Sibship Reconstruction From Genetic Data With Typing Errors

Jinliang Wang[1]

Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom

Manuscript received October 31, 2003
Accepted for publication December 31, 2003

ABSTRACT

Likelihood methods have been developed to partition individuals in a sample into full-sib and half-sib families using genetic marker data without parental information. They invariably make the critical assumption that marker data are free of genotyping errors and mutations and are thus completely reliable in inferring sibships. Unfortunately, however, this assumption is rarely tenable for virtually all kinds of genetic markers in practical use and, if violated, can severely bias sibship estimates as shown by simulations in this article. I propose a new likelihood method with simple and robust models of typing error incorporated into it. Simulations show that the new method can be used to infer full- and half-sibships accurately from marker data with a high error rate and to identify typing errors at each locus in each reconstructed sib family. The new method also improves previous ones by adopting a fresh iterative procedure for updating allele frequencies with reconstructed sibships taken into account, by allowing for the use of parental information, and by using efficient algorithms for calculating the likelihood function and searching for the maximum-likelihood configuration. It is tested extensively on simulated data with a varying number of marker loci, different rates of typing errors, and various sample sizes and family structures and applied to two empirical data sets to demonstrate its usefulness.

KNOWLEDGE of the genealogical relationships among individuals in a population (sample) is important in many research areas in behavioral, ecological, and evolutionary genetics and in conservation biology. It is crucial in studying the social behavior, mating system, and sex and reproductive allocations in social insect and other species (Queller and Strassmann 1998); in managing the conservation of populations of endangered species (Frankham 1995); and in assessing the genetic variation and inheritance of quantitative traits (Lynch and Walsh 1998). In practice, relationships can be estimated easily from pedigree records. Unfortunately, however, detailed pedigree information is rarely available for most natural populations. Genetic markers can be used, instead, to infer the relationships among individuals without pedigree information. Recent developments of highly polymorphic markers, such as microsatellites, have greatly increased the power of relationship inference from markers and enabled many fine-scaled analyses across various species.

Numerous methods have been advanced for inferring relationships among individuals solely from marker information (Blouin 2003). They can be classified broadly into two categories, the pairwise and group approaches. Pairwise approaches infer the relationship of a pair of individuals (dyad) using their marker genotypes, ignoring the other individuals in the sample. Typically, the probability of the marker data of a dyad under a particular relationship is calculated as the likelihood of the relationship, and the inferred relationship is the one with the maximum likelihood (*e.g.*, Thompson 1975; Boehnke and Cox 1997; Epstein *et al.* 2000; McPeek and Sun 2000). These methods can potentially infer any possible relationships in data, although they have difficulties in distinguishing relationships similar in identity-by-descent (IBD) sharing, such as half-sibs, grandparent-grandchild, and avuncular (Epstein *et al.* 2000; McPeek and Sun 2000). Pairwise methods are also simple to implement because all individuals (and their potential impact) other than the pair under consideration are ignored. However, valuable information may be lost in breaking the sampled individuals into pairs and considering each in isolation (Thomas and Hill 2000; Sieberts *et al.* 2002). All individuals in a sample may provide direct and indirect information concerning the relationship of a dyad, especially those closely related to the dyad. In diploid species, for example, exclusion of sibships is impossible for pairs but is possible for trios of individuals using autosomal markers, and more accurate relationship inferences are achieved from trios than from pairs of individuals (Sieberts *et al.* 2002). In addition to a possible loss of power due to the insufficient use of data, pairwise approaches infer relationships directly at the lowest level, between a pair of individuals only. Such pairwise relationships suffice in some instances when they are used, for example, to avoid mating between relatives (Herbinger *et al.* 1995). In most cases, however, knowledge of higher-order relationships is desirable, which requires all the individuals in a sample to be allocated into distinctive genetic groups

[1]*Address for correspondence:* Institute of Zoology, Regent's Park, London NW1 4RY, United Kingdom. E-mail: jinliang.wang@ioz.ac.uk

of a variable size (Smith *et al.* 2001). Further information may be lost in subsequent analyses, such as estimating heritability (Thomas and Hill 2000), if only pairwise relationships are inferred and used. Some methods consider trios of individuals for the candidate relationship of a parental pair and an offspring (Jones and Ardren 2003) or combinations of candidate relationships of full- and half-sibs, unrelated individuals, and identical twins (Sieberts *et al.* 2002). These share essentially the same properties (*e.g.*, fixed group size) as pairwise methods and are thus loosely categorized into pairwise approaches.

Group-likelihood approaches consider all individuals in the entire sample and partition them simultaneously into distinctive genetic groups of variable sizes (Painter 1997; Thomas and Hill 2000, 2002; Smith *et al.* 2001). Currently, they are applicable to a sample of individuals coming from a single cohort consisting of full- and half-sibships only (*e.g.*, tadpoles in a pond). In such circumstances, group approaches are expected to be more powerful than pairwise approaches because the former uses information of the multilocus genotypes of all sampled individuals in assigning them simultaneously into sib groups. Group approaches can also refine allele frequency estimates by accounting for the estimated relationships in a sample, which are then used to improve relationship inference (Thomas and Hill 2000, 2002; Smith *et al.* 2001). Such an iterative procedure is expected to improve estimates of both relationships and allele frequencies from a sample.

The accuracy of both pairwise and group approaches relies heavily on the reliability of marker information used in relationship inference. The exclusion of a given relationship because of its incompatibility with the observed genotypes is legitimate only when the genetic data are perfect. Unfortunately, however, genotype errors can be quite common in practice and are difficult to avoid. Even in the most favorable situation where a large amount of high-quality DNA is available for repeated genotyping under optimized PCR conditions, relationship inference can still suffer from mutations that may occur at a rate as high as 1.4×10^{-2} for microsatellites (Talbot *et al.* 1995). In practice, typing errors may occur frequently, especially when repeated typing is limited or even impossible due to the constraint of DNA amount or typing cost, when the quality of DNA is poor and/or PCR is not optimized. Such typing errors and mutations could have a devastating effect on relationship inference if they are not accounted for. A single scoring error (mutation) at just one locus of an individual may lead to its exclusion from being assigned the correct relationship with others no matter how many other loci of the individual are correctly scored. Some pairwise approaches have been developed to account for typing errors and mutations in inferring parentage (*e.g.*, SanCristobal and Chevalet 1997; Marshall *et al.* 1998; Neff *et al.* 2002) and other relationships (*e.g.*, Douglas *et al.* 2000; Epstein *et al.* 2000). Several empirical studies verified the importance of typing errors in affecting parentage determinations (*e.g.*, Blouin *et al.* 1996; O'Reilly *et al.* 1998).

In contrast, all previous group-likelihood approaches ignored typing errors and mutations completely. This is unfortunate because, on one hand, typing errors are expected to have a much more devastating consequence on relationship inference for group approaches than for pairwise approaches. For the former, a typing error may not only cause the individual carrying it to be incorrectly assigned into a genetic group, but also affect the assignment of the sibs of this individual. When the individual with a typing error is assigned incorrectly to a sib family, it may drag along with it some of its sibs with similar genotypes into the same false family. On the other hand, typing errors can be potentially identified and accounted for more effectively by group approaches than by pairwise approaches. This is because the multilocus genotypes of a group of individuals serve as mutual references and collectively they provide information about both a given relationship and possible typing errors. The larger the group, the more effectively group-likelihood approaches can detect and account for typing errors in relationship determination.

It is possible to accommodate typing errors in marker data in group-likelihood approaches to sibship reconstruction. If a typing error occurs at a locus of an individual and leads to a genotype incompatible with those of its full-sibs, then the likelihood of the full-sib family is zero no matter how many other loci are correctly typed and thus support the full-sib relationship. When typing errors are allowed for in an appropriate model, however, the likelihood of this full-sib family is always greater than zero and the family can be correctly recovered if genotypes at most loci of most individuals support the full-sib relationship even though one or more individuals are incorrectly typed at one or more loci. In this article, I propose two simple models of typing errors and incorporate them into a group-likelihood approach to sibship reconstruction. I use simulations to show that typing errors can cause severe biases in sibship inference if they are ignored. Yet, sibships can be inferred accurately from data in which typing errors occur at high rates, if typing errors are taken properly into account in estimation. I also propose a novel method based on Bayes' theorem to estimate allele frequencies from samples using the inferred relationships, a method to identify typing errors at each locus of each reconstructed family, and a method to infer parental genotypes. The performance of these methods and their robustness to the violation of some assumptions are checked by extensive simulations. Finally, I apply the proposed methods to two empirical data sets to infer the sibship structures and mating systems.

METHODS

First, I briefly review the factors affecting sibship inference from markers. In particular, the effects of genotype errors and how to correct for them in sibship reconstruction are described. Second, the likelihood functions of a nested half-sib family given population allele frequencies are derived for both diploid and haplo-diploid species. Third, an algorithm to search for the maximum-likelihood configuration of sibship structures for an entire sample is presented. Fourth, a method to estimate population allele frequencies from a sample with the reconstructed sibships accounted for is described. Fifth, I propose a method to detect typing errors at each locus within each reconstructed family in data and a method to infer parental genotypes. Last, I describe the simulation procedures employed to generate simulated data sets and the statistics used to measure the accuracy and precision of the proposed methods in inferring relationships, typing errors, and allele frequencies.

Assumptions and models of typing errors: To infer sibships from genetic markers without parental information, several assumptions are necessary in either pairwise- (Smith *et al.* 2001) or group- (Painter 1997; Almudevar and Field 1999; Thomas and Hill 2000, 2002; Smith *et al.* 2001) likelihood approaches.

A sample of individuals is assumed to be taken from a single cohort in a large random-mating population. This assumption implies that the genotype frequencies of the parents of sampled individuals can be calculated from population allele frequencies under Hardy-Weinberg equilibrium and that the probability of a given mating type is just the product of the frequencies of the two parental genotypes. It is possible to relax this assumption and incorporate nonrandom mating into the framework, if information about mating system (*e.g.*, selfing rate) is available.

All genetic markers used in sibship inference are assumed to be neutral, unlinked between loci, and in linkage equilibrium. Each marker locus is assumed to have two or more codominant alleles and to follow Mendelian segregation. All of the observed genotypes in a sample are free of errors and mutations, so that they can be trusted completely in inferring sibships.

Here, I follow previous studies in adopting the above assumptions, except for typing errors. In this study, typing errors are broadly defined as any changes in a genotype that could potentially cause incorrect relationship inference. They can come from the inheritance process (*e.g.*, mutations), the genotyping procedure (*e.g.*, miscalling or allelic dropout), and the course of data analyses (*e.g.*, data entry error). The detailed patterns of changes in genotype are different among various kinds of typing errors and may vary among marker types [*e.g.*, microsatellites *vs.* restriction fragment length polymorphisms (RFLPs)] and samples or studies. Obviously it is difficult or impossible to provide a universal model that reflects the detailed patterns of all kinds of typing errors in all marker data sets. Here I focus on microsatellite markers, which are used widely in ecological, behavioral, and evolutionary studies, and categorize their common typing errors into two classes that are modeled separately.

Class I includes allelic dropouts only. An allelic dropout occurs when PCR fails to amplify one of an individual's two homologous genes (one from each parent) at a locus. If the individual is a heterozygote, then a dropout yields a false homozygote. When dropouts are the sole source of typing errors, an observed heterozygote is always correct but an observed homozygote can be either correct or incorrect. If incorrect, the actual (true) genotype can be any heterozygotes containing the observed allele. For microsatellites, allelic dropouts seem to be the most serious problem (Gagneux *et al.* 1997) and could occur at an extremely high rate with low DNA concentration in PCR (Taberlet *et al.* 1996). Because of their common occurrence and the special error pattern (affecting heterozygotes only), allelic dropouts are considered individually. To account for allelic dropouts in sibship inference, I assume that each of the two alleles in any heterozygote at a locus is equally likely to drop out, at rate ε_1. Ignoring double dropouts at the same locus and individual (which rarely occurs and, if it does, can be easily detected and thus rectified by regenotyping in practice), we obtain the probabilities of $1 - 2e_1$, e_1, and e_1 [where $e_1 = \varepsilon_1/(1 + \varepsilon_1)$] for an actual heterozygote, say A_1A_2, being observed as A_1A_2, A_1A_1, and A_2A_2, respectively. Error rate ε_1 is allowed to vary across loci.

Class II includes all kinds of stochastic typing errors other than allelic dropouts. These errors can come from mutations, false alleles (polymerase errors rendering an allele other than the true one), miscalling (allele identification error), contaminant DNA, and data entry. Systematic typing errors, such as misplacing or admixing samples during DNA extraction, which may cause the entire multilocus genotype of an individual to be erroneous, are excluded. Compared with class I, class II errors are usually less frequent and can affect any homozygous or heterozygous genotype. To account for class II errors in data, I assume that the two homologous genes in any individual genotype at a locus are independently and equally likely to be incorrectly observed, with rate ε_2. I also assume that, for a locus with k codominant alleles, any allele is observed to be any one of the other alleles with an equal probability, $e_2 = \varepsilon_2/(k - 1)$. This error model is similar to that of Sieberts *et al.* (2002) but removes the restriction that only one error per locus per individual is allowed. Similar to ε_1, ε_2 can be variable among loci. Both ε_1 and ε_2 are assumed known for each locus in data.

Hereafter, observed and actual genotypes are distinguished and called phenotypes and genotypes, respec-

tively, for simplicity. Genotypes and phenotypes are denoted by G and R, respectively, when subscripts indexing alleles are used. Phenotypes are also denoted by r when subscripts indexing individuals and families are used.

For a locus with k codominant alleles (denoted by A_w with index $w = 1, \ldots, k$), there are $k(k+1)/2$ possible (ordered) genotypes and phenotypes, $G_{w,x} \equiv A_wA_x$, and $R_{w,x} \equiv A_wA_x$ for $w \le x = 1, \ldots, k$. Taking typing errors of both classes into account, I obtain the transitional probability from a genotype $G_{w,x}$ to a phenotype $R_{u,v}$,

$$\Pr[R_{u,v}|G_{w,x}] = \begin{cases} (1-\varepsilon_2)^2 + e_2^2 - 2e_1(1-\varepsilon_2-e_2)^2 \\ \quad (u=w,\ v=x) \\ e_2(1-\varepsilon_2) + e_1(1-\varepsilon_2-e_2)^2 \\ \quad (u=v=w) \text{ or } (u=v=x) \\ (2-\delta_{u,v})e_2^2 \\ \quad (u\neq w,\ u\neq x,\ v\neq w,\ v\neq x) \\ e_2(1-\varepsilon_2+e_2) \\ \quad \text{(otherwise)} \end{cases} \quad (1)$$

if $G_{w,x}$ is a heterozygote ($w \neq x$), and

$$\Pr[R_{u,v}|G_{w,x}] = \begin{cases} (1-\varepsilon_2)^2 \\ \quad (u=v=w) \\ 2e_2(1-\varepsilon_2) \\ \quad (u=w,\ v\neq w) \text{ or } (v=w,\ u\neq w) \\ (2-\delta_{u,v})e_2^2 \\ \quad (u\neq w,\ v\neq w) \end{cases} \quad (2)$$

if $G_{w,x}$ is a homozygote ($w = x$). In (1) and (2), $\delta_{u,v}$ is Kronecker delta variable with values 1 and 0 when $u = v$ and $u \neq v$, respectively. In deriving the first two equations in (1), I assumed that class II errors occur after class I errors, because the latter are generally more frequent than the former. A reversed sequence of error events leads to a different formulation of but little numerical difference in $\Pr[R_{u,v}|G_{w,x}]$ when ε_1 and ε_2 are not high, as is expected because the probability of both error events occurring to a single-locus genotype is minute.

The likelihood of a putative half-sib family: I assume a population of a dioecious species with one sex monogamous and the other sex polygamous. The polygamous sex can be either males or females. A sample of individuals taken from a single cohort in the population may thus contain full-sib families nested within half-sib families. I derive the likelihood of such a half-sib family for a single locus. If markers are statistically independent, the multilocus likelihood of a putative sib family is simply the product of the single-locus likelihoods.

For a locus with k codominant alleles, denote the population frequency of allele A_w by p_w ($w = 1, \ldots, k$) and the population frequency of the parental diploid genotype $G_{w,x} \equiv A_wA_x$ by $Q_{w,x}$. Under Hardy-Weinberg equilibrium, $Q_{w,x} = (2-\delta_{w,x})p_wp_x$, where $\delta_{w,x}$ is the Kronecker delta variable with values 1 and 0 when $w = x$ and $w \neq x$, respectively.

Consider a putative half-sib family consisting of a number of f putative full-sib families. Suppose, in full-sib family j ($j = 1, \ldots, f$), we observe d_j distinctive phenotypes $\{r_{1,j}, r_{2,j}, \ldots, r_{d_j,j}\}$ with corresponding counts $\{n_{1,j}, n_{2,j}, \ldots, n_{d_j,j}\}$. Under random mating, the probability of the phenotype data of the putative half-sib family (*i.e.*, likelihood) is

$$L = \sum_{w=1}^{k}\sum_{x=w}^{k} Q_{w,x}\prod_{j=1}^{f}\sum_{y=1}^{k}\sum_{z=y}^{k} Q_{y,z}\prod_{i=1}^{d_j}(\Pr[r_{i,j}|G_{w,x},G_{y,z}])^{n_{i,j}}. \quad (3)$$

In (3), the probability of observing an offspring phenotype $r_{i,j}$ ($i = 1, \ldots, d_j$) given parental genotypes $G_{w,x}$ and $G_{y,z}$ is derived from Mendelian segregation:

$$\Pr[r_{i,j}|G_{w,x},G_{y,z}] = \tfrac{1}{4}(\Pr[r_{i,j}|G_{w,y}] + \Pr[r_{i,j}|G_{w,z}] + \Pr[r_{i,j}|G_{x,y}] + \Pr[r_{i,j}|G_{x,z}]). \quad (4)$$

For a given offspring phenotype $r_{i,j} = A_uA_v$, each term on the right side of (4) is calculated by (1) or (2).

Although (1–4) are complete for calculating the likelihood of a half-sib family, they require substantial computation. This problem becomes especially important when Monte Carlo techniques are used in searching for the sibship configuration with the maximum likelihood (below) and the markers are highly polymorphic (k large). Computation can be reduced dramatically by considering just the observed alleles and an "allele" pooled over all the unobserved alleles for a putative family.

Suppose a number of m_j distinctive alleles are observed in the jth full-sib family in a given putative half-sib family. We pool the $k - m_j$ unobserved alleles as allele A_{k+j+1}, whose population frequency is the sum of those of the unobserved alleles. Denote the set of indexes of the m_j observed alleles and the pooled unobserved allele by Ω_j. Similarly, for the entire half-sib family, the set of indexes of the m_0 observed alleles and the single allele (A_{k+1}) pooled over all unobserved ones is denoted by Ω_0. By these arrangements, the likelihood function (3) reduces to

$$L = \sum_{w\in\Omega_0}\sum_{\substack{x\in\Omega_0\\ x\geq w}} Q_{w,x}\prod_{j=1}^{f}\sum_{y\in\Omega_j}\sum_{\substack{z\in\Omega_j\\ z\geq y}} Q_{y,z}\prod_{i=1}^{d_j}(\Pr[r_{i,j}|G_{w,x},G_{y,z}])^{n_{i,j}}. \quad (5)$$

The computational cost of (5) can be a tiny fraction of that of (3) because generally only a small subset of alleles is observed in a sib family.

Note that (3) and (5) are derived for a family with f ($f \geq 1$) full-sibships nested within a half-sibship. Obviously, they apply to pure half-sib and pure full-sib families, which are just two special cases when $f > 1$ and each full-sibship has just one offspring and when $f \equiv 1$, respectively. For a pure full-sib family ($f \equiv 1$ and $\Omega_0 \equiv \Omega_1$), the likelihood computational load can be further reduced by using

$$L = \sum_{w\in\Omega_1}\sum_{\substack{x\in\Omega_1\\ x\geq w}} Q_{w,x} \times \left(\sum_{\substack{z\in\Omega_1\\ z\geq x}} Q_{w,z}(2-\delta_{x,z}) \prod_{i=1}^{d_1} (\Pr[r_{i,1}|G_{w,x},G_{w,z}])^{n_{i,1}} + 2 \sum_{\substack{y\in\Omega_1\\ y\geq w+1}} \sum_{\substack{z\in\Omega_1\\ z\geq y}} Q_{y,z} \prod_{i=1}^{d_1} (\Pr[r_{i,1}|G_{w,x},G_{y,z}])^{n_{i,1}} \right) \quad (5')$$

instead of (5). The computational burden of (5′) is approximately half of that of (5).

Group-likelihood methods can use the phenotype data of individuals in a sample to partition them into sib families without the need for any parental information. However, if parental phenotypes are available, they should be used in sibship inference because they could improve the inference dramatically. Previous group-likelihood methods invariably ignored the use of parental genotypes in sibship inference (*e.g.*, Painter 1997; Almudevar and Field 1999; Thomas and Hill 2000, 2002; Smith *et al.* 2001). Here I present a method to use parental phenotypes in sibship reconstruction, after accounting for their possible typing errors.

Suppose the phenotype of the parent in the monogamous sex of the jth full-sib family is observed as $R_{u,v} \equiv A_uA_v$. Given $R_{u,v}$, the posterior probability of genotype $G_{s,t}$ ($s \leq t \in \Omega_j$) of the parent is calculated as $Q^*_{s,t} = Q_{s,t}\Pr[R_{u,v}|G_{s,t}]/(\sum_{y\in\Omega_j}\sum_{\substack{z\in\Omega_j\\ z\geq y}} Q_{y,z}\Pr[R_{u,v}|G_{y,z}])$ from Bayes' theorem. The sum over all possible genotypes, $\sum_{y\in\Omega_j}\sum_{\substack{z\in\Omega_j\\ z\geq y}} Q_{y,z}$, of that parent in calculating (5) should then be replaced by $\sum_{y\in\Omega_j}\sum_{\substack{z\in\Omega_j\\ z\geq y}} Q^*_{y,z}$. A known phenotype of a parent of the polygamous sex can be treated similarly.

Likelihood of a sib family in haplo-diploid species: For haplo-diploid species, there are two possible scenarios for the hierarchical sibship structure of full-sib families nested within a half-sib family: the polygamous and monogamous sexes are diploid and haploid, respectively, or are haploid and diploid, respectively. Here I consider the first scenario only since the second one can be treated similarly. Assuming sampled offspring are all diploids, the likelihood of a nested half-sib family is

$$L = \sum_{w\in\Omega_0}\sum_{\substack{x\in\Omega_0\\ x\geq w}} Q_{w,x} \prod_{j=1}^{f} \sum_{z\in\Omega_j} p_z \prod_{i=1}^{d_j} (\Pr[r_{i,j}|G_{w,x},G_{y,z}])^{n_{i,j}}. \quad (6)$$

In (6), $Q_{w,x}$ is the probability of the diploid parent's genotype being $G_{w,x} \equiv A_wA_x$ calculated as above, and p_z is the probability of the haploid parent's genotype being $G_z \equiv A_z$, which is the population frequency of allele A_z. The probability of observing the ith distinctive offspring phenotype in the jth full-sib family ($r_{i,j}$) given parental genotypes $G_{w,x}$ and G_z is derived from Mendelian segregation:

$$\Pr[r_{i,j}|G_{w,x},G_z] = \tfrac{1}{2}(\Pr[r_{i,j}|G_{w,z}] + \Pr[r_{i,j}|G_{x,z}]). \quad (7)$$

For a given offspring phenotype $g_{i,j} = A_uA_v$, the right-side terms of (7) are calculated by (1) and (2).

A phenotype of a diploid parent, if available, can be used in sibship inference in the same way as the diploid case. Any phenotype of a haploid parent can also be incorporated in sibship inference after considering its possible typing errors. Suppose the haploid phenotype of the parent of the jth full-sib family is observed as A_u; then the sum over all possible genotypes of this haploid parent in (6), $\Sigma_{z\in\Omega_j} p_z$, should be replaced by

$$\sum_{z\in\Omega_j} \frac{p_z(\delta_{z,u}(1-\varepsilon_2) + (1-\delta_{z,u})e_2)}{p_u(1-\varepsilon_2) + (1-p_u)e_2}$$

derived from Bayes' theorem.

Algorithm for searching the maximum likelihood: Suppose a number of N offspring are sampled and genotyped to infer their relationships. A particular partition of these N offspring into a number of sib families is called a sibship configuration. The total likelihood of a given configuration is the product of single-family likelihoods, each being calculated as shown above. Any prior information about the distribution of sib family sizes in the sample, if available, can be readily incorporated into the likelihood function.

There are many possible configurations even for a small sample size (N). With $N = 10$ and possible relationships constrained to either full-sibs or unrelated, for example, there are still 115,975 possible configurations (Thomas and Hill 2000). In fact, the feasible configurations quickly become too numerous to enumerate with an increasing N. Our task is to search for, through this vast configuration space, the best configuration with the maximum likelihood without considering all the possible configurations. This is accomplished by the algorithm described below, based on the simulated annealing technique (Kirkpatrick *et al.* 1983).

1. Generate an initial configuration by allocating those offspring known to be full-sibs to a full-sib family, those known to be half-sibs to a half-sib family, and those with unknown relationships each to a single half-sib family (Thomas and Hill 2000; Smith *et al.* 2001). Calculate and store the likelihood of each sib family in the initial configuration.
2. Generate a proposal configuration by changing part of the old one. Changes within and between half-sib families are allowed to occur with an equal probability. For a within-half-sib-family change, a full-sib family, F_1, is drawn uniformly from the filled ones (containing at least one individual) in the current configuration. If F_1 is known to be a genuine full-sib family (whether the parents' phenotypes are available or not) from another source of information, then it is replaced by another draw. Repeat this process until a full-sib family, F_1, without prior information is obtained. Then, draw an integer number uni-

formly from [1, $n_{F,1}$] (where $n_{F,1}$ is the number of individuals in F_1) and choose at random that number of individuals from F_1. These chosen individuals are to be moved to another family, F_2, selected at random from the full-sib families (including an empty one with no individual in it) within the half-sib family from which F_1 comes. Like F_1, F_2 must not be a full-sib family known to be genuine from prior information. For a between-half-sib-family change, a half-sib family, H_1, is chosen uniformly from the filled ones. Draw an integer number uniformly from [1, $n_{H,1}$] (where $n_{H,1}$ is the number of filled full-sib families in H_1) and choose at random that number of full-sib families from H_1 to be moved into another half-sib family, H_2, chosen at random from the half-sib families (including an empty one with no individual in it) in the current configuration. Similar to within-half-sib-family changes, both H_1 and H_2 must not be half-sib families known to be authentic from prior information.

3. Calculate the old likelihood (L_{old}) of the parts of the configuration proposed to be changed. For a within-half-sib-family change, L_{old} is the likelihood of the half-sib family from which F_1 and F_2 come. For a between-half-sib-family change, L_{old} is the product of the likelihoods of half-sib families H_1 and H_2.
4. Calculate the new likelihood (L_{new}) of the parts of the proposal configuration that have been changed. For a within-half-sib-family change, L_{new} is the likelihood of the half-sib family that has been altered. For a between-half-sib-family change, L_{new} is the product of the likelihoods of the two half-sib families that have been changed.
5. Determine whether to accept or reject the new configuration. Calculate $\tau = \text{Min}[(L_{new}/L_{old})^{1/T}, 1]$, where T is the annealing temperature governing the rate at which a new configuration is accepted. Generate a random number uniformly distributed between 0 and 1, and compare it with τ. If it is smaller than τ, the new configuration is regarded as successful and is thus accepted; otherwise, the new configuration is rejected and the old one is recovered.
6. Repeat steps 2–5 a sufficiently large number of times. This iterative procedure ensures the likelihood to go uphill in general, but allows it to go downhill occasionally to avoid it being stuck on a local maximum. The probability of a downhill tour is controlled by T, which is decreased as the annealing process proceeds so that a new configuration with a smaller likelihood than the old one becomes less and less frequently accepted. T is set initially at a value of one and reduced in multiplicative steps, each amounting to a 10% decrease. Each new value of T is held constant for $5000N$ reconfigurations or for $100N$ successful reconfigurations, whichever comes first. When efforts to improve configurations (increase likelihood) become sufficiently discouraging, the iterative process is stopped and the best configuration with the maximum likelihood is reported.

Estimating population allele frequency: Sibship reconstruction must use the allele frequencies in the parental population, which are assumed known above. In practice, however, population allele frequencies are generally unavailable and have to be estimated from the sample in which sibships are to be inferred. In other words, usually the only information available is the sample from which we have to deduce population allele frequencies necessary for sibship reconstruction. To better estimate population allele frequencies, sibships in a sample should be taken into account, especially when a sample is dominated by a few large families. Ignoring sibships in a sample leads to overestimates of population frequencies for the alleles present in large families, which results in the likelihoods of large families being too small and those of small families being too large. A possible consequence is that large families tend to split into smaller ones (Thomas and Hill 2000).

Thomas and Hill (2000) used a weighted least-squares approach to estimating allele frequencies, with the sample's family structure accounted for by the relationship matrix based on current inference of sibship structure. Smith *et al.* (2001) proposed a simpler method to estimate allele frequencies using weights inversely proportional to the estimated sibship size. In spirit, the two methods are similar, both weighting the information from a (putative) sib family inversely to its size. The weighted least-squares approach, however, is computationally intensive because the $N \times N$ relationship matrix must be inverted repeatedly over the iterative procedure.

Here, I propose a simple method to estimate allele frequencies of the parental population by using likelihood rather than family size as the weight. Consider, as an example, a half-sib family consisting of f full-sib families in a diploid species and the phenotypes of the $f + 1$ parents are unavailable. The count of an allele, A_u, in parent s (indexed as $s = 0$ for the polygamous parent and $s = 1, \ldots, f$ for the sth parent of the monogamous sex) can be estimated from Bayes' theorem as

$$\hat{c}_{u(s)} = \frac{1}{L}\sum_{w\in\Omega_0}\sum_{\substack{x\in\Omega_0\\ x\geq w}} Q_{w,x}(\delta_{0,s}(\delta_{u,w}+\delta_{u,x}-1)+1) \times \prod_{j=1}^{f}\sum_{y\in\Omega_j}\sum_{\substack{z\in\Omega_j\\ z\geq y}} Q_{y,z}(\delta_{j,s}(\delta_{u,y}+\delta_{u,z}-1)+1)\prod_{i=1}^{d_j}(\Pr[r_{i,j}|G_{w,x},G_{y,z}])^{n_{i,j}}, \quad (8)$$

where L is calculated by (5). For an unobserved allele, A_v, pooled into allele A_{k+s+1}, we first estimate the count ($\hat{c}_{k+s+1(s)}$) of the pooled allele by (8). The count of A_v in the sth ($s = 0, 1, \ldots, f$) parent is then estimated by

$$\hat{c}_{v(s)} = \hat{c}_{k+s+1(s)}\hat{p}_v/\hat{p}_{k+s+1}, \quad (9)$$

where $\hat{p}_v$ and $\hat{p}_{k+s+1}$ are the estimated population frequencies of A_v and A_{k+s+1} before updating. Estimate allele counts for each parent in each putative half-sib family in the current configuration, and update population allele frequencies by the mean of the estimated allele counts across parents. The computational load of (8) and (9) is minimal, because all quantities in them are already known from the calculation of L.

For half-sib families with partially known parental genotypes, and for the case of haplo-diploid species, population allele frequencies are estimated similarly, using Bayes' theorem and the corresponding likelihood functions.

Population allele frequencies are estimated initially by the method from the initial configuration and updated periodically after a certain number of successful reconfigurations. Because of the minimal computational cost of the proposed method, it is possible to update allele frequencies after each reconfiguration. However, it is usually unnecessary to update so frequently because a few improvements on the configuration do not change allele frequencies much (Thomas and Hill 2000).

Identifying possible typing errors: The group-likelihood approach shown above also allows us to identify possible typing errors in data that occurred at each locus within each reconstructed sib family. From the best configuration with the maximum likelihood that is finally rendered by the method, we can calculate, for each family and each locus, the likelihoods considering both class I and II errors (L), class I errors only (L_1, setting $\varepsilon_2 = 0$), class II errors only (L_2, setting $\varepsilon_1 = 0$), and no typing errors (L_3, setting $\varepsilon_1 = \varepsilon_2 = 0$). A likelihood-ratio test can then be carried out to screen the most likely hypothesis. Allelic dropouts are inferred to have occurred at a locus in a family when L_1 calculated for the given locus and family is significantly larger than L_3, for example.

Obviously, not all typing errors are identifiable. If a typing error causes little change in family likelihood, then it is unlikely to be detected. The power of error detection also depends critically on family size. In the extreme case of a family containing just a single individual, it is impossible to ascertain typing errors. Therefore, the typing errors identified should be treated as conservative.

Inferring parental genotypes: From the offspring phenotypes and the reconstructed sibships, we can also infer the parental genotypes using Bayes' theorem. As an example, consider a half-sib family consisting of a number of f full-sib families in a diploid species. The posterior probability of a parental genotype, $G_{u,v}$, given data and the reconstructed sibships, is

$$\Pr[G_{u,v}|\text{data}] = \frac{Q_{u,v}}{L}\prod_{j=1}^{f}\sum_{y=1}^{k}\sum_{z=y}^{k}Q_{y,z}\prod_{i=1}^{d_j}(\Pr[r_{i,j}|G_{w,x},G_{y,z}])^{n_{i,j}} \quad (10a)$$

for the polygamous parent and is

$$\Pr[G_{u,v}|\text{data}] = \sum_{w=1}^{k}\sum_{x=w}^{k}Q_{w,x}\prod_{\substack{j=1\\ j\neq s}}^{f}\sum_{y=1}^{k}\sum_{z=y}^{k}Q_{y,z}\prod_{i=1}^{d_j}(\Pr[r_{i,j}|G_{w,x},G_{y,z}])^{n_{i,j}} \times\left(\frac{Q_{u,v}}{L}\prod_{i=1}^{d_s}(\Pr[r_{i,s}|G_{w,x},G_{u,v}])^{n_{i,s}}\right) \quad (10b)$$

for the sth parent of the monogamous sex, where L is the family likelihood calculated by (3). The maximum-likelihood estimate of a parental genotype is the one with the maximal posterior probability. For a parent with observed phenotypes, its actual genotypes can be inferred similarly.

As is intuitively obvious, parental genotype inference relies heavily on the correctness of sibship reconstruction. It is likely to be inaccurate for an incorrectly reconstructed sib family. Even for a correctly reconstructed family, the inferred parental genotypes are not guaranteed to be correct, especially when family size is small. For a pure full-sib family ($f = 1$) in a diploid species, it is impossible to resolve the male and female parental genotypes no matter how much marker information is available. The parental genotype combination (♀ × ♂), $A_1A_2 \times A_3A_4$, has exactly the same posterior probability as $A_3A_4 \times A_1A_2$, for example. In such situations, one has to genotype one of the parents to infer accurately the genotypes of both. For a pure full-sib family in haplo-diploid species, the power of parental genotype inference is also reduced if the diploid parent is homozygous at a locus. In any case, the reliability of an inferred parental genotype is indicated by its posterior probability. The higher this value is in comparison with those of alternative genotypes, the higher the confidence we have in it.

Simulations: To assess the precision and accuracy of the proposed methods and their robustness when some assumptions are violated, I generate simulated data with known parameters by Monte Carlo, reconstruct sibships from the simulated data by the proposed methods, and measure (below) the fit between the true and estimated sibships. Different combinations of parameter values are used in simulations to check the performance of the methods and to investigate the effects of different factors on the estimation. Full-sib family sizes in a sample are assumed to follow either a Poisson distribution with parameter λ or a negative binomial distribution with parameters ρ (probability of success) and γ (number of successes). For both distributions, families with no offspring are obviously not represented in a sample. The mean and variance of full-sib family sizes in a sample are therefore $\lambda/(1 - e^{-\lambda})$ and $e^{\lambda}(e^{\lambda} - 1 - \lambda)\lambda/(1 - e^{\lambda})^2$, respectively, for the Poisson distribution and $(1 - \rho)\gamma/(\rho(1 - \rho^{\gamma}))$ and $(1 - \rho)\gamma(1 - \rho^{\gamma}(1 + (1 - \rho)\gamma))/(\rho(1 - \rho^{\gamma}))^2$, respectively, for the negative binomial distribution. The number of full-sib families nested within a half-sib family is also assumed to follow a Poisson distribution. For a given family, parental genotypes are generated using population allele frequencies under

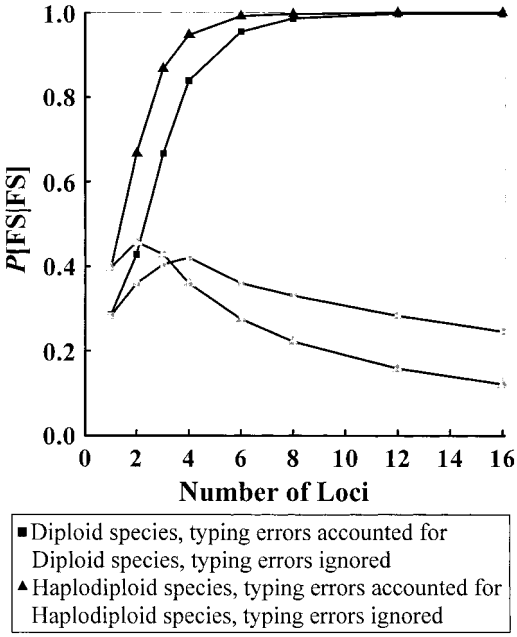

FIGURE 1.—The effect of typing errors on full-sib relationship inference in haplo-diploid and diploid species. A sample of 100 diploid offspring, consisting of full-sib families with a family size in the Poisson distribution of parameter 5, are genotyped for a variable number of loci, each having 10 codominant alleles of equal frequency. Typing errors of both classes occur at rate 0.05 for each locus and individual. For each number of marker loci, 50 data sets are simulated and analyzed comparatively with typing errors either ignored or accounted for in sibship inference. The estimation accuracy is indicated by the proportion of correctly inferred full-sib pairs in the sample, P(FS|FS).

random mating and Hardy-Weinberg equilibrium, and offspring genotypes are generated from parental genotypes following Mendelian segregation. These parents and offspring genotypes are then changed, at a given rate, following the models of class I and II typing errors to give their corresponding phenotypes. The phenotypes are then taken as observed data, which are used in sibship reconstruction. For a given parameter combination, 50 independent data sets are generated and analyzed.

Statistics measuring the performance of the estimation: Several statistics can be employed to measure the fit of the reconstructed to the actual sibships in simulated data. A stringent measurement of the overall fit is the number of full-sib (ξ_{FS}) and half-sib (ξ_{HS}) families being completely recovered relative to the actual numbers in a sample. Obviously $\xi_{FS} = 1$ and $\xi_{HS} = 1$ for a sample containing full-sib families only and half-sib families, respectively, mean the reconstructed sibships are perfect, with no individual being incorrectly assigned a relationship with any other individual. To gain insight into the causes of an imperfect sibship reconstruction ($\xi_{FS} < 1$ or $\xi_{HS} < 1$), I examine the statistic $P(a|b)$, the proportion of dyads assigned relationship a when their actual relationship is b (THOMAS and HILL 2002). In samples containing half-sib families consisting of full-sib families, full-sib (FS), half-sib (HS), and non-sib (NS) relationships are possible so that a, b = FS, HS, or NS.

To assess the accuracy of allele frequency estimates, I use the square root of mean-squared deviation (RMSD) of estimates from true frequencies of all alleles within and between loci. The power of the method for detecting typing errors is measured by the proportion of typing errors being correctly identified (ξ_1 = number of correctly detected errors/total number of detected errors) and the proportion of typing errors being detected (ξ_2 = total number of detected errors/total actual number of errors) across loci and reconstructed families in a sample. For parental genotype inference, I use the average proportion of parental genotypes being correctly inferred ($\overline{\Psi}$) to measure the accuracy of the method. For a single diploid parent, $\Psi \equiv 1$, ½, and 0 if 2, 1, and 0 of its alleles at a locus are correctly inferred, respectively. For a haploid parent, $\Psi \equiv 1$ and 0 when its genotype is correctly and incorrectly inferred, respectively. When a parental genotype is unresolved for a pure full-sib family, Ψ is calculated as the mean of the Ψ values calculated for the two alternative genotypes inferred. $\overline{\Psi}$ is then calculated as the average of Ψ across the two parents of all the sampled individuals and across loci.

RESULTS

Simulation results: *Number of loci:* Assuming typing errors of both classes occur at rate 0.05 at each locus, I generated simulated data that were then analyzed comparatively with typing errors ignored and taken into account, respectively. The proportions of correctly assigned full-sib pairs (P[FS|FS]) are shown in Figure 1 as a function of the number of loci used in estimation. The proportions of correctly assigned unrelated pairs (P[NS|NS]) are not shown because they are always close to 1 regardless of the number of loci and whether typing errors are ignored or not. Typing errors, if ignored in sibship reconstruction, lead to true full-sibs showing sib-incompatible phenotypes and thus to a full-sib family being broken up into several smaller ones. With an increasing number of loci used in sibship inference, both information and noise due to typing errors increase. However, the impact of typing errors overwhelms that of information, and, as a result, sibship inference becomes increasingly inaccurate with an increasing number of marker loci used in estimation. This is understandable because no matter how many loci are correctly typed and thus support a true sib family, it still breaks up into two families in reconstruction if one typing error occurring at a single locus in a single individual leads to a phenotype incompatible with others as sibs. The total multilocus likelihood for a group of individuals as

a sib family is the product of single-locus likelihoods and is zero if a single-locus likelihood is zero. When typing errors are accounted for in estimation, P[FS|FS] increases rapidly with an increasing number of loci and the full-sib relationships of a sample are completely reconstructed ($\xi_{FS} = 1$) once the number of loci is approximately equal to eight. The large impact of typing errors shown in Figure 1 highlights the importance of accounting for typing errors of data in group-likelihood approaches to relationship inference.

Rate of typing errors: The effect of the rate of typing errors in data on relationship inference is shown in Figure 2. When typing errors are ignored in estimation, P[FS|FS] declines rapidly with an increasing rate of typing errors in data. An error rate as low as 0.001, which is possible from mutations alone for microsatellites (*e.g.*, Talbot *et al.* 1995), can affect sibship reconstruction significantly. When typing errors are accounted for in the estimation, however, the estimator becomes very robust and can provide accurate estimates even when typing errors occur at an extremely high rate. In the simulated data, the probabilities of a heterozygous and homozygous genotype at a single locus being incorrectly typed are as high as 0.65 and 0.45, respectively, when $\varepsilon_1 = \varepsilon_2 = 0.256$. Even if in such situations where an observed phenotype is more likely to be erroneous than correct, 78% of the full-sib pairs are correctly identified and 40% of the full-sib families are fully reconstructed for a haplo-diploid species when typing errors are accounted for. In contrast, P[FS|FS] and ξ_{FS} are only 2 and 4%, respectively, when typing errors are ignored in estimation.

The rate of typing errors that the method can tolerate to yield satisfactory estimation depends on the amount of information available from data (number of marker loci and alleles per locus) and actual family sizes. With an increasing amount of marker information and/or family size in data, the method can cope with an error rate >0.256 as shown in Figure 2. In practice, probably no data are so dirty.

From Figures 1 and 2, we can see that typing errors have a greater impact on sibship inference for haplo-diploid species than for diploid species. This is because typing errors result in a larger probability of false exclusion of sibships in haplo-diploid than in diploid species.

Updating allele frequencies: Figure 3 depicts the impact of updating allele frequencies on estimating relationships and allele frequencies. As is expected, the benefit from updating allele frequencies increases with an increasing imbalance (variance) in family sizes (Figure 3A). When allele frequencies are not updated, large families tend to split into smaller ones, resulting in some full-sibs being incorrectly assigned as unrelated. With a variance of family size of 50 in Figure 3, for example, the proportions of full-sib pairs being inferred as unrelated are 2.0 and 6.4% when allele frequencies are and are not updated, respectively. The gain from the updating procedure in estimating allele frequencies also increases with an increasing variance in family size (Figure 3B).

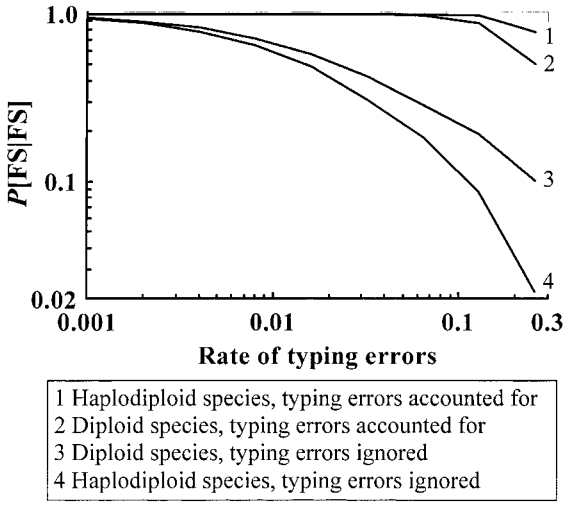

Figure 2.—The effect of the rate of typing errors on full-sib relationship inference. A total of 100 diploid offspring are sampled from a population of a diploid or haplo-diploid species and are genotyped for eight loci, each having 10 co-dominant alleles of equal frequency. The sample consists of full-sib families with a family size in a Poisson distribution with parameter 5. Typing errors of both classes occur at the same rate for each locus and individual. Simulated data are generated assuming various values of error rate in the range ~0.001–0.256 (shown on the *x*-axis) and are analyzed comparatively with typing errors either ignored or taken into account. The estimation accuracy is indicated by the mean (over 50 replicates) proportion of correctly inferred full-sib pairs in the sample, P(FS|FS).

Robustness of the models of typing errors: In the above, the rates of typing errors (ε_1 and ε_2) actually employed in generating simulated data are used in sibship inference. In application, usually ε_1 and ε_2 are unknown but are guessed from prior information or estimated by repeated genotyping (Gagneux *et al.* 1997). How robust the method is to sampling errors of ε_1 and ε_2 is obviously of concern for practical applications. In Figure 4, the data are generated with $\varepsilon_1 = \varepsilon_2 = 0.05$ but are analyzed assuming various values of ε_1 and ε_2. As can be seen, the accuracy of sibship inference (indicated by P[FS|FS] and ξ_{FS}) is quite high even though the assumed values of ε_1 and ε_2 deviate over several orders from their true value (0.05) used in simulation. Full-sibs tend to be assigned as unrelated and unrelated individuals tend to be assigned as full-sibs when the assumed error rate is much smaller and much larger than the actual value, respectively. All such incorrect assignments occur at a very low frequency, however, even if the assumed error rates are many times greater or smaller than the true values. It seems that accurate sibship inference can be obtained using a wildly guessed, rather inaccurate rate of typing errors provided sufficient information is contained in data. A similar conclusion was reached by

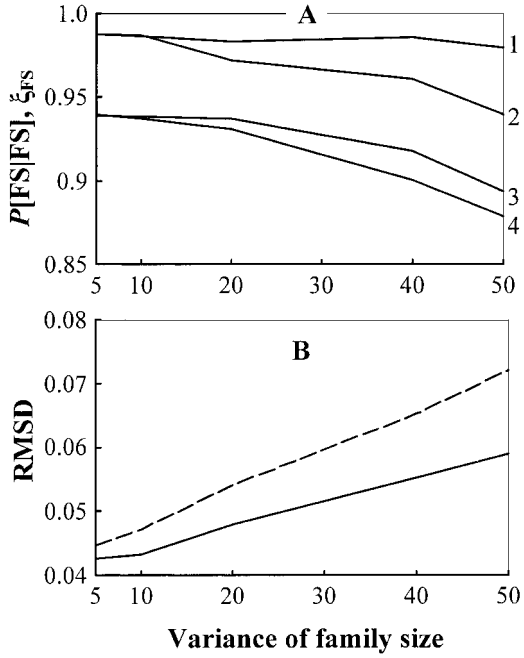

FIGURE 3.—The effect of updating allele frequencies on full-sib relationship and allele frequency estimation. Simulated samples contain full-sib families, with family sizes following a negative binomial distribution with different parameter values to yield the same mean (5) but various variances of family size. Each sample has 100 diploid individuals taken from a haplo-diploid species, and each sampled individual is genotyped for five loci, each having 10 codominant alleles of equal frequency. Typing errors of both classes occur at rate 0.05 for each locus and individual and are accounted for in sibship reconstruction. The simulated data are analyzed comparatively with allele frequencies updated every 1000 successful reconfigurations or left unchanged during the process in searching for the maximum-likelihood configuration. (A) Lines marked by 1 and 2 indicate the mean (over 50 replicates) proportions of full-sib pairs that are correctly assigned [P(FS|FS)] when allele frequencies are updated and not updated, respectively, and lines marked by 3 and 4 indicate the mean proportions of fully recovered full-sib families (ξ_{FS}) when allele frequencies are updated and not updated, respectively. (B) The solid and dashed lines indicate the square roots of mean-squared deviation of estimated from true allele frequencies (RMSD) when allele frequencies are updated and not, respectively.

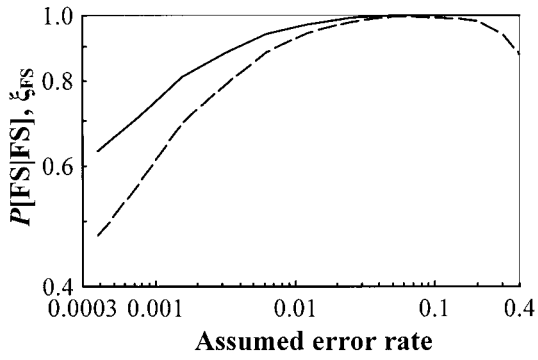

FIGURE 4.—The effect of the error rate assumed in analyses on full-sib relationship inference. A total of 100 diploid offspring, consisting of full-sib families with a family size in a Poisson distribution with parameter 5, are sampled from a population of haplo-diploid species and are genotyped for eight loci, each having 10 codominant alleles of equal frequency. Typing errors of both classes occur at rate 0.05 for each locus and individual, but different values of error rate (shown on the x-axis) are assumed in sibship reconstruction. The solid and dashed lines indicate the mean (over 50 replicates) proportions of full-sib pairs that are correctly assigned [P(FS|FS)] and of fully recovered full-sib families (ξ_{FS}), respectively.

SANCRISTOBAL and CHEVALET (1997) in their pairwise-likelihood inference for parentage and by SIEBERTS *et al.* (2002) in relationship inference from trios of individuals.

The error model assumed that typing errors occur independently across loci within an individual. This assumption can be violated if DNA quality, quantity, or both vary considerably among individuals. When DNA is extracted from noninvasive sources such as hair and feces or from ancient material such as bones and scales, for example, both its quantity and quality can be highly variable among individual samples, resulting in significantly different error rates between individuals (*e.g.*, GAGNEUX *et al.* 1997). Such variation implies that if a typing error occurs at a locus for an individual, then typing errors are more likely to occur at other loci of the same individual than at a locus of an individual taken at random from the sample. To investigate the robustness of the model to such concordant occurrence of errors within an individual, simulated data are obtained by drawing an error rate from a truncated (negative values are discarded) Gaussian distribution with mean 0.05 and standard deviation σ and using it in generating typing errors of both classes across loci for a given individual. Therefore, the larger the value of σ, the higher the intra-individual correlation of error occurrences among loci. The effect of σ on sibship inference is shown in Figure 5, where P[FS|FS] and ξ_{FS} are plotted against σ. As can be seen, the model is robust to moderate levels of variation in typing error rate among individuals. With an increasing value of σ, the proportion of individuals carrying erroneous genotypes at almost all loci increases. For these individuals, it is obviously impossible to correctly assign relationships between one of them and any others in the sample. When $\varepsilon_1 = \varepsilon_2 \geq 0.25$, an individual's phenotype is more likely to be erroneous than correct at each locus. The proportions of such individuals are $\sim$3.3 and 27% for $\sigma = 0.1$ and $\sigma = 0.2$, respectively, in the simulated data sets.

With miscalling or mutations in the single-stepwise model for microsatellites, a typing error usually involves a single tandem repeat change and an allele is more likely to be observed if its size is closer to that of the actual allele. In heterozygotes, larger alleles may be more likely than smaller alleles to drop out. Such size-dependent dropouts may bias allele frequency estima-

tion, which may further affect sibship inference. Families with partially known parental genotypes may suffer a smaller rate of typing errors than families with no parental information available, because in the former case some typing errors may be identified and corrected using the known parent-offspring relationship before the genotype data are analyzed for sibship inference. Simulated data were generated following these error patterns but analyzed using the proposed simple error models. In all cases considered, accurate inference of sibships was obtained (results not shown) when typing error rate was not very high (say, <0.15), indicating that the proposed models of typing errors are quite robust. This is not surprising given the results in Figures 4 and 5.

Hierarchical sibship structures and sample sizes: Figure 6 shows the proportions of the actual full-sib and half-sib pairs being assigned different relationships by the estimator applied to samples of various sizes and composed of full-sib families nested within half-sib families. The assignments of unrelated pairs are almost perfect [P(NS|NS) very close to 1] for various sample sizes and are omitted from the figure. Sibship inference becomes increasingly inaccurate with a decreasing sample size once it becomes very small (<50) and with an increasing sample size once it becomes very large (>800). The former is due to the fact that allele frequencies used in sibship inference are less accurately estimated with a smaller sample size. The latter is caused mainly by the increasing probability with sample size that individuals in a sibship could have, by chance, disparate albeit compatible genotypes and thus the sibship may be split in reconstruction. The magnitude of the effect of sample size on sibship inference depends on the amount of marker information and family size.

Overall, sibships are quite accurately inferred for haplo-diploid species using only five microsatellites, with at least 95% full- and half-sib pairs being correctly inferred when sample size is ~50–800. Even if the sample size is as large as 1600, P(FS|FS) and P(HS|HS) are still >92% in the examples shown in Figure 6. Sibship inference is much less accurate for diploid than for haplo-diploid species, as expected. The contrast is especially evident when the sample size is very large or small. Obviously, the amount of marker information (five microsatellites, each with 10 alleles at equal frequency) is insufficient for accurate sibship inference in diploid species. In Figure 6, P(FS|FS) and P(HS|HS) are 0.51 and 0.38, respectively, when $N = 1600$ for a diploid species. When the number of loci used in the estimation is increased to 10 loci, the corresponding values are 0.97 and 0.98, respectively.

Identifying typing errors and inferring parental genotypes: The proportion of typing errors being correctly detected (ξ_1) by the likelihood method is generally high. For example, $\xi_1 > 0.99$ for various error rates assumed in Figure 2, $\xi_1 \approx 0.91$–0.92 and $\xi_1 > 0.96$ when the number of loci used in sibship inference is ≤2 and >2, respectively, in Figure 1. $\xi_1 > 0.99$ is obtained even though the assumed error rate is several orders larger or smaller than the actual value (Figure 4) or the error model assumptions are violated (Figure 5). On the other hand, the proportion of overall typing errors detected (ξ_2) by the likelihood method is generally low, being <80% in simulations shown in Figures 1–5. This is not surprising because some typing errors cause no or little change in likelihood and are thus not detectable. These results indicate that a typing error identified by the method is highly likely to be genuine, but not all typing errors are identifiable.

The inference of parental genotypes is generally less accurate than sibship inference and typing error detection (ξ_1). This is because it relies on correct sibship reconstruction and sometimes male and female parental genotypes are unresolved for full-sib families. In Figure 1, for example, the proportion of parental genotypes being correctly inferred ($\overline{\Psi}$) is ~80 and 50% for haplo-diploid and diploid species, respectively, when three to six loci are used in estimation. The accuracy of parental genotype inference improves with nested half-sib families, larger family sizes, and more marker information. When eight loci, each having 10 alleles of equal frequency and a typing error rate of 0.05, are used in estimating the relationships of 100 offspring coming from nested half-sib families with the numbers

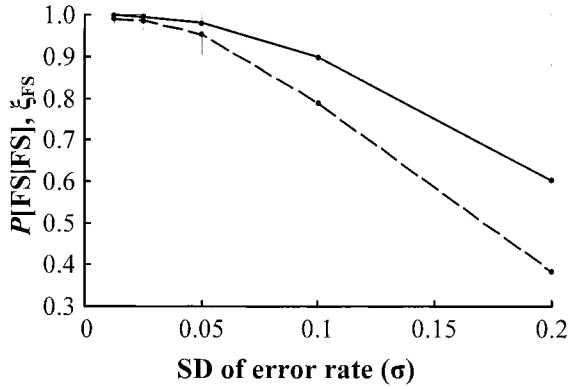

FIGURE 5.—The effect of the dependence between typing error occurrences at different loci within an individual on full-sib relationship inference. A total of 100 diploid offspring, consisting of full-sib families with a family size in a Poisson distribution with parameter 5, are sampled from a population of haplo-diploid species and are genotyped for eight loci, each having 10 codominant alleles of equal frequency. For each sampled individual, an error rate is drawn from a truncated Gaussian distribution with mean 0.05 and standard deviation σ, and typing errors of both classes are simulated to occur at the same sampled rate across loci. The simulated data are analyzed with an assumed error rate of 0.05 for both classes of errors, each individual, and each locus. The solid and dashed lines with error bars indicate the means and standard deviations (over 50 replicates) of the proportions of full-sib pairs that are correctly assigned [P(FS|FS)] and of fully recovered full-sib families (ξ_{FS}), respectively.

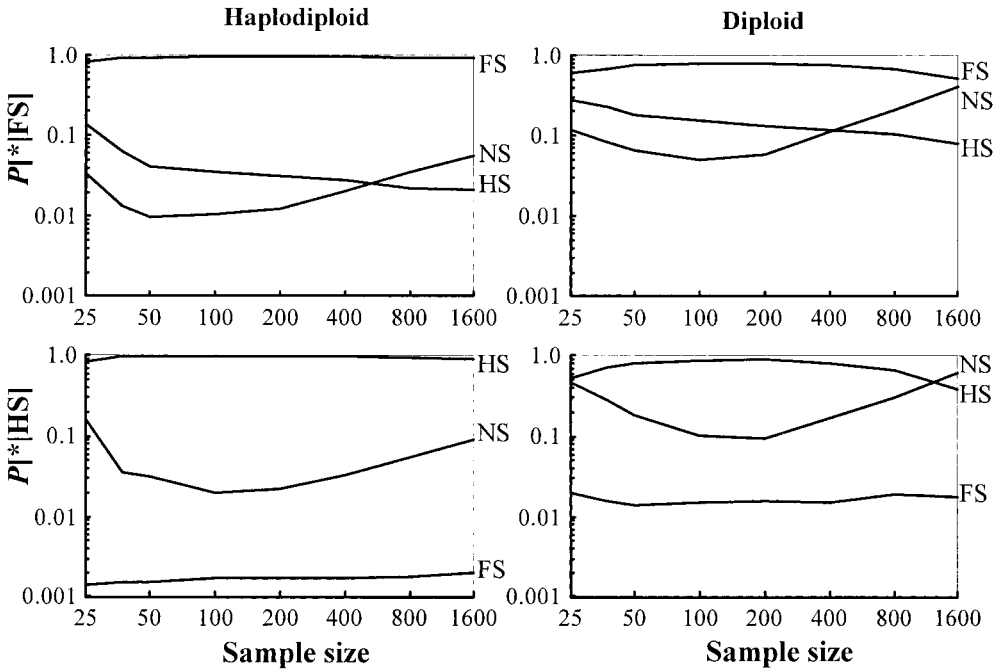

FIGURE 6.—The effect of sample size on relationship inference. Samples of various sizes containing full-sib families nested within half-sib families are simulated, assuming a diploid species or a haplo-diploid species with the polygamous and monogamous parents being diploid and haploid, respectively. The number and size of full-sib families within a half-sib family are drawn from Poisson distributions with parameters 2 and 5, respectively. Each sampled diploid offspring is genotyped for five loci, each having 10 codominant alleles of equal frequency. Typing errors of both classes occur at rate 0.05 for each locus and individual and are accounted for in sibship reconstruction. Lines marked by FS, HS, and NS show the proportions of actual full-sib (top) or half-sib (bottom) pairs being inferred as full-sib, half-sib, and non-sib relationships, respectively.

and sizes of full-sib families being in Poisson distributions with parameters of 3 and 5, respectively, $\overline{\Psi}$ is 96 and 80% for haplo-diploid and diploid species, respectively.

Applications: The method developed in this study has been applied to estimating the number of colonies of two bumble bee species (*Bombus terrestris* and *B. pascuorum*) whose workers visit and use a given foraging site (CHAPMAN *et al.* 2003). As further demonstrations of its usefulness, the method is applied to two empirical data sets.

Analysis on an ant data set: The data set is from a study on the mating frequency of an ant species, *Leptothorax acervorum* (HAMMOND *et al.* 2001). It consists of 377 ant workers (diploid) sampled from 10 known colonies, with each of 6 colonies contributing 45 workers and the remaining 4 colonies contributing 47, 44, 9, and 7 workers to the sample. Each sampled colony is headed by a single (diploid) queen mated with a single (haploid) male. Therefore, the sampled workers are either full-sibs from the same colony or non-sibs from different colonies. These 377 workers are genotyped at up to six microsatellite loci, which have a number of observed alleles varying between 3 and 22. Genotypes at the six loci of nine queens and four of their mates from the 10 sampled colonies were also partially ascertained. The phenotypes of the sampled workers are used alone in estimating the allele frequencies of the population and reconstructing the sibships (colonies) of the sample. The observed parental phenotypes are used only for checking the accuracy of parental genotype inference. The rates for both allelic dropouts and other kinds of errors in this data set are unknown and are assumed to take various values in the analyses. Allele frequencies are updated using reconstructed sibships every 1000 successful reconfigurations.

The likelihood method completely reconstructed the sibships of the sampled 377 workers, using their phenotype information only, without a single worker being assigned an incorrect relationship with any other worker. The 100% successful assignments ($\xi_{FS} = 1$) were obtained with a wide range of possible typing error rates ($\sim$0.001–0.40) assumed in the analyses. However, if typing errors are ignored by setting the error rate as zero, only 6 colonies are fully recovered ($\xi_{FS} = 60\%$) and each of the remaining 4 colonies is split into 2 colonies, resulting in a total number of 14 reconstructed colonies and $P[FS|FS] = 0.96$. The split of the four colonies is due to typing errors. Indeed, a typing error at a single locus in each of the four colonies is identified and reported by the analysis when typing errors are accounted for. Among the four typing errors identified, three can be verified because the observed data show Mendelian inconsistency (*e.g.*, four or more alleles at a locus are observed among workers from a single colony). The other typing error is highly supported by the original data, if Mendelian segregation applies to the locus and colony.

The analysis also inferred the parental genotypes at each locus for each reconstructed family. In total, 67

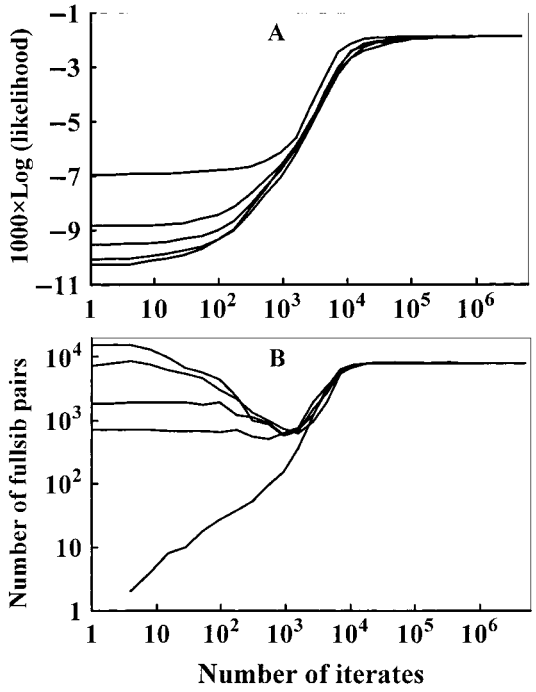

FIGURE 7.—Log-likelihood (A) and number of full-sib pairs (B) as a function of the number of iterates in the analysis of the ant data set. The results are obtained from five replicate runs conducted under identical conditions except that different initial configurations and different seeds for the random number generator are utilized. The lines from bottom to top in the top graph (from top to bottom in the bottom graph) represent runs started from an initial configuration with 5, 10, 40, 100, and 377 full-sib families, each being filled with individuals drawn randomly from the sample of 377 offspring.

single-locus parental phenotypes are available from this data set. If these observed phenotypes are completely correct, the numbers of correctly, incorrectly, and partially (*i.e.*, only one allele correctly inferred for a queen) recovered single-locus parental genotypes are 63, 2, and 2, respectively. The two incorrectly inferred genotypes are at the same locus of a queen and its mate, and the queen is a homozygote. The posterior probability of these two inferred genotypes is 0.54, and that of the alternatively inferred genotypes, which are in full agreement with observations, is 0.46. The two partially recovered parental genotypes occur in the smallest family containing seven offspring in the sample.

The changes in log-likelihood and the number of full-sib pairs as a function of the number of iterates (reconfigurations) during the annealing process are shown in Figure 7 for five independent analyses on the data set. The five replicate runs are carried out in the same conditions using an error rate of 0.05 for both classes of errors at each locus, except that different seeds for the random number generator and different initial configurations are adopted. When typing errors are allowed for at each locus, it is possible to start the simulated annealing algorithm in searching for the maximum-likelihood estimates from an arbitrary initial configuration. As can be seen, all runs converge to the same configuration with the same maximum likelihood after only $\sim 10^6$ iterates, indicating the annealing procedure adopted is powerful and well converged. The same results are obtained assuming different values of error rate in the range of $\sim$0.001–0.4. This differs from SMITH *et al.* (2001) who found a great deal of run-to-run variability in both the maximum likelihood and configuration finally obtained from their likelihood method.

Analysis on a turtle data set: The data set comes from a study on detecting multiple paternity in the Kemp's ridley sea turtle (KICHLER *et al.* 1999). DNA from 26 mother and offspring groups were analyzed at three microsatellite loci to estimate the number of males that mate and contribute to the offspring of each mother. The population allele frequencies at the three loci were ascertained from another larger sample containing individuals with no known close relationships among them. The original analysis of KICHLER *et al.* (1999) inferred the number of males mated with a mother by deducing the number of paternal alleles present in the offspring of the mother. The number of mates thus obtained could be conservative because not all paternal alleles are identifiable. Kichler *et al.* detected three and four paternal alleles in 14 and 1 of the 26 mother-offspring groups, respectively, indicating that at least 58% of female turtles are mated with multiple males. Using a likelihood framework constraining females to matings with either a single male or two males, they obtained the maximum-likelihood estimates that all females are multiply mated, three-quarters of the offspring in a clutch are fathered by a single male, and there are no mutations (typing errors) in data.

KICHLER *et al.*'s (1999) data were reanalyzed by NEFF *et al.* (2002), using their Bayesian method that requires that both the reproductive skew and the number of mates per female be constrained, *a priori*, to some predetermined values. The frequency of multiple mating for females was estimated to be $\sim$70–81%, depending on the particular values of the number of sires contributing to a clutch and the reproductive skew assumed in the analysis.

Assuming an error rate of 0.02 for each locus and individual, I apply the current likelihood method to partition the offspring sampled from each clutch (mother) into full-sib families. Among the 26 sampled females, 5, 10, 8, and 3 females are inferred to have clutches sired by 1, 2, 3, and 4 males, respectively, giving an estimated rate of polyandry of 81%. This estimate is halfway between Kichler *et al.*'s estimates using the paternal allele (58%) and likelihood (100%) methods, and is similar to Neff *et al.*'s estimates ($\sim$70–81%). Among the 21 multiply mated mothers, the average proportion of offspring contributed by a single dominant male is 58%, which is lower than Kichler *et al.*'s likelihood estimate (75%). This is expected because the number of poten-

tial fathers for a clutch of offspring is limited to a maximum of two in their analysis. Under this constraint, offspring from three or more fathers must be allocated to two of them. Assuming an offspring whose father is not inferred is equally likely to be assigned to the two alternative false fathers, then reproductive skew must be overestimated due to the constraint. From the current analysis, 11 (52%) of the 21 polygamous females are mated with three or four males.

The current likelihood analysis has not detected any typing errors in this data set, and analysis assuming a nil error rate gives essentially the same results as that assuming an error rate of 0.02. This is in agreement with Kichler *et al.*'s conclusion that mutations are not important as the cause of multiple paternal alleles detected from the sample.

The current likelihood analysis partitioned the offspring from each mother into full-sib groups without constraining the number of mates and the reproductive skew. The results can be used in further analyses, such as calculating the effective number of mates per female; inferring the distribution of full-sib family sizes within a half-sib family; inferring sperm competition (*e.g.*, Jones and Clark 2003); and estimating the relationships between the number of sires and clutch size, female size, or age, etc.

DISCUSSION

Genotyping errors can occur to almost all kinds of markers in practical use. Their impact varies greatly among different analyses. When allele frequencies rather than genotypes are used in analyses, such as those in estimating population size and migration rate from temporal samples (Wang and Whitlock 2003), we expect little effect of typing errors, except when they are extremely frequent so that changes in allele frequency incurred by them are substantial compared with those caused by other factors (*e.g.*, drift, migration, and sampling). When individual genotypes are employed in analyses such as estimating relatedness and relationships, the consequences of ignoring typing errors critically depend on whether exclusion of an estimate based on genotypes exists in the analyses. For relatedness estimation and sibship inference in diploid species by pairwise likelihood approaches, no estimate is excluded for any possible combination of genotypes, and therefore typing errors should have a small effect in such analyses. In contrast, both parentage inference in pairwise-likelihood approaches and sibship reconstruction in group-likelihood approaches involve exclusions of a particular relationship due to its incompatibility with some genotype combinations and are thus badly affected by typing errors in marker data. Several methods accounting for typing errors have been developed in inferring parentage (*e.g.*, Marshall *et al.* 1998) and other relationships (*e.g.*, Douglas *et al.* 2000; Epstein *et al.* 2000), but this is the first attempt made to incorporate typing errors in group-likelihood approaches to inferring sibships.

Allowing for typing errors in group-likelihood approaches not only improves relationship inference dramatically, but also enables the detection of such errors in data so that one can regenotype those genotypes identified as incorrect. Simulations show that not all typing errors are detectable, and therefore only conservative estimates of errors can be obtained. A typing error identified by the method, however, is highly likely to be genuine, even though relationships are poorly reconstructed due to insufficient marker information and small family size. For the ant data set, the four identified typing errors are all verifiable by checking the original marker data and known colony structures. With large families and sufficient marker information in a sample such as the ant data set, the likelihood method acts as a reliable error detector to pinpoint possible typing errors at each locus in each reconstructed family. Unlike previous methods inferring relationships and typing errors jointly from marker data (*e.g.*, Douglas *et al.* 2000; Epstein *et al.* 2000; Sieberts *et al.* 2002), the current one can consider any number of individuals in a family so that some typing errors, which do not cause Mendelian inconsistency, may become apparent and thus are identified when the family is large. For example, we may observe 10 genotypes of A_1A_1 and one genotype of A_1A_2 at a diploid locus in a putative full-sib family. Obviously the data are in Mendelian consistency, but are highly unlikely if the sibship is true. The present method can also identify such kinds of typing errors. In practice, error rate can be estimated by regenotyping. However, even regenotyping is impossible due to resource constraints; the investigator usually still has some prior information (estimate) of it from previous studies or literature. Simulations (Figure 4) show that the current likelihood method is quite robust to sampling errors of error rate. To be safe, it is better to assume a small error rate in the analysis when we have no information about the reliability of data.

The calculation of family likelihood largely determines the overall computational load of group-likelihood approaches. This is because family-likelihood function involves summing over all possible parental genotype combinations and is thus not trivial in computation. Furthermore, it must be calculated repeatedly in searching for the maximum-likelihood configuration of the entire sample. When relationships are restricted to either full-sibs or unrelated, family likelihood for a single locus can be calculated by one of several polynomial functions of allele frequencies. These polynomial functions are used by Painter (1997) and Smith *et al.* (2001) and can save a substantial amount of computational time. Unfortunately, however, it is difficult, if not impossible, to derive such polynomial functions when half-sibs or typing errors are included in data. Thomas and Hill (2000, 2002) proposed a general and efficient

method to calculate the likelihood of a full-sib or half-sib family from marker data free of typing errors. By choosing a random offspring and assigning one of its alleles to one parent and the other allele to the other parent, their method requires summing over only $k(k+1)/2$ terms [instead of $k^2(k+1)^2/4$], each term being a product of several factors, for the likelihood calculation of a diploid full-sib family and a single locus with k alleles. In comparison, my method needs summing over $(m+1)(m+2)(m^2+3m+4)/8$ terms if m distinctive alleles are observed in a full-sib family and all the $k-m$ unobserved alleles are pooled (see Equation 5a). For a true full-sib family, $m \leq 4$ and can be much smaller than k for typical microsatellites. When $m = 1$, 2, 3, and 4, and when approximately $k \geq 3$, 6, 10, and 15, respectively, my method is more efficient than Thomas and Hill's method. More importantly, their method is not applicable to accounting for typing errors in data.

My simulated annealing algorithm in searching for the maximum-likelihood configuration is well behaved and converged as verified by the analyses of both simulated and real data sets. In simulations, I calculated the likelihood of the true configuration and compared it with the maximum likelihood of the best configuration found for each replicate. For all simulations conducted that are only partly shown in Figures 1–6, the maximum likelihood is always not smaller than the likelihood of the true configuration. Convergence is well evidenced even for very large samples (say, 1600 individuals). Multiple runs on a single data set (*e.g.*, Figure 7) using independent random number series and different initial configurations give identical results. All these suggest the convergence of the proposed algorithm. The computational time required by the current algorithm is determined mainly by sample size, family structure (full- or nested half-sib family), and number of marker loci. For each of the two empirical data sets, the analysis takes ~15 min on a Pentium4 PC.

Mating patterns (monogamous *vs.* polygamous) and reproductive allocations (skew) in social insects and other species are of interest in the fields of ecology, evolutionary genetics, and conservation. Numerous methods have been developed (*e.g.*, Harshman and Clark 1998; Kichler *et al.* 1999; Pedersen and Boomsma 1999; Neff *et al.* 2002; Jones and Clark 2003) to estimate the number of sires and their reproductive skews from the marker genotypes of a brood of offspring and their mother. Most of these methods require constraining either the number of sires or the reproductive skew to estimate both. Group-likelihood methods (Thomas and Hill 2002; this study) make it possible to partition sampled offspring into full-sib families nested within half-sib families, using their genotypes solely or together with the information about brood structure and maternal genotypes. Without any constraint to the quantities of interest, such as actual and effective numbers of mates (paternity) and reproductive skew, they can then be calculated naturally from the partitions.

Currently, all group-likelihood methods constrain the potential relationships in a sample to full-sibs and unrelated (*e.g.*, Painter 1997; Smith *et al.* 2001) or full-sibs, half-sibs, and unrelated (Thomas and Hill 2002; this study). Can these methods infer sibships in data sets containing relationships other than sibships and unrelated? Some simulations show that my group-likelihood method applies to data containing background relationships ignored by the method. The power of the method could be reduced substantially, however. For the case of haplo-diploid species shown in Figure 1, for example, P[FS|FS]s are 99.3 and 100% when typing errors are accounted for and 6 and 10 loci are used in estimation, respectively. However, if half of the mothers of the sampled 100 offspring are from a single full-sib family (thus ~24% pairs of the sampled offspring assumed to be unrelated are actually first cousins), P[FS|FS]s are reduced to 79.1 and 94.9% when 6 and 10 loci are used in estimation, respectively. Similar results are obtained for nested half-sib families and for diploid species. It seems that the method is applicable to sibship inference from data containing unaccounted relationships not too close in relatedness coefficient to sibships and having a reasonable amount of marker information.

Although my group-likelihood approach is based on the models of typing errors commonly found in microsatellites, it can also use other codominant markers (such as allozymes and proteins, single-nucleotide polymorphisms, and RFLPs) in sibship reconstruction. In such cases, allelic dropouts may be omitted and only class II typing errors need to be considered. Given the robustness of the error models as verified by both simulated and empirical data, any small deviations in error patterns between different markers should have little effect on the power of the method. Obviously, these less polymorphic markers necessitate more loci to achieve the same power of inference as microsatellites. In Figure 1, P[FS|FS] is 0.955 for a diploid species when 6 loci, each with 10 alleles of an equal frequency and each with an error rate of 0.05, are used in sibship inference. When 20, 30, and 40 loci, each having 2 alleles of an equal frequency and an error rate of 0.0190, 0.0131, 0.0095, respectively, to give the same probability of an erroneous multilocus genotype, are used in inferring sibships, P[FS|FS]s are 0.674, 0.897, and 0.970, respectively. It is difficult to determine exactly how many loci are necessary to achieve a certain level of power in sibship inference. In parentage analysis, one can use the exclusion probability of a marker to quantify its ability to exclude a random individual from paternity. Such a probability depends solely on the number and frequencies of alleles for a given locus and can be calculated for multiple loci (*e.g.*, Gerber *et al.* 2000). In practice, such probabilities help to choose marker loci and to determine the power of parentage inference.

Although a similar exclusion probability can be defined and calculated for sibship inference, it would be highly dependent on both size and structure (full- and half-sibs) of sib families, in addition to marker properties. For a given marker, higher accuracy of sibship assignment is obtained from samples with larger sibship sizes. In diploid species, exclusion of sibships is impossible for pairs of individuals or for diallelic markers. Furthermore, the inclusion of typing errors makes the calculation of exclusion probability even more problematic. As a rough guide, one may characterize the amount of information from a marker locus by the heterozygosity in sibship inference.

With slight modification, my group-likelihood method can use dominant markers, such as random amplified polymorphic DNA and amplified fragment length polymorphism, separately or in conjunction with codominant markers in sibship inference. These dominant markers are less informative but less expensive than microsatellites and have been applied to parentage analysis of many animal and plant species (see Gerber *et al.* 2000 and many references therein). In contrast, they are invariably ignored in previous group-likelihood approaches to sibship reconstruction. Consider a dominant locus with two alleles, A and a. There are two possible phenotypes, the dominant one (denoted by D) with two possible genotypes AA and Aa and the recessive one (denoted by d) with one possible genotype aa. Assuming the class II error model with error rate e, we can obtain the transitional probability of an actual genotype (AA, Aa, or aa) to an observed phenotype (D or d):

$$\begin{aligned}
\Pr[D|AA] &= 1 - e^2 \\
\Pr[D|Aa] &= 1 - e + e^2 \\
\Pr[D|aa] &= 2e - e^2 \\
\Pr[d|AA] &= e^2 \\
\Pr[d|Aa] &= e(1 - e) \\
\Pr[d|aa] &= (1 - e)^2.
\end{aligned}$$

These transitional probabilities should be used instead of (1) and (2) in the same family-likelihood functions as codominant markers in sibship reconstruction.

One limitation of current group-likelihood methods is that they allow for a single sex being multiply mated. Extension to the more general case of polygamy for both sexes is planned. It should also be possible to infer other relationships (such as identical twins and cousins) jointly with full-sib, half-sib, and unrelated relationships in a sample by the same group-likelihood framework. It should be noted, however, that some inferences made by the current likelihood model are conditional on point estimates of some (nuisance) parameters. For example, sibships are inferred conditional to fixed typing error rates while typing errors and parental genotypes are inferred conditional to an estimated sibship structure. A full Bayesian approach to the joint estimation of all the parameters in the model would account for the uncertainty related to point estimates of each nuisance parameter and allow for the incorporation of any prior information about the parameters.

A software package, COLONY, implementing the likelihood method described in this article, is available for free download from http://www.zoo.cam.ac.uk/ioz/software.htm.

I thank Andrew Bourke and Rob Hammond for providing me with their original data and Andrew Bourke, Rob Hammond, Bill Hill, Bill Jordan, and two anonymous referees for constructive comments on earlier versions of this manuscript.

LITERATURE CITED

Almudevar, A., and C. Field, 1999 Estimation of single-generation sibling relationships based on DNA markers. J. Agric. Biol. Environ. Stat. **4:** 136–165.

Blouin, M. S., 2003 DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. TREE **18:** 503–511.

Blouin, M. S., M. Parsons, V. Lacaille and S. Lotz, 1996 Use of microsatellite loci to classify individuals by relatedness. Mol. Ecol. **5:** 393–401.

Boehnke, M., and N. J. Cox, 1997 Accurate inference of relationships in sib-pair linkage studies. Am. J. Hum. Genet. **61:** 423–429.

Chapman, R. E., J. Wang and A. F. G. Bourke, 2003 Genetic analysis of spatial foraging patterns and resource sharing in bumble bee pollinators. Mol. Ecol. **12:** 2801–2808.

Douglas, J. A., M. Boehnke and K. Lange, 2000 A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. Am. J. Hum. Genet. **66:** 1287–1297.

Epstein, M. P., W. L. Duren and M. Boehnke, 2000 Improved inference of relationship for pairs of individuals. Am. J. Hum. Genet. **67:** 1219–1231.

Frankham, R., 1995 Conservation genetics. Annu. Rev. Genet. **29:** 305–327.

Gagneux, P., C. Boesch and D. S. Woodruff, 1997 Microsatellite errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hairs. Mol. Ecol. **6:** 861–868.

Gerber, S., S. Mariette, R. Streiff, C. Bodenes and A. Kremer, 2000 Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. Mol. Ecol. **9:** 1037–1048.

Hammond, R. L., A. F. G. Bourke and M. W. Bruford, 2001 Mating frequency and mating system of the polygynous ant, *Leptothorax acervorum.* Mol. Ecol. **10:** 2719–2728.

Harshman, L., and A. G. Clark, 1998 Inference of sperm competition from broods of field-caught Drosophila. Evolution **52:** 1334–1341.

Herbinger, C. M., R. W. Doyle, E. R. Pitman, D. Paquet, K. A. Mesa *et al.*, 1995 DNA fingerprint based analysis of paternal and maternal effects on offspring growth and survival in communally reared rainbow trout. Aquaculture **137:** 245–256.

Jones, A. G., and W. R. Ardren, 2003 Methods of parentage analysis in natural populations. Mol. Ecol. **12:** 2511–2523.

Jones, B., and A. G. Clark, 2003 Bayesian sperm competition estimates. Genetics **163:** 1193–1199.

Kichler, K., M. T. Holder, S. K. Davis, R. Márquez-M and D. W. Owens, 1999 Detection of multiple paternity in the Kemp's ridley sea turtle with limited sampling. Mol. Ecol. **8:** 819–830.

Kirkpatrick, S., C. D. Gelatt and M. P. Vecchi, 1983 Optimization by simulated annealing. Science **220:** 671–680.

Lynch, M., and J. B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, MA.

Marshall, T. C., J. Slate, L. E. B. Kruuk and J. M. Pemberton, 1998 Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. **7:** 639–655.

McPeek, M. S., and L. Sun, 2000 Statistical tests for detection of misspecified relationships by use of genome-screen data. Am. J. Hum. Genet. **66:** 1076–1094.

Neff, B. D., T. E. Pitcher and J. Repka, 2002 A Bayesian model for assessing the frequency of multiple mating in nature. J. Hered. **93:** 406–414.

O'Reilly, P. T., C. Herbinger and J. M. Wright, 1998 Analysis of parentage determination in Atlantic salmon (*Salmo salar*) using microsatellites. Anim. Genet. **29:** 363–370.

Painter, I., 1997 Sibship reconstruction without parental information. J. Agric. Biol. Environ. Stat. **2:** 212–229.

Pedersen, J. S., and J. J. Boomsma, 1999 Multiple paternity in social Hymenoptera: estimating the effective mate number in single-double mating populations. Mol. Ecol. **8:** 577–587.

Queller, D. C., and J. E. Strassmann, 1998 Kin selection and social insects. BioSciences **48:** 165–175.

SanCristobal, M., and C. Chevalet, 1997 Error tolerant parent identification from a finite set of individuals. Genet. Res. **70:** 53–62.

Sieberts, S. K., E. M. Wijsman and E. A. Thompson, 2002 Relationship inference from trios of individuals, in the presence of typing error. Am. J. Hum. Genet. **70:** 170–180.

Smith, B. R., C. M. Herbinger and H. R. Merry, 2001 Accurate partition of individuals into full-sib families from genetic data without parental information. Genetics **158:** 1329–1338.

Taberlet, P., S. Griffin, B. Goossens, S. Questiau, V. Manceau *et al.*, 1996 Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Res. **24:** 3189–3194.

Talbot, C. C., Jr., D. Avramopoulos, S. Gerken, A. Chakravarti, J. A. Armour *et al.*, 1995 The tetranucleotide repeat polymorphism D21S1245 demonstrates hypermutability in germline and somatic cells. Hum. Mol. Genet. **4:** 1193–1199.

Thomas, S. C., and W. G. Hill, 2000 Estimating quantitative genetic parameters using sibships reconstructed from marker data. Genetics **155:** 1961–1972.

Thomas, S. C., and W. G. Hill, 2002 Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. Genet. Res. **79:** 227–234.

Thompson, E. A., 1975 The estimation of pairwise relationships. Ann. Hum. Genet. **39:** 173–188.

Wang, J., and M. C. Whitlock, 2003 Estimating effective population size and migration rates from genetic samples over space and time. Genetics **163:** 429–446.

Communicating editor: J. B. Walsh