A simple and rapid method for calculating identity-by-descent matrices using multiple markers

Ricardo Pong-Wong, Andrew George, John Woolliams, Chris Haley

Genet. Sel. Evol. 33 (2001) 453–471

Original article

A simple and rapid method for calculating identity-by-descent matrices using multiple markers

Ricardo PONG-WONG*, Andrew Winston GEORGE, John Arthur WOOLLIAMS, Chris Simon HALEY

Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, UK

(Received 1st September 2000; accepted 9 April 2001)

Abstract – A fast, partly recursive deterministic method for calculating Identity-by-Descent (IBD) probabilities was developed with the objective of using IBD in Quantitative Trait Locus (QTL) mapping. The method combined a recursive method for a single marker locus with a method to estimate IBD between sibs using multiple markers. Simulated data was used to compare the deterministic method developed in the present paper with a stochastic method (LOKI) for precision in estimating IBD probabilities and performance in the task of QTL detection with the variance component approach. This comparison was made in a variety of situations by varying family size and degree of polymorphism among marker loci. The following were observed for the deterministic method relative to MCMC: (i) it was an order of magnitude faster; (ii) its estimates of IBD probabilities were found to agree closely, even though it does not extract information when haplotypes are not known with certainty; (iii) the shape of the profile for the QTL test statistic as a function of location was similar, although the magnitude of the test statistic was slightly smaller; and (iv) the estimates of QTL variance was similar. It was concluded that the method proposed provided a rapid means of calculating the IBD matrix with only a small loss in precision, making it an attractive alternative to the use of stochastic MCMC methods. Furthermore, developments in marker technology providing denser maps would enhance the relative advantage of this method.

IBD / QTL mapping / genetic relationships / marker assisted selection

1. INTRODUCTION

Recently, variance component approaches have been suggested for use in the detection of quantitative trait loci (QTL) that affect continuous traits. In this approach, the QTL effect associated with each individual is considered as a random effect with a covariance structure proportional to the identity-by-descent (IBD) probability at the QTL position. The main advantages of

* Correspondence and reprints
E-mail: ricardo.pong-wong@bbsrc.ac.uk

this approach are its potential use for data from outbred populations with a complex pedigree structure and its robustness to violations in the assumptions of the underlying genetic model. Its implementation in QTL studies involves both estimating of breeding values due to an identified QTL (*e.g.* [4]) and mapping them using linked markers (*e.g.* [5–7, 13, 15]).

The corner-stone of this methodology lies in the estimation of the covariance matrix for the QTL. Several methods have been suggested in the literature. Wang *et al.* [14] presented a recursive approach to estimate IBD probabilities, but it accounts only for one marker. Almasy and Blangero [1] used a regression approach where the IBD status at the markers are used to calculate the IBD at a given locus. However, the regression coefficients used in the IBD calculation can become very difficult to estimate in a complex pedigree. Lately, MCMC methods are beginning to be used (*e.g.* [11]). The advantage of these methods is that they can be used in very complex pedigree structures. However, MCMC methods suffer from the fact that they can be very slow and, in some situations, convergence is difficult to diagnose and may not even be achieved.

In this study, a fast partly recursive deterministic method was developed to calculate IBD probabilities. This methodology was compared with a stochastic method (*i.e.* LOKI, [11]) to assess in estimating the IBD matrix and performance in the QTL detection task.

2. METHOD

2.1. Deterministic calculation of the IBD matrix

2.1.1. Assumptions and model

Let us assume a chromosome has N marker loci with known positions and recombination rates as expected from the Haldane mapping function. All individuals are assumed to have a known genotype at each marker locus, but their haplotype phases may not be known with absolute certainty for all markers.

2.1.2. Gametic IBD matrix (G)

The gametic IBD matrix (**G**) is a matrix which contains the IBD probabilities between the two gametes (the paternally and the maternally inherited alleles) of an individual with themselves and the gametes of all other individuals. The IBD probability between a pair of gametes is the probability of these alleles being the same gamete originating from a common ancestor in the base population.

*2.1.3. Construction of the gametic IBD matrix (***G***)*

The methodology used here for constructing the gametic IBD matrix is a mixture of a recursive algorithm for general pedigree structure proposed by Wang *et al.* [14] and a method to estimate IBD between sibs suggested by Knott and Haley [9]. Given a single marker locus, Wang *et al.* [14] showed that IBD value at a linked position on the genome can be estimated recursively from ancestors to descendants. Assessing the inheritance pattern at the marker locus, the probability of descent of a gamete from parent to offspring (PDQ (probability of descent); using the same notation as in [14]) is calculated for the position on the genome in question. The IBD probability between a gamete of an individual and an ancestor's gamete is, then, estimated as a function of the probability of descent and the IBD probability of an ancestor's gamete with the two gametes of the parent. However, the direct implementation of this approach using multiple marker loci requires knowledge of the haplotype phases for the closest informative marker bracket of each individual. If the true phases are uncertain, the IBD estimation should be integrated over all possible phases across the whole population. In the present paper, in order to avoid complicated calculation of the haplotype probabilities, the closest informative marker bracket which is known with absolute certainty is used.

Thus the IBD probability between the gamete of individual i inherited from parent x (A_i^x) and the gamete of an ancestor j inherited from parent y (A_j^y), conditional on the linked marker genotypes (**M**), is equal to:

$$P\left(\mathrm{A}_i^x \equiv \mathrm{A}_j^y|\mathbf{M}\right) = P\left(\mathrm{A}_j^y \equiv \mathrm{A}_x^{\mathrm{p}}|\mathbf{M}\right) * PDQ\left(\mathrm{A}_i^x \Leftarrow \mathrm{A}_x^{\mathrm{p}}|\mathbf{M}\right) + P\left(\mathrm{A}_j^y \equiv \mathrm{A}_x^{\mathrm{m}}|\mathbf{M}\right) * PDQ\left(\mathrm{A}_i^x \Leftarrow \mathrm{A}_x^{\mathrm{m}}|\mathbf{M}\right) \quad (1)$$

where $P\left(\mathrm{A}_j^y \equiv \mathrm{A}_x^{\mathrm{p}}|\mathbf{M}\right)$ and $P\left(\mathrm{A}_j^y \equiv \mathrm{A}_x^{\mathrm{m}}|\mathbf{M}\right)$ are the IBD probabilities between gamete A_j^y and the paternal ($\mathrm{A}_x^{\mathrm{p}}$) and maternal ($\mathrm{A}_x^{\mathrm{m}}$) gametes of parent x, respectively. $PDQ\left(\mathrm{A}_i^x \Leftarrow \mathrm{A}_x^{\mathrm{p}}|\mathbf{M}\right)$ and $PDQ\left(\mathrm{A}_i^x \Leftarrow \mathrm{A}_x^{\mathrm{m}}|\mathbf{M}\right)$ are the probability of gamete A_i^x of the individual i, being the same as gamete $\mathrm{A}_x^{\mathrm{p}}$ or $\mathrm{A}_x^{\mathrm{m}}$ of parent x, respectively. Following the same terminology of Wang *et al.* [14], these terms are referred to as the probability of descent from parent to offspring.

Note that the two gametes for each individual defined above are the paternal and the maternal inherited alleles, while this is slightly different in the definition of Wang *et al.* [14], in which they represent the alleles linked to the marker alleles regardless of which parents they were inherited from.

2.1.4. Probability of descent of the gamete (PDQ)

The *PDQ* of a gamete is the probability that the gamete (say A_i^x) of an individual inherited from one of its parents (say x) is either the parent's paternal (say $\mathrm{A}_x^{\mathrm{p}}$) or maternal (say $\mathrm{A}_x^{\mathrm{m}}$) gamete. When the parent is not inbred, the *PDQ*

Table I. Probability of gamete (A_i^x) from offspring i being the same as the paternal (A_x^p) or the maternal (A_x^m) gamete of parent x, given the nearest informative flanking marker haplotype inherited from parent x.

Marker descent [(a)]			
M1	M2	$PDQ\left(A_i^x \Leftarrow A_x^p \mid \mathbf{M}\right)$	$PDQ\left(A_i^x \Leftarrow A_x^m \mid \mathbf{M}\right)$
P	P	$(1-\theta_1)(1-\theta_2)/(1-\theta)$	$\theta_1\theta_2/(1-\theta)$
P	M	$(1-\theta_1)\theta_2/\theta$	$\theta_1(1-\theta_2)/\theta$
M	P	$\theta_1(1-\theta_2)/\theta$	$(1-\theta_1)\theta_2/\theta$
M	M	$\theta_1\theta_2/(1-\theta)$	$(1-\theta_1)(1-\theta_2)/(1-\theta)$
P	–	$(1-\theta_1)$	θ_1
M	–	θ_1	$(1-\theta_1)$
–	P	$(1-\theta_2)$	θ_2
–	M	θ_2	$(1-\theta_2)$
–	–	0.5	0.5

θ_1, θ_2, θ: recombination rate between the first marker and the QTL, the second marker and the QTL and between the two markers, respectively (assuming the Haldane mapping function).

[(a)] P: the individual inherited the paternal allele from the parent. M: inherited the maternal allele. –: uninformative marker.

is the same as the IBD between the individual's gamete and its parent's gamete. The probability of descent of a gamete is calculated conditional on the closest marker bracket inherited from the parent in question and on their distance from the position where IBD status is calculated. The *PDQ* given the inherited markers are presented in Table I.

2.1.5. Inference of marker haplotype phases without uncertainty

In this study, the inference of marker haplotype phase (*i.e.* knowledge of which parents they were inherited from) to determine which marker allele the individual inherited from each parent is carried out using only the marker genotype of the individual and its parents. By using this information only, the individual's marker haplotype phase can be inferred with absolute certainty in three different cases: (i) when the individual's marker genotype is homozygous, (ii) when one of the parents is a homozygote, and (iii) when the individual inherited an allele only present in one of the parents' marker genotypes. This process results in partial knowledge of the haplotype phase where the phases of some of the marker loci will be unknown. Additionally, this approach does not infer phases in the base population individuals as their parents and their genotypes

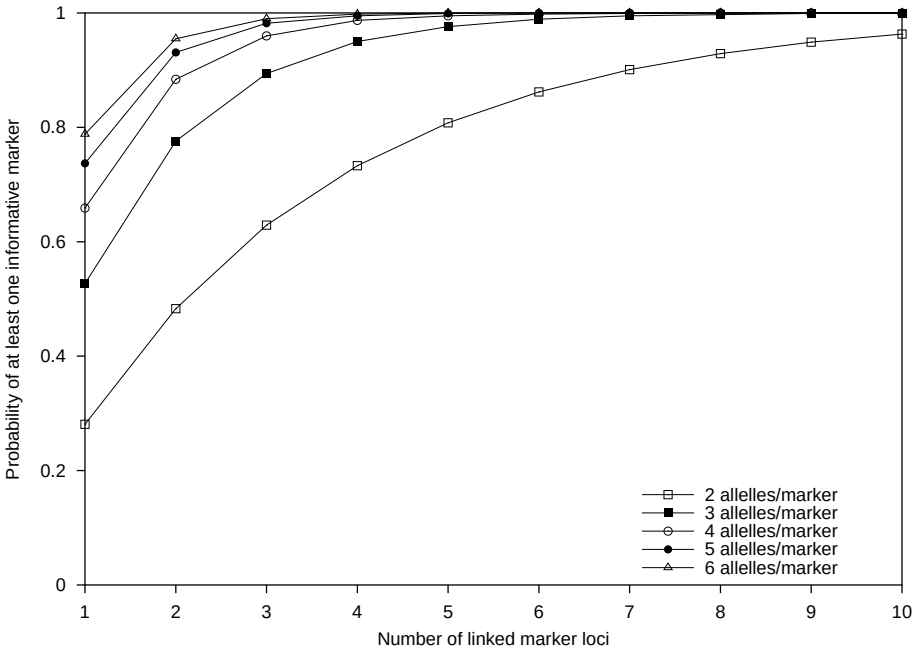

Figure 1. Proportion of individuals expected to have at least one informative marker locus to calculate their probability of descent (*PDQ*) assuming a different number of linked marker loci and a different number of alleles per locus at equal frequency.

are unknown. As phases cannot possibly be inferred with absolute certainty for all marker loci, the probability of descent is not necessarily estimated using the closest informative marker bracket. Thus the estimation of the IBD matrix is not, strictly speaking, calculated using all the information contained on the marker genotypes. Therefore, the IBD matrix calculated here is not expected to be the same as those obtained with stochastic and other exact methods where all information of the marker genotype is used in the calculation. Because the sub-optimal utilisation on the information on the markers results in a IBD matrix with less precision, it may be called an approximation.

Thus an informative marker locus which can be used to calculate the *PDQ* of one of the gametes of an individual is one for which the phase is known in both the offspring and the parent, and the parent is heterozygous for the locus in question. Figure 1 shows the expected proportion of individuals which should have at least one informative marker locus given the number of marker loci around the position of the genome where the IBD is estimated. Assuming equal gene frequencies for all alleles, three marker loci with three or more alleles is sufficient for 80% of all individuals to be expected to have an informative marker to estimate their *PDQ*.

When a marker genotype is missing for a given individual, the method does not attempt to reconstruct the genotype (unless the genotype can be inferred with absolute certainty given its parents' and offspring's genotypes). The missing genotype is, simply, assumed to be uninformative, and the next suitable marker locus is sought to calculate the *PDQ*. In the extreme case when the individual is completely untyped, the *PDQ* values are calculated to be 0.5 (the same as when all marker loci are completely uninformative for the individual in question). It is then important to notice that, although the handling of missing genotypes is operationally possible with the proposed method, any information on IBD which may exist in the missing genotype (due to inferences from relatives' genotype) is not used during the calculation. Hence, this method would not be recommended for pedigrees where a large proportion of the individuals are untyped, but more complicated approaches for retrieving information from missing marker genotypes may be the preferred method of choice.

2.1.6. Estimation of IBD between sibs

As the method of inferring haplotype phases used here cannot be used to infer those of the base individuals, the *PDQ* for their offspring cannot be estimated and the gamete is considered to have an equal probability of being the paternal or maternal gamete of the parent in question. In order to overcome this problem, the IBD probability between sibs whose common parent is a base individual was estimated using the method proposed by Knott and Haley [9]. This method does not require knowledge of the parental haplotype phase but only that of the sibs themselves. The IBD probability between sibs using the method of Knott and Haley [9] given the inherited marker allele is given in Table II. Figure 2 shows the expected proportion of sib pairs which would have at least one informative marker locus given the number of marker loci surrounding the position of the genome where the IBD is estimated. Assuming equal gene frequencies for all alleles, three marker loci with three or more alleles is sufficient to expect 80% of all sibs to have at least one informative marker to estimate their IBD. As the estimation of IBD proposed by Knott and Haley [9] does not use information about haplotype phases, it is expected to be less accurate than when using the recursive approach.

2.1.7. Protocol for estimating the gametic IBD matrix

The protocol to follow for constructing the gametic IBD matrix is as follows:

1. Reconstruct marker haplotype phases for all possible markers given the individual and its parents' marker genotypes.
2. Calculate the IBD recursively starting from gametes with the oldest ancestors to the youngest descendants:

Table II. Probability of identity by descent between sibs at the gamete inherited from the common parent which is not inbred [(a)].

IBD state at flanking markers		
M_1	M_2	IBD [(b)]
1 [(c)]	1	$\big((1-\theta_1)^2+\theta_1^2\big)\big((1-\theta_2)^2+\theta_2^2\big)/\big((1-\theta)^2+\theta^2\big)$
1	0	$\big((1-\theta_1)^2+\theta_1^2\big)\big((1-\theta_2)\theta_2\big)/\big((1-\theta)\theta\big)$
0	1	$\big((1-\theta_1)\theta_1\big)\big((1-\theta_2)^2+\theta_2^2\big)/\big((1-\theta)\theta\big)$
0	0	$4\big((1-\theta_1)\theta_1\big)\big((1-\theta_2)\theta_2\big)/\big((1-\theta)^2+\theta^2\big)$
1	– [(d)]	$\big((1-\theta_1)^2+\theta_1^2\big)$
0	–	$2\big((1-\theta_1)\theta_1\big)$

[(a)] from Knott and Haley [9]. [(b)] Formula is the IBD probability assuming the common parent is non-inbred (IBD_n). If the parent is inbred (*i.e.* the two gametes of the parent have non-zero IBD probability), the total IBD probability between the sibs' gametes is equal to $F + (1 - F)*IBD_n$, where F is the IBD between the parent's gametes and IBD_n the value as in the formula given in the above table. θ_1, θ_2, θ, recombination rate between the first marker and the QTL, the second marker and the QTL and between the two markers, respectively (assuming the Haldane mapping function.) [(c)] 1/0: both sibs inherited the same/different marker allele from the parent. [(d)] no informative marker found (the parent is a homozygote or inheritance in sibs is unknown.)

(a) the diagonal of the IBD matrix is 1 (*i.e.* a gamete is always 100% IBD with itself);
(b) if the individual is from the base population, its IBD probability between its gametes and previous individuals (ancestors) is zero (no calculation is done);
(c) if the individual is not from the base population:
   - calculate the probability of descent for each gamete (paternal and maternal) given the closest informative marker bracket with a known haplotype phase;
   - use formula (1) to calculate IBD probability between the paternal gamete and gametes of previous ancestors of which IBD probabilities with the individual's father has already been calculated. Repeat the same for the maternal gamete;
   - if IBD probability is to be calculated between two gametes that originated from a common parent (*i.e.* individuals are sibs) use the formulae given by Knott and Haley [9] including offspring of base animals for which *PDQ* cannot be estimated.

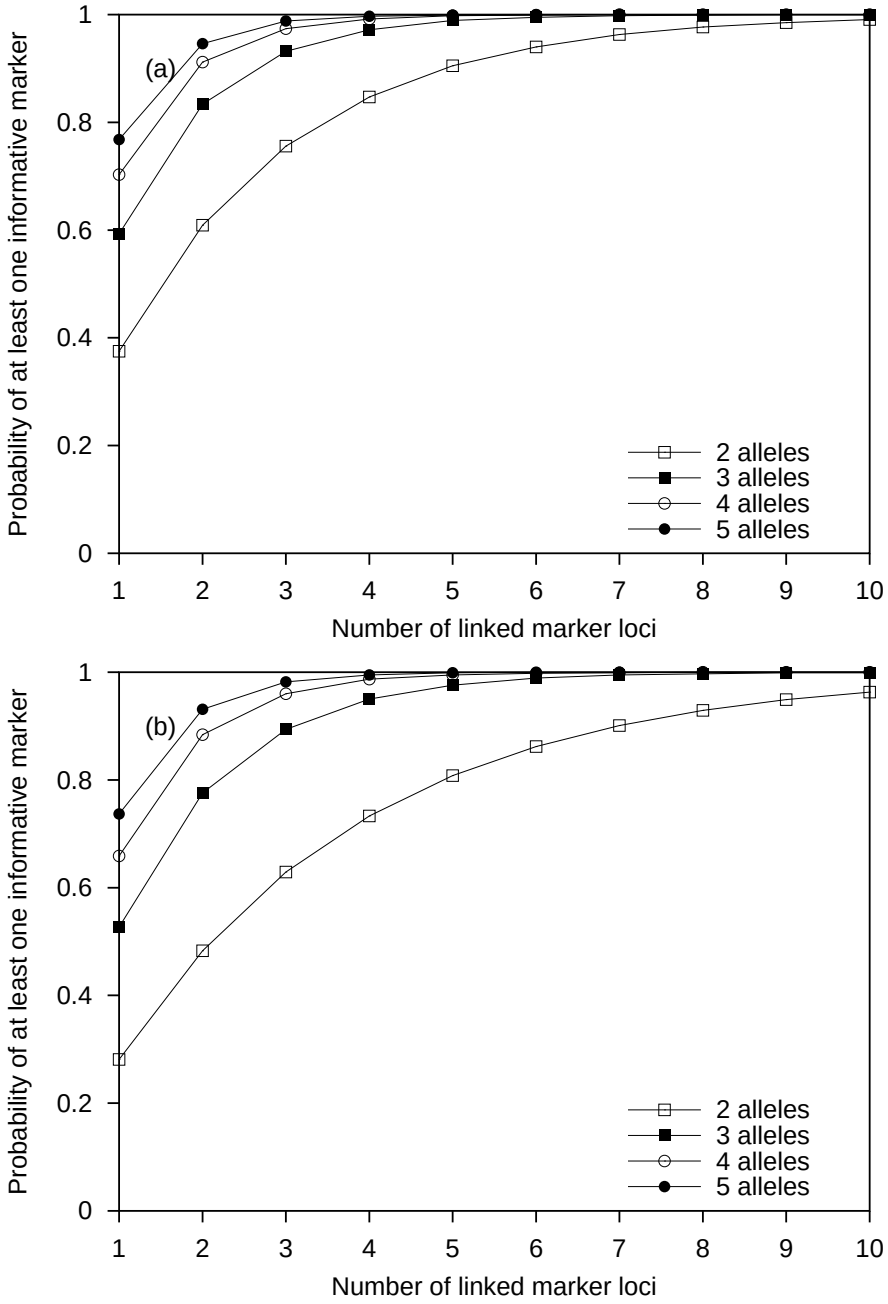

Figure 2. Proportion of full- (a) and half- (b) sib pairs expected to have at least one informative marker locus to calculate their IBD probabilities assuming a different number of linked marker loci and a different number of alleles per locus at equal frequency.

*2.1.8. IBD matrix at the Individual level (**Q**)*

The gametic IBD matrix (**G**) contains the IBD probabilities of both gametes of an individual independently. Thus the overall IBD status of an individual with others in the pedigree is represented by two different rows (one for each gamete) of **G**, meaning that its rank is twice the number of individuals. The overall IBD status of an individual may, however, be expressed by joining together the IBD of both gametes. The resulting matrix would be the IBD matrix at the individual level (**Q**) in which a single row represents the overall IBD status of an individual. The **Q** matrix is calculated by a linear transformation of the matrix **G** [12] equal to:

$$\mathbf{Q} = \frac{1}{2}\mathbf{K}\mathbf{G}\mathbf{K}' \tag{2}$$

where $\mathbf{K} = \mathbf{I}^*[1, 1]$, **I** is an identity matrix of equal rank as the number of individuals and * denotes the Kronecker product of two matrices.

Thus the IBD matrix (**Q**) is a matrix in which the element in row *i* and column *j* contains the overall IBD status between individuals *i* and *j*. Hence, the definition of IBD in matrix **Q** is not the probability of all gametes among two individuals being inherited from a common ancestor, but is twice the coefficient of coancestry among them. The elements of **Q** are therefore directly related to probabilities but are strictly not probabilities themselves as they can be greater than 1 for inbred individuals and are equal to 2 for completely inbred individuals. Note that the Wright numerator relationship matrix commonly used to model the polygenic effects (usually denoted **A**) is equivalent to the **Q** matrix with either no markers or completely uninformative markers.

2.2. Comparison with a stochastic method to calculate IBD

The value of the present method was assessed by comparing the resulting IBD matrix with a method which implements an MCMC approach to calculate the IBD between individuals. The comparison between these two methods was done in two different areas: (i) a direct comparison on the IBD values themselves; and (ii) a comparison on the effects of using these matrices for QTL mapping *via* a variance component approach.

2.2.1. MCMC method to estimate the IBD matrix

The MCMC method employed in this paper is LOKI [11]. LOKI is a freely available software package capable of calculating IBD probabilities between individuals given marker genotypes in complex pedigree structures. It is capable of handling substantial missing marker information. The package has been intensively tested with respect to QTL mapping by George *et al.* [5], hence our familiarity and preference for this package.

2.2.2. QTL mapping comparison

The objective of this section was to compare the results of the QTL mapping when using the IBD matrix obtained with the deterministic method and the IBD matrix obtained with the stochastic method. The methodology for mapping QTLs *via* the variance component approach used here has been described by George *et al.* [5].

A description for using the variance component approach in QTL mapping has been given elsewhere (*e.g.* [5–7, 13, 15]). Essentially, it consists of performing an interval mapping approach in which each position across the chromosome is tested for the presence of the QTL [10]. For the variance component approach, each position is evaluated through a REML analysis fitting the QTL effect with a covariance structure equal to the expected IBD matrix at the position in question (given the genotypes of linked marker loci). Then, the statistical profile needed to test for the presence of the QTL is constructed using the maximum log-likelihood value obtained from the REML analyses.

The comparison was done using simulation. In each replicate, the IBD matrices for several positions across the chromosome were calculated using both the deterministic and MCMC methods. They were later used in the REML analysis at each position.

Simulated population

Pedigree structure: The data were simulated assuming two different pedigree structures which were extracted from pig and sheep experimental populations to illustrate the type of complexity commonly expected in populations of these species. For the pig data set the size of the population was 500 individuals consisting of 35 related families spanning over five generations. The average full-sib family size was 14.3 with an average inbreeding coefficient of 4.5% (maximum 17%). In contrast, the sheep data set was composed of 500 individuals consisting of 269 families spanning over four generations. The average full-sib family size was 1.8 and there was no inbreeding. No selection was assumed in both types of data sets.

Parameters used: A 60 cM chromosome was simulated assuming four marker loci located at positions 0, 20, 40 and 60 cM. Each marker was assumed to have 2, 3 or 8 alleles at equal frequencies in the base population. A biallelic additive QTL was simulated to be located at position 35 cM. The gene frequency of the QTL was 0.5 and its additive effect chosen to be 13.4 (*i.e.* the total variance of the QTL was 90). Additionally, the trait was also assumed to be affected by a polygenic effect completely unlinked to the chromosome where the QTL was positioned, and an environmental random effect. The variances of both the polygenic and environmental effects were assumed to be 300 and 500, respectively (*i.e.* the QTL effect accounted for approximately 10% of the total variance of 890).

3. RESULTS

IBD values

Figure 3 (a and b) shows a scatter plot of the IBD values for all individuals calculated with the stochastic and deterministic methods for a single replicate of the pig (3a) and sheep (3b) data-type assuming eight and three alleles per marker locus, respectively. The number of non-zero elements in the matrix for the replicates shown in the figure are over 87 000 and 14 000 for the pig and the sheep data set, respectively. There were fewer non-zero elements in the sheep data as there was no inbreeding in the pedigree. The IBD values estimated with the deterministic method are generally close to those calculated with LOKI (values lying on the diagonal are found when the estimated IBD values are the same with both methods). The correlation between the estimated IBD probabilities was above 0.90 for all the replicates tested in both the pig and the sheep type data set (results not shown). However, for the replicate of the sheep data set (Fig. 3b), other trends are evident. For instance, in some cases, the deterministic method calculated the IBD between a pair of individuals to be 0.25 while the stochastic method resulted in an IBD probability ranging from 0 to 0.5. This may be explained by the fact that the deterministic method can only use information from informative markers with a known haplotype phase. Thus if an individual has no informative marker loci with a known phase, the IBD probability between that individual and another one calculated *via* the deterministic method is its expectation, which is equivalent to the IBD from the numerator relationship matrix calculated using pedigree information only. This trend is more marked in the replicate of sheep data than in the pig data due to fact that the latter provides an example with more informative marker loci. The higher the number of alleles at the marker loci, the greater the chance of an individual having informative marker loci with a known phase (see Fig. 1).

An interesting observation is that some of the IBD matrices calculated with the deterministic method were non-positive definite. This characteristic was more common in the pig data where it was seen for every test position along the chromosome. However, negative eigenvalues (which characterise a non-positive definite matrix) were always few and small in magnitude either for type of data sets or degree of polymorphism of the marker loci. The IBD matrices calculated with LOKI were always positive definite.

QTL mapping

The IBD matrix obtained with the deterministic method was used to map a QTL using a variance component approach in simulated data. The same simulated data were also analysed using the IBD matrix obtained with LOKI and their results were used as a benchmark to assess the value of the deterministic

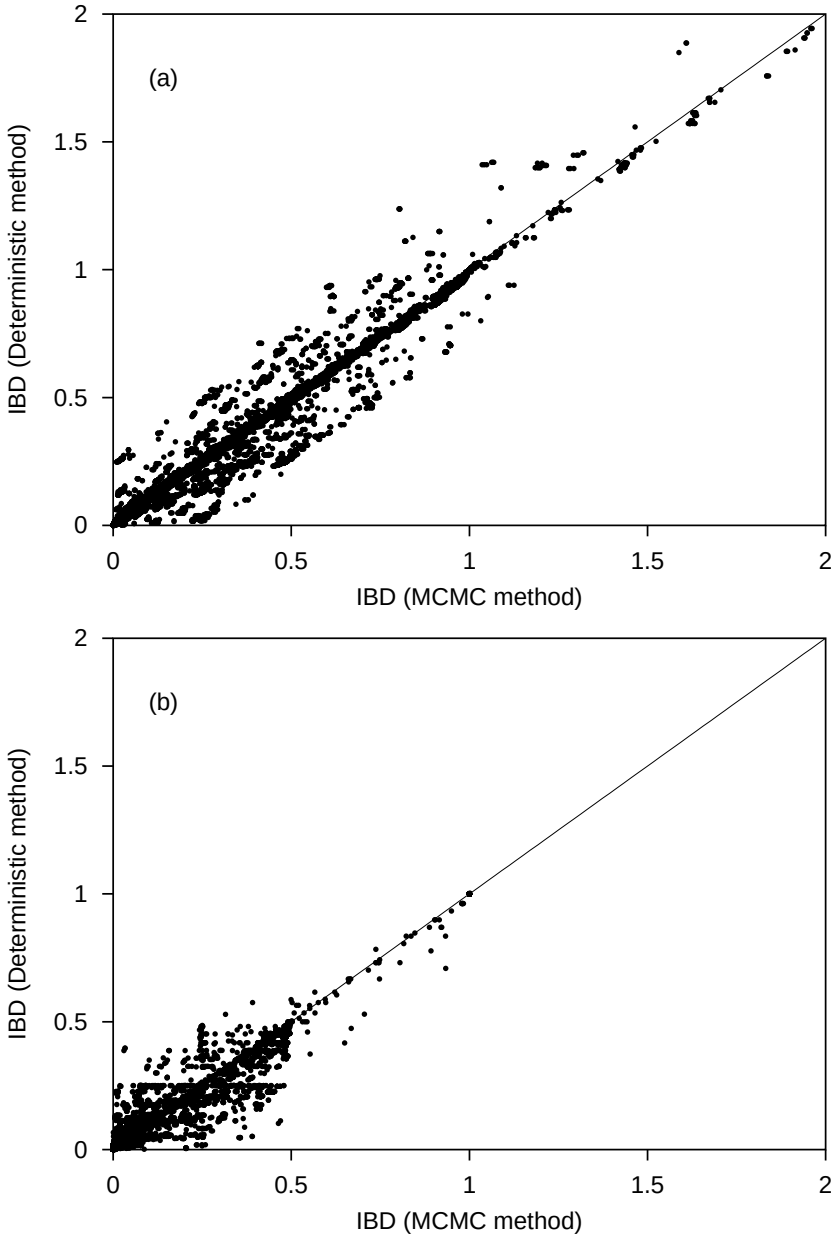

Figure 3. Scatter plot of the non-zero elements of the IBD matrix for position 30 cM calculated for one replicate of the pig (a) and sheep (b) data type using LOKI and the deterministic method. The number of alleles per marker loci assumed for the pig and the sheep data set were eight and three, respectively. The number of non-zero elements plotted in the graphs are over 87 000 and 14 000 for the pig and sheep replicates, respectively.

method. The various situations studied involved the sheep- and the pig-type data sets (to assess the effect of family size) and assuming 8, 3 or 2 alleles per marker locus (to assess the impact of the degree of polymorphism in the marker loci). Since the purpose of this study was to assess the effect of using the deterministic method to calculate IBD probabilities, the criterion of assessment was how similar were the results of both methods.

Figure 4 summarises the results from the interval mapping of the "sheep" data using the IBD matrix calculated with the MCMC and the deterministic methods. The QTL profiles are the average from 50 independent replicates assuming either eight or three alleles per marker locus. The variation between replicates was very similar for both methods. The use of the IBD matrix obtained with the deterministic method resulted in a lower test statistic across most positions in the chromosome than when the IBD matrix was calculated with the MCMC method. However, these differences were small and the shape of the curve remained the same regardless of the method used to obtain the IBD matrix.

The results from the analysis of the pig data is shown in Figure 5. The population structure of the pig data provided more power than the sheep data set for mapping QTL and this was reflected in a higher QTL profile regardless of the method used to estimate **Q**. However, the deterministic and MCMC methods were similar in terms of trend. The QTL profile obtained *via* the deterministically estimated IBD matrix was slightly lower than when the MCMC method was used.

In order to perform a more stringent test, a QTL mapping analysis was performed assuming the pig population structure consisted of only two alleles per marker loci. This population structure with large family sizes and markers exhibiting low polymorphism was expected to yield a large divergence between the results of the analyses using the IBD matrices estimated either with the deterministic or the stochastic methods. The lowly polymorphic marker loci would result in less informative markers with fewer marker phases known with absolute certainty. Similarly, the larger family sizes resulted in LOKI achieving a better estimation of uncertain haplotype phases thus retrieving more information to calculate the IBD matrix. Figure 6 shows the results for this type of population assuming the QTL explained 20% of the total variance. As it can be seen, the results of the deterministic method showed a similar behaviour to the previous situations. The test statistic values were only slightly lower, but the shape of the curve remained similar to that when using LOKI.

4. DISCUSSION

In the present study, a deterministic approach to calculate the IBD at a position on the genome using multiple linked marker genotypes was proposed. This method combines the recursive method of Wang *et al.* [14] for a single

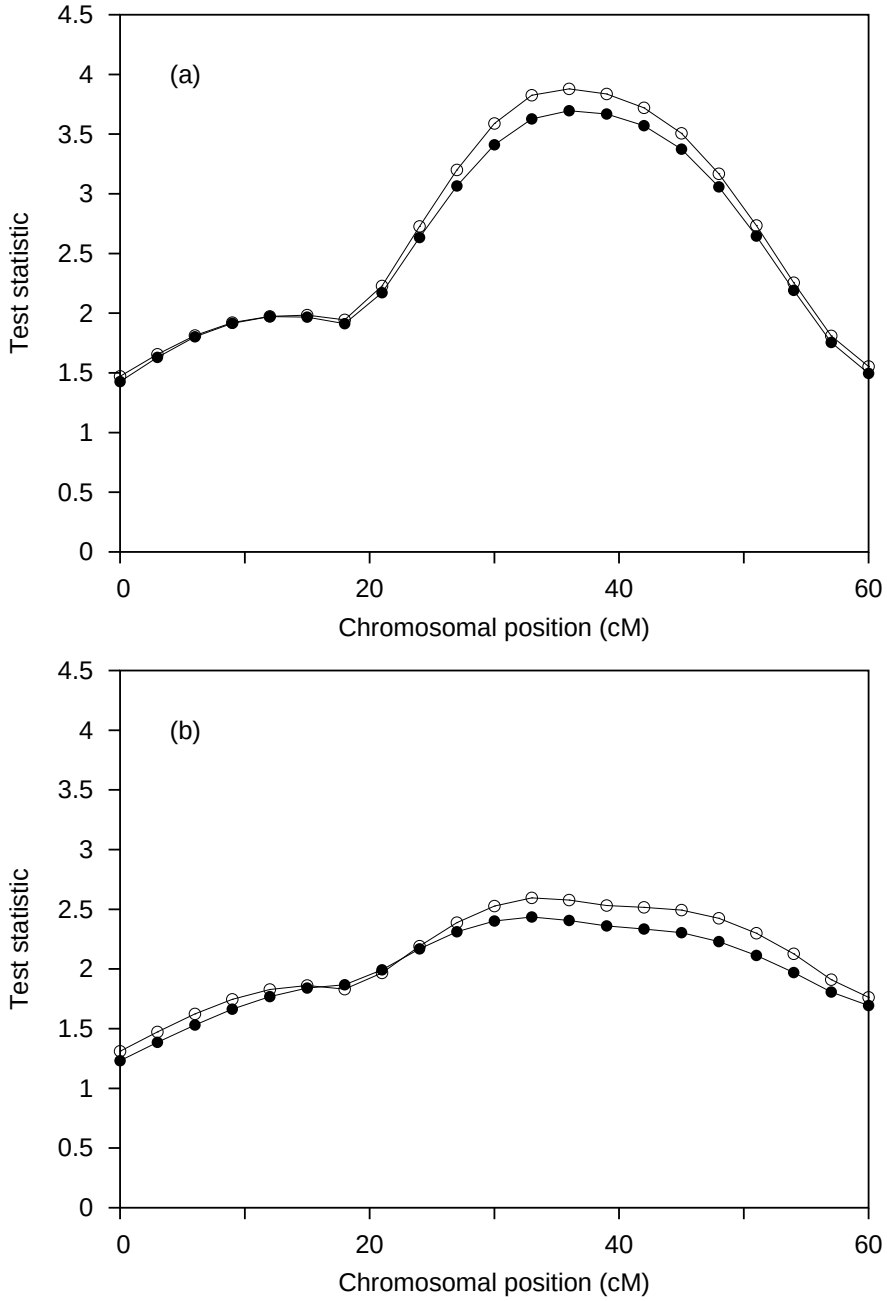

Figure 4. Average profile of the test statistic obtained from the analysis of the sheep data set assuming eight (a) and three (b) alleles per marker locus when the IBD matrix was calculated using LOKI (open circle) or the deterministic method (close circle). The results are the average over 50 independent replicates.

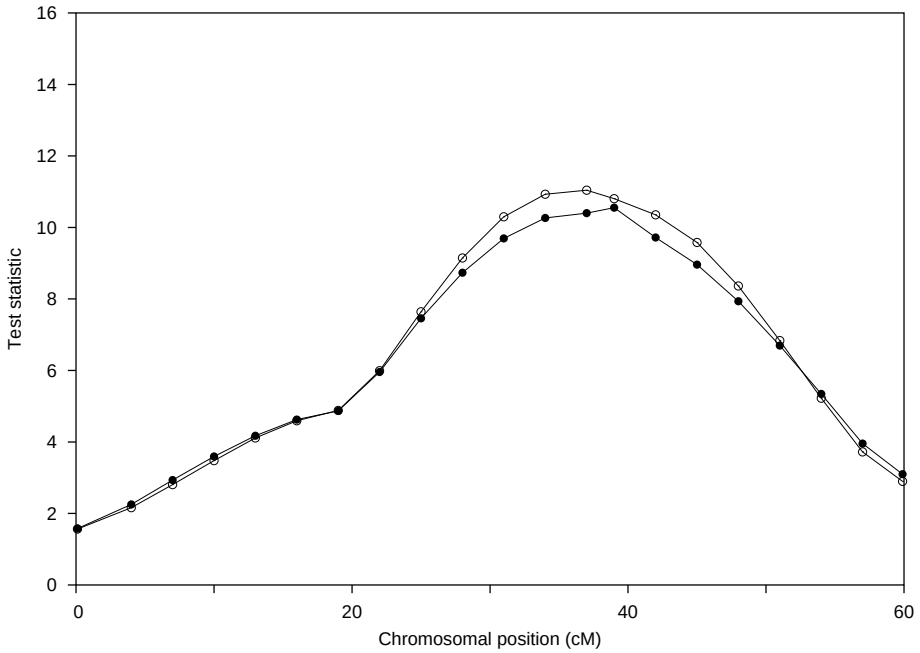

Figure 5. Average profile of the test statistic obtained from the analysis of the pig data set assuming eight alleles per marker locus when the IBD matrix was calculated using LOKI (open circle) or the deterministic method (close circle). The results are the average over 50 independent replicates.

marker locus with the method of Knott and Haley [9] for estimating IBD between sibs. The method provides an alternative to the use of MCMC and is an order of magnitude quicker. For example on a pedigree of 500, the proposed method took 13% of the time of LOKI.

The deterministic method proposed here only infers the haplotype phases of those marker loci which can be known with absolute certainty. Although this inference is very easy, as it only requires to assess the genotypes of the individual itself and those of its parents, not all individuals would have informative markers with a known phase. Thus in the present method the IBD is calculated recursively between most individuals with the exception of sib pairs for which a pair-wise calculation is done. Because the haplotype phases that are not known with absolute certainty are excluded from the calculations, the calculation of the IBD matrix can be done much quicker than when using the MCMC methods.

The consequence of using only haplotypes known with certainty is that the IBD matrix derived using the deterministic method will not necessarily be the same as the one obtained with an MCMC approach. This is because

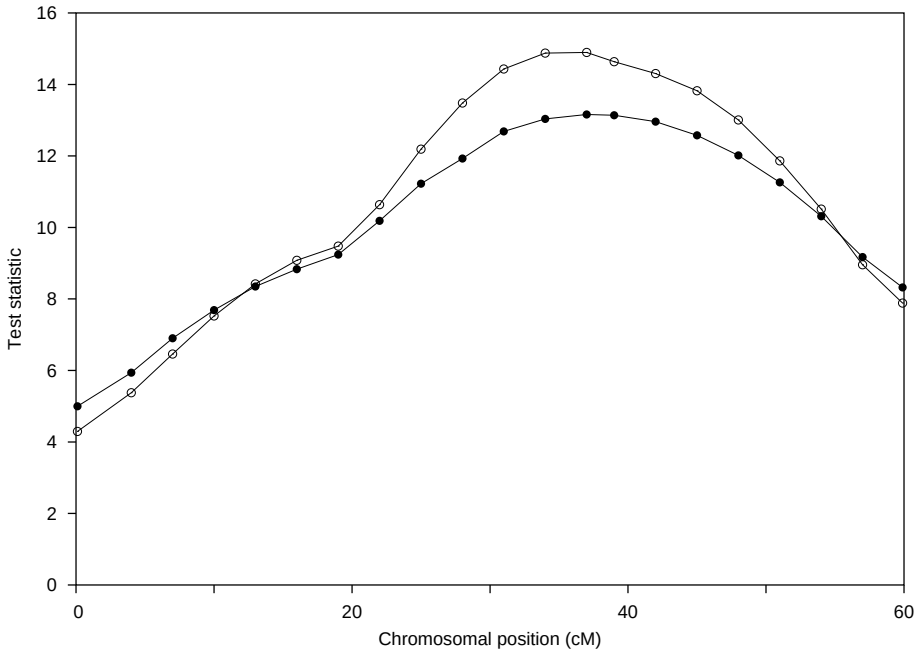

Figure 6. Average profile of the test statistic obtained from the analysis of the pig data set assuming two alleles per marker locus when the IBD matrix was calculated using LOKI (open circle) or the deterministic method (close circle). The results are the average over 50 independent replicates and the QTL was assumed to explain 20% of the total variance.

the MCMC methods are capable of extracting information from haplotypes inferred with uncertainty. Thus the gain in speed and simplicity achieved by the deterministic method is made at the cost of sub-optimal use of some of the information contained in the marker genotypes. For instance, in an MCMC method, the genotype information from the offspring would be used to calculate the parent's haplotype, which would result in a more precise estimate of the IBD between the parent and other individuals.

Nevertheless, this difference in the IBD calculation appears to have a very small effect in QTL mapping *via* a variance component approach. The results from using the IBD matrix calculated with the deterministic method showed that the test statistics were only slightly smaller than when using the MCMC method to calculate the IBD matrix. Moreover, the shape of the profile for the test statistic across the entire chromosome was the same using the two methods. This implies that the most likely position for a QTL is predicted to be in the same place regardless of the methodology used to calculate the IBD matrix. The estimates of variance explained by the QTL were also very similar regardless of the method used to calculate IBD.

These results were consistent across all the cases studied here, even for those situations where the MCMC method was expected to be significantly better. As the MCMC method would utilise partially known haplotypes better, it was expected that the MCMC would yield better results when (i) the markers were less informative, (ii) family size increased, and (iii) there were less dense markers. Thus intuitively, as these phenomena became more prevalent we might have then expected to observe a gradual divergence in the results. We tested the first two of these assumptions with a marker density of 20 cM, and surprisingly, the results from the pig-type data set assuming only two alleles per marker locus showed that the differences were not as large as had been expected. This result is very encouraging since we would expect the next generation of markers (*e.g.* single nucleotide polymorphism) to produce maps of much higher density. In these circumstances we would expect the deterministic method proposed here to suffer less from the sub-optimal use of some marker information. Moreover, under these circumstances (*i.e.* large number of closely linked markers) the MCMC methods may face mixing problems and will place increasing demands on computer time.

Although it was not part of the objective of this study, these results demonstrate the robustness of the variance component approach for mapping QTLs. The interval mapping results varied little, even in those cases where the IBD matrix was expected to diverge the most. One explanation of this observation is that the estimation of the QTL effect in a variance component approach may be largely determined by the IBD status of only a few types of relationships. It could be speculated that the IBD status of sibs or other few close relatives is the most important information and both methods may be calculating them very similarly. Therefore, these results prompt the need for a closer study on where the information comes from when mapping QTLs using the variance component approach. If this were known the deterministic method could be extended to retrieve more information when calculating the IBD from the relevant relationships, although at some cost to its simplicity.

Improving the method to utilise more information may be done by reconstructing uncertain haplotype phases for a selective group of individuals in the population. For instance, when the pedigree contains large families, the parents' haplotype may be inferred with a reasonable degree of accuracy by assessing only their offspring whose haplotype is known with absolute certainty. Additionally, haplotype inference of final progeny may also be considered as they are relatively easy to carry out. This approach of inferring the haplotype phase of a selective number of individuals would increase the precision of the IBD matrix with a very low impact on the simplicity and speed of the method.

One observation made during the course of the study was that the IBD matrix calculated with the deterministic method may not be positive definite, *i.e.* some

contrasts among individuals may be estimated to have negative QTL variances. The cause for this characteristic appears to be due to the calculation of the IBD for sibs in a pair-wise fashion. As different marker brackets may be used when calculating the IBD value between different pairs of sibs, the IBD may be slightly inconsistent, resulting in a non-positive definite IBD matrix. An analogy of this problem can be found when estimating variance components of several traits, where non-positive definite covariance matrices can be obtained if they are estimated independently using several bi-variate analyses.

This characteristic is theoretically inconsistent with a mixed linear model approach but the observation was not serious for several reasons. First, a close study of the different IBD matrices calculated across the replicates and positions showed that for the particular situations studied here, only a few negative eigenvalues (symptomatic of the problem) were observed, all of very small magnitude. Secondly, and possibly as a result of the former, the occurrences appeared to have little impact on the results. However, if circumstances were identified where a larger proportion of the eigenvalues were negative and large, a solution could be obtained from bending the IBD matrix. This was suggested by Hayes and Hill [8] in the context of estimating genetic correlations with multiple traits. A more detailed study is required to establish the circumstances under which the problem of negative eigenvalues becomes large enough to have a substantial impact.

Benefits from using the deterministic method can be seen in a variety of applications. The deterministic method allows our QTL analysis to be performed far quicker. Therefore, previously intractable permutation tests [2] for obtaining empirical thresholds become a reality. A further application arising from simplicity is the use of IBD matrices to constrain the rates of loss of genetic variation either in one region or genome-wide; Fernandez *et al.* [3] suggest that the use of the expected numerator relationship matrix limits their application, however, simple substitution of the expected for the observed relationship matrix (*i.e.* an IBD matrix as calculated here) overcomes these limitations.

The simplicity and speed of the described method is an attractive alternative to more complex MCMC methods. However, there are certain situations where MCMC methods such as LOKI would prove to be the preferred choice. For instance, when the linkage map is sparse and a substantial proportion of the individuals in the pedigree have missing genotypes for all marker loci, the amount of information from which the IBD can be calculated is very scarce. In these circumstances, LOKI would be able to utilise the available information and the loss when using the deterministic method may be substantial. However, dense marker maps are becoming more of a reality with the development of DNA marker technology, and their richness of information is increasingly favouring a deterministic approach for the calculation of IBD.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge funding from the Ministry of Agriculture, Fisheries and Food (UK) and the Biotechnology and Biological Sciences Research Council (BBSRC) and Dr. S.C. Heath for generously allowing the use of his LOKI software. We would like to thank Drs. P.M. Visscher and S.A. Knott for their comments.

REFERENCES

[1] Almasy L., Blangero J., Multipoint quantitative-trait linkage analysis in general pedigrees, Am. J. Hum. Genet. 62 (1998) 1198–1211.
[2] Churhill G., Doerge R., Empirical threshold values for quantitative trait mapping, Genetics 138 (1994) 963–971.
[3] Fernandez B., Santiago E., Toro M.A., Caballero A., Effect of linkage on the control of inbreeding in selection programmes, Genet. Sel. Evol. 32 (2000) 249–264.
[4] Fernando R.L., Grossman M., Marker-assisted selection using best linear unbiased prediction, Genet. Sel. Evol. 21 (1989) 467–477.
[5] George A.W., Visscher P.M., Haley C.S., Mapping quantitative trait loci in complex pedigrees: a two step variance component approach, Genetics 156 (2000) 2081–2092.
[6] Goldgar D.E., Multiple point analysis of human quantitative genetic variation, Am. J. Hum. Genet. 47 (1990) 957–967.
[7] Grignola F.E., Hoeschele I., Tier B., Mapping quantitative trait loci in outcross populations *via* residual maximum likelihood. I. Methodology, Genet. Sel. Evol. 28 (1996) 479–490.
[8] Hayes J.F., Hill W.G., Modification of estimates of parameters in the construction of genetics selection indices ("bending"), Biometrics 37 (1981) 483–493.
[9] Knott S.A., Haley C.S., Simple multiple-marker sib-pair analysis for mapping quantitative trait loci, Heredity 81 (1998) 48–54.
[10] Lander E.S., Botstein D., Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, Genetics 121 (1989) 185–199.
[11] Thompson E.A., Heath S.C., Estimation of conditional multilocus gene identity among relatives. In: Seillier-Moseiwitch F., Donnelly P., Waterman M. (Eds.), Statistics in Molecular Biology, Springer-Verlag IMS lecture note series, New York, 1999.
[12] Van Arendonk J.A.M., Tier B., Kinghorn B.P., Use of multiple genetic markers in prediction of breeding values, Genetics 137 (1994) 319–329.
[13] Visscher P.M., Haley C.S., Heath S.C., Muir W.J., Blackwood D.H.R., Detecting QTLs for uni and bipolar disorder using variance component method, Psychiatr. Genet. 9 (1999) 75–84.
[14] Wang T., Fernando R.L., Van der Beek S., Grossman M., Van Arendonk J.A.M., Covariance between relatives for a marked quantitative locus, Genet. Sel. Evol. 27 (1995) 251–274.
[15] Xu S., Atchley W.R., A random model approach to interval mapping of quantitative trait loci, Genetics, 141 (1995) 1189–1197.