantially to 0·039.

Although the general tendency is clear from Table 1, the curves of the mean *F* ratio (Fig. 1A) and Table 2 (which presents the mean *F*-ratio at the QTL position) are more demonstrative and show directly how much the results of the search for a QTL might be affected by the use of either OH or MCMC approximation. Fig. 1A shows that the mean *F* ratios increase with the number of alleles. It can be seen from Table 2 that, at the QTL location (50 cM), the mean *F* ratio based on the QTL coefficients calculated using MCMC is higher than the *F* ratio resulting from QTL coefficients calculated using OH for all cases. When the number of common alleles is high (ten or five), there is only a small increase in the *F* ratio calculated at the 50 cM point based on MCMC over that calculated based on OH (factors of 1·006 and 1·012, respectively). In Fig. 1A, the curves resulting from OH and MCMC are nearly indistinguishable. When the number of common alleles is small, the *F* ratio is increased by a factor of 1·124 when the MCMC approximation is used. This 10% increase in the *F* ratio might have an important influence on the power and so MCMC-based QTL coefficients should be chosen in this situation.

We also studied the dependence of approximation accuracy and the frequency of missing marker genotypes. 100 data sets were simulated for scenarios assuming the frequency of missing markers to be either 0·05 or 0·2, and assuming two, five or ten equally frequent alleles common for both parental lines at the marker loci.

As expected, Table 1 indicates that, in all cases with missing data, the efficiency of a method to predict QTL genotypes is lower: the higher the frequency of missing markers, the lower the precision. Table 1 also

https://doi.org/10.1017/S0016672301005432 Published online by Cambridge University Press

shows that the MCMC approximation gives better results than OH: the prediction is always better when QTL coefficients are calculated using MCMC. Moreover, as the number of alleles decreases and the frequency of missing markers increases, the difference between the precision of MCMC and OH approximations increases in favour of MCMC. From Table 1, it follows that the situation with two marker alleles and a frequency of missing genotypes of 0·2 is crucial for the OH approximation: the mean distance between true and estimated QTL genotypes is approaching 0·55, as expected in a situation with a lack of marker information; at the same time, the distance resulting from MCMC is considerably smaller. This indicates that MCMC should be preferred in situations where the frequency of missing marker genotypes is high and especially if the number of marker alleles is low and the allelic frequency distribution is similar in both parental lines.

Fig. 1B,C shows the influence of the frequency of missing marker genotypes and number of marker alleles on the mean *F* ratio obtained by OH and MCMC approximations. A comparison of Fig. 1A–C and the columns of Table 2 demonstrates that, when data is missing, the mean *F* ratio decreases, irrespective of the approximation used: the higher the frequency, the lower the mean *F* ratio. Remarkably, the difference between the *F* ratios calculated by OH and MCMC approximations increases in favour of MCMC as the frequency of missing data increases.

At the QTL location point, the MCMC-based *F* ratios increase over the OH-based ratios by factors of 1·007, 1·040 and 1·245 when the number of alleles is ten, five and two, respectively, and the frequency of missing data is 0·05. When the frequency of missing data is 0·2, the increase of the mean *F* ratios calculated using MCMC becomes even higher (increasing by factors of 1·065, 1·085 and 1·377 for ten, five and two marker alleles, respectively).

It is worth noting that, when there are no missing marker genotypes or the frequency is 0·05, the mean *F* ratios resulting from the analysis of the data assuming ten alleles are higher than the ratios resulting from the analysis of data assuming five alleles, which in turn are higher than the two-allele ratios. When the frequency of missing data is 0·2, the two curves resulting from MCMC assuming ten or five alleles exhibit the highest values (Fig. 1C).

The results indicate that, when the frequency of missing data is high or even moderate, the MCMC approximation for calculation of QTL coefficients is superior to the OH approximation. The MCMC approximation should certainly be preferred when the frequency of missing marker genotypes is high, the number of marker alleles is low and the allelic distributions overlap substantially in both populations.

Table 1. *Mean ± standard deviation of the distance between the true and the estimated probability distributions at the QTL location of QTL genotypes by number of marker alleles, frequency of missing marker genotypes and approximation used*

Number of marker alleles	Frequency of missing marker genotypes								
	0			0·05			0·2		
	Approximation			Approximation			Approximation		
	OH	MCMC	*D* (OH − MCMC)[a]	OH	MCMC	*D* (OH − MCMC)[a]	OH	MCMC	*D* (OH − MCMC)[a]
10	0·201 ± 0·024	0·199 ± 0·023	0·002	0·221 ± 0·026	0·213 ± 0·022	0·008	0·271 ± 0·035	0·241 ± 0·023	0·030
5	0·235 ± 0·034	0·227 ± 0·032	0·008	0·253 ± 0·037	0·235 ± 0·033	0·018	0·320 ± 0·039	0·275 ± 0·029	0·045
2	0·418 ± 0·050	0·379 ± 0·059	0·039	0·437 ± 0·048	0·392 ± 0·057	0·045	0·478 ± 0·034	0·420 ± 0·046	0·058

[a] *D*(OH − MCMC) is the difference between the mean distances calculated by the OH and MCMC approximations.

https://doi.org/10.1017/S0016672301005432 Published online by Cambridge University Press

https://doi.org/10.1017/S0016672301005432 Published online by Cambridge University Press

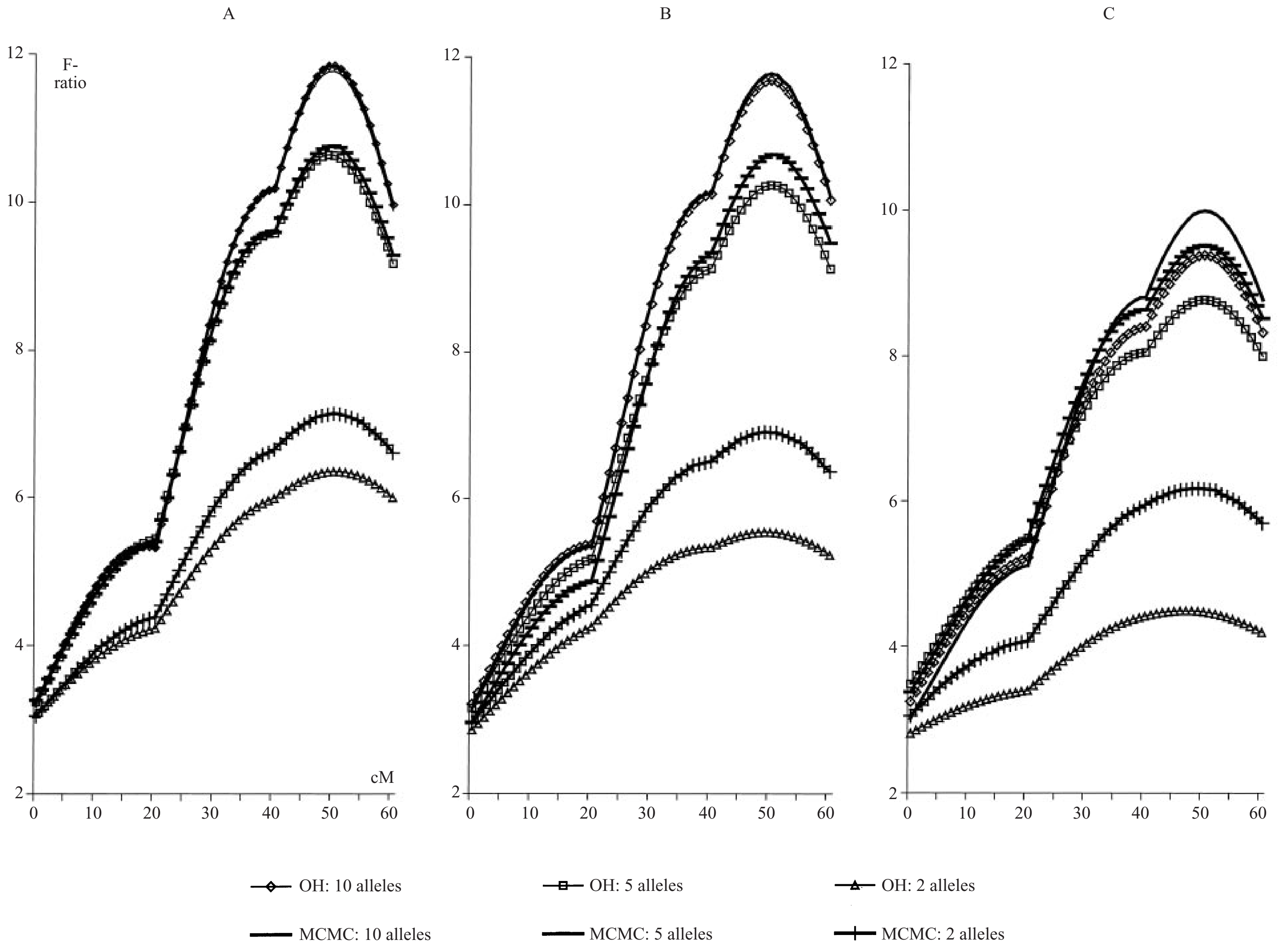

Fig. 1. Mean *F* ratio by number of marker alleles segregating at equal frequency in parental populations, the approximation used and the frequency of missing marker genotypes. Four marker loci are located at 0 cM, 20 cM, 40 cM and 60 cM along the chromosome, and a QTL is located at 50 cM. (A) No missing marker genotypes. (B) Frequency of missing marker genotypes: 0·05. (C) Frequency of missing marker genotypes: 0·20.

Table 2. *Mean* F*-ratio at the QTL location (50 cM) by number of marker alleles, frequency of missing marker genotypes and approximation used*

Number of marker alleles	Frequency of missing marker genotypes								
	0			0·05			0·2		
	Approximation			Approximation			Approximation		
	OH	MCMC	$F_{MCMC} \div F_{OH}$	OH	MCMC	$F_{MCMC} \div F_{OH}$	OH	MCMC	$F_{MCMC} \div F_{OH}$
10	11·756	11·830	1·006	11·685	11·771	1·007	9·385	9·993	1·065
5	10·624	10·749	1·012	10·270	10·680	1·040	8·774	9·522	1·085
2	6·349	7·138	1·124	5·544	6·901	1·245	4·481	6·169	1·377

4. Discussion

The method originally proposed by Haley *et al*. (1994) for the calculation of probabilities of line origin of alleles is not well suited to the case when all information that can possibly be used by a simultaneous consideration of all animals in a pedigree and all marker genotypes is intended to be used. However, when the markers are highly polymorphic and/or the allelic distribution between lines is sufficiently different, and the number of missed marker genotypes is small, the benefits from the speed and simplicity of Haley's method should normally outweigh the benefits from slower and more complex methods that make use of complete information. An example is the case of two lines in which different marker alleles are fixed and there is no missing data. Here, Haley's method uses all marker information for calculating QTL coefficients (Haley *et al*., 1994; Haley & Knott, 1992).

However, a practical experiment might deviate from the ideal situation. The frequency of missing marker genotypes might be substantially higher than zero, the distributions of marker allele frequencies might overlap and the number of different marker alleles might be small in outbred lines. In these situations, a method that uses all information should be preferred. The implementation of an exact likelihood method using all available information, although theoretically possible, is in practice too expensive in relation to the time needed for computations (Sobel & Lange, 1996; Heath, 1997). Therefore, approximate methods should be used. We propose a modification of the regression method of Haley using QTL coefficients, which are calculated by means of the Markov chain Monte Carlo method. The advantage of this is that it uses all available information about these coefficients. We have studied the extent to which deviations from the ideal situation (missing marker genotypes, similarity of allelic distribution in parental lines) affect the relative operational characteristics of the original Haley approach and MCMC approximations.

In all situations considered, the MCMC approximation to calculating QTL coefficients exhibits a better precision than the original approximation proposed by Haley and colleagues. Use of the MCMC allows better prediction of QTL genotypes. The mean *F* ratios are generally higher if MCMC rather then OH approximation is used. The benefits from the use of MCMC approximations are most evident if the frequency of missing marker genotypes is high and/or the distribution of allelic frequencies is similar in both parental lines while the number of marker alleles is small.

In this study, we considered one 60 cM-length chromosome with four marker loci, located at 0 cM,

https://doi.org/10.1017/S0016672301005432 Published online by Cambridge University Press

20 cM, 40 cM and 60 cM. We assumed that the distribution of the frequency of marker alleles is equal in both parental lines. Scenarios studied differed in the assumed number of marker alleles in parental lines and the frequency of missing data. We demonstrated that, as the number of marker alleles segregating at equal frequency in both parental populations decreases, the precision and power of analysis also decreases, regardless of the method used to calculate QTL coefficients. The same is true for the frequency of missing marker genotypes: as this frequency grows, the precision and power of the analysis is reduced. Remarkably, the precision and power of Haley's original approach decreases faster than those of the MCMC approach as the number of marker alleles decreases. This occurs because the line origin of alleles of a genotype of an F_1 heterozygous offspring whose parents have the same heterozygous genotypes cannot be recovered within the framework of OH approximation, whereas the MCMC approximation attempts to recover the line origin by the joint use of information from all animals and flanking markers. When the number of marker alleles is small and these alleles are common in both parental lines, any types of markers can be used if the intercrossed populations under analysis are derived from a common ancestral population with low number of marker alleles (see for example Brockmann *et al*., 1998).

We have also demonstrated that, when the frequency of missing markers increases, the precision and the power of the OH approach decreases faster than those of MCMC. This occurs because missing marker genotypes lead to situations in which the OH approximation fails to recover the population origin of marker alleles in F_2 progeny, whereas this does not occur under MCMC. For example, if a couple of grandparents include one with a missing genotype and one that is heterozygous, and their offspring has the same heterozygous genotype, this F_1 offspring is considered to be uninformative by the OH approach. At the same time, the line origin of alleles of this F_1 individual might be recovered by the joint consideration of its sibs and half-sibs. Furthermore, even if individuals in parental and F_1 generations are not genotyped at all, MCMC has the potential to recover the line origin of alleles in F_2 provided that the pedigree structure is complex enough.

It is common for there to be several missing marker genotypes in experimental data. We have demonstrated that, even if the frequency of missing data is small (5 %), the use of MCMC leads to an increase of the mean *F* ratio from 1 % to 25 % in situations that differ in the frequency distribution of marker alleles. As the frequency of missing data increases, the benefits from the use of MCMC become more and more evident. When the frequency of missing data is 20 %, the increase in the mean *F* ratio is between 7 % and 38 %. Theoretically, the use of MCMC should be most beneficial if, for some grandparents and/or parents, there is no marker information at all.

The major advantage of the OH method is its computational speed and relative simplicity. Although the calculation of QTL coefficients at 61 points for 1100 iterates of MCMC in our simulation study required 2–6 min on a 650 MHz Pentium III, the OH algorithm required only a few seconds. However, we demonstrated that, in some situations, the use of MCMC approximation to calculate QTL coefficients might help to improve the results of gene hunting significantly. Even with current computational facilities, it is feasible to calculate QTL coefficients genome-wide for several hundred individuals using several hundred thousand MCMC iterations. We suppose that, with further development of computational facilities, the MCMC approximation will more and more become the method of choice.

The original software implementing the algorithm of Haley *et al.* (1994) is restricted to the analysis of three-generation pedigrees in which all individuals are genotyped and only F_2 progeny are phenotyped. Although this type of data is typical for livestock (Andersson *et al*., 1994; Knott *et al*., 1998; Brockmann *et al*., 1998; de Koning *et al*., 1999; Jeon *et al*., 1999; de Koning *et al*., 2000), several other experimental designs are possible. Based on the LOKI v.2.3 program, we have created a package of programs that calculate MCMC-based QTL coefficients. The software is not restricted to three-generation outbred pedigrees coming from an intercross experiment. Rather, it can deal with any type of pedigree from a cross between individuals from two populations. The software is available via anonymous access from http://mga.bionet.nsc.ru/SOFT/.

We are grateful to S. Heath for the software and his useful comments. Y.S.A. acknowledges financial support from RFBR (01-04-49518, 01-04-48875), a short-term research grant from the German Academic Exchange Service (DAAD) and a young scientists support grant from SD RAS.

References

Aggrey, S. E., Yao, J., Sabour, M. P., Lin, C. Y., Zadworny, D., Hayes, J. F. & Kuhnlein, U. (1999). Markers within the regulatory region of the growth hormone receptor gene and their association with milk-related traits in Holsteins. *Journal of Heredity* **90**, 148–151.

Andersson, L., Haley, C. S., Ellegren, H., Knott, S. A., Johansson, M., Andersson, K., Andersson-Eklund, L., Edfords-Lilja, I., Fredholm, M., Hansson, I., Hakansson, J. & Lundstrom, K. (1994). Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science* **263**, 1771–1774.

Aulchenko, Y. S. (2000). The system for analysis of complex traits. In *Proceedings of the Young Scientists Conference* (Vol. 2), pp. 26–28. Novosibirsk: The Siberian Division of the Russian Academy of Sciences.

https://doi.org/10.1017/S0016672301005432 Published online by Cambridge University Press

Aulchenko, Y. S., Oda, S., Rogatcheva, M. B., Borodin, P. M. & Axenovich, T. I. (1998). Inheritance of litter size at birth in the house musk shrew (*Suncus murinus*, Insectivora, Soricidae). *Genetical Research* **71**, 65–72.

Brockmann, G. A., Haley, C. S., Renne, U., Knott, S. A. & Schwerin, M. (1998). Quantitative trait loci affecting body weight and fatness from a mouse line selected for extreme high growth. *Genetics* **150**, 369–381.

Cannings, C., Thompson, E. A. & Skolnik, M. H. (1978). Probability function on complex pedigrees. *Advances in Applied Probability* **10**, 26–61.

Daw, E. W., Heath, S. C. & Wijsman, E. M. (1999). Multipoint oligogenic analysis of age-at-onset data with applications to Alzheimer disease pedigrees. *American Journal of Human Genetics* **64**, 839–851.

de Koning, D. J., Janss, L. L. G., Rattink, A. P., van Oers, P. A. M., de Vries, B. J., Groenen, M. A. M., van der Poel, J. J., de Groot, P. N., Brascamp, E. W. & van Arendonk, J. A. M. (1999). Detection of quantitative trait loci for backfat thickness and intramuscular fat content in pigs (*Sus scrofa*). *Genetics* **152**, 1679–1690.

de Koning, D. J., Rattink, A. P., Harlizius, B., van Arendonk, J. A. M., Brascamp, E. W. & Groenen, M. A. M. (2000). Genome-wide scan for body composition in pigs reveals important role of imprinting. *Proceedings of the National Academy of Sciences of the USA* **97**, 7947–7950.

Elston, R. C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21**, 523–542.

Guo, S. W. & Thompson, E. A. (1992). A Monte Carlo method for combined segregation and linkage analysis. *American Journal of Human Genetics* **51**, 1111–1126.

Guo, S. W. & Thompson, E. A. (1994). Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics* **50**, 417–432.

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.

Heath, S. C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics* **61**, 748–760.

Heath, S. C. (2000). *Loki 2.3. A Package for Multipoint Linkage Analysis on Large Pedigrees Using Reversible Jump Markov Chain Monte Carlo*. Documentation for the package can be ordered at the following address: Simon C. Heath, Memorial Sloan–Kettering Cancer Center, Department of Human Genetics, 1275 York Avenue, New York, NY 10021 or downloaded as the file loki_doc.ps within ftp://ftp.u.washington.edu/pub/user-supported/pangaea/PANGAEA/Loki/loki_2.3.tar.gz

Jeon, J. T., Carlborg, O., Tornsten, A., Giuffra, E., Amarger, V., Chardon, P., Andersson-Eklund, L., Andersson, K., Hansson, I., Lundstrom, K. & Andersson, L. (1999). A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the *IGF2* locus. *Nature Genetics* **21**, 157–158.

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

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics* **58**, 1347–1363.

Lander, E. & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the USA* **84**, 2363–2367.

Perez-Enciso, M. & Varona, L. (2000). Quantitative trait loci mapping in F_2 crosses between outbred lines. *Genetics* **155**, 391–405.

Riquet, J., Coppieters, W., Cambisano, N., Arranz, J. J., Berzi, P., Davis, S. K., Grisart, B., Farnir, F., Karim, L., Mni, M., Simon, P., Taylor, J. F., Vanmanshoven, P., Eagenaar, D., Womack, J. E. & Georges, M. (1999). Fine mapping of quantitative trait loci by identity by descent in outbred populations: application to milk production in dairy cattle. *Proceedings of the National Academy of Sciences of the USA* **96**, 9252–9257.

Rogatcheva, M. B., Oda, S., Axenovich, T. I., Aulchenko, Y. S., Searle, J. B. & Borodin, P. M. (1998). Chromosomal segregation and fertility in Robertsonian chromosomal heterozygotes of the house musk shrew (*Suncus murinus*, Insectivora, Soricidae). *Heredity* **81**, 335–341.

Sobel, E. & Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *American Journal of Human Genetics* **58**, 1323–1337.

Tagliaro, C. H., Franco, M. H. L. P. & Meincke, W. (1993). Biochemical polymorphism and genetic relationships among Landrace, Large White and Duroc pigs from southern Brazil. *Review of Brazilian Genetics* **16**, 671–678.

Thompson, E. A. & Heath, S. C. (1999). Estimation of conditional multilocus gene identity among relatives. In F. Seillier-Moiseiwitsch (editor), *Statistics in Molecular Biology and Genetics* – Selected proceedings of a 1997 Joint AMS–IMS–SIAM Summer Conference on Statistics in Molecular Biology. Volume 33 of *IMS Lecture Notes – Monograph Series*. Institute of Mathematical Statistics, Hayward, California.

Thompson, E. A. (2000). MCMC estimation of multi-locus genome sharing and multipoint gene location scores. *International Statistical Review* **68**, 53–73.

Weir, B. S. (1990). *Genetic Data Analysis*. Sunderland, Mass.: Sinauer Associates, Inc.

Yosida, T. H. (1982). Cytogenetical studies on Insectivora. II. Geographical variation of chromosomes in the house shrew, *Suncus murinus* (Soricidae), in East, Southeast and Southwest Asia, with a note on the karyotype evolution and distribution. *Japan Journal of Genetics* **57**, 101–111.

https://doi.org/10.1017/S0016672301005432 Published online by Cambridge University Press