

A Monte Carlo algorithm for computing the IBD matrices using incomplete marker information

Y Mao and S Xu

Department of Botany & Plant Sciences, University of California, Riverside, CA 92521-0124, USA

Identity-By-Descent (IBD) is a general measurement of the relationship between two groups of genes. If the two groups consist of two homologous genes, one from each individual, the IBD is called the coancestry between the two individuals. Coancestry is an important concept in both population and quantitative genetics. It is the probability that both genes are copies of the same gene in the genealogy. The average coancestry value at a random locus in a population reflects the level of population diversity, effective population size, the level of inbreeding and other attributes. Coancestry is also the building block for the covariance structure used to estimate the additive genetic variance component for a quantitative trait. There are many other types of IBD matrices, depending on the natures of the genes included in each group, and these IBD matrices vary from locus to

locus. Molecular markers distributed along the genome provide information that can be used to infer these locus-specific IBD matrices. As a result, we can estimate and test the variance components of a quantitative trait contributed by these loci using the inferred IBD matrices. In this study, we develop the concept of locus-specific epistatic IBD matrices and a Monte Carlo method to infer these IBD matrices. The method is suitable for large pedigrees with arbitrary complexity and various levels of missing marker information. With these locus-specific IBD matrices, we are ready to search for quantitative trait loci along the genome in complicated pedigrees.

Heredity (2005) 94, 305–315. doi:10.1038/sj.hdy.6800564
Published online 15 September 2004

Keywords: complex pedigree; descent graph; hidden Markov model; identical-by-descent; quantitative trait locus

Introduction

Identity-By-Descent (IBD) is an important concept in both population and quantitative genetics. It describes the relationship between two groups of genes. A gene, as opposed to an allele or a locus, is the DNA segment that is copied from parents to offspring. Two genes are IBD, if one is a copy of the other or they are both copies of the same ancestral gene.

The concept of IBD was introduced by Cotterman (1940), and has since been widely developed. It has been applied to the problem of four genes in two individuals by Li and Sacks (1954), Jacquard (1972), Nadot and Vaysseix (1973) and Denniston (1974), and to four arbitrary genes by Cockerham (1971).

In diploid organisms, each individual carries two genes per locus. The IBD between two individuals becomes complicated because this IBD is decomposed into four IBD coefficients between two groups of single genes and one IBD coefficient between two groups of double genes. The average value of the four single-gene IBD coefficients is called the coancestry between the two individuals and the two-gene IBD coefficient is called the fraternity between the two individuals (Lynch and Walsh, 1998). There are a total of 15 possible detailed IBD coefficients for the two individuals and nine condensed IBD coefficients if the IBD coefficients of similar nature are combined (Jacquard, 1974; Cockerham,

1980). For details, see Chapter 7 of Lynch and Walsh (1998) or Sobel *et al* (2001). The problem becomes considerably complex when dealing with more than four genes, largely because the number of IBD states grows rapidly with the number of genes involved. For example, the number of IBD states for six (ordered) genes is 203, and for eight genes there are 4140 IBD states (Thompson, 1974). A complete discussion of this problem may also be found in Whittemore and Halpern (1994).

Coancestry and fraternity are important terms used for estimating the additive and dominance genetic variance components in quantitative genetics. If each group contains two nonallelic genes of the same individual, then the IBD between the two individuals measures the probability that the two individuals share the same haplotype (ie, share two nonallelic genes jointly). This IBD has not been formally named in the literature, but is an important term used to estimate the additive-by-additive genetic variance component for quantitative traits. If one group contains two allelic genes and one nonallelic gene, all from the same individual, the IBD between two individuals measures the probability that the two individuals share three such defined genes jointly in the genealogy. Again, there is no formal name for this IBD, but it is used to estimate the dominance-by-additive genetic variance component. A similar IBD is used to estimate the additive-by-dominance variance component. Finally, if each group contains four genes of two loci from the same individual, the IBD between two individuals defines the probability that the two individuals share IBD genotypes for both loci in the genealogy. This IBD is used to estimate

the dominance-by-dominance variance component. The two-loci IBD probabilities have also been used by many other authors, see for example, Weir and Cockerham (1969), Weir *et al* (1980), Weir and Hill (1980) and Whitlock *et al* (1993).

IBD coefficient is actually defined as a discrete event because each of the two groups contains a finite number of genes. Since the genotypes of individuals are not observable, the IBD event is replaced by the probability. However, if we actually observe the genotypes, we may already know whether the two relatives, for example, sibs, do actually share the same IBD genotype or not. If they do, the fraternity is 1 and, if not, the fraternity is 0. Therefore, the conditional IBD when the genotypes are observed is different from the expected IBD when the genotypes are not observed, and these conditional IBD values vary from locus to locus. If we can actually observe the genotypes of all individuals for all loci in a pedigree, we may be able to know the IBD matrices for all loci (all pairwise IBD values between all individuals in the pedigree). These IBD matrices can be used to estimate various types of variance components contributed by individual loci. This is the IBD-based variance component method for quantitative trait loci (QTL) mapping (Almasy and Blangero, 1998; Yi and Xu, 2000).

Segregation patterns of molecular markers allow inference of the IBD matrices at an arbitrary genome location, and thus we can scan the entire genome to search for QTL. Haseman and Elston (1972) first introduced the concept of conditional IBD between sib pairs given observed marker information. Goldgar (1990) and Schork (1993) developed methods to estimate the genetic material shared by sibs for an interval flanked by two markers. Fulker and Cardon (1994) combined the sib-pair regression method of Haseman and Elston (1972) with the interval mapping of Lander and Botstein (1989) to systematically scan the genetic variance for the entire genome. Fulker *et al* (1995) eventually extended the interval mapping into multipoint mapping, where they used all linked markers simultaneously to infer the IBD of an arbitrary position. This method has utilized information from all markers. However, the multipoint method itself is a multiple regression approach where they treated the IBD of a putative position as the dependent variable and the IBDs of the markers as independent variables. The drawback of the multiple regression method is that it only gives the expectation, not the distribution, of the IBD. In addition, the method only uses the expected IBD values, instead of the IBD distributions, of the markers, which proves to be suboptimal. Kruglyak and Lander (1995) also developed a multipoint method to infer the IBD values of sib pair. This new method takes advantage of the Markov chain property of markers along the genome and proves to be optimal. Since an IBD matrix is proportional to a covariance matrix, it should be semipositive definite. A semipositive definite matrix is loosely defined as a matrix with non-negative determinant. This property is required in computing likelihood functions. The method of Kruglyak and Lander (1995) utilizes all sibs simultaneously in a full-sib family, which warrants the semipositive definite property of the IBD matrix. The increased information of this method is gained at the price of extensive computing, in both the memory and speed. The largest family size that the method can handle

is about 10, which is feasible for nuclear families, but not so for large pedigrees.

Wang *et al* (1995, 1998) developed a method to infer the conditional IBD matrix for large pedigrees using one or two markers at a time. The method can efficiently handle pedigrees of virtually unlimited size. Unfortunately, it does not use all markers simultaneously and thus is not optimal in that regard. The multipoint method of Kruglyak and Lander (1995) is feasible for small pedigrees, but application to even moderate pedigrees appears to be difficult because of the heavy computing load. Almasy and Blangero (1998) recently developed a multipoint method that is efficient for large pedigrees with arbitrary complexity. This multipoint method has been incorporated into a program called Solar, which is so far the most practical software package for human pedigree analysis. However, the computing efficiency is gained at the price of optimality because the multipoint method they adopted is not based on the hidden Markov model; rather, it is an extension of the sib-pair regression method of Fulker *et al* (1995) to arbitrary relative pairs. The regression method for estimating IBD values does not use the distribution of marker IBD; instead, it uses the expectation of the marker IBD. Furthermore, the pairwise IBD values are estimated independently, although they should be estimated jointly because members within a pedigree are not independent.

Epistasis occurs when the gene action of one locus is influenced by that of another locus. Epistatic effects can occur among several loci, but interactions involving three or more loci are difficult to interpret and estimate. In a two-loci model with two alleles per locus, there are nine possible genotypes and the total genetic variance can be partitioned into eight components of variance (Cockerham, 1980). The eight components of variance include the additive and dominance variances of both loci and their interactions, additive by additive, additive by dominance, dominance by additive, and dominance by dominance.

The purpose of this study is to develop a Monte Carlo method to calculate various IBD matrices for arbitrarily complicated large pedigrees with optimal utilization of marker information. These IBD matrices can be used for candidate gene evaluation and QTL mapping under the variance component model.

Methods

The pedigree with 22 individuals shown in Figure 1 is used to illustrate the proposed method. There are six founders, identified as 1, 2, 3, 6, 7, and 12 in the pedigree. Genotypic data were simulated for 10 markers located at positions 0, 5, 12, 15, 25, 36, 40, 51, 75, and 90 cM along a chromosome. The founder alleles of each locus were sampled from 10 equally frequent alleles. The two targeted positions evaluated are 10 and 45 cM, that is, between markers two and three for the first position and between markers seven and eight for the second position. Marker genotypes of individuals 1, 2, 4, 5, 8, 9, 10, and 11 were removed from the data set and replaced by missing values (Table 1).

Legal descent graph

Descent graph can be represented by inheritance indicators, which consists of a binary vector (ie a vector

with zeros and ones as entries) describing the inheritance pattern of a pedigree. For a pedigree with N members, the inheritance indicator at a certain locus is written as

$$\mathbf{v} = (p_1, m_1, p_2, m_2, \dots, p_N, m_N) \quad (1)$$

whose coordinates describe the outcome of the paternal and maternal meioses giving rise to the N members in the pedigree. Specifically, $p_i=0$ or 1, according to whether the paternal or maternal allele of the father has been transmitted to the i th member of the pedigree. In the same way, m_i carries the same information for the corresponding maternal meiosis. Thus, the inheritance indicator completely specifies which of the $2n$ distinct founder alleles are inherited by each member, where n is the number of founders in the pedigree. However, the true inheritance indicators are not observable, and it is only possible to observe the linkage information at genotyped marker loci. If no chromatid interference is

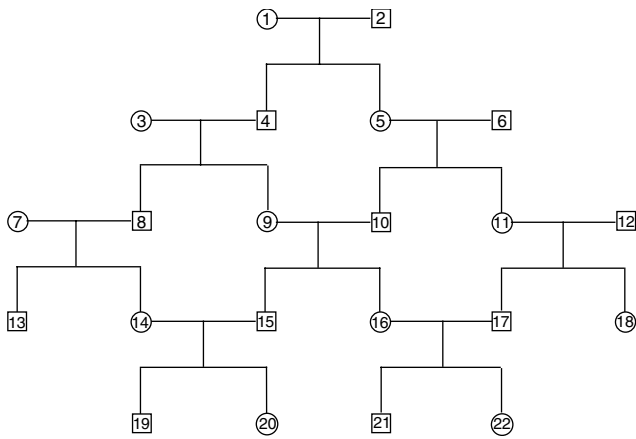


Figure 1 A complex pedigree with 22 members including six founders.

assumed between markers, the inheritance indicators at all m genotyped loci constitute a hidden Markov chain.

Sampling legal descent graphs for markers

To calculate the IBD values, the first step is to sample the allelic inheritance configuration for each member of a pedigree at the locus of interest. The allelic inheritance configurations of all individuals in a pedigree represent the graph of gene transmission from the founders to all the descents, called the descent graph (Sobel and Lange, 1996). If a marker is fully informative, the inheritance indicators can be inferred with 100% certainty from the observed marker genotypes. However, most markers will not be fully informative and we must sample the descent graph using the conditional probabilities given observed marker information. Since the inheritance indicator depends on which of the four possible genotypes each individual takes for each marker, we only need to sample genotypes for all the markers across the genome. We actually take the multipoint approach to calculate the joint probability of marker genotypes and sample the joint genotype from this probability. We take advantage of the Markov property of the marker arrangement along the chromosomes and simulate the joint sample sequentially from one end of the chromosome to the other end. In the appendix, we describe the sampling process under the following different situations: (1) no missing marker genotypes and linkage phases are known, (2) no missing marker genotypes but the linkage phases are unknown, and (3) incomplete marker information with some missing marker genotypes and the linkage phases are unknown.

Sampling descent graph of an arbitrary locus

Given the descent graph of markers, we are ready to sample the descent graphs at some targeted loci of

Table 1 Tabular forms of the pedigree and the incomplete marker data for the complicated pedigree with 22 members

Member ID	Sex	Father ID	Mother ID	Genotypes of markers located at										
				0 cM	5 cM	12 cM	15 cM	25 cM	36 cM	40 cM	51 cM	75 cM	90 cM	
1	F	0	0	**	**	**	**	**	**	**	**	**	**	**
2	M	0	0	**	**	**	**	**	**	**	**	**	**	**
3	F	0	0	7 9	4 8	7 8	3 5	2 5	4 10	5 7	4 4	6 10	3 5	
4	M	2	1	**	**	**	**	**	**	**	**	**	**	**
5	F	2	1	**	**	**	**	**	**	**	**	**	**	**
6	M	0	0	5 5	5 6	3 5	3 6	2 9	4 5	3 7	2 5	4 10	3 10	
7	F	0	0	7 9	2 4	8 9	2 9	5 8	7 7	9 9	3 5	1 8	1 7	
8	M	4	3	**	**	**	**	**	**	**	**	**	**	**
9	F	4	3	**	**	**	**	**	**	**	**	**	**	**
10	M	6	5	**	**	**	**	**	**	**	**	**	**	**
11	F	6	5	**	**	**	**	**	**	**	**	**	**	**
12	M	0	0	8 8	2 7	3 7	3 5	7 9	4 9	1 6	4 5	3 8	3 10	
13	M	8	7	7 7	2 8	8 9	5 9	5 5	7 10	5 9	4 5	1 6	3 7	
14	F	8	7	7 9	4 8	8 8	2 5	5 8	7 10	5 9	3 4	6 8	1 3	
15	M	10	9	5 7	2 8	4 5	7 8	2 2	3 7	8 10	4 8	2 10	3 5	
16	F	10	9	5 10	2 9	3 4	3 7	2 2	4 5	3 7	2 4	3 4	2 10	
17	M	12	11	5 8	5 7	3 7	3 3	2 9	5 9	3 6	2 5	5 8	2 10	
18	F	12	11	5 8	2 7	3 7	3 3	2 9	7 9	6 10	5 6	8 10	10 10	
19	M	15	14	5 7	2 8	5 8	5 8	2 5	7 10	5 10	4 8	2 6	3 3	
20	F	15	14	7 7	8 8	4 8	5 7	2 5	3 10	5 8	4 4	8 10	1 5	
21	M	17	16	8 10	7 9	4 7	3 7	2 9	4 5	3 7	4 5	3 8	10 10	
22	F	17	16	8 10	7 9	4 7	3 3	2 9	5 9	3 6	2 5	4 8	10 10	

Members with parents 0 are recognized as founders, and * indicates missing allele.

biological interest using the sampled descent graph of markers. The descent graphs of these targeted loci are then used to calculate the IBD matrices for variance component analysis. There are two ways to sample the descent graph of the targeted loci. One approach is to insert the targeted loci into the existing marker map as virtual markers. The genotypes of these virtual markers are filled with missing values. We then sample the joint descent graph of the real and the virtual markers simultaneously using the method described above. The other approach is to sample the descent graph of these targeted loci using the sampled descent graphs of flanking markers, provided that the recombination process in meiosis exhibits no interference. Given the descent graph of markers, the descent graph of an arbitrary locus can be sampled one individual at a time. The inheritance indicators of any individual are sampled from the four probabilities representing the four genotypic configurations, denoted by (0,0), (1,0), (0,1), and (1,1). For example, (0,1) means that the individual has inherited the maternal allele of the father and the paternal allele of the mother. Let $M=i$, for $i=1,2,3,4$, be a discrete variable representing the four genotypic configurations of the left marker in the aforementioned order and $N=j$, for $j=1,2,3,4$, be the corresponding variable for the right marker. Denote the corresponding variable for the locus bracketed by the two markers by $U=k$, for $k=1,2,3,4$. The conditional distribution of U is

$$P(U=k|M=i, N=j) = \frac{P(M=i|U=k)P(N=j|U=k)}{\sum_{k'=1}^4 P(M=i|U=k')P(N=j|U=k')} \quad (2)$$

where $P(M=i|U=k)$ and $P(N=j|U=k)$ are the marker genotypic probabilities conditional on the genotype of the intermediate locus and can be found from the transition matrices given by Xu (1998). The probability $P(U=k|M=i, N=j)$ will be used to simulate a particular genotypic configuration from which the inheritance indicators are automatically given.

The inheritance indicators of all members in the pedigree, when considered together, can link each one of the $2N$ alleles in the pedigree to one of the $2n$ founder alleles, where N is the size of the pedigree and n is the number of founders. Therefore, each individual is connected to the founder alleles by two lines, one for the paternal allele and the other for the maternal allele. The two lines can be represented by two vectors, \mathbf{Z}^p and \mathbf{Z}^m , each with $2n$ elements. Each vector has one and only one non-zero element and the non-zero element equals one. The position of this non-zero element points to the position of the particular allele arranged in the list of the $2n$ founder alleles. For example, if the paternal allele of a member is traced back to the l th founder allele and the maternal allele of it is traced back to the k th founder allele, then $z_l^p = z_k^m = 1$ and all other elements of the two vectors are zero. These link vectors are used to calculate the conditional IBD matrices.

Calculating the IBD matrices

The IBD matrices are better described in the context of variance component analysis. Therefore, we will first describe the linear mixed effect model and then introduce the concept of IBD matrices as a function of

the inheritance indicators. We will consider a single locus first and then consider two loci. Each of the $2n$ founder alleles is assigned an allelic value in the unit of a quantitative trait (a_k for the k th founder allele) and each pair of the founder alleles is assigned a dominance interaction effect ($d_{kk'}$ for the combination of the k th and k' th founder alleles). The phenotypic value of individual i in the pedigree is then described by the following model:

$$y_i = \mathbf{X}_i^T \mathbf{b} + (\mathbf{Z}_i^p + \mathbf{Z}_i^m)^T \mathbf{a} + (\mathbf{Z}_i^p \otimes \mathbf{Z}_i^m)^T \mathbf{d} + e_i \quad (3)$$

where \mathbf{X}_i is a $q \times 1$ known design matrix, \mathbf{b} is a $q \times 1$ vector for nongenetic fixed effects (eg effects of location, year, age, etc), \mathbf{a} is a $(2n) \times 1$ vector for the founder allelic effects, \mathbf{Z}_i^p is a $(2n) \times 1$ vector indicating which of the $2n$ founder alleles has been passed to individual i through its father, \mathbf{Z}_i^m is defined similarly but through its mother, \otimes represents the Kronecker product of two matrices (also called direct product), the superscript T means matrix transposition, \mathbf{d} is a $(2n)^2 \times 1$ vector for the dominance effects defined in the founder population, and e_i is the residual error with an assumed $N(0, \sigma^2)$ distribution. Note that \mathbf{Z}_i^p and \mathbf{Z}_i^m are highly sparse because each matrix has only one nonzero element. Assume that \mathbf{a} is multivariate normal with $\mathbf{a} \sim N(\mathbf{0}, \sigma_A^2 \mathbf{I}_{2n})$ and so is \mathbf{d} with $\mathbf{d} \sim N(\mathbf{0}, \sigma_D^2 \mathbf{I}_{(2n)^2})$, where σ_A^2 and σ_D^2 are called the allelic and dominance variances, respectively. The covariance between y_i and y_j is

$$\text{Cov}(y_i, y_j) = E[(\mathbf{Z}_i^p + \mathbf{Z}_i^m)^T (\mathbf{Z}_j^p + \mathbf{Z}_j^m)] \sigma_A^2 + E[(\mathbf{Z}_i^p \otimes \mathbf{Z}_i^m)^T (\mathbf{Z}_j^p \otimes \mathbf{Z}_j^m)] \sigma_D^2 + \delta_{ij} \sigma^2 \quad (4)$$

where δ_{ij} , the Kronecker delta, is defined as $\delta_{ij} = 1$ for $i=j$ and $\delta_{ij} = 0$ for $i \neq j$. Therefore, various IBD values can be defined as follows (Yi and Xu, 2000):

$$\pi_{A(ij)} = E[(\mathbf{Z}_i^p + \mathbf{Z}_i^m)^T (\mathbf{Z}_j^p + \mathbf{Z}_j^m)] \quad (5)$$

is the IBD for the additive effect ($0 \leq \pi_{A(ij)} \leq 4$) and

$$\pi_{D(ij)} = E[(\mathbf{Z}_i^p \otimes \mathbf{Z}_i^m)^T (\mathbf{Z}_j^p \otimes \mathbf{Z}_j^m)] = E[(\mathbf{Z}_i^p \otimes \mathbf{Z}_i^m)^T (\mathbf{Z}_j^p \otimes \mathbf{Z}_j^m)] \quad (6)$$

is the IBD for the dominance effect ($0 \leq \pi_{D(ij)} \leq 1$ and $\pi_{D(ii)} = 1$), where $\mathbf{Z}_i^p \otimes \mathbf{Z}_i^m = (\mathbf{Z}_i^p \otimes \mathbf{Z}_i^m)^T$, etc.

Now, let us consider two loci with arbitrary linkage relationship. We need to use one additional subscript for all the variables defined earlier to index the locus, for example, \mathbf{Z}_{i1}^p and \mathbf{Z}_{i1}^m denote the design matrices for locus one, and \mathbf{a}_1 and \mathbf{d}_1 denote the genetic effects for locus one. The two-loci model becomes

$$y_i = \mathbf{X}_i^T \mathbf{b} + e_i + (\mathbf{Z}_{i1}^p + \mathbf{Z}_{i1}^m)^T \mathbf{a}_1 + (\mathbf{Z}_{i1}^p \otimes \mathbf{Z}_{i1}^m)^T \mathbf{d}_1 + (\mathbf{Z}_{i2}^p + \mathbf{Z}_{i2}^m)^T \mathbf{a}_2 + (\mathbf{Z}_{i2}^p \otimes \mathbf{Z}_{i2}^m)^T \mathbf{d}_2 + [(\mathbf{Z}_{i1}^p + \mathbf{Z}_{i1}^m) \otimes (\mathbf{Z}_{i2}^p + \mathbf{Z}_{i2}^m)]^T \mathbf{i}_{aa} + [(\mathbf{Z}_{i1}^p + \mathbf{Z}_{i1}^m) \otimes (\mathbf{Z}_{i2}^p \otimes \mathbf{Z}_{i2}^m)]^T \mathbf{i}_{ad} + [(\mathbf{Z}_{i1}^p \otimes \mathbf{Z}_{i1}^m) \otimes (\mathbf{Z}_{i2}^p + \mathbf{Z}_{i2}^m)]^T \mathbf{i}_{da} + [(\mathbf{Z}_{i1}^p \otimes \mathbf{Z}_{i1}^m) \otimes (\mathbf{Z}_{i2}^p \otimes \mathbf{Z}_{i2}^m)]^T \mathbf{i}_{dd} \quad (7)$$

where \mathbf{i}_{aa} , \mathbf{i}_{ad} , \mathbf{i}_{da} and \mathbf{i}_{dd} represent the additive-by-additive, additive-by-dominance, dominance-by-additive and dominance-by-dominance effects, respectively.

The covariance between y_i and y_j is

$$\begin{aligned} \text{Cov}(y_i, y_j) &= E[(\mathbf{Z}_{1i}^p + \mathbf{Z}_{1i}^m)^T (\mathbf{Z}_{1j}^p + \mathbf{Z}_{1j}^m)] \sigma_{1A}^2 \\ &\quad + E[(\mathbf{Z}_{1i}^p \otimes \mathbf{Z}_{1i}^m)^T (\mathbf{Z}_{1j}^p \otimes \mathbf{Z}_{1j}^m)] \sigma_{1D}^2 \\ &\quad + E[(\mathbf{Z}_{2i}^p + \mathbf{Z}_{2i}^m)^T (\mathbf{Z}_{2j}^p + \mathbf{Z}_{2j}^m)] \sigma_{2A}^2 \\ &\quad + E[(\mathbf{Z}_{2i}^p \otimes \mathbf{Z}_{2i}^m)^T (\mathbf{Z}_{2j}^p \otimes \mathbf{Z}_{2j}^m)] \sigma_{2D}^2 \\ &\quad + E\{[(\mathbf{Z}_{1i}^p + \mathbf{Z}_{1i}^m) \otimes (\mathbf{Z}_{2i}^p + \mathbf{Z}_{2i}^m)]^T \\ &\quad \times [(\mathbf{Z}_{1j}^p + \mathbf{Z}_{1j}^m) \otimes (\mathbf{Z}_{2j}^p + \mathbf{Z}_{2j}^m)]\} \sigma_{AA}^2 \\ &\quad + E\{[(\mathbf{Z}_{1i}^p + \mathbf{Z}_{1i}^m) \otimes (\mathbf{Z}_{2i}^p \otimes \mathbf{Z}_{2i}^m)]^T \\ &\quad \times [(\mathbf{Z}_{1j}^p + \mathbf{Z}_{1j}^m) \otimes (\mathbf{Z}_{2j}^p \otimes \mathbf{Z}_{2j}^m)]\} \sigma_{AD}^2 \\ &\quad + E\{[(\mathbf{Z}_{1i}^p \otimes \mathbf{Z}_{1i}^m) \otimes (\mathbf{Z}_{2i}^p + \mathbf{Z}_{2i}^m)]^T \\ &\quad \times [(\mathbf{Z}_{1j}^p \otimes \mathbf{Z}_{1j}^m) \otimes (\mathbf{Z}_{2j}^p + \mathbf{Z}_{2j}^m)]\} \sigma_{DA}^2 \\ &\quad + E\{[(\mathbf{Z}_{1i}^p \otimes \mathbf{Z}_{1i}^m) \otimes (\mathbf{Z}_{2i}^p \otimes \mathbf{Z}_{2i}^m)]^T \\ &\quad \times [(\mathbf{Z}_{1j}^p \otimes \mathbf{Z}_{1j}^m) \otimes (\mathbf{Z}_{2j}^p \otimes \mathbf{Z}_{2j}^m)]\} \sigma_{DD}^2 + \delta_{ij} \sigma^2 \\ &= \pi_{1A(ij)} \sigma_{1A}^2 + \pi_{1D(ij)} \sigma_{1D}^2 \\ &\quad + \pi_{2A(ij)} \sigma_{2A}^2 + \pi_{2D(ij)} \sigma_{2D}^2 \\ &\quad + \pi_{AA(ij)} \sigma_{AA}^2 + \pi_{AD(ij)} \sigma_{AD}^2 \\ &\quad + \pi_{DA(ij)} \sigma_{DA}^2 + \pi_{DD(ij)} \sigma_{DD}^2 + \delta_{ij} \sigma^2 \end{aligned} \tag{8}$$

where

$$\pi_{kA(ij)} = E[(\mathbf{Z}_{ki}^p + \mathbf{Z}_{ki}^m)^T (\mathbf{Z}_{kj}^p + \mathbf{Z}_{kj}^m)], \quad k = 1, 2 \tag{9}$$

$$\begin{aligned} \pi_{kD(ij)} &= E[(\mathbf{Z}_{ki}^p \otimes \mathbf{Z}_{ki}^m)^T (\mathbf{Z}_{kj}^p \otimes \mathbf{Z}_{kj}^m)] \\ &= E[(\mathbf{Z}_{ki}^{pT} \mathbf{Z}_{ki}^m) (\mathbf{Z}_{kj}^m \mathbf{Z}_{kj}^p)], \quad k = 1, 2 \end{aligned} \tag{10}$$

$$\begin{aligned} \pi_{AA(ij)} &= E\{[(\mathbf{Z}_{1i}^p + \mathbf{Z}_{1i}^m) \otimes (\mathbf{Z}_{2i}^p + \mathbf{Z}_{2i}^m)]^T \\ &\quad \times [(\mathbf{Z}_{1j}^p + \mathbf{Z}_{1j}^m) \otimes (\mathbf{Z}_{2j}^p + \mathbf{Z}_{2j}^m)]\} \\ &= E[(\mathbf{Z}_{1i}^p + \mathbf{Z}_{1i}^m)^T (\mathbf{Z}_{1j}^p + \mathbf{Z}_{1j}^m) \\ &\quad \times (\mathbf{Z}_{2i}^p + \mathbf{Z}_{2i}^m)^T (\mathbf{Z}_{2j}^p + \mathbf{Z}_{2j}^m)], \end{aligned} \tag{11}$$

$$\begin{aligned} \pi_{AD(ij)} &= E\{[(\mathbf{Z}_{1i}^p + \mathbf{Z}_{1i}^m) \otimes (\mathbf{Z}_{2i}^p \otimes \mathbf{Z}_{2i}^m)]^T \\ &\quad \times [(\mathbf{Z}_{1j}^p + \mathbf{Z}_{1j}^m) \otimes (\mathbf{Z}_{2j}^p \otimes \mathbf{Z}_{2j}^m)]\} \\ &= E[(\mathbf{Z}_{1i}^p + \mathbf{Z}_{1i}^m)^T (\mathbf{Z}_{1j}^p + \mathbf{Z}_{1j}^m) \\ &\quad \times (\mathbf{Z}_{2i}^p \mathbf{Z}_{2i}^m)^T (\mathbf{Z}_{2j}^p \mathbf{Z}_{2j}^m)], \end{aligned} \tag{12}$$

$$\begin{aligned} \pi_{DA(ij)} &= E\{[(\mathbf{Z}_{1i}^p \otimes \mathbf{Z}_{1i}^m) \otimes (\mathbf{Z}_{2i}^p + \mathbf{Z}_{2i}^m)]^T \\ &\quad \times [(\mathbf{Z}_{1j}^p \otimes \mathbf{Z}_{1j}^m) \otimes (\mathbf{Z}_{2j}^p + \mathbf{Z}_{2j}^m)]\} \\ &= E[(\mathbf{Z}_{1i}^p \mathbf{Z}_{1i}^m)^T (\mathbf{Z}_{1j}^p \mathbf{Z}_{1j}^m) \\ &\quad \times (\mathbf{Z}_{2i}^p + \mathbf{Z}_{2i}^m)^T (\mathbf{Z}_{2j}^p + \mathbf{Z}_{2j}^m)], \end{aligned} \tag{13}$$

$$\begin{aligned} \pi_{DD(ij)} &= E\{[(\mathbf{Z}_{1i}^p \otimes \mathbf{Z}_{1i}^m) \otimes (\mathbf{Z}_{2i}^p \otimes \mathbf{Z}_{2i}^m)]^T \\ &\quad \times [(\mathbf{Z}_{1j}^p \otimes \mathbf{Z}_{1j}^m) \otimes (\mathbf{Z}_{2j}^p \otimes \mathbf{Z}_{2j}^m)]\} \\ &= E[(\mathbf{Z}_{1i}^p \mathbf{Z}_{1i}^m)^T (\mathbf{Z}_{1j}^p \mathbf{Z}_{1j}^m) \\ &\quad \times (\mathbf{Z}_{2i}^p \mathbf{Z}_{2i}^m)^T (\mathbf{Z}_{2j}^p \mathbf{Z}_{2j}^m)]. \end{aligned} \tag{14}$$

In matrix notation, let $\mathbf{Y} = (y_1, y_2, \dots, y_N)^T$ be the vector of phenotypic values. The variance–covariance matrix of \mathbf{Y} is

$$\begin{aligned} \text{Var}(\mathbf{Y}) = \mathbf{V} &= \mathbf{\Pi}_{1A} \sigma_{1A}^2 + \mathbf{\Pi}_{1D} \sigma_{1D}^2 \\ &\quad + \mathbf{\Pi}_{2A} \sigma_{2A}^2 + \mathbf{\Pi}_{2D} \sigma_{2D}^2 \\ &\quad + \mathbf{\Pi}_{AA} \sigma_{AA}^2 + \mathbf{\Pi}_{AD} \sigma_{AD}^2 \\ &\quad + \mathbf{\Pi}_{DA} \sigma_{DA}^2 + \mathbf{\Pi}_{DD} \sigma_{DD}^2 + \mathbf{I} \sigma^2 \end{aligned} \tag{15}$$

where $\mathbf{\Pi}_{1A}$, $\mathbf{\Pi}_{1D}$, $\mathbf{\Pi}_{2A}$, $\mathbf{\Pi}_{2D}$, $\mathbf{\Pi}_{AA}$, $\mathbf{\Pi}_{AD}$, $\mathbf{\Pi}_{DA}$ and $\mathbf{\Pi}_{DD}$ are the IBD matrices, whose elements are given by equations (9)–(14).

Note that the IBD matrices are the expected values of some functions of the descent graph. Some of the functions are linear and some are quadratic or higher order power functions. The explicit expressions of the expectations for higher power functions are difficult to derive. Instead, we adopt the Monte Carlo method to approximate the expectation. A nice property of the Monte Carlo approximation is that as the Monte Carlo samples increase the approximation will approach the expectation (following the law of large numbers).

Results

The Monte Carlo method requires individuals to be entered into a table representing the pedigree in the chronological order. In this example, the identification numbers (ids) are already arranged in this order and thus members are entered into the pedigree according to their ids. The first column stores the individual ids, the second column stores the sex of the individuals, and the third and fourth columns store the ids of the two parents. The ids of the parents for founders are not known and thus they are entered into the table as 0, see Table 1 for details.

The method was split into the following five steps, which were repeated 10 000 times:

Step 1: Missing genotypes imputation: For any given marker locus, impute the missing genotypes to generate a legal genotype sample that is compatible with the observed marker genotypes.

Step 2: Randomly determine the phases for the founders at the first locus, then determine the allelic order for all the individuals at the first locus, and obtain the inheritance vector \mathbf{v}_1 for all the nonfounder individuals at the first marker.

Step 3: Sample the inheritance indicators sequentially from locus 2 to locus m (the last marker).

Step 4: Sample the inheritance indicators for the two targeted loci at positions 10 and 45 cM, and then convert the inheritance indicators of these two loci into the design vectors, \mathbf{Z}_{ki}^p and \mathbf{Z}_{ki}^m for $k = 1, 2$ and $i = 1, \dots, N$.

Step 5: Calculate the eight IBD matrices.

The 10 000 replicates took about 20 min in our Pentium PC. After the 10 000 repeated simulations were complete, we obtained the average value for each of the eight IBD matrices.

When we simulated the markers of the pedigree, we also simulated the allelic inheritances of the two targeted loci and recorded their descent graphs, and thus the true IBD matrices of the two loci. The Monte Carlo estimated IBD matrices are quite close to the true values (data not shown, see Tables 2–4).

Additional simulations were conducted in order to gain an understanding of its statistical properties. To measure the performance of the approach, we focused on one diagonal element and one off-diagonal element of the IBD matrices. We used the sample mean, bias, and standard deviation to give an intuitive measure of an estimator's effectiveness. From Tables 2 and 3, we can see that the bias becomes negligible quickly as the sample size increases, and the MSE decreases quickly accordingly.

Comparison with Markov chain Monte Carlo (MCMC) method

MCMC can be used to calculate the IBD additive matrix conditional on multiple markers, when marker phases are not known, and using all available information. The MCMC method employed in this paper is SIMWALK (Sobel and Lange, 1996). SIMWALK is a freely available software package capable of calculating IBD additive matrix on any size of pedigree. It uses MCMC and simulated annealing algorithms to perform these multi-point analyses. It is capable of handling substantial missing marker information. Sobel *et al* (2001) test the performance of the MCMC method implemented in SIMWALK.

The matrices calculated by our method and the MCMC method were compared directly to the matrix containing the true IBD matrices, which were known from the simulations in this study. The criterion for comparison was the mean square error (MSE) (Sorensen *et al*, 2002):

$$\text{MSE}_{(\text{calc})} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (q_{ij}^{(\text{calc})} - q_{ij}^{(\text{true})})^2$$

Table 2 Mean estimates (ave), bias, and standard deviation (std) of the 15th diagonal element of the IBD matrices calculated from 10 000 repeated simulations.

Sample size	π_{1A}	π_{2A}	π_{1D}	π_{2D}	π_{AA}	π_{AD}	π_{DA}	π_{DD}
1000								
ave	2.0033	2.5644	1	1	5.1368	2.0033	2.5644	1
bias	0.0033	0.0084	0	0	0.0243	0.0033	0.0084	0
std	0.0816	0.1002	0	0	0.3168	0.0816	0.1002	0
2000								
ave	2.0032	2.5633	1	1	5.1353	2.0032	2.5633	1
bias	0.0032	0.0073	0	0	0.0228	0.0032	0.0073	0
std	0.0799	0.0997	0	0	0.3149	0.0799	0.0997	0
4000								
ave	2.0028	2.5614	1	1	5.1296	2.0028	2.5614	1
bias	0.0028	0.0054	0	0	0.0171	0.0028	0.0054	0
std	0.0748	0.0988	0	0	0.3089	0.0748	0.0988	0
6000								
ave	2.0027	2.5607	1	1	5.1280	2.0027	2.5607	1
bias	0.0027	0.0047	0	0	0.0155	0.0027	0.0047	0
std	0.0730	0.0984	0	0	0.3073	0.0730	0.0984	0
8000								
ave	2.0025	2.5578	1	1	5.1200	2.0025	2.5578	1
bias	0.0025	0.0017	0	0	0.0075	0.0025	0.0015	0
std	0.0707	0.0969	0	0	0.2964	0.0707	0.0969	0
10 000								
ave	2.0000	2.5560	1	1	5.1130	2.0000	2.5560	1
bias	0.0000	0.0000	0	0	0.0005	0.0000	0.0000	0
std	0.0000	0.0965	0	0	0.2930	0.0001	0.0965	0

where n is the number of individuals, $\mathbf{Q}^{(\text{true})} = (q_{ij}^{(\text{true})})_{n \times n}$ is the true IBD matrix, and $\mathbf{Q}^{(\text{calc})} = (q_{ij}^{(\text{calc})})_{n \times n}$ is the calculated IBD matrix from either our Monte Carlo method or the MCMC method.

The MSEs of the eight IBD matrices calculated by the Monte Carlo method and MCMC method versus the corresponding true IBD matrices are given in Table 4, where case 1 is the case that the two targeted positions are also markers, that is, all individuals are typed at the positions, case 2 is the case that there are no genotypes at the two targeted positions and there is no missing marker information, and case 3 is the general case that there are no genotypes at the two targeted positions and with incomplete marker information. Our Monte Carlo method resulted in a reasonable MSE that is very close to MCMC method. As expected, MSE increased as the marker information content decreased.

Discussion

Heath (1997) developed an MCMC method for mapping QTL in arbitrarily complex pedigrees with substantial missing marker information. Implementation of the multi-site segregation sampler is via a program called Loki. Since the program was released, Loki has been modified for IBD probability calculation. George *et al* (2000) have incorporated this program into their two-step method of QTL mapping under the mixed model methodology. The Monte Carlo method developed here differs from that of Heath (1997) in two aspects: (1) Heath's method is an MCMC approach in which the sample of the next round is dependent on that of the previous round, whereas our method is a Monte Carlo

Table 3 Mean estimates (ave), bias, and standard deviation (std) of the (8, 20)th off-diagonal element of the IBD matrices calculated from 10000 repeated simulations

Sample size	π_{1A}	π_{2A}	π_{1D}	π_{2D}	π_{AA}	π_{AD}	π_{DA}	π_{DD}
1000								
ave	1.9933	1.0057	0.9924	0.0080	2.0048	0.0160	0.9968	0.0080
bias	-0.0067	0.0057	-0.0076	0.0080	0.0048	0.0160	-0.0032	0.0080
std	0.0814	0.1140	0.0866	0.0891	0.1410	0.0983	0.1079	0.0391
2000								
ave	1.9934	1.0057	0.9925	0.0075	2.0047	0.0148	0.9970	0.0073
bias	-0.0066	0.0057	-0.0075	0.0075	0.0047	0.0148	-0.0030	0.0073
std	0.0808	0.1043	0.0863	0.0863	0.1223	0.0911	0.1050	0.0353
4000								
ave	1.9937	1.0051	0.9926	0.0074	2.0040	0.0146	0.9973	0.0072
bias	-0.0063	0.0051	-0.0074	0.0074	0.0040	0.0146	-0.0027	0.0072
std	0.0793	0.1015	0.0857	0.0857	0.1168	0.0897	0.1026	0.0347
6000								
ave	1.9937	1.0047	0.9927	0.0072	2.0030	0.0143	0.9976	0.0071
bias	-0.0063	0.0047	-0.0073	0.0072	0.0030	0.0143	-0.0024	0.0071
std	0.0788	0.1009	0.0853	0.0846	0.1145	0.0882	0.1019	0.0340
8000								
ave	1.9940	1.0045	0.9933	0.0069	2.0025	0.0136	0.9980	0.0067
bias	-0.0060	0.0045	-0.0067	0.0069	0.0025	0.0136	-0.0020	0.0067
std	0.0773	0.0980	0.0814	0.0826	0.1106	0.0841	0.0999	0.0319
10000								
ave	1.9947	1.0030	0.9940	0.0065	2.0000	0.0130	0.9982	0.0065
bias	-0.0053	0.0030	-0.0060	0.0065	0.0000	0.0130	-0.0018	0.0065
std	0.0728	0.0930	0.0773	0.0804	0.1000	0.0808	0.0938	0.0304

Table 4 Comparisons of MSE of the eight IBD matrices calculated by our Monte Carlo method and MCMC method versus the true IBD matrices for the complicated pedigree with 22 individuals

Case	Method	Π_{1A}	Π_{2A}	Π_{1D}	Π_{2D}	Π_{AA}	Π_{AD}	Π_{DA}	Π_{DD}
1	MC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	MCMC	0.0331	0.0000	—	—	—	—	—	—
2	MC	0.0026	0.0001	0.0003	0.0000	0.0011	0.0000	0.0003	0.0000
	MCMC	0.0331	0.0001	—	—	—	—	—	—
3	MC	0.0851	0.0456	0.0090	0.0016	0.0592	0.0063	0.0053	0.0000
	MCMC	0.0750	0.0487	—	—	—	—	—	—

Case 1: Both loci are genotyped; case 2: Neither locus is genotyped with full marker information; case 3: Neither locus is genotyped with incomplete marker information.

method in which samples are independent, and (2) our method calculates the epistatic IBD matrices which allow mapping of epistatic QTL. Of course, the method of Heath (1997) can be further extended for calculating the epistatic IBD matrices because these IBD matrices are functions of the inheritance indicators. However, these functions were not given in Heath (1997) and they are introduced in this study for the first time. The multipoint method of Almasy and Blangero (1998) is potentially capable of calculating the epistatic IBD matrices through the Hadamard products of the locus-specific IBD matrices, that is, $\Pi_{AA} = \Pi_{1A} \odot \Pi_{2A}$, $\Pi_{AD} = \Pi_{1A} \odot \Pi_{2D}$, $\Pi_{DA} = \Pi_{1D} \odot \Pi_{2A}$ and $\Pi_{DD} = \Pi_{1D} \odot \Pi_{2D}$. However, these relationships hold only when the two loci are not linked in the same chromosome. There have been no general formulas available for incorporating the linkage parameters. The Monte Carlo method introduced here is general for two loci with any linkage relationship.

The IBD matrices provide the machinery for partitioning the total phenotypic variance into variance components. These variance components can be estimated and tested using the maximum likelihood (ML) or restricted maximum likelihood (REML) method. The SAS procedure, PROC MIXED, has the ability to directly incorporate these matrices for estimation of variance components (Littell *et al*, 2000). Therefore, estimation of variance components, especially the dominance and epistatic variance components, from arbitrarily complicated pedigree with incomplete marker information is now possible and realistic. Using complicated pedigrees to separate the nonadditive variance components from the additive components is more efficient because the nonadditive IBD matrices tend to have more nonzero elements in complicated pedigrees than in simple pedigrees. However, current methods for computing the nonadditive IBD matrices are not capable of handling

complicated pedigrees with high degree of inbreeding. The Monte Carlo method developed in this study provides such a tool.

MCMC is a powerful tool to use all available information for calculating additive IBD matrices in complex pedigrees. However, for very tight linkage, for example, with very dense marker maps, the mixing properties of existing MCMC methods deteriorate. In addition, convergence of MCMC is difficult to diagnose. Our Monte Carlo method can be used as an alternative when convergence of MCMC cannot be achieved. Furthermore, our method can calculate all the eight IBD matrices and may, therefore, play a role in searching for QTL along the genome in complicated pedigrees.

A computer program written in MATLAB is available to implement the Monte Carlo method for IBD matrix calculation. The program is free to academic researchers and can be downloaded from our website: www.statgen.ucr.edu.

Acknowledgements

The work was supported by the National Institutes of Health Grant R01-GM55321 and the USDA National Research Initiative Grants Program 00-35300-9245 to SX.

References

- Almasy L, Blangero J (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* **62**: 1198–1211.
- Almudevar A, Field C (1999). Estimation of single generation sibling relationships based on DNA markers. *J Agric Biol Environ Statist* **4**: 136–165.
- Cockerham CC (1971). Higher-order probability functions of identity of alleles by descent. *Genetics* **69**: 235–246.
- Cockerham CC (1980). Random and fixed effects in plant genetics. *Theor Appl Genet* **56**: 119–131.
- Cotterman CW (1940). A calculus for statistical genetics. PhD Thesis, Ohio State University. Ballonoff P (ed) (1975). (Published in *Genetics and Social Structure*. Bench-mark papers in Genetics, Dowden, Hutchinson and Ross).
- Denniston C (1974). An extension of the probability approach to genetic relationships, one locus. *Theor Popul Biol* **6**: 58–75.
- Fulker DW, Cardon LR (1994). A sib-pair approach to interval mapping of quantitative trait loci. *Am J Hum Genet* **54**: 1092–1103.
- Fulker DW, Cherny SS, Cardon LR (1995). Multipoint interval mapping of quantitative trait loci, using sib pairs. *Am J Hum Genet* **56**: 1224–1233.
- George AW, Visscher PM, Haley CS (2000). Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics* **156**: 2081–2092.
- Goldgar DE (1990). Multipoint analysis of human quantitative genetic variation. *Am J Hum Genet* **47**: 957–967.
- Haseeman JK, Elston RC (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* **2**: 3–19.
- Heath S (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* **61**: 748–760.
- Jacquard A (1972). Genetic information given by a relative. *Biometrics* **28**: 1101–1114.
- Jacquard A (1974). *The Genetic Structure of Populations*. Springer-Verlag: New York.
- Kruglyak L, Lander ES (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* **57**: 439–454.
- Lander ES, Botstein D (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.
- Li CC, Sacks L (1954). The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics* **10**: 347–360.
- Littell RC, Pendergast J, Natarajan R (2000). Modelling covariance structure in the analysis of repeated measures data. *Stat Med* **19**: 1793–1819.
- Lynch M, Walsh B (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates: Sunderland, MA, USA.
- Nadot R, Vaysseix G (1973). Apparentement et identité. Algorithme de calcul des coefficients d'identité. *Biometrics* **29**: 347–359.
- O'Connell JR, Weeks DE (1999). An optimal algorithm for automatic genotype elimination. *Am J Hum Genet* **56**: 1733–1740.
- Schork NJ (1993). Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. *Am J Hum Genet* **53**: 1306–1319.
- Sobel E, Lange K (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* **58**: 1323–1337.
- Sobel E, Sengul H, Weeks DE (2001). Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Hum Hered* **52**: 121–131.
- Sorensen AC, Pong-Wong R, Windig JJ, Woolliams JA (2002). Precision of methods for calculating identity-by-descent matrices using multiple markers. *Genet Sel Evol* **34**: 557–579.
- Thompson EA (1974). Gene identities and multiple relationships. *Biometrics* **30**: 667–680.
- Wang T, Fernando RL, Grossman M (1998). Genetic evaluation by best linear unbiased prediction using marker and trait information in a multibreed population. *Genetics* **148**: 507–515.
- Wang T, Fernando RL, van der Beek S, Grossman M, van Arendonk JAM (1995). Covariance between relatives for a marked quantitative trait locus. *Genet Sel Evol* **27**: 251–274.
- Weir BS, Avery PJ, Hill WG (1980). Effect of mating structure on variation in inbreeding. *Theor Popul Biol* **18**: 396–429.
- Weir BS, Cockerham CC (1969). Inbreeding with two linked loci. *Genetics* **63**: 711–742.
- Weir BS, Hill WG (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**: 477–488.
- Whitlock MC, Phillips PC, Wade MJ (1993). Gene interaction affects the additive genetic variance in subdivided populations with migration and extinction. *Evolution* **47**: 1758–1769.
- Whittemore AS, Halpern J (1994). Probability of gene identity by descent: computation and applications. *Biometrics* **50**: 109–117.
- Xu S (1998). Iteratively reweighted least squares mapping of quantitative trait loci. *Behav Genet* **28**: 341–355.
- Yi N, Xu S (2000). Bayesian mapping of quantitative trait loci under the identity-by-descent-based variance component model. *Genetics* **156**: 411–422.

Appendix A1. Sample legal descent graph for markers with incomplete information

Phase known

Phase known means that the parental marker haplotypes are known. In this case, the marker genotype of one progeny is independent of that of another progeny. Therefore, the probability of marker genotype can be calculated one individual at a time. We start sampling the genotype of the first marker and then the second marker conditional on the first marker and so on, taking advantage of the Markov property. The observed form of

each marker is used to make sure that the genotype sampled is compatible with the data. This approach of sampling is a hidden Markov model approach, with the marker genotype as the hidden true state and the form of the marker as the observed state.

Let us consider a parent-offspring trio and define the father's marker type by a_1b_1 and the mother's marker type by c_1d_1 , where a_1 and b_1 are the paternal and maternal allele types of the father, respectively, and c_1 and d_1 are the paternal and maternal allele types of the mother, respectively. Let (e_1f_1) be the marker type of the progeny in question. The subscript 1 means the first marker of the chromosome under investigation. The parentheses used for the marker type of the progeny mean that the phase or order of the alleles is unknown. The probabilities of the four genotypes are calculated as follows. The four possible configurations of the progeny generated from this mating type are a_1c_1 , a_1d_1 , b_1c_1 and b_1d_1 . We now compare (e_1f_1) with the four configurations. A particular configuration, say k , will be assigned $s_k = 1$ if e_1f_1 or f_1e_1 matches this configuration, otherwise, $s_k = 0$. When all the four s_k 's have been assigned a value, we normalize these s_k 's to obtain

$$p_k = \frac{s_k}{\sum_{k'} s_{k'}} \quad \text{for } k = 1, \dots, 4 \quad (\text{A1})$$

These are the probabilities of the four marker genotypes conditional on the observed marker forms. From these probabilities, we can simulate a realized genotypic configuration. Taking the same approach, we can obtain complete inheritance indicators for all members of the pedigree at the first marker. So far, we have obtained the vector of inheritance indicators of the first locus for the entire pedigree, denoted by \mathbf{v}_1 as defined in equation (1), where the subscript indicates the locus. Conditional on the inheritance indicators of the first marker, we now proceed with the sampling of the configurations at the second marker. Let us denote the observed forms of the parental markers by a_2b_2 and c_2d_2 , and the marker phenotype of the progeny by (e_2f_2) for the second marker. Again, we can check the compatibility of the progeny type with the four possible configurations, a_2c_2 , a_2d_2 , b_2c_2 and b_2d_2 , and define for s_k for $k=1, \dots, 4$ in the same manner as the first marker. We now calculate the probability of the four marker genotypes. This time, the sampling is conditioned on the first marker state using the equation given below:

$$p_k = \frac{s_k \Pr(M_2 = k | M_1 = l)}{\sum_{k'} s_{k'} \Pr(M_2 = k' | M_1 = l)} \quad (\text{A2})$$

where M_1 and M_2 denote the genotypes of the first and second markers, respectively, and each can take a value between 1 and 4, depending on which of the four genotypic configurations has been taken by the individual. The probability $\Pr(M_2 = k | M_1 = l)$ is the transition probability between the two markers and the value is a function of the recombination fraction between the two markers, which is found from the 4×4 transition matrix given by Xu (1998). The probability given in Equation (3) is used to sample the genotype of the second marker. The inheritance indicators for all members in the pedigree, \mathbf{v}_2 , are obtained in the same way. Given \mathbf{v}_2 , we are now ready to sample the genotypic configuration of the third

marker \mathbf{v}_3 . The process continues until all markers have been sampled for all individuals, and we now have a complete descent graph for each of the m markers, $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$.

Phase unknown

The phase of all the founders at the first marker of a chromosome is irrelevant to the genetic analysis. Therefore, we can arbitrarily assign a phase of the marker type to all the founders at the first marker. The genotype of each nonfounder at the first marker is sampled using the method described above. We now need to sample the genotypes of the second marker for all individuals and the founder marker linkage phase simultaneously. Since there are two different phases for each founder at the second marker, totally there are 2^n different phases for n founders at the second marker. All the 2^n different phases are evaluated and the probability of each phase is calculated. These probabilities are then used to sample a particular phase.

Under each phase, we work out the four possible genotypic configurations in each parent-offspring trio. For example, if the phase for a parent-offspring trio is $b_2a_2 \times c_2d_2$, the four possible genotype configurations of a progeny will be b_2c_2 , b_2d_2 , a_2c_2 and a_2d_2 . Using the genotype of the first marker already sampled and the compatibility of the type of the second marker (e_2f_2) to the four configurations, we can calculate the probabilities of the genotypes at the second marker. With these probabilities, we sample a particular genotype. The procedure of sampling the genotypic configuration for the second marker is applied to all members in the pedigree. Given \mathbf{v}_1 and \mathbf{v}_2 , the inheritance indicators of the first and the second markers, respectively, for all individuals in the pedigree, we can count the total number of recombinations, denoted by

$$h = \|\mathbf{v}_2 - \mathbf{v}_1\|_1$$

where $\|\cdot\|_1$ is the vector 1-norm. This number is obtained conditional on the particular marker linkage phase in the founders under investigation. For n founders, there are a total of 2^n possible phases and each one of them will be evaluated in the same way. Eventually, we will obtain 2^n such numbers, denoted by h_i for $i = 1, \dots, 2^n$. The probability of the i th phase is

$$q_i = \frac{(1-r)^{2N-h_i} r^{h_i}}{\sum_{k=1}^{2^n} (1-r)^{2N-h_k} r^{h_k}} \quad \forall i = 1, \dots, 2^n \quad (\text{A3})$$

where r is the recombination fraction between markers one and two. These probabilities are used to sample the phase. The inheritance indicators of the second marker, \mathbf{v}_2 , under this sampled phase will be accepted along with the phase. The process continues until all $\mathbf{v}_1, \dots, \mathbf{v}_m$ have been sampled for all m loci.

Missing markers

Missing markers are common in pedigree analysis. O'Connell and Weeks (1999) give a genotype-elimination algorithm that uses all individuals in the pedigree, but does not work for pedigrees with more than a few loops. In this subsection, we describe our algorithm of imputing the missing genotypes and sampling a legal genotypic configuration. The algorithm starts from the descents to

the founders (bottom-up) to replace each missing marker genotype by a legal genotype.

First we decompose the pedigree into individual nuclear families or sibships, ordered in reverse pedigree order (ie, progeny before parents), then we impute the missing parental marker genotypes on the basis of the observed genotypes in each nuclear family sequentially. Two situations have to be distinguished: (1) when both parental genotypes are missing and (2) when one parental genotype is missing but the other parental genotype has been typed.

When both parental genotypes are missing: We begin by the set of observed genotypes among a sibship. We regard two sets of observed genotypes among a sibship

to be equivalent if one can be obtained from the other through an appropriate substitution of allele from the set of population alleles. There are 14 equivalence classes determined by the mechanism of Mendelian inheritance (Almudevar and Field, 1999). Table A1, which is based on Table 11 in Almudevar and Field (1999), lists the possible parental genotypes of the interested parent for each of the 14 equivalence classes of observed genotypes among a sibship.

When one parental genotype is missing: Almudevar and Field (1999) do not distinguish between families for which both parental genotypes are missing and families with only one missing parental genotype. To impute parental genotypes, however, such partial information

Table A1 Imputation parental genotype pairs for 14 equivalence offspring genotype classes, when both parental genotypes are missing

Type	Given offspring genotypes	Possible genotypes pair of parents
A1	{<a,d>,<b,d>,<b,c>,<a,c>}	{(<a,b>,<c,d>), (<c,d>,<a,b>)}
A2	{<a,d>,<b,c>,<a,c>}	
B	{<a,d>,<b,c>}	{(<a,b>,<c,d>), (<c,d>,<a,b>), (<a,c>,<b,d>), (<b,d>,<a,c>)}
C1	{<a,b>,<b,c>,<a,c>,<b,b>}	
C2	{<a,b>,<b,c>,<b,b>}	{(<a,b>,<c,b>), (<c,b>,<a,b>)}
C3	{<a,c>,<b,c>,<b,b>}	
C4	{<a,c>,<b,b>}	
D	{<a,b>,<b,c>,<a,c>}	{(<a,b>,<c,b>), (<c,b>,<a,b>), (<b,a>,<c,a>), (<c,a>,<b,a>), (<a,c>,<b,c>), (<b,c>,<a,c>)} {(<a,c>,<x,b>), (<x,b>,<a,c>): x ∈ A} ∪{(<a,b>,<b,c>), (<b,c>,<a,b>)}
E	{<a,b>,<b,c>}	{(<a,b>,<a,b>)}
F1	{<a,a>,<b,b>,<a,b>}	
F2	{<a,a>,<b,b>}	
G	{<a,b>,<b,b>}	{(<a,b>,<x,b>), (<x,b>,<a,b>): x ∈ A}
H	{<a,b>}	{(<a,x>,<b,y>), (<b,x>,<a,y>): x,y ∈ A}
I	{<a,a>}	{(<a,x>,<a,y>): x, y ∈ A}

Here, **A** is the population set of alleles. We use <a,b> to denote an unordered genotype, (<a,b>,<c,d>) to denote unordered genotype pair of the parent, that is, where <a,b> is the paternal unordered genotype and <c,d> is the maternal unordered genotype.

Table A2 Imputation of the parental genotype for the equivalent offspring genotype classes, when only one parental genotype is missing

Type	Given offspring genotypes	Given spouse genotype	Possible parental genotype
A1	{<a,d>,<b,d>,<b,c>,<a,c>}	<a,b> <c,d>	<c,d> <a,b>
A2	{<a,d>,<b,c>,<a,c>}		Same as A1
B	{<a,d>,<b,c>}	<a,b> <c,d> <a,c> <b,d>	<c,d> <a,b> <b,d> <a,c>
C1	{<a,b>,<b,c>,<a,c>,<b,b>}	<a,b> <c,b>	<c,b> <a,b>
C2	{<a,b>,<b,c>,<b,b>}		Same as C1
C3	{<a,c>,<b,c>,<b,b>}		Same as C1
C4	{<a,c>,<b,b>}		Same as C1
D	{<a,b>,<b,c>,<a,c>}	<a,b> <a,c> <b,c>	{<a,c>,<b,c>} {<a,b>,<b,c>}
E	{<a,b>,<b,c>}	<a,c> <y,b> (y ∈ A, y ≠ a,c) <a,b> <b,c>	{<a,b>,<a,c>} {<x,b>: x ∈ A} <a,c> {<b,c>,<a,c>}
G	{<a,b>,<b,b>}	<a,b> <y,b> (y ∈ A, y ≠ a)	{<a,b>,<a,c>} {<x,b>: x ∈ A}
H	{<a,b>}	<a,x> (x ∈ A, x ≠ b) <b,x> (x ∈ A, x ≠ a) <a,b>	<a,b> {<b,y>: y ∈ A} {<a,y>: y ∈ A} {<a,y>, <b,y>: y ∈ A}

Here **A** is the population set of alleles. We use <a,b> to denote an unordered genotype. Spouse genotype gives no more information for classes F1, F2 and I, and hence not list here.

can be taken into account. Table A2, which is organized analogously to Table A1, lists all parental mating types for the 14 equivalence classes of observed genotypes among a sibship.

We use the following three steps to impute the missing genotypes.

- (1) Determine the possible alleles set a missing genotype at each locus can take. If the paternal (maternal) parent is typed, then the possible paternal (maternal) allele set is the paternal (maternal) parent's genotype at the locus, otherwise it is the population set of alleles of the pedigree at the locus.
- (2) Using Table A1, infer a missing genotype at each locus in each parent, conditional on genotypes in offspring in each parent–offspring trio. It should be noticed that the possible missing genotype set of the parent should be compatible with the alleles set obtained in step 1.
- (3) Using Table A2, infer a missing genotype at each locus in each parent, conditional on genotypes in spouse and offspring in each parent–offspring trio. The possible missing genotype set of the parent

should be compatible with the alleles set obtained in step 1.

It is common that several candidate marker genotypes may be compatible with the data and we randomly choose one. Conditional on the simulated legal genotype, we proceed with replacing missing genotypes of members in the upper level of the pedigree until all the missing genotypes have been filled with legal genotypes. We then sample the linkage phases and the inheritance indicator vectors conditional on the current legal genotypes. Since the sampling processes are independent, the next cycle of sampling may start with completely different legal genotypes. If the Monte Carlo sample is sufficiently large, all possible legal genotypes may be tried. Fortunately, many different legal genotypes will lead to the same IBD matrices, and thus there is little concern about not trying the exhaustive list of legal genotypes. The descent graphs of multiple linked markers are then used to sample the descent graph of an arbitrary locus or the joint descent graph of two arbitrary loci.