Genet. Res., Camb. (2002), **79**, *pp.* 247–258. *With* 4 *figures.* © 2002 Cambridge University Press
DOI: 10.1017/S0016672302005645 Printed in the United Kingdom

A method for computing identity by descent probabilities and quantitative trait loci mapping with dominant (AFLP) markers

MIGUELPÉREZ-ENCISO[1]* AND ODILEROUSSOT[2]
[1] *Station d'Amélioration Génétique des Animaux, INRA, BP 27, 31326 Castanet-Tolosan, France*
[2] *Laboratoire de Génétique Cellulaire, INRA, BP 27, 31326 Castanet-Tolosan, France*

(*Received 13 August 2001 and in revised form 12 December 2001 and 21 February 2002*)

Summary

Amplified fragment length polymorphisms (AFLPs) are a widely used marker system: the technique is very cost-effective, easy and rapid, and reproducibly generates hundreds of markers. Unfortunately, AFLP alleles are typically scored as the presence or absence of a band and, thus, heterozygous and dominant homozygous genotypes cannot be distinguished. This results in a significant loss of information, especially as regards mapping of quantitative trait loci (QTLs). We present a Monte Carlo Markov Chain method that allows us to compute the identity by descent probabilities (IBD) in a general pedigree whose individuals have been typed for dominant markers. The method allows us to include the information provided by the fluorescent band intensities of the markers, the rationale being that homozygous individuals have on average higher band intensities than heterozygous individuals, as well as information from linked markers in each individual and its relatives. Once IBD probabilities are obtained, they can be combined into the QTL mapping strategy of choice. We illustrate the method with two simulated populations: an outbred population consisting of full sib families, and an F_2 cross between inbred lines. Two marker spacings were considered, 5 or 20 cM, in the outbred population. There was almost no difference, for the practical purpose of QTL estimation, between AFLPs and biallelic codominant markers when the band density is taken into account, especially at the 5 cM spacing. The performance of AFLPs every 5 cM was also comparable to that of highly polymorphic markers (microsatellites) spaced every 20 cM. In economic terms, QTL mapping with a dense map of AFLPs is clearly better than microsatellite QTL mapping and little is lost in terms of accuracy of position. Nevertheless, at low marker densities, AFLPs or other biallelic markers result in very inaccurate estimates of QTL position.

1. Introduction

Microsatellites are the state-of-the-art markers due to their high information content. Nevertheless, a microsatellite genetic map is expensive to develop, and mapping efforts have been concentrated on the economically most important species. Dense genetic maps for microsatellites are thus available for these species but are unlikely to be developed for all species. More recently, single nucleotide polymorphisms (SNPs) have tended to be favoured for humans as a systematic effort to uncover all genetic variation (Wang *et al.*, 1998). Such analysis requires large-scale genotyping systems. Alternatively, other techniques can provide markers; among them fragment length polymorphisms (AFLPs; Vos *et al.*, 1995) is a very cost-effective, easy and rapid technique that reproducibly generates hundreds of markers. Unlike microsatellites it needs no prior knowledge of the genome under study nor prior development. Moreover, since AFLP is a multilocus technique (a single PCR reaction allows the genotyping of several markers), it is a cheap alternative to microsatellites: genotyping cost per individual per marker is lowered by 10- to 20-fold.

AFLPs are used for broad applications (for a review see Mueller & Wolfenbarger, 1999), including establishing or enriching maps, mainly in plants (Kuiper, 1998; Alonso-Blanco *et al.*, 1998; Cho *et al.*,

* Corresponding author. Tel: +33 5 61 28 51 82. Fax: +33 5 61 28 53 53. e-mail: mperez@toulouse.inra.fr

1998; Vuylsteke *et al*., 1999; Klein *et al*., 2000) but also in animal species (Ajmone-Marsan *et al*., 1997; Herbergs *et al*., 1999; Knorr *et al*., 1999), and for quantitative trait locus (QTL) studies (Vanhaeringen *et al*., 2001).

The main drawback of AFLPs is their biallelic and dominant nature. Thus, in addition to a much lower information content than microsatellites due to the small number of alleles, new methodological problems are posed due to their dominant behaviour. AFLPs have been used to identify loci affecting traits of interest, but they have primarily been concerned with dichotomous traits, mainly disease resistance, where presence or absence of bands was correlated with phenotype category (e.g. Cervera *et al*., 1996; Jin *et al*., 1998; Lu *et al*., 1998). The fact of markers being dominant makes the analysis of quantitative traits more problematic, because the trait cannot be categorized into discrete classes. For instance, Nandi *et al*. (1997) identified QTLs related to submergence tolerance in rice using AFLPs but they divided a 9-category trait into two groups and they used recombinant inbred lines, where no heterozygous individuals are expected. Otsen *et al*. (1996) also used recombinant inbred lines in a QTL mapping project via AFLPs. Jiang & Zeng (1997) derived a method to map QTLs in crosses between inbred lines, assuming that all markers have alternative fixed alleles in each line. To our knowledge, there is no statistically sound method that allows us to carry out a QTL analysis with dominant markers in a general pedigree and uses all linkage and marker information.

Here we present a Bayesian Monte Carlo Markov Chain (MCMC) method that allows us to compute the identity by descent probabilities (IBD) in any general pedigree in which individuals have been typed for dominant markers. The method allows us to include the information provided by the fluorescent band intensities of the markers, the rationale being that homozygous individuals have on average higher band intensities than heterozygous individuals. Piepho & Koch (2000) and Jansen *et al*. (2001) also studied how peak density can be applied to discriminate between dominant homozygous and heterozygous genotypes, but they did so by considering only the distribution of peak densities within each marker. Here we show how to combine peak density with linkage and pedigree information using all markers simultaneously. We illustrate the method with simulated data. The simulated data are also employed to compare the performance of QTL mapping with AFLPs versus codominant markers such as SNPs or microsatellites. This work was inspired by a continuing project in quail aimed at identifying genes with an effect on tonic immobility, and in which AFLPs have been chosen because there is no published genetic map for that species.

2. Materials and methods

(i) *Computation of identity by descent probabilities*

The approach employed here is a generalization of that in Pérez-Enciso *et al*. (2000). In short the method consists of iterating successively over three steps: genotype and phase sampling, crossover sampling, and assessment of the identity by descent status at predetermined genome positions. In that work, markers were assumed to be additive and thus only the phases of genotyped individuals plus the genotypes of untyped individuals were sampled. Here biallelic markers with dominant 'D' and recessive allele 'R' are assumed. We assume that only genotypes 'RR' are identified unambiguously, and correspond to no band amplification. The first step consists of sampling the ordered marker genotypes. By ordered genotype we mean that the two heterozygous genotypes, 'RD' and 'DR', are treated as distinct, corresponding to the two possible phases. The distances between markers are assumed to be known, and the Haldane mapping function is supposed to hold. Let G_{ij} be the ordered genotype of individual i at marker j; the genotypes are sampled from

$$\begin{aligned} \mathrm{p}(G_{ij}\,|\,\mathbf{M},\,\mathbf{G},\,h_{ij}) &\propto \mathrm{p}(\mathbf{G}\,|\,G_{ij},\,\mathbf{M})\ \mathrm{p}(h_{ij}\,|\,G_{ij})\ \mathrm{p}(G_{ij}) \\ &= [a]\times[b]\times[c], \end{aligned} \tag{1}$$

where $\mathbf{M}$ is the available marker information, consisting of the unordered genotypes of typed individuals. Typically $\mathbf{M}$ consists of records 'RR', 'D0' and '00', where '0' stands for missing allele. $\mathbf{G}$ is the set of ordered genotypes other than the one considered (G_{ij}), and h_{ij} is the band intensity recorded from the fluorescent reader for individual i ($i = 1, n$, the number of individuals) and marker j. The first term $[a]$ represents the linkage information from current genotype configurations, as fully described previously (Pérez-Enciso *et al*., 2000). The sampling algorithm for dominant markers is presented in the Appendix.

The second term $[b] = \mathrm{p}(h_{ij}\,|\,G_{ij})$ conveys the information provided by the band intensities as a result of PCR amplification. These depend on the individual's genotype, but can also be affected by a number of systematic environmental effects such as gel, DNA extraction or PCR amplification, and run conditions. If we assume that the residual variances are independent across markers, we can model each marker band separately. Thus, using a hierarchical Bayesian approach for marker j, we have:

$$\mathbf{h}_j = \mathbf{Z}_j\boldsymbol{\mu}_j + \mathbf{X}_j\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \tag{2}$$

where $\mathbf{h}_j$ is the vector containing the band intensities for marker j, $\boldsymbol{\mu}_j$ is a two-dimensional vector with the mean band intensity of individuals having genotype 'DD' (μ_{j1}) and 'DR' or 'RD' (μ_{j2}), $\boldsymbol{\beta}_j$ is a vector with the remaining fixed effects, $\mathbf{Z}_j$ is a $n\times 2$ incidence

https://doi.org/10.1017/S0016672302005645 Published online by Cambridge University Press

matrix where the first column is 1 if G_{ij} is equal to 'DD', 0 otherwise. Symmetrically, the second column contains 1's when G_{ij} is 'DR' or 'RD', 0 otherwise. Note that the $\mathbf{Z}_j$ are stochastic matrices; they change from iteration to iteration for those individuals whose genotype is uncertain. Finally, $\mathbf{X}_j$ is an incidence matrix relating band intensities to fixed effects, and $\boldsymbol{\varepsilon}_j$ is the residuals' vector which are independently identically distributed as $N(0, \sigma^2)$. Different variances can be accommodated within genotype, as described below. Each marker can certainly be influenced by different factors, e.g. there can be markers whose bands are more stable and less dependent on environmental conditions than others. Thus the model (2) fitted is not necessarily the same for every marker. For any marker, a model with the minimum number of parameters affecting band intensities should be chosen. The Bayes factor between two competing models in (2) can be computed in order to choose the more reasonable model, although in practice a visual inspection of the plot of the residuals $\boldsymbol{\varepsilon}$ obtained with the two models may suffice. Now consider the usual least square equations (the marker subscript j is omitted in the following equations for the sake of clarity):

$$\begin{bmatrix} \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{X} \\ \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}'\mathbf{h} \\ \mathbf{X}'\mathbf{h} \end{bmatrix},$$

or $\mathbf{C}\ \mathbf{r} = \mathbf{d}$. It is well established (e.g. Wang *et al.*, 1993) that the posterior conditional distribution of any element r_k of $\mathbf{r} = [\boldsymbol{\mu}, \boldsymbol{\beta}]'$ is

$$p(r_k \mid \mathbf{Z}, \mathbf{X}, \mathbf{h}, \mathbf{r}_{-k}, \sigma^2) = \text{Normal} \left(d_k - \sum_{h=1, h \neq k}^{m} c_{kh} r_h, \sigma^2 / c_{kk} \right), \quad (3)$$

where m is the dimension of $\mathbf{C}$, c_{kh} is the element khth of $\mathbf{C}$, d_k is kth $\mathbf{d}$ element. In the simplest case, the only effect included in (2) is the genotype, and the posterior distribution of μ_1 is

$$p(\mu_1 \mid \mathbf{Z}, \mathbf{h}, \sigma^2) = N[\mathbf{Z}_1' \mathbf{h}/(\mathbf{Z}_1' \mathbf{Z}_1), \sigma^2/(\mathbf{Z}_1' \mathbf{Z}_1)],$$

where $\mathbf{Z}_1$ is the first column of $\mathbf{Z}$, and $\mathbf{Z}_1' \mathbf{Z}_1$ is the current number of individuals with genotype 'DD'. Similarly for μ_2. The underlying model assumes that dominant homozygotes have a mean band intensity that is necessarily higher than heterozygotes, i.e. $\mu_1 \geqslant \mu_2$. This is accomplished within the sampling process simply by setting $\mu_1 = \mu_2$ if μ_1 is sampled to be smaller than μ_2. This is equivalent to disregarding band information in that iteration. An alternative is to discard the whole iteration. Usual Bayesian theory dictates that the posterior distribution of the residual variance is an inverted chi-squared when we assume conditional normal distributions in (2). Assuming a non-informative prior of the type $1/\sigma$ for scale parameters (Wang *et al.*, 1993),

$$p(\sigma^2 \mid \mathbf{Z}, \mathbf{X}, \mathbf{h}, \mathbf{b}, \boldsymbol{\mu}) = (\mathbf{h} - \mathbf{Z}\boldsymbol{\mu} - \mathbf{X}\mathbf{b})' (\mathbf{h} - \mathbf{Z}\boldsymbol{\mu} - \mathbf{X}\mathbf{b})\ \chi_n^{-2}, \quad (4)$$

where χ_n^{-2} stands for an inverted chi-squared distribution with n degrees of freedom, the number of band observations. It is not necessarily implicit in model (2) that the variance σ^2 is the same within both dominant and heterozygote individuals. In fact, it has been commonly observed that the variance within heterozygous individuals is smaller than within the dominant homozygote (Jansen *et al.*, 2001). A heteroskesdastic model can be accommodated and the sampling should be carried out from

$$p(\sigma_1^2 \mid \mathbf{Z}_1, \mathbf{X}, \mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\mu}) = (\mathbf{h} - \mathbf{Z}_1\boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{h} - \mathbf{Z}_1\boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta})\ \chi_{n1}^{-2}$$

$$p(\sigma_2^2 \mid \mathbf{Z}_2, \mathbf{X}, \mathbf{h}, \boldsymbol{\beta}, \boldsymbol{\mu}) = (\mathbf{h} - \mathbf{Z}_2\boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{h} - \mathbf{Z}_2\boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta})\ \chi_{n2}^{-2},$$

with σ_k^2 being the error variance within genotype k, and $n_k = \mathbf{Z}_k' \mathbf{Z}_k$. This strategy avoids the need for a square root transformation as proposed by Jansen *et al.* (2001) to allow for heterogeneous variances within genotypic classes.

Finally, the third term in (1), $[c]$, is the prior probability of the genotypes. They should be supplied by the user; in some instances they will be known, as in crosses between inbred lines, or they can be estimated from the data assuming Hardy–Weinberg frequencies, i.e. p(R) can be set equal to the square root of the frequency of recessive individuals identified. It can be shown that this term vanishes for non-founder individuals, as the sampling is conditional on parents' genotypes.

Once the ordered genotypes ($\mathbf{G}$) are obtained, crossover locations are generated according to flanking genotype information as described previously (Pérez-Enciso *et al.*, 2000). In short, first it is identified whether the number of recombination events is odd or even. An even number of crossovers (including zero) must have occurred between consecutive informative markers if both alleles in the offspring have the same paternal haplotype origin, and an odd number of crossovers must have occurred if each allele in the offspring comes from different parental haplotypes, i.e. at least one crossover must have occurred during the meiosis. In the absence of interference, as assumed throughout this study, the number of recombinants follows a censored Poisson distribution. The location of the crossovers is assigned at random within the appropriate marker interval. Subsequently, the IBD state between all individuals is assessed at any desired number of genome positions. The method iterates and the IBD probabilities are computed as the frequency

https://doi.org/10.1017/S0016672302005645 Published online by Cambridge University Press

of the IBD states over replicates. In practice, the IBD states at the predefined genome positions are added up by the program in a matrix of dimension $n \times n \times$ number of positions; once the MCMC chain ends, the IBD probabilities are obtained dividing by the number of iterations.

In summary, once all variables have been initialized a complete MCMC round consists of sampling G_{ij} for all i and j using (1), sampling $\boldsymbol{\mu}$, $\boldsymbol{\beta}$ and σ^2 from (3) and (4) for those markers whose band intensities are considered, and finally sampling crossover locations and computing IBD status.

(ii) *QTL mapping*

A systematic assessment of the usefulness of AFLP markers for QTL detection is beyond the scope of this paper, but we consider it useful at least to compare in a limited setting the performance of AFLP versus SNP and microsatellite mapping. Computing the IBD probabilities is the cornerstone of any QTL mapping procedure. Once these probabilities are obtained, a number of QTL mapping strategies can be followed, depending on the genetic material available or on the statistical procedure of choice. For the outbred data set simulated (see below), it seems appropriate to model the founder QTL effects as random effects. Suppose the following linear model for a quantitative trait $\mathbf{y}$ applies,

$$\mathbf{y} = \mathbf{Wb} + \mathbf{q} + \mathbf{u} + \mathbf{e},$$

where $\mathbf{b}$ is the vector containing the fixed effects affecting the trait, $\mathbf{q}$ is a vector containing the random QTL effects (dimension n), $\mathbf{u}$ is an infinitesimal genetic effect, $\mathbf{e}$ is vector with residuals, and $\mathbf{W}$ is an incidence matrix. Further assume $\mathbf{q} \sim \mathrm{N}(0, \mathbf{Q}\sigma_q^2)$, $\mathbf{u} \sim \mathrm{N}(0, \mathbf{A}\sigma_u^2)$, and $\mathbf{e} \sim \mathrm{N}(0, \sigma_e^2)$, where $\mathbf{Q}$ is the $n \times n$ matrix containing the IBD probabilities at the QTL position, $\mathbf{A}$ is the usual additive relationship matrix, $\mathbf{I}$ is the identity matrix, σ_q^2 is the QTL variance, σ_u^2 is the infinitesimal genetic variance and σ_e^2 is the environmental variance.

Bink *et al.* (1998) used a Bayesian analysis conditional on the IBD matrix $\mathbf{Q}$ to estimate σ_q^2. Here we have preferred to include $\mathbf{Q}$ in the IBD likelihood approach described by Goldgar (1990) and as implemented by Pérez-Enciso & Varona (2000). The frequentist strategy provides an estimate of the QTL position at a reasonable computing cost by fitting and maximizing the likelihood every few centimorgans. In contrast, the estimation of the QTL position by the IBD Bayesian approach requires the computation of the posterior probability of the QTL position (Sillanpaa *et al.*, 1998; Bink *et al.*, 2000) or that the Bayes factor be obtained at each position analysed. Any of these options is far more computer-intensive than the likelihood profile. Varona *et al.* (2001, their Fig. 5) showed that there was a very good correlation between the Bayes factor profile and the likelihood ratio profile along the chromosome. The likelihood maximized here was

$$L = -1/2\ [\mathrm{Constant} + \log|\mathbf{V}| + (\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb})], \quad (5)$$

where $\mathbf{V} = \mathbf{Q}\sigma_q^2 + \mathbf{A}\sigma_u^2 + \mathbf{I}\sigma_e^2$. A simplex algorithm was employed for maximization; $\mathbf{Q}$ consisted of the means of the posterior distributions of the relationship coefficients between individuals. Note that this is a two-step approach: first we obtain $\mathbf{Q}$, and then we use $\mathbf{Q}$ in the linearized likelihood (5). We have found that this approach is quite robust in a variety of settings but more sophisticated and theoretically sound methods exist, where the IBD state and the QTL position are estimated jointly (e.g. Yi & Xu, 2000).

Further, an F_2 cross between inbred lines was also simulated (see below). In this instance, the IBD probabilities of the F_2 or backcross individuals can be utilized in a regression-type strategy (Haley & Knott, 1992; see Satagopan *et al.*, 1996 for a Bayesian approach). Here we used the model

$$\mathbf{y} = \mathbf{Wb} + \mathbf{c}\ a + \mathbf{e},$$

where $\mathbf{c}$ is a vector containing the probabilities, for each individual, of having received both alleles from one line minus the probability of having received both alleles from the other line, and a is the additive allelic effect. As before, we computed the likelihood profile of the model including the QTL over the model without the QTL effect.

(iii) *Simulated data*

We simulated two populations. The first, outbred, population consisted of 50 full sib families, each family consisting of 10 full sibs. The second population was an F_2 cross between divergent lines, with 10 male and 20 female founders and 500 F_2 individuals in total. In both populations, a single chromosome of 60 cM was considered. Several marker configurations were analysed in the outbred population: (0) 4 microsatellites located every 20 cM; (1) 4 AFLPs at identical positions; (2) 4 SNPs at identical positions; (3) 13 AFLPs spaced every 5 cM; and (4) 13 SNPs at the same positions. Only configurations 0, 3 and 4 were studied in the F_2 cross. SNPs provide the maximum information that can ever be conveyed by AFLPs and thus provide a useful benchmark to test AFLP performance. In order to interpret the comparisons in a more straightforward manner, we first simulated a data set with 13 microsatellites, 6 alleles each. The allele frequencies were equal in the outbred population. In the F_2 cross, two extreme situations

https://doi.org/10.1017/S0016672302005645 Published online by Cambridge University Press

were considered: either the allele frequencies were the same in both parental lines, or they were fixed for alternative alleles. The data sets with the 13 microsatellites were not analysed; they were used for the sole purpose of generating the corresponding marker files. Microsatellite alleles coded as '1' through '3' ('4' through '6') were collapsed into SNP allele '1' ('2'), and SNP genotypes '21' and '22' were collapsed into AFLP genotype '20', where '2' is the dominant allele and '0' stands for missing allele. Data files for the first three cases, those with only 4 markers, were obtained by removing the markers not considered. Finally, we simulated a 'peak height' file that contained the band intensities for each individual and marker. The band intensities were distributed as N(10, 1) and N(12, 1) for heterozygous and dominant homozygous genotypes, respectively. The only effect used to simulate band intensity was the genotype, i.e. we used the underlying model (2): $\mathbf{h}_j = \mathbf{Z}_j\,\boldsymbol{\mu}_j + \boldsymbol{\varepsilon}_j$. The same distribution was used for all markers. A normal mixture is unimodal when the difference between means of the two component distributions is less than or equal to 2 standard deviations, approximately (Titterington *et al*., 1985). Some authors (Piepho & Koch, 2000; Jansen *et al*., 2001) have studied the case where the band intensity distribution was clearly multimodal, making genotype classification relatively easy. Our own experience (see discussion below and Fig. 3) is much less optimistic in that respect, and thus we used here a conservative guess for the underlying band intensity distributions.

We studied the distribution of the relationship coefficient between full sib pairs in the first, outbred, population for each data set at position 12 cM. If we knew the IBD state with certainty, the relationship coefficient between full sibs would be distributed with mean 0·5 and variance 0·125 if the parents are not inbred and unrelated. When the IBD state is not known, the mean should be very similar but the variance will decrease, as the method 'regresses' the parameter towards the mean, here 0·5. Thus, in general terms, the larger the variance, the higher the power of the method to discriminate between alternative IBD states. The MCMC chain was run 10 times for each of 2000 iterations; a total of 20000 iterations was thus employed. Previous work has shown that the results did not change significantly with larger chains.

Finally, we also simulated a phenotypic data set using the complete marker information. The trait was determined by a single QTL at position 12 cM. The founder allelic effects were sampled from a normal distribution N(0, 0·25) in the outbred population. In the F_2 cross, the parental lines had alternative QTL alleles fixed and the allelic genetic effect was 0·5. In either population, the genic action was additive, and there was no polygenic effect. The performance record was generated adding the two transmitted allelic effects plus a random normal environmental deviate N(0, 1). The likelihood was fitted every 2 cM, and the ratio versus the likelihood maximized under a model without the QTL effect was obtained at each position.

(iv) *Real data*

In this work, we have used real data (quail) only to illustrate the variety of band density distributions that can be obtained. AFLP markers were generated by α*Taq*I/*Eco*RI (New England Biolabs) restriction of quail genomic DNA, followed by adapters ligation using T4 DNA ligase (Appligene). Two selective amplification steps with *Taq* polymerase (Gibco BRL) were then performed on GeneAmp 9700 (Perkin-Elmer) thermocyclers. Samples were run on DNA Analyzer 3700 (Perkin-Elmer) and patterns analysed with Genescan 3·5 and Genotyper 3·6 softwares (Abi Prism, Perkin-Elmer).

Results and discussion

The main results on the performance of the method proposed are in Figs. 1 and 2. Fig. 1 is a plot of the relationship coefficient (ρ) between all possible pairs of full sibs in the simulated outbred population. The expected profile when the IBD is known is a discrete distribution with points 0, 0·5 and 1 with frequencies

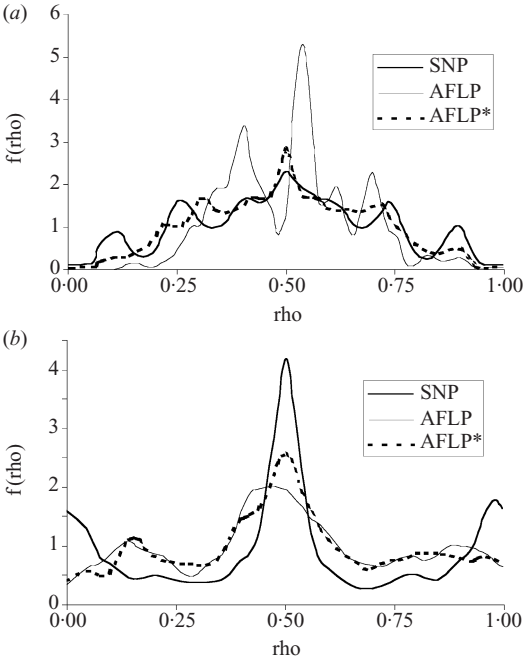

Fig. 1. Distribution of the mean relationship coefficients between all possible pairs of full sibs at position 12 cM in the outbred population. The graph compares the results with SNPs (thick line), AFLPs (thin line), AFLPs and using the intensity band information (dashed line). (*a*) Four markers at positions 0, 20, 40 and 60 cM; (*b*) 13 markers every 5 cM.

https://doi.org/10.1017/S0016672302005645 Published online by Cambridge University Press

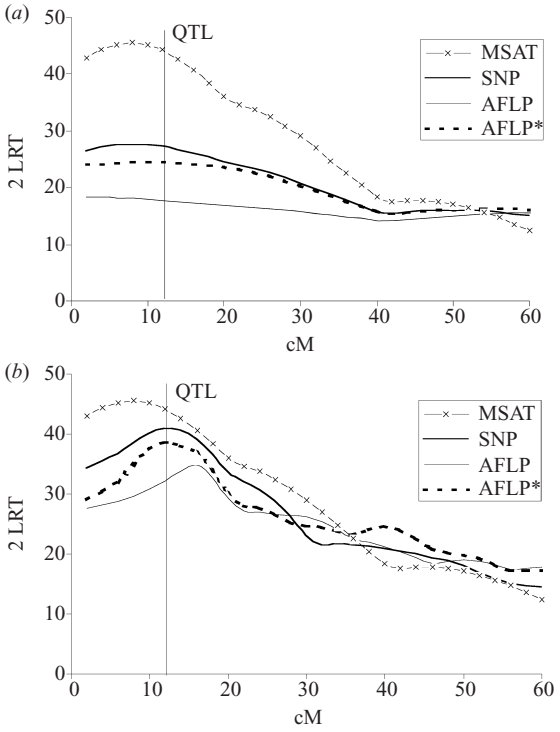

Fig. 2. Outbred population: likelihood ratio profile at 2 cM steps with different marker types – microsatellites (crossed line), SNPs (thick line), AFLPs (thin line), and AFLPs and using the intensity band information (dashed line). (*a*) Markers located every 20 cM; (*b*) markers located every 5 cM. In both (*a*) and (*b*) the microsatellite profile corresponds to a 20 cM spacing. The vertical bar shows the QTL position (12 cM).

0·25, 0·5 and 0·25 respectively. However, when there is uncertainty the relationship coefficient between two individuals is itself a random variable, with its corresponding posterior distribution. We used the mean of the posterior distribution in Fig. 1. It can be seen that a sparse map of AFLPs provides little discrimination about the IBD status between individuals if band density is not utilized. This is illustrated by the low variance of the relationship coefficients (Table 1), a consequence of the fact that a large proportion of the values are centred around the prior value, 0·50 (Fig. 1). In contrast, using band intensities results in a profile similar to that of SNPs. The ragged profile of the distributions is due to the different recombination events and the distance between the point at which the relationship coefficient is estimated and the closest markers (for a discussion see Pérez-Enciso *et al.*, 2000). When marker density is higher the influence of using band intensities decreases (Fig. 1*b*, Table 1). This is because adjacent markers provide more information as the distance between them decreases.

The ability of the method to discriminate between dominant homozygotes and heterozygotes depends critically on whether the mean band intensities (μ_1 and μ_2) and the band error variance (σ^2) are correctly estimated. Table 2 presents the main parameters of the corresponding posterior distributions. As expected, the parameter estimates are better at a high marker density, although the estimates are reasonable in the two scenarios studied. A more accurate estimation of the individual band intensity can be obtained by reamplifying each sample and using the mean instead of a single measurement. This strategy should also provide a better estimate of the error variance and help to assess to what extent genetic heterogeneity exists across genotypes or experimental protocol factors. In practice, however, this is a realistic option only for a limited number of markers in, say, more detailed mapping after a first genome scan.

Certainly, the most relevant issue for evaluating the method proposed is its performance for QTL mapping. A 2 cM scan was performed using the IBD approach described above and the results are shown in Fig. 2 for the two marker spacings studied in the outbred population. Note that the likelihood ratio would be significant in all cases, assuming that the statistic is distributed as a mixture of chi-squared distributions under the null hypothesis, $1/2\chi^2_0+1/2\chi^2_1$. In consequence, all approaches would be powerful enough to identify that a QTL is located in the chromosome. Of course the power is maximum with highly informative markers (microsatellites) and minimum with dominant markers when the band intensity information is not used. Marker informativity and dominance has, nevertheless, a dramatic impact on the error of the QTL position estimate at low marker density (Fig.

Table 1. *Distribution statistics of the relationship coefficient between pairs of full sibs*

	20 cM		5 cM	
Marker spacing	Mean	Variance	Mean	Variance
Microsatellite	0·501	0·080	–	–
SNP	0·501	0·044	0·503	0·094
AFLP	0·502	0·022	0·507	0·053
AFLP[a]	0·502	0·036	0·506	0·065

[a] Using band intensity information.

https://doi.org/10.1017/S0016672302005645 Published online by Cambridge University Press

Table 2. *Main statistics for the posterior distribution of the mean intensity differences between genotypes,* $f(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|\mathbf{M}, \mathbf{h})$, *and the error variance* $f(\sigma^2|\mathbf{M}, \mathbf{h})$. *The true values were* $\mu_1-\mu_2 = 2$, *and* $\sigma^2 = 1$. *For shortness, only the average statistics of all 4 and 13 markers is presented*

	$f(\mu_1-\mu_2\|\mathbf{M}, \mathbf{h})$		$f(\sigma^2\|\mathbf{M}, \mathbf{h})$	
No. of markers	Mean	Standard deviation	Mean	Standard deviation
4	2·53	0·25	0·72	0·24
13	1·80	0·39	1·02	0·36

1*a*). If we use the coarse rule of thumb that a 95% confidence interval corresponds to a drop in likelihood ratio of about 4 units, the widths of the confidence interval are 14, 20, 28 and 36 cM when we employ microsatellites, SNPs, AFLPs using band intensities, and AFLPs, respectively. Note that these figures are used only for the sake of comparison between methods, and are not intended to be the exact confidence intervals. Rather we argue that the QTL position confidence interval is proportional to the curvature of the likelihood ratio profile around the maxima.

The picture changes dramatically when we compare the marker performance at a higher density for the diallelic markers. Fig. 2*b* shows the results when AFLPs and SNPS are located every 5 cM, where the microsatellite plot is repeated for completeness. It is clear that power increases and that the confidence interval is reduced compared with Fig. 2*a*, but the most important result is that the plots corresponding to SNPs and AFLPs when band intensity is used are very similar. This strongly suggests that AFLPs can have an almost additive behaviour at moderate to high marker densities if the information from band intensities is utilized. This is an important result, as the difference in the cost of developing AFLPs versus SNPs or microsatellites is several orders of magnitude.

As a general guideline, it can be conjectured that only QTLs of moderate to large effect will be detected using a sparse map of dominant markers, even if band density is employed. For those QTLs actually identified, the user should be aware that the position will be estimated very loosely. In contrast, as marker density increases and provided that band intensity can be taken into account, the performance of AFLPs can be remarkably similar to that of additive markers. In relative terms, the usefulness of including the band intensity is larger when markers are sparsely located. Nevertheless, using band intensity always results in a sharper QTL position irrespective of the marker spacing (Fig. 2). A thorough assessment of the power for QTL mapping at different AFLP spacing and QTL effects remains to be done.

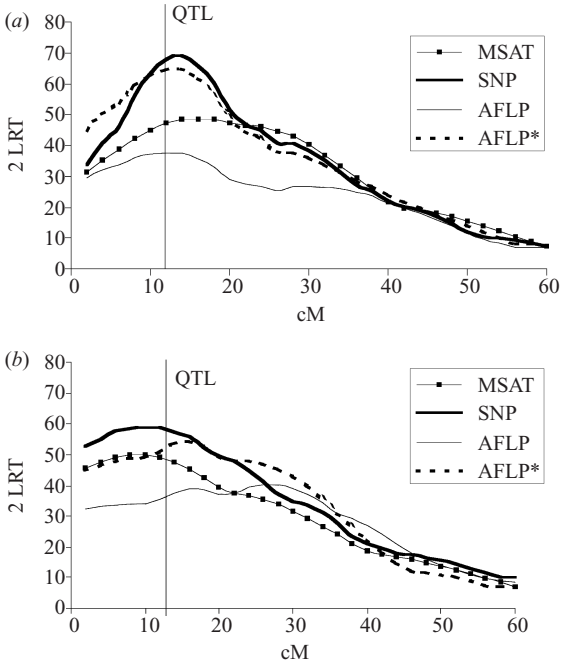

Fig. 3. F_2 cross population: likelihood ratio profile at 2 cM steps with different marker types – microsatellites (crossed line), SNPs (thick line), AFLPs (thin line), and AFLPs and using the intensity band information (dashed line). (*a*) The two parental lines do not share any common allele; (*b*) the two lines have the same allele frequencies for every marker. The vertical bar shows the QTL position (12 cM). The microsatellites are spaced every 20 cM, the remaining markers, every 5 cM.

The results for the F_2 data are in Fig. 3. Microsatellites were spaced every 20 cM, and SNPs and AFLPs every 5 cM. Here we studied the effect of differences in allele frequencies between lines on QTL detection. Fig. 3*a* shows the results in the ideal situation, when all markers are fixed for alternative alleles in each breed, whereas Fig. 3*b* represents the least favourable case: no difference in allele frequencies between breeds. A real data set will lie somewhere in between. Recall that in both cases the two lines did have different QTL alleles fixed. As expected, the likelihood ratios were higher when markers were completely informative with respect to line origin

https://doi.org/10.1017/S0016672302005645 Published online by Cambridge University Press

(Fig. 3*a*) than when they were not (Fig. 3*b*). But it is more interesting to observe that the AFLP profile using band densities is almost indistinguishable from the SNP profile when marker alleles are fixed (Fig. 3*a*) and they are clearly more 'peaked' around the QTL than the microsatellite profile, suggesting a more accurate QTL localization. This is because, in a regression-type analysis, we are only interested in the breed origin of each allele; if the marker alleles are fixed, it does not matter whether there are two or more alleles, so the better performance is due exclusively to using a more dense map with AFLPs than with microsatellites. The important result here is that, as in Fig. 2, the AFLPs using band densities and SNPs had a very similar behaviour. The most unfavourable case is shown in Fig. 3*b*; here the higher the number of alleles, the better the ability to always distinguish the allele origin. Nevertheless, the maximum likelihood ratio with AFLPs using band density was higher than that with microsatellites, suggesting at least a similar power. A more thorough study would be required to assess this. It can be seen that, in all cases, the performance of AFLP markers diminishes dramatically if band intensity is not taken into account. In agreement with the results in Fig. 2, the curves are very flat as a sign that the QTL will be located rather loosely.

Despite the evident advantages of the method presented here, the user should also be aware of the potential dangers of MCMC methods. The most common one is that the chain gets 'stuck' in a set of genotypic configurations such that the whole space of possible genotypes is not sampled. For instance, if two parents are sampled to be homozygous, there is only one possible genotype of offspring, irrespective of other available information. In this case it is said that the chain is 'reducible'. This risk increases when there are missing genotypes and when a single genotype is updated at a time, which was the strategy followed here. In order to minimize the risk of reducibility, a simple procedure is to run separate chains with different starting points and use the results from all chains. Diagnosing convergence can also be a difficult issue and a wide number of approaches are available (Cowles & Carlin, 1996). Here we ran 10 chains of 2000 iterations each. We did not find any noticeable difference in terms of the distribution of IBD probabilities compared with when we ran a smaller number of restarts. We also computed the autocorrelation in IBD probabilities every successive iterate for two full sibs. The values varied from 0·01 to 0·15 for most of the full sib pairs studied, suggesting that the effective number of iterations was close to the actual number of iterations performed. Nevertheless, complex pedigrees with a significant number of missing genotypes require more sophisticated genotype strategies than the single update used here. The reader is referred to, for example, Sobel & Lange (1996) or Heath (1997) for alternative sampling schemes.

Here we have assumed that the genetic map was known without error for either codominant or AFLP markers. It can be argued that this is unfair since maps with codominant markers will be more accurate than AFLP maps and, as a result, QTL studies will be necessarily more accurate with codominant markers. In practice, however, AFLP maps can also be obtained with reasonable accuracy in a QTL mapping population, as the number of individuals required to construct a genetic map is much smaller than the number required for localizing QTLs. In fact, a single family may suffice if the recombinants can be identified. The key parameter to control is the marker order, rather than the precision in the recombinant fraction itself; thus a series of markers can be discarded if the lod score between different orders is very similar. Given the high polymorphism uncovered by the AFLP system, it is almost guaranteed that a reasonable coverage of each linkage group will be attained. In addition, some of the AFLP markers will have an almost codominant behaviour (e.g. like those in the bottom row of Fig. 4, described below). These AFLPs will serve as 'anchors' for the remaining markers. In our specific experiment, the CRIMAP software is being used (Green *et al*., 1990) to build a quail AFLP map (Roussot *et al*., in preparation).

The approach presented here improves upon previous work aimed at rendering codominant the AFLPs markers (Piepho & Koch, 2000; Jansen *et al*., 2001). The main, but fundamental, advantage of our work is that all information is taken into account simultaneously as shown in equation (1). The probability of an individual's genotype is modified not only by its band intensity as in Piepho & Koch (2000) or in Jansen *et al*. (2001) but also by the adjacent markers of the individual and its relatives. As shown in Fig. 2, there can be an important gain for the practical purpose of QTL mapping when closely linked markers are considered. Furthermore, by using an MCMC Bayesian approach, the uncertainty about all parameters is considered jointly and the misclassification error is built into the model. Here there is no need to classify genotypes, as Jansen *et al*. (2001) suggested, in categories like 'no dominant homozygote' or 'no recessive homozygote'. Nevertheless, these can be incorporated into the model by constraining the sample space to a subset of genotypes without much change in the method. In some cases, the improvement by using pedigree information may be dramatic. For instance, consider the case that two parents with dominant genotype produce a recessive genotype offspring; it follows immediately that both parents are necessarily heterozygotes. This allows, in turn, parameters σ^2 and μ in (3) and (4) to be accurately estimated. If, in contrast, a family produces only

https://doi.org/10.1017/S0016672302005645 Published online by Cambridge University Press

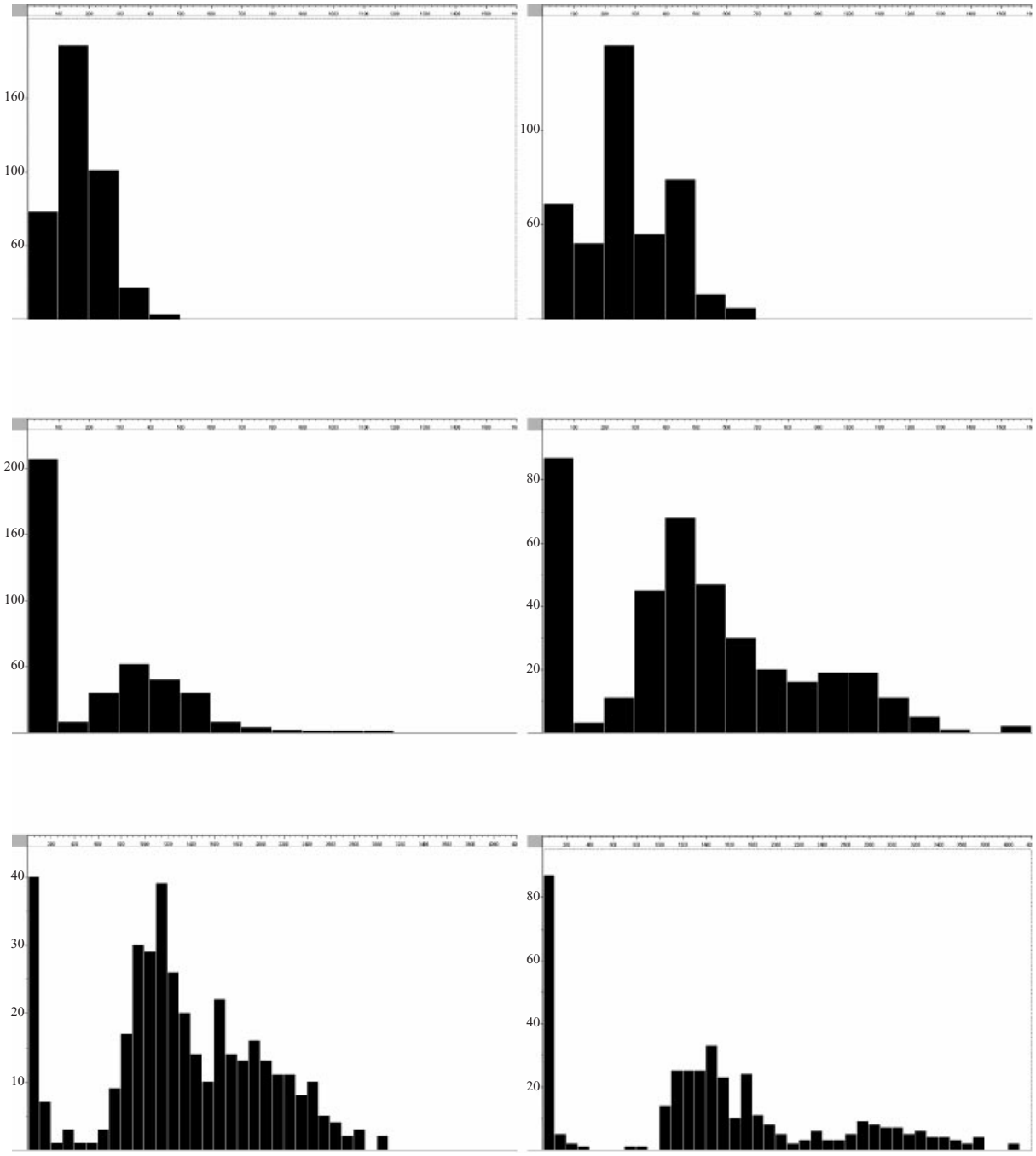

Fig. 4. Representative examples of AFLP marker density distributions for six markers in quail, with 360 individuals. Markers are generated by α*Taq*I/*Eco*RI restriction of quail genomic DNA, followed by adapters ligation and two selective amplification steps with various primer combinations. Patterns are analysed with Genotyper 3.6 software, and the marker density distribution is displayed in the histogram window of Genotyper 3.6. The *x*-axis (shown along the top) represents peak density and the *y*-axis, the number of individuals. The scale on the *x*-axis varies because of different peak densities found according to each marker.

offspring with the dominant genotype, at least one of the parents is likely to be homozygote dominant. The exact probabilities are taken into account automatically in our method via the algorithm described in the Appendix.

The performance of the approaches of Piehpo & Koch (2000) and of Jansen *et al.* (2001) depend heavily on the properties of a multimodal distribution, which clearly facilitates genotype classification by simply inspecting the distribution. But multimodality is unlikely to hold for all AFLP markers. Fig. 4 illustrates this point by showing the distribution of six quail AFLP markers. These graphs are representative of most of the distribution shapes that we observed and they provide clues about further statistical refinements. The first feature worth noticing is the

https://doi.org/10.1017/S0016672302005645 Published online by Cambridge University Press

wide variety of distributions; in view of this variability the user should check for each marker whether the normal mixture is adequate or whether a certain transformation is warranted. In this work we have retained the simple mixture of normals but, provided that other distributions are deemed more appropriate, these could be incorporated into the model and the sampling distributions (3) and (4) should be changed accordingly. If a no standard distribution is obtained, one needs to resort to Metropolis-Hastings sampling, instead of the more straightforward Gibbs sampling employed here. Piepho & Koch (2000), for instance, explored the Box-Cox transformations. Fig. 4 indicates that modelling AFLP band intensities is indeed a challenging research topic.

Further notice in Fig. 4 that for a number of markers there is a continuum between absence and presence of a band for a number of markers (top two figures). As a result, it may be appropriate to modify our model to account also for uncertainty in scoring homozygote recessive genotypes, as also suggested by Jansen *et al*. (2001). In a three-mixture model, some of the genotypes with a low peak density could in fact be recessive homozygotes or heterozygotes. The matrix **Z** in (2) now becomes a three-column matrix and the vector **h** will contain three parameters, corresponding to the three density means for each genotype. Nevertheless, most of the markers did exhibit a clear distinct peak at no band amplification, as shown in the middle and bottom rows of Fig. 4. The graph in the left middle row may correspond to the case simulated here, i.e. a unimodal distribution. The remaining middle right and bottom row graphs show different multimodal distributions. Only about 10% of the markers assayed did exhibit clear distinct peaks and can be considered to behave as almost codominant markers (e.g. bottom right graph). In an additional 8% of the markers, the distribution looks multimodal but genotype classes overlapped so that genotype assignment cannot be done unambiguously by eye (e.g. middle right graph).

In conclusion, we have shown that using the band intensity together with all pedigree and marker information can decrease most of the disadvantages of dominant markers, and that their behaviour can be very much like that of additive biallelic markers (SNPs), at least for the practical purpose of QTL mapping. In practice, there exists a mixture of marker types, say both microsatellites and AFLPs, for most species. The theory developed allows us to combine all marker information in an optimum way. Band intensity information can be included or not, or included only partially if the user finds that some marker information is more reliable than others, depending on the goodness of fit of a mixture of normals to the real data (e.g. Fig. 4). Finally, other QTL mapping strategies can be envisaged. Hansen *et al*. (2001) argued that a strict linkage disequilibrium mapping can not be carried out with AFLPs due to its dominant nature. But the same principles outlined here can be employed to carry out proper linkage disequilibrium mapping with dominant markers. For instance, first the haplotypes can be inferred from pedigree information, and second a measure of disequilibrium, say D′, is computed between the markers and the trait. By iterating on this sampling scheme a MCMC estimate of D′ between dominant markers and a categorical trait can be obtained.

We thank Catherine Beaumont for encouragement to work on this topic and discussions, and both referees for their useful comments. O.R.'s PhD project is funded in part by Aventis Nutrition Animale; work was funded by 'Bureau des Ressources Génétiques' (France), project no. 20 (2001–2002).

Appendix. Pseudocode for genotype sampling

Consider that we are sampling a genotype at marker j from founder individual i of sex $k(=1, 2)$ and thus we condition on haplotype k of each of its n offspring. Take j' to be the first informative marker to the 'right' of marker j, i.e. the closest upward marker for which individual i is heterozygous. In the following we denote the genotype at locus j, individual i by G_{ij}, and the allele at the kth haplotype by G_{ijk}. The pseudocode for sampling the ijth genotype is as follows:

Initialize

$\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DD}) = \mathrm{p}(G_{ij} = \mathrm{DD})$
$\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DR}) = \mathrm{p}(G_{ij} = \mathrm{DR})$
$\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{RD}) = \mathrm{p}(G_{ij} = \mathrm{RD})$
$\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{RR}) = \mathrm{p}(G_{ij} = \mathrm{RR})$

For $o = 1, n$ {
  If $(G_{oj'k} = G_{ij'1})$ then $\Rightarrow$ allele $G_{i,j',1}$ transmitted
    if $(G_{ojk} = \mathrm{D})$ then
      $\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DD}) = \mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DD}) * 1$
      $\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DR}) = \mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DR}) * (1-r)$
      $\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{RD}) = \mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{RD}) * r$
      $\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{RR}) = 0$
    elseif $(G_{ojk} = \mathrm{R})$ then
      $\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DD}) = 0$
      $\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DR}) = \mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DR}) * r$
      $\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{RD}) = \mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{RD}) * (1-r)$
      $\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{RR}) = \mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{RR}) * 1$
    endif
  elseif $(G_{oj'\mathrm{k}} = G_{ij'2})$ then $\Rightarrow$ allele $G_{i,j',2}$ transmitted
    if $(G_{ojk} = \mathrm{D})$ then
      $\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DD}) = \mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DD}) * 1$
      $\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DR}) = \mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{DR}) * r$
      $\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{RD}) = \mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{RD}) * (1-r)$
      $\mathrm{p}(\mathbf{G}_{-} \mid G_{ij} = \mathrm{RR}) = 0$

https://doi.org/10.1017/S0016672302005645 Published online by Cambridge University Press

elseif (G_{ojk} = R) then

$p(\mathbf{G}_{-} | G_{ij} = DD) = 0$

$p(\mathbf{G}_{-} | G_{ij} = DR) = p(\mathbf{G}_{-} | G_{ij} = DR) * (1 - r)$

$p(\mathbf{G}_{-} | G_{ij} = RD) = p(\mathbf{G}_{-} | G_{ij} = RD) * r$

$p(\mathbf{G}_{-} | G_{ij} = RR) = p(\mathbf{G}_{-} | G_{ij} = RR) * 1$

endif

endif

}

where o is an offspring subindex, there are n offspring in total, and r is the recombination fraction between the marker sampled and the closest informative marker. The first initialization step is carried out for each marker independently, without regard to the linkage information. Information from the closest right marker is similarly included. Once all offspring for a given parent are processed, the prior and band information are combined via (1), the probabilities are standardized and the genotype is sampled. Offspring genotypes are subsequently sampled conditional on parents' genotypes. Of course, only compatible genotypes are considered and G_{ij} is sampled only if it can not be ascertained unambiguously, e.g. the probability $p(G_{ij} = RR)$ is not considered unless that genotype is missing. If a genotype is missing, it is sampled (simulated) conditionally on current genotypic configuration from parents and offspring. If an individual i has parents and offspring genotyped, its genotype probabilities are modified by both sources of information, i.e. $p(G_i | \mathbf{G}_{-}) \propto p(G_i | \mathbf{G}_{\text{parents}}) p(\mathbf{G}_{\text{offspring}} | G_i)$.

References

Ajmone-Marsan, P., Valentini, A., Cassandro, M., Vecchiotti-Antaldi, G., Bertoni, G. & Kuiper, M. (1997). AFLP markers for DNA fingerprinting in cattle. *Animal Genetics* **28**, 418–426.

Alonso-Blanco, C., Peeters, A. J., Koornneef, M., Lister, C., Dean, C., van den Bosch, N., Pot, J. & Kuiper, M. T. (1998). Development of an AFLP based linkage map of Ler, Col and Cvi *Arabidopsis thaliana* ecotypes and construction of a Ler/Cvi recombinant inbred line population. *Plant Journal* **14**, 259–271.

Bink, M. C. A. M., Janss, L. L. G. & Quaas, R. L. (2000). Markov chain Monte Carlo for mapping a quantitative trait locus in outbred populations. *Genetical Research* **75**, 231–241.

Bink, M. C. & Van Arendonk, J. A. (1998). Detection of quantitative trait loci in outbred populations with incomplete marker data. *Genetics* **151**, 409–420.

Cervera, M. T., Gusmao, J., Steenackers, M., Peleman, J., Storme, V., Vanden Broeck, A., Van Montagu, M. & Boerjan, W. (1996). Identification of AFLP molecular markers for resistance against *Melampsora larici-populina* in *Populus*. *Theoretical and Applied Genetics* **93**, 733–737.

Cho, Y. G., Mc Couch, S. R., Kuiper, M., Kang, M. R., Pot, J., Groenen, J. T. M. & Eun, M. Y. (1998). Integrated maps of AFLP, SSLP and RFLP markers using a recombinant inbred population of rice (*Oryza sativa* L.). *Theoretical and Applied Genetics* **97**, 370–380.

Cowles, M. K. & Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* **91**, 883–904.

Goldgar, D. E. (1990). Multiple point analysis of human quantitative genetic variation. *American Journal of Human Genetics* **47**, 957–967.

Green, P., Falls, K. & Crooks, S (1990). Documentation for CRIMAP. Unpublished mimeo. *http://biobase.dk/Embnetut/Crimap/*

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.

Hansen, M., Kraft, T., Ganestam, S., Säll, T. & Nilsson, N. O. (2001). Linkage disequilibrium mapping of the bolting gene in sea beet using AFLP markers. *Genetical Research* **77**, 61–66.

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

Herbergs, J., Siwek, M., Crooijmans, R. P., Van der Poel, J. J. & Groenen, M. A. (1999). Multicolour fluorescent detection and mapping of AFLP markers in chicken (*Gallus domesticus*). *Animal Genetics* **30**, 274–285.

Jansen, R. C., Geerlings, H., Van Oeveren, A. J. & Van Schaik, R. C. (2001). A comment on codominant scoring of AFLP markers. *Genetics* **158**, 925–926.

Jiang, C. & Zeng, Z. B. (1997). Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* **101**, 47–58.

Jin, H., Domier, L. L., Kolb, F. L. & Brown, C. M. (1998). Identification of quantitative trait loci for tolerance to barley yellow dwarf virus in oat. *Phytopathology* **88**, 410–405.

Klein, P. E., Klein, R. R., Cartinhour, S. W., Ulanch, P. E., Dong, J., Obert, J. A., Morishige, D. T., Schlueter, S. D., Childs, K. L., Ale, M. & Mullet, J. E. (2000). A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. *Genome Research* **10**, 789–807.

Knorr, C., Cheng, H. H. & Dodgson, J. B. (1999). Application of AFLP markers to genome mapping in poultry. *Animal Genetics* **30**, 28–35.

Kuiper, M. T. (1998). Building a high-density genetic map using the AFLP technology. *Methods in Molecular Biology* **82**, 157–171.

Lu, Z. X., Sosinski, B., Reighard, G. L., Baird, W. V. & Abbott, A. G. (1998). Construction of a genetic linkage map and identification of AFLP markers for resistance to root-knot nematodes in peach rootstocks. *Genome* **41**, 199–207.

Mueller, U. G. & Wolfenbarger, L. L. (1999). AFLP genotyping and fingerprinting. *Trends in Ecology and Evolution* **14**, 389–394.

Nandi, S., Subudhi, P. K., Senadhira, D., Manigbas, N. L., Sen-Mandi, S. & Huang, N. (1997). Mapping QTLs for submergence tolerance in rice by AFLP analysis and selective genotyping. *Molecular and General Genetics* **255**, 1–8.

Otsen, M., den Bieman, M., Kuiper, M. T., Pravenec, M., Kren, V., Kurtz, T. W., Jacob, H. J., Lankhorst, A. & van Zutphen, B. F. (1996). Use of AFLP markers for gene mapping and QTL detection in the rat. *Genomics* **37**, 289–294.

Pérez-Enciso, M. & Varona, L. (2000). Quantitative trait loci mapping in F2 crosses between outbred lines. *Genetics* **155**, 391–405.

Pérez-Enciso, M., Varona, L. & Rothschild, M. (2000).

Computation of identity by descent probabilities conditional on DNA markers via a Monte Carlo Markov Chain method. *Genetics*, *Selection*, *Evolution* **32**, 467–482.

Piepho, H. P. & Koch, G. (2000). Codominant analysis of banding data from a dominant marker system by normal mixtures. *Genetics* **155**, 1459–1468.

Satagopan, J. M., Yandell, B. S., Newton, M. A. & Osborn, T. C. (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144**, 805–816.

Sillanpaa, M. J. & Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**, 1373–1388.

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.

Titterington, D. M., Smith, A. F. M. & Markov, U. E. (1985). *Statistical Analysis of Finite Mixtures*. New York: Wiley.

Vanhaeringen, W. A., Denbieman, M., Gillissen, G. F., Lankhorst, A. E., Kuiper, M. T. R., Vanzutphen, L. F. M. & Vanlith, H. A. (2001). Mapping of a QTL for serum HDL cholesterol in the rabbit using AFLP technology. *Journal of Heredity* **92**, 322–326.

Varona, L., Garcia-Cortes, L. A. & Pérez-Enciso, M. (2001). Bayes factors for detection of quantitative trait loci. *Genetics*, *Selection*, *Evolution* **33**, 133–152.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M. *et al.* (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **23**, 4407–4414.

Vuylsteke, M., Mank, R., Antonise, R., Bastiaans, E., Senior, M. L., Stuber, C. W., Melchinger, A. E., Lübberstedt, T., Xia, X. C., Stan, P., Zabeau, M. & Kuiper, M. (1999). Two high-density AFLP(R) linkage maps of *Zea mays* L.: analysis of distribution of AFLP markers. *Theoretical and Applied Genetics* **99**, 921–935.

Wang, C. S., Rutledge, J. J. & Gianola, D. (1993). Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genetics*, *Selection*, *Evoution* **25**, 41–62.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M. & Lander, E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082.