DOI: 10.1534/genetics.108.089912

Identity-by-Descent Estimation and Mapping of Qualitative Traits in Large, Complex Pedigrees

Mark Abney[1]

Department of Human Genetics, University of Chicago, Chicago, Illinois 60637

Manuscript received April 4, 2008
Accepted for publication May 2, 2008

ABSTRACT

Computing identity-by-descent sharing between individuals connected through a large, complex pedigree is a computationally demanding task that often cannot be done using exact methods. What I present here is a rapid computational method for estimating, in large complex pedigrees, the probability that pairs of alleles are IBD given the single-point genotype data at that marker for all individuals. The method can be used on pedigrees of essentially arbitrary size and complexity without the need to divide the individuals into separate subpedigrees. I apply the method to do qualitative trait linkage mapping using the nonparametric sharing statistic S_{pairs}. The validity of the method is demonstrated via simulation studies on a 13-generation 3028-person pedigree with 700 genotyped individuals. An analysis of an asthma data set of individuals in this pedigree finds four loci with *P*-values $<10^{-3}$ that were not detected in prior analyses. The mapping method is fast and can complete analyses of ~150 affected individuals within this pedigree for thousands of markers in a matter of hours.

COMPUTATION of identical-by-descent (IBD) allele sharing between related individuals is a necessary ingredient in many methods for linkage mapping of complex traits. Typically, IBD allele sharing is used either directly to assess whether affected individuals are sharing more at a locus than expected under the null hypothesis or as a component in the covariance matrix in a variance component model. A number of algorithms for computing IBD exactly exist (*e.g.*, Elston and Stewart 1971; Lander and Green 1987; Kruglyak *et al.* 1996; Fishelson and Geiger 2002); however, these methods become computationally infeasible when pedigrees are very large and complex. Under such circumstances approximate methods become necessary, whether Markov chain Monte Carlo (Thompson *et al.* 1993; Sobel and Lange 1996; Heath 1997) or regression based (Fulker *et al.* 1995; Almasy and Blangero 1998). Even these methods, however, have difficulty when the pedigree is very deep with many generations of individuals with no data.

In humans, very deep, and possibly complex, pedigrees often arise in conjunction with genetic studies of isolated populations. Isolated populations are commonly thought to have characteristics that may prove advantageous for mapping (Wright *et al.* 1999; Peltonen *et al.* 2000; Escamilla 2001; Shifman and Darvasi 2001; Service *et al.* 2006), yet may require specialized statistical methods to both properly leverage these advantages and provide a valid test for the presence of a trait-influencing gene (Bourgain and Genin 2005). Large pedigrees also arise in other animal systems where breeding is carefully controlled. For example, there is interest in methods that are applicable to complex pedigrees for both livestock (Thallman *et al.* 2001) and dogs (Sutter and Ostrander 2004).

What I present here is a rapid computational method for estimating, in large complex pedigrees, the probability that pairs of alleles are IBD given the single-point genotype data at that marker for all individuals. Because the method is very fast, it can easily be used on genomewide data with many thousands of markers on hundreds of related individuals. It can be used directly to do linkage mapping with affected individuals using the S_{pairs} statistic or to compute approximate multipoint probabilities both for alleles being IBD, using regression-based approaches (*e.g.*, Almasy and Blangero 1998), and for alleles being homozygous by descent (HBD) using a hidden Markov model (HMM) (Abney *et al.* 2002). Here, I describe this computational method and its application to qualitative trait linkage analysis. Although computing S_{pairs} is straightforward, in principle, a number of challenges must be overcome in creating a practical and valid mapping method for very large, and possibly complex, pedigrees. In particular, it is common in studies involving large pedigrees to have one, or a few, pedigrees, making the asymptotic distribution of the test statistic, which is appropriate when there are many independent pedigrees, not necessarily applicable. Also, the allele-frequency distribution may have a major influence on the test statistic when

[1]*Address for correspondence:* Department of Human Genetics, University of Chicago, 920 E. 58th St., Chicago, IL 60637.
E-mail: abney@genetics.uchicago.edu

inheritance information is obscured by missing data. Unfortunately, the relevant allele-frequency distribution is that in the founders of the pedigree, which in large pedigrees may be many generations earlier than the sampled individuals. As a result, estimation of the founder allele-frequency distribution from the sampled data can result in a large bias of the conditional expected sharing statistic. The difficulties posed by not knowing the true allele-frequency distribution can be largely overcome through the use of simulations, but the capacity to do many simulations requires a computationally efficient method, particularly when a large number of markers are involved. Below, I describe the theoretical basis of the IBD estimation method, the approximations used, and how it differs from earlier methods that take a similar approach (Wang *et al.* 1995; Davis *et al.* 1996). I then show its application to single-point linkage mapping using S_{pairs} and how the difficulties mentioned above are solved. The appendix describes how to use the IBD estimation method to obtain multipoint estimates of HBD by modifying the HMM of Abney *et al.* (2002).

METHODS

IBD estimation

The objective is to compute the probability of two alleles being IBD given all available genotype data at that locus and the entire, unbroken pedigree. The method is based on the recursive strategy suggested by Wang *et al.* (1995) and Davis *et al.* (1996). In both of these studies the probability is computed in a manner analogous to the recurrence relation for kinship coefficients, $\phi_{\mathscr{A}\mathscr{B}} = \frac{1}{2}\phi_{\mathscr{M}\mathscr{B}} + \frac{1}{2}\phi_{\mathscr{F}\mathscr{B}}$, where $\mathscr{M}$ and $\mathscr{F}$ are the mother and the father of individual $\mathscr{A}$, and individual $\mathscr{B}$ is not a descendant of $\mathscr{A}$. The equivalent recurrence equation when there are genotype data at the locus, as given by Wang *et al.* (1995) and Davis *et al.* (1996), is valid only when there are no missing genotypes and is

$$\begin{aligned}\Pr(A_1 \equiv B_1 \mid G) = {} & \Pr(A_1 \leftarrow M_1 \mid G)\Pr(M_1 \equiv B_1 \mid G) \\ & + \Pr(A_1 \leftarrow M_2 \mid G)\Pr(M_2 \equiv B_1 \mid G) \\ & + \Pr(A_1 \leftarrow F_1 \mid G)\Pr(F_1 \equiv B_1 \mid G) \\ & + \Pr(A_1 \leftarrow F_2 \mid G)\Pr(F_2 \equiv B_1 \mid G), \quad (1)\end{aligned}$$

where $\Pr(A_i \leftarrow M_j|G)$ is the probability that the ith allele from $\mathscr{A}$ was inherited from the jth allele of $\mathscr{A}$'s mother, given the observed genotype data; and $A_i \equiv B_j$ means allele A_i is IBD with allele B_j. This equation is applied repeatedly until the founders of the pedigree are reached and boundary conditions are used to obtain the probability. When there are no individuals with missing genotypes the method is both fast and returns exact probabilities.

Unfortunately, as recognized by both Wang *et al.* (1995) and Davis *et al.* (1996), Equation 1 is not valid when there are missing genotypes in the data. Although neither study formulated a version of Equation 1 that holds under missing data conditions, they each suggested approaches for this case. The most recent version of SimIBD (Davis *et al.* 1996) uses a Monte Carlo procedure where, for each realization, a random genotype is assigned to each missing genotype, and the recursive algorithm is applied. The final probability is the average of the probabilities computed at each Monte Carlo realization. In contrast, Wang *et al.* (1995) suggest two different possibilities. In the first, when the recursive algorithm encounters an individual who has a missing genotype, the relevant inheritance probability [*e.g.*, $\Pr(A_1 \leftarrow M_1|G)$] is computed by summing over all possible genotypes for the missing data weighted by the probability of the genotype given the observed genotypes. To simplify the computation, one can use the probability of the genotype given only the genotypes of close relatives rather than all observed genotypes. The second possibility is to find the genotype configuration for all individuals with missing genotypes that has the highest probability and apply the recursive algorithm to that configuration. Finding the highest-probability genotype configuration, however, can be computationally demanding if there are many missing genotypes.

A common situation when analyzing large pedigrees is to have several generations of the pedigree completely untyped. None of the above strategies are entirely sufficient in such a situation. The problem is that there is little information in the untyped portion of the genealogy from which to infer the genotype probability distribution in those individuals. Simulating over valid genotype configurations can then be time consuming and, possibly, inaccurate. Summing over all possible genotypes, on the other hand, may be computationally impracticable.

The approach I propose relies on classifying individuals into two groups, *A* and *S*. An *S* individual is someone who either is genotyped or has at least one ancestor who is genotyped, while an *A* individual is someone who is not among the *S* group. Note that by this definition the *S* group may contain individuals for whom no data were actually collected. I also define a set of individuals called "quasi-founders," where each quasi-founder is either an *S* individual with both parents in the *A* group or an *A* individual with a spouse in *S*. A version of recurrence Equation 1, reformulated to hold true even under missing data, is applied to the *S* individuals until the quasi-founders are reached, at which point boundary conditions are employed to determine the final probability. This allows one to avoid using the recurrence equation over those generations with no genotype data, thereby speeding up the computation significantly. Furthermore, additional computational efficiency is gained by applying approximations designed specifically to work well when the rate of missing genotype data in *S* is reasonably low (*e.g.*, <20%). Note that this

constraint on the rate of missing genotype data in S still allows for potentially many generations of untyped individuals in A.

The algorithm is described in four parts. First, I describe the general form of Equation 1 and how to use this as a recurrence relation by updating the genotype information the probabilities are conditional on. I then show how the conditional probability of two alleles being IBD given some genotype information should be expressed when the allelic type of either of those alleles is unknown. This provides a general expression that can be applied recursively to compute the IBD probability. Applying this expression requires computing transmission probabilities in the presence of missing data. I derive an equation for calculating this probability and describe the approximations made to assure computational efficiency. Finally, the recursive algorithm is completed by specifying the boundary conditions to the recurrence equations.

Recurrence rules: The following notation is used throughout the remainder of this article. Individuals are indicated with uppercase script characters (*e.g.*, $\mathscr{A}, \mathscr{B}, \mathscr{M}, \mathscr{F}$), while the true genotype of, for instance, individual $\mathscr{A}$ comprises two random variables (A_1, A_2), where the alleles A_1 and A_2 may each take on one of the L possible allelic types at the locus, $\nu_1, \ldots, \nu_L$. Note that the ordering of A_1 and A_2 is arbitrary. Throughout, I assume that the pattern of missing genotype data is noninformative and that there is no genotyping error. Hence, observed allelic types indicate the true underlying genotype whereas the event that a genotype is missing provides no information, by itself, on the true genotype. Furthermore, the entire analysis is done conditional on the pattern of missing data. I let G represent the genotype information, which, as I show below, will grow during the course of the algorithm. Then, $G = g^r$ represents the information at the rth stage, where g^r is a vector with two elements for each quasi-founder and for each person in S, where the element $g^r_{\mathscr{A},1} = \nu_i$ if the allelic type of the first allele of individual $\mathscr{A}$ at the rth stage is known to be ν_i (*i.e.*, $A_1 = \nu_i$ at stage r) or is equal to zero if unknown. The vector g^0, then, has elements representing all directly observed genotype data or data that can be inferred without ambiguity. The vector $1_{\mathscr{A}}$ has the same length as g^r with entries equal to zero at all locations except for the two elements representing the alleles of $\mathscr{A}$. Then, for instance, $1_{\mathscr{A}} \cdot q^r$, where $\cdot$ is the inner product, is a vector with entries for $\mathscr{A}$ equal to the corresponding entries of g^r and all other entries equal to zero.

To extend Equation 1 to the case when some genotypes are missing, first note that it includes conditional probabilities for descent events involving only one allele at a time from $\mathscr{A}$ (*e.g.*, $\{A_1 \leftarrow M_1\}$, $\{A_2 \leftarrow M_2\}$, etc.). In fact, if A_1 came from the mother, for instance, then A_2 must have come from the father. A version of Equation 1 that includes the descent events for the other allele and is true even with missing data is

$$\begin{aligned}
&\Pr(A_1 \equiv B_1 \mid G)\\
&\quad = \Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_1 \mid G)\Pr(M_1 \equiv B_1 \mid A_1 \leftarrow M_1, A_2 \leftarrow F_1, G)\\
&\qquad + \Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_2 \mid G)\Pr(M_1 \equiv B_1 \mid A_1 \leftarrow M_1, A_2 \leftarrow F_2, G)\\
&\qquad + \Pr(A_1 \leftarrow M_2, A_2 \leftarrow F_1 \mid G)\Pr(M_2 \equiv B_1 \mid A_1 \leftarrow M_2, A_2 \leftarrow F_1, G)\\
&\qquad + \Pr(A_1 \leftarrow M_2, A_2 \leftarrow F_2 \mid G)\Pr(M_2 \equiv B_1 \mid A_1 \leftarrow M_2, A_2 \leftarrow F_2, G)\\
&\qquad + \Pr(A_1 \leftarrow F_1, A_2 \leftarrow M_1 \mid G)\Pr(F_1 \equiv B_1 \mid A_1 \leftarrow F_1, A_2 \leftarrow M_1, G)\\
&\qquad + \Pr(A_1 \leftarrow F_1, A_2 \leftarrow M_2 \mid G)\Pr(F_1 \equiv B_1 \mid A_1 \leftarrow F_1, A_2 \leftarrow M_2, G)\\
&\qquad + \Pr(A_1 \leftarrow F_2, A_2 \leftarrow M_1 \mid G)\Pr(F_2 \equiv B_1 \mid A_1 \leftarrow F_2, A_2 \leftarrow M_1, G)\\
&\qquad + \Pr(A_1 \leftarrow F_2, A_2 \leftarrow M_2 \mid G)\Pr(F_2 \equiv B_1 \mid A_1 \leftarrow F_2, A_2 \leftarrow M_2, G), \qquad (2)
\end{aligned}$$

where $G = g^r$ for all terms. This equation is valid as long as $\mathscr{B}$ and $\mathscr{A}$ are not the same individual and $\mathscr{B}$ is not a descendant of $\mathscr{A}$. If $\mathscr{A}$ and $\mathscr{B}$ are the same individual the equation becomes

$$\begin{aligned}
&\Pr(A_1 \equiv A_2 \mid G)\\
&\quad = \Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_1 \mid G)\Pr(M_1 \equiv F_1 \mid A_1 \leftarrow M_1, A_2 \leftarrow F_1, G)\\
&\qquad + \Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_2 \mid G)\Pr(M_1 \equiv F_2 \mid A_1 \leftarrow M_1, A_2 \leftarrow F_2, G)\\
&\qquad + \Pr(A_1 \leftarrow M_2, A_2 \leftarrow F_1 \mid G)\Pr(M_2 \equiv F_1 \mid A_1 \leftarrow M_2, A_2 \leftarrow F_1, G)\\
&\qquad + \Pr(A_1 \leftarrow M_2, A_2 \leftarrow F_2 \mid G)\Pr(M_2 \equiv F_2 \mid A_1 \leftarrow M_2, A_2 \leftarrow F_2, G). \qquad (3)
\end{aligned}$$

Unlike Equation 1, Equations 2 and 3 are not strictly recurrence equations because the IBD probabilities on the right-hand side have additional descent conditions not present in the left-hand side probability. In the case of no missing genotype data, the equations may be applied recursively by noting that terms such as $\Pr(M_1 \equiv B_1 | A_1 \leftarrow M_1, A_2 \leftarrow F_1, G = g^r) = \Pr(M_1 \equiv B_1 | G = g^r)$ on the right-hand side of Equations 2 and "expanding" these terms using the appropriate recurrence relation. Also, the descent probabilities $\Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_1 | G = g^r)$, etc., are easily tabulated on the basis of the possible genotype configurations of $\mathscr{A}$, $\mathscr{M}$, and $\mathscr{F}$. When there are missing genotypes, it is still possible to employ a recursive method based on Equations 2 and 3 by updating G with the genotype information provided by the descent events (*e.g.*, $A_1 \leftarrow M_1, A_2 \leftarrow F_1$). To show this I describe the application of the updating scheme to the first term on the right-hand side of Equation 2, but the arguments apply equally well to all terms on the right-hand side. First, focus on the conditional IBD probability $\Pr(M_1 \equiv B_1 | A_1 \leftarrow M_1, A_2 \leftarrow F_1, G)$, keeping in mind that Equation 2 holds only when $\mathscr{B}$ is neither $\mathscr{A}$ nor a descendant of $\mathscr{A}$. If $\mathscr{A}$ has a known genotype but either $\mathscr{F}$ or $\mathscr{M}$ does not, then the additional information from the conditions $A_1 \leftarrow M_1$ and $A_2 \leftarrow F_1$ must be included in the probability calculation. If, for example, $\mathscr{M}$ and $\mathscr{A}$ have known genotypes, $\mathscr{F}$ does not, and $A_2 = \nu_l$, then $\Pr(M_1 \equiv B_1 | A_1 \leftarrow M_1, A_2 \leftarrow F_1, G = g^r) = \Pr(M_1 \equiv B_1 | F_1 = \nu_l, G = g^r) = \Pr(M_1 \equiv B_1 | G = g^{r+1})$, where g^{r+1} is identical to g^r, but with component $g^{r+1}_{\mathscr{F},1} = \nu_l$. The subsequent applications of the recurrence Equation 2 from this term must be done conditional on $G = g^{r+1}$ rather than $G = g^r$. Note that this implies that the computations must allow for the case of a partially known genotype, as F_2 is known but F_1 may not be.

In general, then, the IBD probabilities $\Pr(A_1 \equiv B_1 \mid G = g^r)$ and $\Pr(A_1 \equiv A_2 \mid G = g^r)$ are conditional on both the observed genotype data and the additional

genotype information that results from the previous application of recurrence Equations 2 and 3. Even with the additional information, however, it is possible for A_1 or B_1 to be unknown. In this case, the probabilities must be written as a sum over the allelic types for the unknown alleles before the recurrence equations are applied. So, if A_1 is unknown and $g^r_{\mathscr{B},1} = \nu_k$,

$$\begin{aligned}\Pr(A_1 \equiv B_1 \mid G = g^r) &= \sum_{i,j} \Pr(A_1 \equiv B_1 \mid G = g^r, A_1 = \nu_i, A_2 = \nu_j) \\ &\quad \cdot \Pr(A_1 = \nu_i, A_2 = \nu_j \mid G = g^r) \\ &= \Pr(A_1 \equiv B_1 \mid G = g^r, A_1 = \nu_k) \Pr(A_1 = \nu_k \mid G = g^r). \end{aligned} \tag{4}$$

Note that in this equation $\Pr(A_1 \equiv B_1 \mid G = g^r, A_1 = \nu_k) = \Pr(A_1 \equiv B_1 \mid G = g^{r+1})$. Then, to compute the probability $\Pr(A_1 \equiv B_1 \mid G = g^r)$ one applies Equation 2 to the right-hand side of Equation 4, obtaining

$$\begin{aligned}\Pr(A_1 \equiv B_1 \mid G = g^r) &= \Pr(A_1 = \nu_k \mid G = g^r) \\ &\quad \cdot [\Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_1 \mid G = g^{r+1}) \\ &\quad \cdot \Pr(M_1 \equiv B_1 \mid A_1 \leftarrow M_1, A_2 \leftarrow F_1, G = g^{r+1}) + \ \ldots]. \end{aligned} \tag{5}$$

In general, if both A_1 and B_1 are unknown, the summation would be over all possible values of both A_1 and B_1. Doing such a sum would require applying the recursive algorithm to all terms in the sum, which can be computationally expensive even when there are few alleles. When the missing genotype rate is low this will occur infrequently, and instead of summing both A_1 and B_1 over all alleles, both alleles are left as unknown and Equation 5 reduces to Equation 2. This strategy is, in effect, equivalent to computing the probabilities conditional only on genotype information ancestral to $\mathscr{A}$ and $\mathscr{B}$ (and the other alleles of $\mathscr{A}$ and $\mathscr{B}$ if known) and, in practice, generally serves as a very good approximation.

Equation 5 provides a general recurrence equation that may be applied recursively to determine the probability that A_1 and B_1 are IBD, as long as $\mathscr{B}$ and $\mathscr{A}$ are not the same individual and $\mathscr{B}$ is not a descendant of $\mathscr{A}$. If $\mathscr{A}$ and $\mathscr{B}$ are the same individual, Equation 3 may be generalized similarly. To compute the IBD probability, it is necessary to determine the transmission probabilities $\{\Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_1 \mid G = g^{r+1}), \ldots\}$ in the presence of missing data. The derivation of these probabilities and the approximations made are described in Equations 6–10. Although only the probability $\Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_1 \mid G = g^{r+1})$ is computed here, extending this derivation to the other transmission probabilities is straightforward.

To compute this probability we must consider the case where A_1 or A_2 may be unknown. If $g^{r+1}_{\mathscr{A},2} = 0$, one must sum over possible values of A_2,

$$\begin{aligned}&\Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_1 \mid G = g^{r+1}) \\ &\quad = \sum_i \Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_1 \mid G = g^{r+1}, A_2 = \nu_i) \\ &\qquad \cdot \Pr(A_2 = \nu_i \mid G = g^{r+1}) \end{aligned} \tag{6a}$$

$$\begin{aligned}&\approx \sum_i \Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_1 \mid G = [1_{\mathscr{M}} + 1_{\mathscr{F}} + 1_{\mathscr{A},1}] \\ &\qquad \cdot g^{r+1}, A_2 = \nu_i) \Pr(A_2 = \nu_i \mid G = g^{r+1}) \end{aligned} \tag{6b}$$

$$\approx \Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_1 \mid G = [1_{\mathscr{M}} + 1_{\mathscr{F}} + 1_{\mathscr{A},1}] \cdot g^{r+1}), \tag{6c}$$

where Equation 6b is exact instead of approximate if the genotypes of $\mathscr{M}$ and $\mathscr{F}$ are known, and where Equation 6c results from approximating $\Pr(A_2 = \nu_i \mid G = g^{r+1}) \approx \Pr(A_2 = \nu_i \mid G = [1_{\mathscr{M}} + 1_{\mathscr{F}} + 1_{\mathscr{A},1}] \cdot g^{r+1})$. Approximation (6c) allows one to compute the probability without performing a sum over all alleles by assuming that the sum over all allelic types is approximated by the conditional probability with A_2 unknown. Note that when A_2 is known and equal to, for instance, ν_j, Equation 6c becomes $\Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_1 \mid G = [1_{\mathscr{M}} + 1_{\mathscr{F}} + 1_{\mathscr{A}}] \cdot g^{r+1})$ and is exact if $g^{r+1}_{\mathscr{M}}$ and $g^{r+1}_{\mathscr{F}}$ are known. Equation 6c says that we may approximate the conditional probability of transmission events $A_1 \leftarrow M_1$, $A_2 \leftarrow F_1$ given all the genotype data with the probability given just the genotype data of $\mathscr{A}$, $\mathscr{M}$, and $\mathscr{F}$. In fact, as described below, these probabilities will be computed using the genotype data from first-degree relatives. In computing the conditional probability of these transmission events I assume $g^{r+1}_{\mathscr{A}} = (x_{A_1}, x_{A_2})$, where x_{A_1} and x_{A_2} are allowed to equal either the unknown state 0 or one of the known allelic types $\nu_1, \ldots, \nu_L$. This allows us to replace $1_{\mathscr{A},1} \cdot g^{r+1}$ with $1_{\mathscr{A}} \cdot g^{r+1}$ in Equation 6c.

The probability in Equation 6c may be computed using Bayes' rule,

$$\Pr[A_1 \leftarrow M_1, A_2 \leftarrow F_1 \mid G = (1_{\mathscr{M}} + 1_{\mathscr{F}} + 1_{\mathscr{A}}) \cdot g^{r+1}] = \frac{\Pr(g^{r+1}_{\mathscr{A}} = (x_{A_1}, x_{A_2}) \mid A_1 \leftarrow M_1, A_2 \leftarrow F_1, G = \tilde{g})}{\sum_{i,j=1}^{2} [\Pr(g^{r+1}_{\mathscr{A}} = (x_{A_i}, x_{A_j}) \mid A_1 \leftarrow M_i, A_2 \leftarrow F_j, G = \tilde{g}) + \Pr(g^{r+1}_{\mathscr{A}} = (x_{A_i}, x_{A_j}) \mid A_1 \leftarrow F_j, A_2 \leftarrow M_i, G = \tilde{g})]}, \tag{7}$$

where $\tilde{g} = (1_{\mathscr{M}} + 1_{\mathscr{F}}) \cdot g^{r+1}$, the observed genotypes, at the $r + 1$ step, of $\mathscr{M}$ and $\mathscr{F}$ only. Consider the numerator of this equation,

$$\begin{aligned}&\Pr(g^{r+1}_{\mathscr{A},1} = x_{A_1}, g^{r+1}_{\mathscr{A},2} = x_{A_2} \mid A_1 \leftarrow M_1, A_2 \leftarrow F_1, G = \tilde{g}) \\ &\quad = \Pr(g^{r+1}_{\mathscr{A},1} = x_{A_1} \mid g^{r+1}_{\mathscr{A},2} = x_{A_2}, A_1 \leftarrow M_1, A_2 \leftarrow F_1, G = \tilde{g}) \\ &\qquad \cdot \Pr(g^{r+1}_{\mathscr{A},2} = x_{A_2} \mid A_2 \leftarrow F_1, G = \tilde{g}). \end{aligned} \tag{8}$$

The second probability on the right-hand side is

$$
\begin{aligned}
&D(A_2, F_1)\\
&\quad = \Pr(g^{r+1}_{\mathscr{A},2} = x_{A_2} \mid A_2 \leftarrow F_1, g^{r+1}_{\mathscr{F},1} = x_{F_1}, g^{r+1}_{\mathscr{F},2} = x_{F_2}, g^{r+1}_{\mathscr{M}} = (x_{M_1}, x_{M_2}))\\
&\quad = \begin{cases} 1 & x_{A_2}, x_{F_1} \text{ are known and IBS}\\ 0 & x_{A_2}, x_{F_1} \text{ are known and not IBS}\\ 1 & x_{A_2} \text{ is unknown}\\ \Pr(g^{r+1}_{\mathscr{F},1} = x_{A_2} \mid g^{r+1}_{\mathscr{F},2} = x_{F_2}, g^{r+1}_{\mathscr{M}} = (x_{M_1}, x_{M_2})) & x_{A_2}\text{known, } x_{F_1}\text{unknown.}\end{cases}
\end{aligned}
\tag{9}
$$

Although the probability $\Pr(g^{r+1}_{\mathscr{F},1} = x_{A_2} | g^{r+1}_{\mathscr{F},2} = x_{F_2}, g^{r+1}_{\mathscr{M}} = (x_{M_1}, x_{M_2}))$ in Equation 9 is conditional only on the known values of $g^{r+1}_{\mathscr{F},2}$ and $g^{r+1}_{\mathscr{M}}$, I improve the missing data approximations in Equations 6b and 6c by instead computing this probability conditional on the observed genotypes of all first-degree relatives of $\mathscr{F}$.

To complete the computation of Equation 8 one needs to determine the probability $\Pr(g^{r+1}_{\mathscr{A},1} = x_{A_1} | g^{r+1}_{\mathscr{A},2} = x_{A_2}, A_1 \leftarrow M_1, A_2 \leftarrow F_1, G = \tilde{g})$. This probability is $D(A_1, M_1)$, but in the case where x_{A_1} is known and x_{M_1} is unknown, the two conditions $g^{r+1}_{\mathscr{A},2} = x_{A_2}$ and $A_2 \leftarrow F_1$ result in the conditional probability on the right-hand side of Equation 9 becoming $\Pr(g^{r+1}_{\mathscr{M},1} = x_{A_1} | g^{r+1}_{\mathscr{M},2} = x_{M_2}, g^{r+1}_{\mathscr{F}} = (x_{F_1}, x_{F_2}), g^{r+1}_{\mathscr{F},1} = x_{A_2})$, when $x_{F_1} = 0$ and $x_{A_2} \neq 0$. Hence, $\Pr(g^{r+1}_{\mathscr{A},1} = x_{A_1} | g^{r+1}_{\mathscr{A},2} = x_{A_2}, A_1 \leftarrow M_1, A_2 \leftarrow F_1, G = \tilde{g}) = D(A_1, M_1)$, and combining this with Equations 9, 8, and 7, Equation 6c becomes

$$
\begin{aligned}
&\Pr(A_1 \leftarrow M_1, A_2 \leftarrow F_1 \mid G = g^{r+1})\\
&\quad \approx \frac{D(A_1, M_1)D(A_2, F_1)}{\sum_{i,j=1}^{2} D(A_1, M_i)D(A_2, F_j) + D(A_1, F_j)D(A_2, M_i)}.
\end{aligned}
\tag{10}
$$

Equation 10 provides a rapid means of computing the necessary descent probabilities, even in the presence of missing data, and may be applied to the recurrence relation of Equation 5 to find the probability of any two alleles being IBD. This is effective because the approximations in Equation 10 were chosen to minimize computation while maintaining accuracy under the assumption that the fraction of individuals in S with missing genotypes is not large. This strategy works well in practice because markers are usually included in a genetic analysis only if the rate of missing genotype data is low.

Computing Equation 9 may require knowing the conditional genotype distribution of $\mathscr{M}$ if $g^{r+1}_{\mathscr{M}}$ is unknown. Although this distribution is computed conditional on the genotypes of first-degree relatives, it also depends on the relatedness of $\mathscr{M}$ and $\mathscr{F}$ (which is zero in outbred individuals). The current implementation of the algorithm to compute this probability ignores the relatedness of $\mathscr{M}$ and $\mathscr{F}$ because, when there are a fair amount of genotype data among the first-degree relatives, this has a relatively small effect, unless $\mathscr{M}$ and $\mathscr{F}$ are very closely related. The algorithm for obtaining the conditional genotype distribution for an individual is to descend through the pedigree, one generation at a time, beginning with the quasi-founders, and for each person encountered who has missing genotype data, compute the genotype-frequency distribution given the genotypes of the parents, the spouse, and the offspring. If the person is a quasi-founder, the computation is done conditional only on the offspring and population allele frequencies. If the person is not a quasi-founder, but one or both of the parents are untyped, the computation is done conditional on the genotype-frequency distribution already computed for that parent. If an offspring is untyped, that offspring is ignored in the computation. To propagate information from more distant relatives when close relatives are untyped, additional iterations could be done, as needed. Experience suggests, however, that when the rate of missing genotypes is fairly low ($< \sim 20\%$ in the set S), a single pass through the pedigree is sufficient to obtain reasonably accurate estimates of the conditional genotype distributions in individuals with missing genotype data.

The general form of Equation 10 is particularly useful in that it naturally provides a framework to include effects such as mutation and genotyping error. In particular, while normally $D(A_1, M_1)$ is the zero or one indicator function when x_{A_1} and x_{M_1} are known, this may be generalized to a nontrivial function of the observed genotypes. Selecting a particular model for mutation or genotyping error where the observed genotypes depend on the true underlying genotypes would allow one to devise a more general form for $D(A_1, M_1)$ than what is given in Equation 9.

Boundary conditions: The recursive algorithm is completed by specifying boundary conditions that are applied to Equations 2 and 3 when $\mathscr{A}$ and $\mathscr{B}$ are quasi-founders. I assume below that the boundary conditions are applied at step t of the recurrence. The first rule is a general condition that applies to all sample individuals.

Boundary condition 1: For any allele A_i of individual $\mathscr{A}$, $\Pr(A_i \equiv A_i \mid G = g^t) = 1$.

For the next two boundary conditions p_i is the frequency of allele i in the founding population. To properly account for missing allele information I encode a missing allele as 0 and use the convention $p_0 = 1$. I also define the following two functions,

$$
\delta'_{rs} = \begin{cases} 1 & r = s \text{ or } r = 0 \text{ or } s = 0\\ 0 & \text{otherwise,}\end{cases}
$$

and

$$
\eta(r, s) = \begin{cases} r & s = 0 \text{ or } r = s\\ s & r = 0\\ \text{undefined} & \text{otherwise.}\end{cases}
$$

For notational convenience, label the alleles of $\mathscr{A}$'s genotype $(g^t_{\mathscr{A},1} = i, g^t_{\mathscr{A},2} = j)$ and the alleles of $\mathscr{B}$'s genotype $(g^t_{\mathscr{B},1} = k, g^t_{\mathscr{B},2} = l)$, where any of i, j, k, l may be unknown (*i.e.*, there may be partial or no genotype information on one or both individuals). The conditional probabilities given the genotypes of $\mathscr{A}$ and $\mathscr{B}$ in boundary conditions 2 and 3 are exact, even in the

presence of missing data, but approximate the conditional probability given all the genotype data.

Boundary condition 2: Let $f_{\mathscr{A}}$ be the inbreeding coefficient of individual $\mathscr{A}$. When $\mathscr{A}$ is a quasi-founder,

$$\Pr(A_1 \equiv A_2 \mid G = g^t) \approx \Pr(A_1 \equiv A_2 \mid G = g^t_{\mathscr{A}}) = \frac{p_{\eta(i,j)} f_{\mathscr{A}} \delta'_{ij}}{p_{\eta(i,j)} f_{\mathscr{A}} + (1 - f_{\mathscr{A}}) p^2_{\eta(i,j)}}.$$

When the alleles are from two separate quasi-founders $\mathscr{A}$ and $\mathscr{B}$, one needs the condensed identity coefficients (Jacquard 1974) for the pair. Without loss of generality, I consider only the probability that A_1 and B_1 are IBD.

Boundary condition 3: When $\mathscr{A}$ and $\mathscr{B}$ are quasi-founders,

$$\Pr(A_1 \equiv B_1 \mid G = g^t) \approx \Pr[A_1 \equiv B_1 \mid G = (g^t_{\mathscr{A}}, g^t_{\mathscr{B}})] = \sum_{r=1}^{9} \Pr(i \equiv k \mid G = (g^t_{\mathscr{A}}, g^t_{\mathscr{B}}), S_r) \cdot \frac{\Pr[G = (g^t_{\mathscr{A}}, g^t_{\mathscr{B}}) \mid S_r]\Delta_r}{\sum_s \Pr[G = (g^t_{\mathscr{A}}, g^t_{\mathscr{B}}) \mid S_s]\Delta_s},$$

where S_r is condensed identity state r and Δ_r is its unconditional probability.

Furthermore, the probabilities in the sum are

r	$\Pr[i \equiv k \mid G = (g^t_{\mathscr{A}}, g^t_{\mathscr{B}}), S_r]$	$\Pr[G = (g^t_{\mathscr{A}}, g^t_{\mathscr{B}}) \mid S_r]$
1	δ'_{ik}	$\delta'_{ij}\delta'_{ik}\delta'_{il}\delta'_{jk}\delta'_{jl}\delta'_{kl}\, p_{i'}$ $i' = \eta[\eta(\eta[i,j],k)]$
2	0	$\delta'_{ij}\delta'_{kl}\, p_{i'}\, p_{k'}$ $i' = \eta(i,j),\ k' = \eta(k,l)$
3	$\delta'_{ij}\delta'_{ik}\frac{\delta'_{i'k}p_{\eta(i',k)}p_l}{\delta'_{i'k}p_{\eta(i',k)}p_l + \delta'_{i'l}p_{\eta(i',l)}p_k}$	$\frac{1}{2}(2-\delta'_{kl})(\delta'_{i'k}p_{\eta(i',k)}p_l + \delta'_{i'l}p_{\eta(i',l)}p_k)$ $i' = \eta(i,j)$
4	0	$(2-\delta'_{kl})\delta'_{ij}\, p_{i'}\, p_k p_l \quad i' = \eta(i,j)$
5	$\delta'_{kl}\delta'_{ik}\frac{\delta'_{k'i}p_{\eta(k',i)}p_j}{\delta'_{k'i}p_{\eta(k',i)}p_j + \delta'_{k'j}p_{\eta(k',j)}p_i}$	$\frac{1}{2}(2-\delta'_{ij})(\delta'_{k'i}p_{\eta(k',i)}p_j + \delta'_{k'j}p_{\eta(k',j)}p_i)$ $k' = \eta(k,l)$
6	0	$(2-\delta'_{ij})\delta'_{kl}\, p_{k'}\, p_i p_j$ $k' = \eta(k,l)$
7	$\frac{\delta'_{ik}\delta'_{il}p_{\eta(i,k)}p_{\eta(j,l)}}{\delta'_{ik}\delta'_{jl}p_{\eta(i,k)}p_{\eta(j,l)} + \delta'_{il}\delta'_{jk}p_{\eta(i,l)}p_{\eta(j,k)}}$	$\frac{1}{2}(2-\delta'_{ij})(2-\delta'_{kl})(\delta'_{ik}\delta'_{jl}p_{\eta(i,k)}p_{\eta(j,l)} + \delta'_{il}\delta'_{jk}p_{\eta(i,l)}p_{\eta(j,k)})$
8	$\frac{\delta'_{ik}p_{\eta(i,k)}p_j p_l}{\delta'_{ik}p_{\eta(i,k)}p_j p_l + \delta'_{il}p_{\eta(i,l)}p_j p_k + \delta'_{jk}p_{\eta(j,k)}p_i p_l + \delta'_{jl}p_{\eta(j,l)}p_i p_k}$	$\frac{1}{4}(2-\delta'_{ij})(2-\delta'_{kl})$ $\cdot(\delta'_{ik}p_{\eta(i,k)}p_j p_l + \delta'_{il}p_{\eta(i,l)}p_j p_k + \delta'_{jk}p_{\eta(j,k)}p_i p_l + \delta'_{jl}p_{\eta(j,l)}p_i p_k)$
9	0	$(2-\delta'_{ij})(2-\delta'_{kl})p_i p_j p_k p_l.$

Frequency estimation: When $\mathscr{A}$ and $\mathscr{B}$ are not just quasi-founders but actual founders, the boundary conditions reduce to those given in Wang *et al.* (1995) and Davis *et al.* (1996). Both boundary conditions 2 and 3 depend on the allele frequencies in the *founding population*, unless all the founders are genotyped. These frequencies may be estimated either from the founding population itself, if available, or through the genotypes of those individuals in the pedigree. With deep pedigrees, where the founders lived many generations ago, it may be impossible to obtain accurate estimates of the allele frequencies in the founding population. Unless this population is large enough that genetic drift will have had a negligible effect, sampling from the present-day population from which the pedigree originated might not provide accurate estimates of the founder-allele frequencies. Instead, allele frequencies are often estimated from the genotyped study sample.

The current method uses the best linear unbiased estimator of allele frequencies of McPeek *et al.* (2004) to obtain estimates from the genotyped study sample. The estimator of the frequency p_ν for allele ν is given by

$$\hat{p_\nu} = (1^t K^{-1} 1) 1^t K^{-1} y_\nu,$$

where 1 is a vector with each element equal to one, K is a matrix with elements $K_{ij} = 2\phi_{ij}$ with ϕ_{ij} being the kinship coefficient between individuals i and j, and y_ν is the vector whose ith element equals one-half the number of alleles of type ν in individual i.

The danger of using the same population both for estimating frequencies and for computing sharing is that the sharing estimate will now be negatively biased. This is a general phenomenon when estimating IBD sharing among individuals and can be understood as follows. When related individuals share an allele identical by state there is some probability that this allele is also shared IBD. The rarer this allele is in the founding population, the higher the probability is that these two copies were inherited from a common ancestral allele and are IBD. Conversely, an allele shared between individuals that was fairly common in the founders is relatively unlikely to be IBD. Because of genetic drift, the frequency of an allele in the typed individuals will vary from its frequency in the founding population, and, when IBD is estimated for individuals sharing this allele, the estimate will be either too high or too low depending on whether the allele frequency has drifted down or up, respectively. Underestimating the probability will, on average, happen more often because common alleles, where underestimation most often takes place, are necessarily more frequent than rare alleles. Without some adjustment for this effect, $\Pr(A_1 \equiv B_1 \mid G)$ will tend to be below its expected value. The approach taken here is to use simulations to estimate the amount of bias when the allele frequencies are estimated. This is discussed in more detail below.

Linkage mapping

I implement a nonparametric, affecteds-only mapping method by using the sharing statistic S_{pairs}. For affected pair k (composed of individuals $\mathscr{A}$ and $\mathscr{B}$),

$$S_{\text{pairs},k} = \left(\sum_{i,j=1}^{2} 1_{\mathscr{A}_i \equiv \mathrm{B}_j}\right) - 4\phi_{\mathscr{A}\mathscr{B}},$$

$$S_{\text{pairs}} = \sum_{k=1}^{N_{\text{pairs}}} S_{\text{pairs},k},$$

where 1_E is 1 if E is true and 0 when E is false and $\phi_{\mathscr{A}\mathscr{B}}$ is the kinship coefficient between $\mathscr{A}$ and $\mathscr{B}$. Here, the summation for S_{pairs} is done over all possible pairs,

including when $\mathscr{A}$ and $\mathscr{B}$ represent the same individual, in which case $4\phi_{\mathscr{A}\mathscr{B}} = 2(1 + f_{\mathscr{A}})$ and $S_{\text{pairs},k} = 2(1_{a_1 \equiv a_2} - f_{\mathscr{A}})$, where $f_{\mathscr{A}}$ is the inbreeding coefficient of $\mathscr{A}$. Under the null hypothesis of no linkage, we have

$$E(S_{\text{pairs},k}) = 0$$

and

$$E(S_{\text{pairs}}) = 0.$$

The definition of $S_{\text{pairs},k}$ relies on knowing the true, unobserved IBD state of the alleles. Instead, we use its expectation given the available genotype data at that locus G,

$$\hat{S}_{\text{pairs},k} = E(S_{\text{pairs},k} \mid G) = \left[\sum_{i,j=1}^{2} \Pr(A_i \equiv B_j \mid G)\right] - 4\phi_{\mathscr{A}\mathscr{B}},$$

$$\hat{S}_{\text{pairs}} = \sum_{i=1}^{N_{\text{pairs}}} \hat{S}_{\text{pairs},k}.$$

Linkage mapping may be undertaken by ascertaining a sample of affected individuals and computing $\hat{S}_{\text{pairs},k}$ for every possible pair of individuals at a locus. Under the null hypothesis of no linkage $\hat{S}_{\text{pairs}}$ will form a distribution with a mean of zero, whereas under the alternative of linkage there should be excess sharing and $\hat{S}_{\text{pairs}}$ will generally be positive. Hence, to test for linkage one would do a one-sided test of $\hat{S}_{\text{pairs}}$ and reject the null hypothesis if $\hat{S}_{\text{pairs}} > t$ for some value of t.

Significance: A major challenge of the method is determining the proper statistical significance of the estimated sharing $\hat{S}_{\text{pairs}}$ at a marker. Three problems, in particular, must be overcome: (1) Computation of an exact likelihood-ratio statistic, as done in Kong and Cox (1997), is computationally prohibitive; (2) with only a single large pedigree, asymptotic theory for the distribution of $\hat{S}_{\text{pairs}}$ might not hold and, hence, using a normal distribution approximation, as in Kruglyak *et al.* (1996), might not be valid; and (3) obtaining an empirical distribution for $\hat{S}_{\text{pairs}}$ for each marker, conditional on that marker's allele frequency, could entail analysis of 10^2–10^5 simulated data sets *per locus*, depending on the degree of evidence for linkage at the locus. (The varying number of simulations per locus results from a time-saving strategy of using a limited number of simulations to screen out potentially interesting markers for which more accurate empirical distributions are then found.) This may be computationally impractical for data sets with many markers.

A potential solution is to develop a more computationally efficient variation on approach (3), in which the screening step is based on an appropriate transformation of the empirical distribution from a single completely informative marker. Because a completely informative marker allows immediate identification of which alleles are IBD, it is possible to rapidly compute the distribution of S_{pairs}. A first-pass P-value could then be estimated at each locus by comparing $\hat{S}_{\text{pairs}}$ at the marker against this distribution. For markers showing strong evidence of linkage, the more accurate and time-consuming approach (3) could be used. The difficulty with this approach is twofold. First, the variance of S_{pairs} is much larger than the variance of $\hat{S}_{\text{pairs}}$. The less informative the marker, the greater the difference in the variances. For essentially any marker the P-value obtained by comparing $\hat{S}_{\text{pairs}}$ to the distribution of S_{pairs} is far too conservative to be useful. Second, when allele frequencies are estimated from the sample $\hat{S}_{\text{pairs}}$ is negatively biased. Without some form of correction the P-value becomes even more conservative.

With a fast enough method, simulations can be used to estimate the bias and variance at a marker and correct the estimate of $\hat{S}_{\text{pairs}}$. Note that the number of simulations needed to correct the bias and variance at a marker would typically be much less than the number needed to assess the empirical P-value at the marker by method (3). To do this, for each simulation the founders are given marker alleles according to the allele-frequency distribution estimated in the real data set. The allele frequencies for each simulated data set are computed from the same individuals who are typed in the real data set, and these frequencies are used to compute $\hat{S}^{\text{s}}_{\text{pairs}}$ for that simulation. The mean value over the simulated $\hat{S}^{\text{s}}_{\text{pairs}}$ is used as an estimator for the bias, $\hat{b} = \sum_s \hat{S}^{\text{s}}_{\text{pairs}}/N_{\text{sim}}$, where N_{sim} is the number of simulations. The statistic for that marker is adjusted by this bias. Although this procedure can dramatically reduce the bias, it cannot completely eliminate it, even as N_{sim} becomes very large, because estimates of the bias are themselves biased. This comes about because the estimated frequencies are, in general, larger than the true frequencies (*e.g.*, rare alleles are often lost) and simulating founder genotypes using these estimated frequencies results in the simulated estimates of $\tilde{S}_{\text{pairs}}$ being slightly more biased than the actual data. The final result is a slight overcorrection. The ideal solution would be to use allele frequencies determined not from the study sample itself but from the population from which the founders originated.

The distribution problem is solved by assuming that the distributions of S_{pairs} and $\hat{S}_{\text{pairs}}$ have the same shape, but different variance. Normalizing both S_{pairs} and $\hat{S}_{\text{pairs}}$ to have a variance of one would allow one to compare $\hat{S}_{\text{pairs}}$ to the distribution of S_{pairs}. I estimate the variance of $\hat{S}_{\text{pairs}}$ from the values of $\hat{S}^{\text{s}}_{\text{pairs}}$ obtained from the bias estimation procedure. A new statistic is defined as $\tilde{S}_{\text{pairs}} = (\hat{S}_{\text{pairs}} - \hat{b})/\sqrt{\text{Var}(\hat{S}^{\text{s}}_{\text{pairs}})}$. The statistic $\tilde{S}_{\text{pairs}}$ can now be compared directly against the empirical distribution obtained for $S_{\text{pairs}}/\sqrt{\text{Var}(S_{\text{pairs}})}$ to obtain an approximate P-value. The amount of computation per

marker now includes a single analysis of the real data and analysis of N_{sim} simulated data sets with allele frequencies and marker informativeness based on the real marker data. Note that in the case where the founder-allele frequencies are known, adjusting for the bias is unnecessary, but simulations are still needed to estimate $\mathrm{Var}(\hat{S}^{s}_{pairs})$. The procedure of adjusting the bias and normalizing the variance has the benefit that now $\tilde{S}_{pairs}$ (and, hence, the approximate *P*-value) is directly comparable across all markers, regardless of markers' allele-frequency distributions. This is unlike $\hat{S}_{pairs}$, where the distribution depends significantly on the informativeness and allele frequencies of the marker.

In practice, the assumption that the shapes of the distribution for S_{pairs} and $\hat{S}_{pairs}$ are identical does not always hold for very large values of the statistics, where *P*-values are very small. In this case, the approximate *P*-value obtained using the above method tends to be highly conservative. Estimates for very small *P*-values can be obtained empirically by doing many simulations under the null hypothesis, using either the given or the estimated allele frequencies. Although many simulations are needed, the method is fast enough to be able to determine the empirical distribution of a marker under the null hypothesis for a small number of markers.

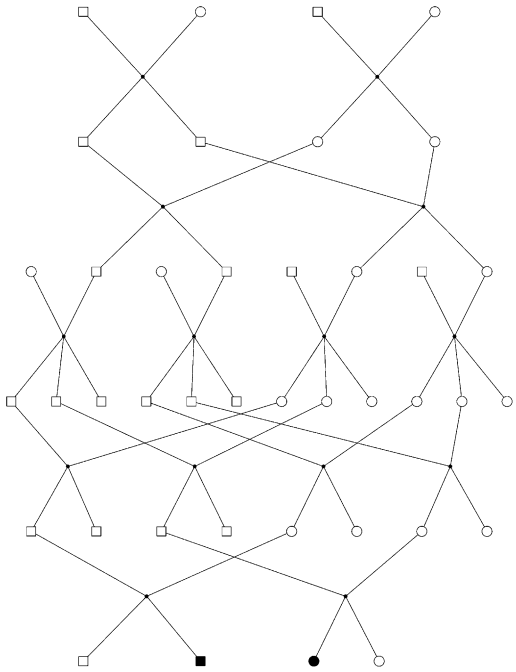

FIGURE 1.—Pedigree used to compute the accuracy of the IBD estimation method. The number of alleles shared IBD was estimated for the pair of individuals (solid symbols).

RESULTS

Simulations: Simulations were done to test the accuracy and validity of the proposed methods. Genotype data for a designated set of individuals were obtained by assigning the founders of a pedigree a genotype at a single locus on the basis of given allele frequencies and allowing these alleles to "drop" through the pedigree according to Mendelian inheritance. Only a single locus (*i.e.*, not multiple linked markers) was simulated. Three different types of markers were used in the simulations. The first was a single-nucleotide polymorphism (SNP) having two alleles of equal frequency. The second marker represented a microsatellite and had five equifrequent alleles. The third marker was meant to be representative of a "super marker," where each allele is actually a haplotype of tightly linked SNPs. The strategy of using haplotypes as alleles of a marker provides a straightforward means of using multipoint data to increase information about the inheritance process. Hence, the marker has 21 alleles with allele frequencies given by 2×0.2, 0.12, 0.07, 0.06, 0.05, 15×0.02. This allele-frequency distribution was chosen as being approximately representative of haplotype distributions for 90-kb segments on the basis of HAPMAP data for Caucasians.

Using both the method proposed here and an exact computation as done by MERLIN (ABECASIS *et al.* 2002), I checked the accuracy of the proposed method by estimating the number of alleles shared IBD for the two individuals in the pedigree shown by solid symbols in Figure 1. For each of the three types of markers, 1000 simulations were done with the genotypes from the top two generations removed and 10% of the remaining genotypes randomly set to be missing. For the purposes of the proposed method, then, all individuals in the third generation from the top were considered quasi-founders. Both exact and approximate estimates of the number of alleles shared IBD were computed for each simulation with the results shown in Figure 2. The plots show good agreement between the approximate and the exact computations with the square root of the mean squared error being 0.024, 0.049, and 0.044 for the SNP, microsatellite, and haplotype markers, respectively.

The above methods were applied to simulated data sets in a large, complex pedigree taken from the Hutterite population. The Hutterites are an isolated religious sect that originated in the Tyrolean Alps during the 1500s and now largely reside in the northern United States and western Canada (MANGE 1964; HOSTETLER 1974). The pedigree used here is an extension of the pedigree described in ABNEY *et al.* (2000), comprising 3028 individuals in 13 generations. To assess the efficacy of the bias-correcting and scaling procedures, simulations were done with 700 individuals' assigned genotypes. These 700 are the same sample that is currently being genotyped using the Affymetrix 500k SNP array. From this group of 700, 148 have been diagnosed with bronchial hyperresponsiveness and were labeled as "affected." Figures 3 and 4 show the histograms of S_{pairs} and $\tilde{S}_{pairs}$ from 10,000 simulations with the haplotype marker before and after the bias-

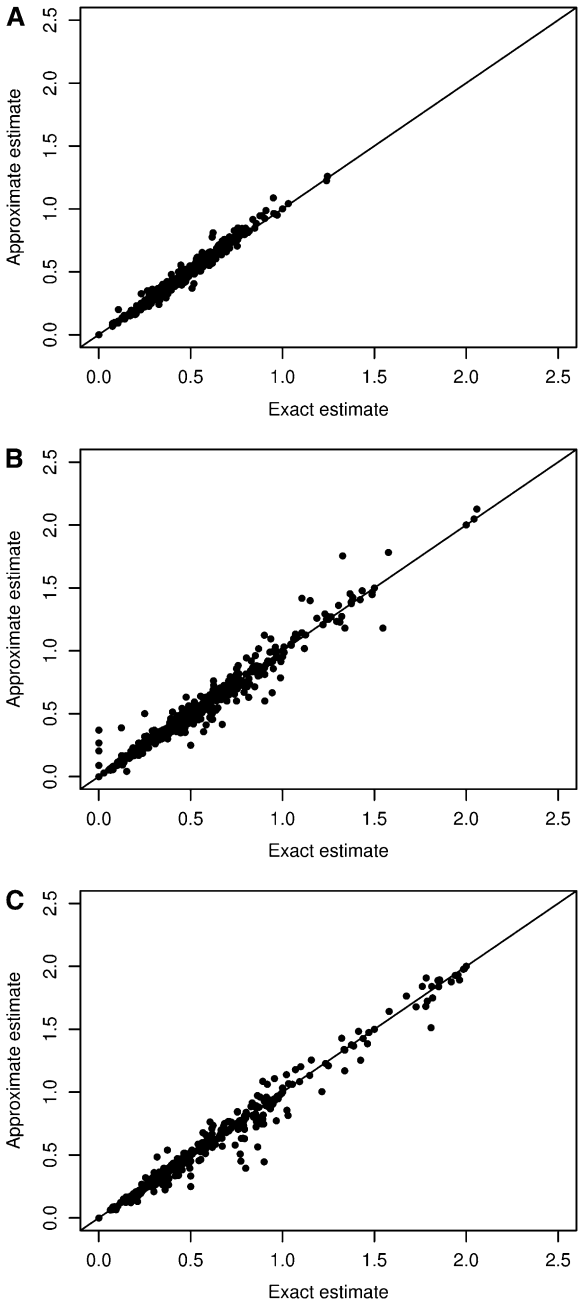

FIGURE 2.—Approximate *vs.* exact estimates of the number of alleles shared IBD for the pair of individuals in Figure 1 (solid symbols) for a (A) the SNP marker, (B) the microsatellite marker, and (C) the haplotype marker.

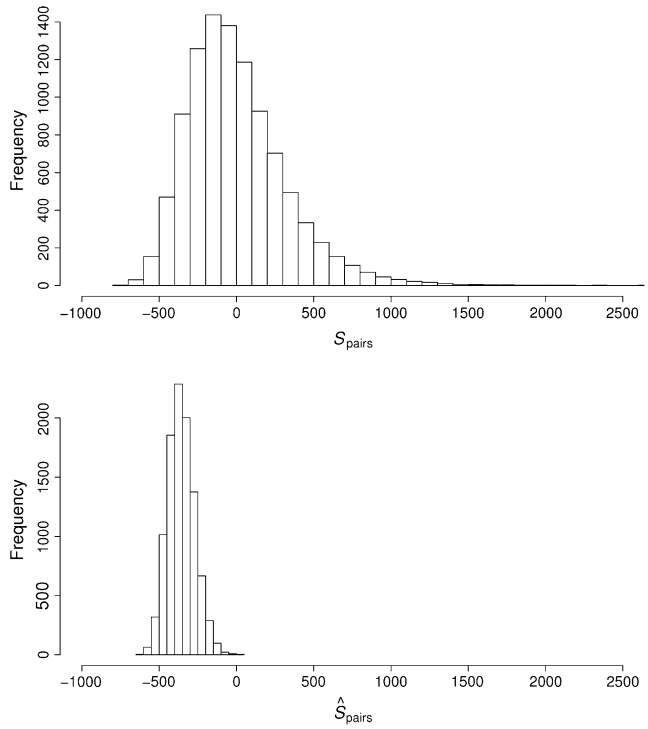

FIGURE 3.—Histogram of the true S_{pairs} statistic (top) and of the $\hat{S}_{pairs}$ statistic as computed from simulated data (bottom).

correction and scaling procedure was applied. The distributions in Figure 4 are fairly well matched over much of the range of statistic values, indicating that the scaled distribution of S_{pairs} can be used as a reference distribution from which approximate *P*-values can be obtained. It is also evident that S_{pairs} has a heavier tail at very large values, resulting in generally conservative estimates for very small approximate *P*-values.

The main utility of the bias-correction and scaling procedure is to put all markers in the study, regardless of informativeness, on a common scale over which they may be compared. This allows one to select the best markers to follow up with an empirical analysis by choosing those with the largest $\tilde{S}_{pairs}$ (or, equivalently, those with the smallest approximate *P*-value). Figure 5 compares the distributions of $\tilde{S}_{pairs}$ for the microsatellite and SNP markers over 10,000 simulations. The two distributions show good agreement, suggesting $\tilde{S}_{pairs}$ is a reliable measure for selecting the markers that have the largest amount of IBD sharing. In Figure 6 the distribution of $\tilde{S}_{pairs}$ for the haplotype marker is compared against the distributions of the microsatellite and SNP markers. In this case, where the allele-frequency distributions of the markers are radically different, the distributions do not compare as well, although the microsatellite marker distribution is notably closer to the haplotype marker distribution than is the SNP marker distribution. Most genetic mapping studies, however, tend to have markers with allele-frequency distributions that are relatively similar (*e.g.*, all SNPs, all microsatellites, or possibly a mixture of the two), suggesting that $\tilde{S}_{pairs}$ will generally be an effective measure of sharing across markers. It is worth noting that the haplotype marker allele-frequency distribution, with many alleles with small frequencies, presents the most challenging scenario when trying to estimate the bias and scale of the distributions. Inevitably, genetic drift within the sampled population results in a number of lost alleles and frequencies within the current population that may not be very representative of the

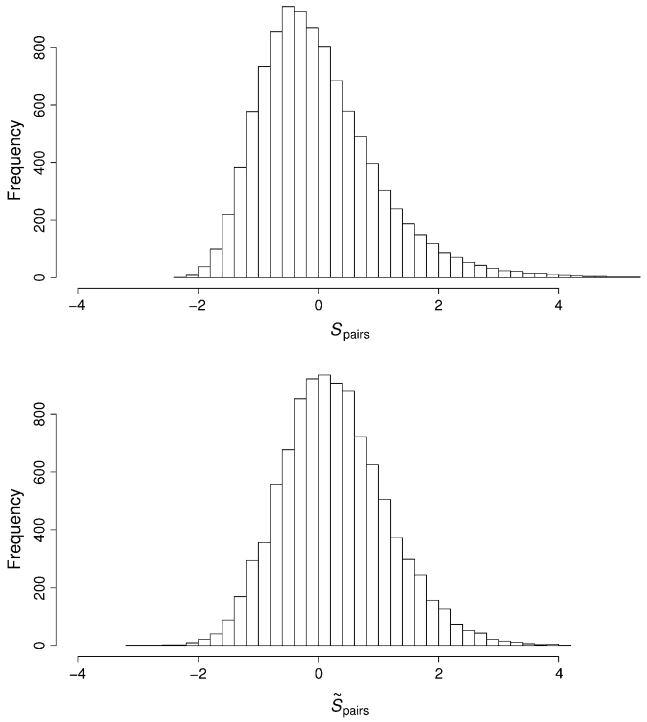

FIGURE 4.—Histogram of the true S_{pairs} statistic (top) and of the bias-corrected statistic $\tilde{S}_{pairs}$ as computed from simulated data (bottom). Both distributions have been scaled to have equal variance.

frequencies in the founding population. This results in inefficiency in the bootstrapping method used and a less accurate estimate of the bias of $\tilde{S}_{pairs}$.

To check the type-I error rate of the empirical *P*-value estimates, 5000 replicates were generated for each marker type in the Hutterite pedigree. In each replicate 910 individuals in the bottom three to four generations of the pedigree were assigned genotypes and analyses were done

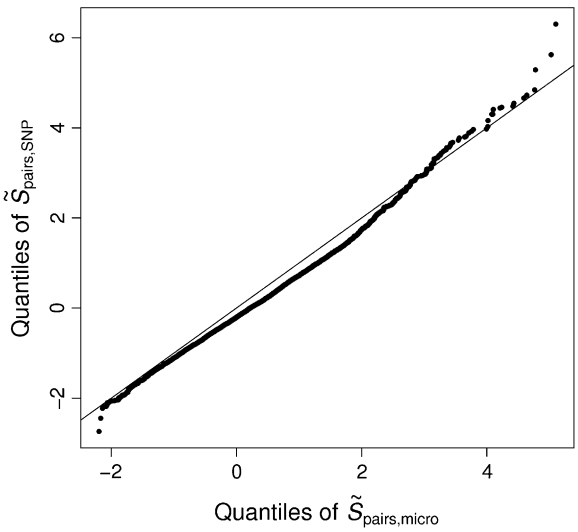

FIGURE 5.—QQ plot of 10,000 values of $\tilde{S}_{pairs}$ for the SNP marker (vertical axis) and the microsatellite marker (horizontal axis).

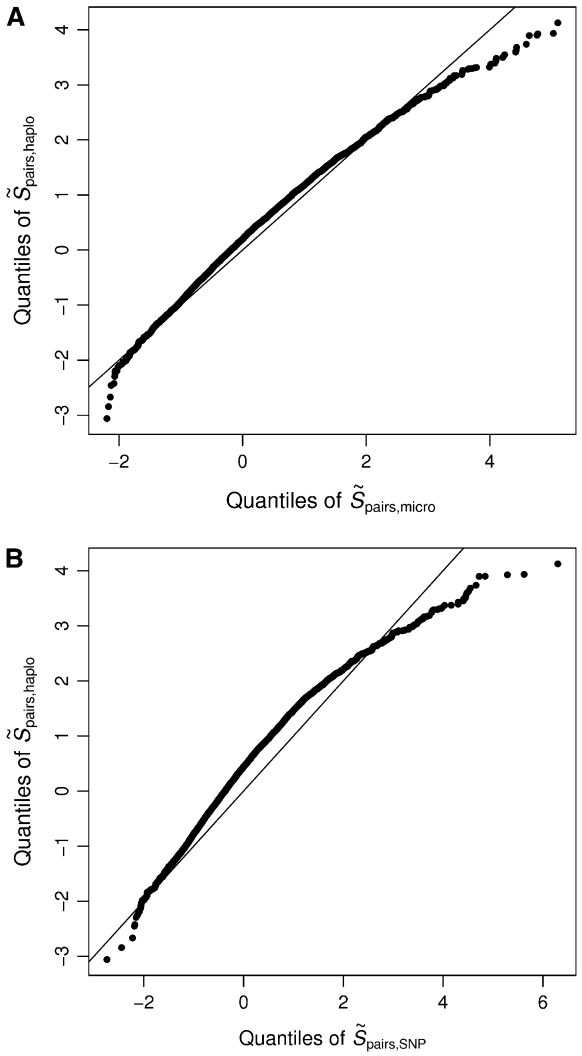

FIGURE 6.—QQ plot of 10,000 values of $\tilde{S}_{pairs}$ for the haplotype marker *vs.* (A) the microsatellite marker and (B) the SNP marker.

with 71 individuals assigned as affected, the same individuals as were previously diagnosed with asthma in this population (OBER *et al.* 2000). In each replicate the empirical *P*-value of the marker was estimated on the basis of 5000 simulations. Two analyses were done within each replicate, one where 10% of the genotype data was removed at random to emulate missing data and another with no missing genotypes in the sample individuals. Results are in Table 1 and, when there are no missing data, show good agreement with the nominal type I error. With 10% of data missing, the empirical *P*-values are slightly conservative. This results from the empirical distribution being computed under simulations with no missing data. Although this results in slightly conservative *P*-values, the computation time can be significantly reduced when doing 10,000–100,000 simulations. Optionally, with data sets that can be analyzed quickly, one can condition on a particular set of individuals having missing data to obtain more accurate empirical *P*-values.

TABLE 1

Empirical type I error based on 5000 replicates with different rates of missing genotypes

		Type I error at a nominal type I error of		
Marker type	Missing rate	0.05	0.01	0.005
SNP	0	0.059	0.013	0.0054
	0.1	0.052	0.0102	0.0034
Microsatellite	0	0.043	0.0084	0.003
	0.1	0.038	0.0076	0.0028
Haplotype	0	0.049	0.009	0.0032
	0.1	0.040	0.0066	0.0018

Asthma data set: A genome screen for asthma using 563 markers, including both microsatellites and SNPs, that was previously published (Ober *et al.* 2000) was reanalyzed using the method described here. In the previous analysis, the pedigree was divided into 20 subpedigrees with all inbreeding loops trimmed. A single-point, semiparametric method was used where a likelihood was maximized over parameters representing the penetrances, disease-susceptibility allele frequency, and recombination frequency. A likelihood-ratio χ^2 was obtained by comparing to a maximized likelihood where the recombination frequency was set to 0.5. Because of the size of the subpedigrees, each marker was analyzed separately. This analysis resulted in five markers with *P*-values <0.001 with the smallest *P*-value being 0.0002.

The genome screen data were analyzed using the methods described here and approximate *P*-values were obtained for each marker. Four markers had approximate *P*-values <0.01 and these were selected for follow-up with empirical *P*-value estimates. Empirical *P*-values were estimated on the basis of 100,000 simulations for each marker. Results for these four markers are shown in Table 2. Three of the four markers had smaller *P*-values than the smallest *P*-value of the previous analysis with two being much more significant. There was no overlap between these markers and the five markers reported previously. Approximate *P*-values for the five markers identified earlier ranged from 0.13 to 0.65. Of the four markers identified with the new method all had *P*-values >0.05 in the old analysis except for D5S1505, which had a *P*-value of 0.0441. That the markers found to have significant linkage in the two studies are different is not surprising, considering the very different methods used. The previous analysis finds markers that show some evidence for linkage under some genetic model in at least some of the subpedigrees. Although this may, it does not necessarily indicate higher than expected levels of IBD sharing among affecteds when looking at the pedigree as a whole, which is what the current method detects.

I also assessed the computation time required to do the analyses. These were accomplished on an AMD Opteron 252 at 2.6 GHz, running Linux 2.6. The exact computation of the number of alleles shared IBD, as done by MERLIN, for 1000 simulations, took 9.4×10^4, 3930, and 2880 sec for the SNP, microsatellite, and haplotype markers, respectively. In contrast, the approximate computations took 0.4, 0.2, and 1.5 sec for the same data sets. The empirical *P*-value estimates, using the asthma sample, took 41, 9.5, and 7 hr to estimate *P*-values for 1000 markers, for the SNP, microsatellite, and haplotype markers, respectively, where the estimates were based on 5000 replicates for each marker. That is, 5×10^6 estimates of IBD sharing were computed over all 2556 pairs, averaging 0.03, 0.0068, and 0.0052 sec per $\hat{S}_{\text{pairs}}$ estimate. When analyzing the real genotype data for the asthma sample, where markers consisted of both microsatellites and SNPs of varying informativeness, each estimate of $\hat{S}_{\text{pairs}}$ took $\sim$0.074 sec. This corresponds to obtaining approximate *P*-values in a genome screen of 1000 markers, with bias estimated using 50 simulations for each marker, in $\sim$1 hr. In contrast, obtaining empirical *P*-values based on 100,000 replicates took $\sim$2 hr per marker.

TABLE 2

Empirical *P*-value estimates for the four markers with the smallest approximate *P*-values in the asthma data set

Marker	Chromosome	Distance from p terminus (cM)	*P*-value
D4S405	4	57	0.00035
D5S1505	5	130	0[a]
D12S1042	12	49	0.00013
ATA41E04	16	11	3×10^{-5}

[a] Less than 3×10^{-5} at 95% confidence.

DISCUSSION

Linkage mapping in very large and complex pedigrees has been a computationally daunting task. The methods proposed here are a significant step forward in making the estimation of IBD sharing feasible in pedigrees of even extremely large size and complexity. This is possible, in part, by focusing on computing pairwise IBD probabilities, rather than computing the entire distribution of IBD sharing among all individuals, and by implementing a single-point method. Also, a great deal of computational effort is saved by doing computations only on the set of quasi-founders and their descendants and using precomputed values of the identity coefficients to determine the sharing probabilities among the quasi-founders. This allows for efficient computation even when there are many generations of untyped individuals at the top of the pedigree. Although determining the identity coefficients in the quasi-founders can also be computationally challenging if the pedigree is very large, recently developed methods

can accomplish this very efficiently (*e.g.*, identity coefficients for all pairs in the genotyped sample from the pedigree described above can be done in <1 day) (M. Abney, unpublished data). The most significant weakness of this approach is the single-point nature of the method. However, being able to do a single-point computation is often an integral part of a more general multipoint method. Indeed, as described in the appendix, I have used the method described here to extend the multipoint HBD method (Abney *et al.* 2002) to use the genotype data from all individuals in the pedigree. Nevertheless, even a single-point method of estimating IBD can prove useful. In particular, as markers become more informative the amount of information gained from a multipoint method decreases. Tightly linked SNPs, for instance, can be combined into haplotypes that can be treated as alleles of a highly informative super marker. Such a super marker would require inferring haplotypes from SNPs that have not recombined in the pedigree. This can be a challenging task in and of itself, but a number of methods have been proposed recently to address this problem (O'Connell 2000; Windig and Meuwissen 2004; Zhang *et al.* 2005; Baruch *et al.* 2006; Albers *et al.* 2007).

A direct application of the method is linkage mapping for qualitative traits. In this case, one computes the $\tilde{S}_{\text{pairs}}$ statistic for affected individuals and determines significance from a simulated distribution. There are a number of challenges inherent in accomplishing this, using pedigrees such as the one described here. A major difficulty is determining the distribution of the statistic $\tilde{S}_{\text{pairs}}$. Because studies involving large pedigrees often are restricted to a single, or a few, pedigrees rather than many independent ones, the usual central limit theorem argument may not lead to an accurate approximation of the null distribution. Without a theoretical basis for the nature of the distribution, we must simulate to determine its characteristics under the null. Because an empirical distribution for a marker depends on the characteristics of that particular marker (*e.g.*, founder allele frequencies, rates of missing data, etc.), one would, in principle, need to determine the null distribution for each marker in the study. This approach is problematic not only when there are many markers, but also for any single marker if the founder allele frequencies are not known. Here, I propose an alternative approach that may be used even when the number of markers is large. The distribution for a marker with perfect information is determined through simulation and scaled to have a mean of zero and a variance of one. The score $\hat{S}_{\text{pairs}}$ at a marker is then shifted—to remove bias—and scaled to also have a variance of one, on the basis of a limited number of simulations. If the shape of the perfect marker distribution matches that of the real marker, then this provides a computationally efficient means of obtaining a reference distribution for $\tilde{S}_{\text{pairs}}$ and an approximate *P*-value. Although the approximate *P*-values obtained in this manner are fairly accurate over much of the 0–1 interval, the *P*-values tend to become increasingly conservative as they approach zero. Nevertheless, being able to obtain rapid estimates of the *P*-value allows one to select those "best" markers for which to obtain a more accurate, empirical *P*-value from a large number of simulations. Note that when allele frequencies are estimated from the data, additional error beyond the Monte Carlo uncertainty of simulations is introduced due to the bootstrap nature of simulating data conditional on the estimated allele frequencies. If at all possible, then, it is best to obtain population allele frequencies that are independent of the data. A further complication of the need to do simulations to obtain a *P*-value is the difficulty of including genotyping error. Although it is straightforward to include genotyping error in the computation for IBD, when estimating significance, genotyping error would also have to be included in the simulations. An accurate *P*-value, then, necessitates simulating from a genotyping error model that is representative of the actual errors in the data. Because determination of this model may be difficult, at present the best strategy is to take any necessary precautions to ensure that errors in the genotype data are minimized.

Although the current method obtains *P*-values that are smaller than those done in an earlier analysis with different methods, these results do not directly address the question of the relative power of the two approaches. The merits of the proposed method, however, can be evaluated in light of the findings of the Genetic Analysis Workshop 12, where the asthma data set described here was analyzed by a number of different research groups. As summarized by Chapman and Wijsman (2001, pp. S222–S223), none of the groups were able to perform linkage analysis on the entire, intact pedigree, including those who used Markov chain Monte Carlo or regression-based approaches that are normally capable of handling large genealogies. They also suggest that two lessons could be learned from the results as a whole: "First, the results were sensitive to the method used to simplify the pedigree." They conclude that "minimal simplification of the pedigree is desirable. In general, stronger linkage signals came from data sets that used larger sub-pedigrees." Second, they conclude that "the method used to simplify the pedigree may be more important than the exact method of analysis used." The method described here is unique and advantageous in that it is capable of performing linkage mapping without any pedigree simplification. This should not only maximize the power but also avoid the difficulties of both creating appropriate subpedigrees and interpreting the results when multiple pedigree splittings have been analyzed. A power comparison with other methods involving possible pedigree simplification strategies (Falchi *et al.* 2004; Brocklebank *et al.* 2007; Liu *et al.* 2008) and genetic models would, nevertheless, be informative and will be pursued in future work.

In addition to linkage mapping for a qualitative trait, the proposed IBD estimation method is directly applicable to quantitative trait linkage mapping based on variance components (Amos 1994; Almasy and Blangero 1998). There, the covariance due to a quantitative trait locus (QTL) is modeled by the IBD sharing between all pairs of individuals in the pedigree. The method described here would allow rapid determination of the QTL covariance matrix for pedigrees that have been too large to consider as a whole. The variance component mapping strategy is, in many ways, considerably more straightforward than the S_{pairs} linkage mapping approach. Because the hypothesis test is not based explicitly on the degree of sharing estimated, the problems associated with determining the proper distribution of the statistic (*e.g.*, bias) are avoided. Although bias will still exist in the IBD estimates, if allele frequencies are estimated, the amount of bias for any particular pair will generally be small. In contrast, the bias when computing $\hat{S}_{\text{pairs}}$ can be large because it is the sum over very many pairs, each of which has a small amount of bias. Furthermore, an empirical distribution for the sharing statistic need not be constructed to determine statistical significance. The consequence is that it is not necessary to do the simulations at each marker to determine the bias and variance of the statistic.

The utility of a method that addresses a computationally difficult problem depends critically on the time it takes to solve the problem, particularly when many analyses and simulations are necessary. The above methods have been implemented in freely available software coded in C (available at http://www.genes.uchicago.edu/abney.html) that can analyze the data presented here with, as yet, unprecedented speed. The analyses accomplished here, using both simulated and real data, show that it is possible to analyze large data sets (~700 genotyped individuals, ~150 affecteds) in a large pedigree (~3000 individuals) on a timescale of hours for thousands of markers. As far as I am aware, there are no other available methods that can perform linkage mapping with this quantity of data on such a large, unbroken pedigree in any reasonable length of time. With the methods and software tools introduced here researchers with large, complex pedigrees will be able to leverage their genetic data to a degree that was not possible before.

I thank Carole Ober for allowing me to use the Hutterite pedigree and asthma data and Don Conrad for his assistance in evaluating haplotype frequencies in the HAPMAP data. This work was supported by National Institutes of Health grant HG002899.

LITERATURE CITED

Abecasis, G. R., S. S. Cherny, W. O. Cookson and L. R. Cardon, 2002 Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. **30:** 97–101.

Abney, M., M. S. McPeek and C. Ober, 2000 Estimation of variance components of quantitative traits in inbred populations. Am. J. Hum. Genet. **66:** 629–650.

Abney, M., C. Ober and M. S. McPeek, 2002 Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. Am. J. Hum. Genet. **70:** 920–934.

Albers, C. A., T. Heskes and H. J. Kappen, 2007 Haplotype inference in general pedigrees using the cluster variation method. Genetics **177:** 1101–1116.

Almasy, L., and J. Blangero, 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. **62:** 1198–1211.

Amos, C. I., 1994 Robust variance-components approach for assessing genetic linkage in pedigrees. Am. J. Hum. Genet. **54:** 535–543.

Baruch, E., J. I. Weller, M. Cohen-Zinder, M. Ron and E. Seroussi, 2006 Efficient inference of haplotypes from genotypes on a large animal pedigree. Genetics **172:** 1757–1765.

Baum, L. E., 1972 An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities **3:** 1–8.

Bourgain, C., and E. Genin, 2005 Complex trait mapping in isolated populations: Are specific statistical methods required? Eur. J. Hum. Genet. **13:** 698–706.

Brocklebank, D., J. Gayan and L. R. Cardon, 2007 Novel combinatorial optimisation methods to partition large pedigrees for genetic analysis. The American Society of Human Genetics Annual Meeting, October 25, 2007, San Diego.

Chapman, N. H., and E. M. Wijsman, 2001 Introduction: linkage analyses in the Hutterites. Genet. Epidemiol. **21**(Suppl. 1): S222–S223.

Davis, S., M. Schroeder, L. R. Goldin and D. E. Weeks, 1996 Nonparametric simulation-based statistics for detecting linkage in general pedigrees. Am. J. Hum. Genet. **58:** 867–880.

Elston, R. C., and J. Stewart, 1971 A general model for the genetic analysis of pedigree data. Hum. Hered. **21:** 523–542.

Escamilla, M. A., 2001 Population isolates: their special value for locating genes for bipolar disorder. Bipolar Disord. **3:** 299–317.

Falchi, M., P. Forabosco, E. Mocci, C. C. Borlino, A. Picciau *et al.*, 2004 A genomewide search using an original pairwise sampling approach for large genealogies identifies a new locus for total and low-density lipoprotein cholesterol in two genetically differentiated isolates of sardinia. Am. J. Hum. Genet. **75:** 1015–1031.

Fishelson, M., and D. Geiger, 2002 Exact genetic linkage computations for general pedigrees. Bioinformatics **18**(Suppl. 1): S189–S198.

Fulker, D. W., S. S. Cherny and L. R. Cardon, 1995 Multipoint interval mapping of quantitative trait loci, using sib pairs. Am. J. Hum. Genet. **56:** 1224–1233.

Heath, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. **61:** 748–760.

Hostetler, J. A., 1974 *Hutterite Society.* Johns Hopkins University Press, Baltimore.

Jacquard, A., 1974 *The Genetic Structure of Populations.* Springer-Verlag, New York.

Kong, A., and N. J. Cox, 1997 Allele-sharing models: Lod scores and accurate linkage tests. Am. J. Hum. Genet. **61:** 1179–1188.

Kruglyak, L., M. J. Daly, M. P. Reeve-Daly and E. S. Lander, 1996 Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet. **58:** 1347–1363.

Lander, E. S., and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.

Liu, F., A. Kirichenko, T. Axenovich, C. van Duijn and Y. Aulchenko, 2008 Anapproach for cutting large and complex pedigrees for linkage analysis. Eur. J. Hum. Genet. **16:** 854–860.

Mange, A. P., 1964 Growth and inbreeding of a human isolate. Hum. Biol. **36:** 104–133.

McPeek, M. S., X. Wu and C. Ober, 2004 Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics **60:** 359–367.

Ober, C., A. Tsalenko, R. Parry and N. J. Cox, 2000 A second-generation genomewide screen for asthma-susceptibility alleles in a founder population. Am. J. Hum. Genet. **67:** 1154–1162.

O'Connell, J. R., 2000 Zero-recombinant haplotyping: applications to fine mapping using snps. Genet. Epidemiol. **19**(Suppl. 1): S64–S70.

Peltonen, L., A. Palotie and K. Lange, 2000 Use of population isolates for mapping complex traits. Nat. Rev. Genet. **1:** 182–190.
Rabiner, L. R., 1989 A tutorial on hidden Markov-models and selected applications in speech recognition. Proc. IEEE **77:** 257–286.
Service, S., J. DeYoung, M. Karayiorgou, J. L. Roos, H. Pretorious *et al.*, 2006 Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat. Genet. **38:** 556–560.
Shifman, S., and A. Darvasi, 2001 The value of isolated populations. Nat. Genet. **28:** 309–310.
Sobel, E., and K. Lange, 1996 Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am. J. Hum. Genet. **58:** 1323–1337.
Sutter, N. B., and E. A. Ostrander, 2004 Dog star rising: the canine genetic system. Nat. Rev. Genet. **5:** 900–910.
Thallman, R. M., G. L. Bennett, J. W. Keele and S. M. Kappes, 2001 Efficient computation of genotype probabilities for loci with many alleles: Ii. Iterative method for large, complex pedigrees. J. Anim. Sci. **79:** 34–44.
Thompson, E. A., S. Lin, A. B. Olshen and E. M. Wijsman, 1993 Monte Carlo analysis on a large pedigree. Genet. Epidemiol. **10:** 677–682.
Wang, T., R. Fernando, S. Vanderbeek, M. Grossman and J. Vanarendonk, 1995 Covariance between relatives for a marked quantitative trait locus. Genet. Sel. Evol. **27:** 251–274.
Windig, J., and T. Meuwissen, 2004 Rapid haplotype reconstruction in pedigrees with dense marker maps. J. Anim. Breed. Genet. **121:** 26–39.
Wright, A. F., A. D. Carothers and M. Pirastu, 1999 Population choice in mapping genes for complex diseases. Nat. Genet. **23:** 397–404.
Zhang, K., F. Sun and H. Zhao, 2005 Haplore: a program for haplotype reconstruction in general pedigrees without recombination. Bioinformatics **21:** 90–103.

Communicating editor: R. W. Doerge

APPENDIX: MULTIPOINT HBD ESTIMATION

When the above algorithm is applied to the two alleles in a single individual, it provides an estimate of the probability of that individual being HBD conditional on the entire pedigree and all ancestral genotype data at that locus. A previous method (Abney *et al.* 2002) computed this probability given the pedigree and the multipoint genotype data of that individual, while the ancestral genotype data were ignored. Here I describe how to extend the HMM in Abney *et al.* (2002) to include ancestral genotype data.

An HMM is normally defined in the following way (Baum 1972; Rabiner 1989). Let

$$\alpha_k(i) = \Pr(O_1, \ldots, O_k, Q_k = i)$$

$$\beta_k(i) = \Pr(O_{k+1}, \ldots, O_M \mid Q_k = i),$$

where O_k are the observed genotypes at marker k, Q_k is the true HBD state (*i.e.*, HBD or not HBD) at marker k, and M is the number of genotyped markers. These variables follow the recurrence formulas

$$\alpha_{k+1}(i) = \sum_j \alpha_k(j) T_{ji} \Pr(O_{k+1} \mid Q_{k+1} = i)$$

$$\beta_k(i) = \sum_j T_{ij} \Pr(O_{k+1} \mid Q_{k+1} = j) \beta_{k+1}(j),$$

where T_{ij} is the transition probabilities between states i and j. The probability of HBD at marker t is

$$\Pr(Q_t = \text{HBD} \mid O_1, \ldots, O_M) = \frac{\alpha_t(\text{HBD})\beta_t(\text{HBD})}{\sum_i \alpha_t(i)\beta_t(i)}. \tag{A1}$$

Note that the quantity $\Pr(O_{k+1} \mid Q_{k+1} = i)$ in the recurrence equations is the probability of all the genotype data at the marker given the HBD state of a single individual. To use the single-point algorithm that now includes ancestral genotype data, the HMM is modified so that the probabilities found in the recurrence formulas are rewritten using Bayes' rule,

$$\Pr(O_k \mid Q_k = i) = \frac{\Pr(Q_k = i \mid O_k)\Pr(O_k)}{\Pr(Q_k = i)}.$$

The probability $\Pr(Q_k = I \mid O_k)$ is computed as described in the methods section, and $\Pr(Q_k)$ is the inbreeding coefficient. We do not need to compute the unconditional probability of the observed genotype data at the marker $\Pr(O_k)$ because it cancels out in Equation A1. The result is a method for computing multipoint HBD given all of an individual's ancestral genotype data and the entire pedigree. These probabilities can then be used in the homozygosity mapping method discussed in Abney *et al.* (2002).