Heredity (2013) 111, 165–174

www.nature.com/hdy

ORIGINAL ARTICLE

An improvement on the maximum likelihood reconstruction of pedigrees from marker data

J Wang

Many methods have been proposed to reconstruct the pedigree of a sample of individuals from their multilocus marker genotypes. These methods, like those in other fields of statistical inferences, may suffer from both type I (falsely related) and type II (falsely unrelated) errors. In sibship reconstruction, type I errors come from the spurious fusion of two or more small sibships into a single sibship, and type II errors originate from the spurious splitting of a large sibship into two or more small sibships. In this study I investigate the tendencies of both types of errors made by the likelihood methods in sibship reconstruction, using both analytical and simulation approaches. I propose an improvement on the likelihood methods to reduce sibship splitting, and thus type II errors by downscaling the number of inferred siblings sharing the same genotype at a locus. Simulations are then conducted to compare the accuracy of the original and improved likelihood methods in sibship reconstruction of a large sample of individuals in full-sib families of the same small size, the same large size and highly variable sizes, using a variable number of loci with a variable number of alleles per locus. The methods were also applied to the analysis of a salmon data set. I show that my scaling scheme prevents effectively the splitting of large sibships, and reduces type II errors greatly with little increase in type I errors. As a result, it improves the overall accuracy of sibship assignments, except when sibships are expected to be uniformly small or marker information is unrealistically scarce.
Heredity (2013) **111,** 165–174; doi:10.1038/hdy.2013.34; published online 24 April 2013

Keywords: full-sibs; half-sibs; parentage; genetic markers; maximum likelihood; pedigree

INTRODUCTION

Pedigrees delineate the genealogical relationships of individuals and are invaluable in many research fields such as molecular ecology, conservation biology, forensics and human medicine (see reviews of Blouin, 2003; Pemberton, 2008; Jones and Wang, 2010). When breeding records are unavailable, incomplete or unreliable, the pedigree of a sample of individuals can be reconstructed solely from the genetic marker data of the individuals, using a number of statistical methods. These methods are invariably based on Mendel's laws of inheritance, but differ greatly in the statistical treatment of the marker data (Wang, 2012). The simplest methods calculate the likelihoods of different candidate relationships for a pair (for example, Marshall *et al.*, 1998; Epstein *et al.*, 2000; McPeek and Sun, 2000) or trio (for example, Sieberts *et al.*, 2002) of individuals, and choose the relationship with the maximum likelihood as the estimate. They are simple to implement and capable of modeling linkage among markers and inferring multiple types of relationships. However, these pairwise or triowise methods fail to use the valuable marker information efficiently. Any relatives of the focal dyad or trio provide vital information (Wang, 2007) about the relationship of the dyad or trio, but are simply ignored in the estimation procedure. Furthermore, the inferred pairwise or triowise relationships may be incompatible when they are assembled for an entire sample of individuals (Wang, 2012).

More sophisticated methods infer the relationships of all individuals in a sample simultaneously to use the marker information efficiently. This is feasible for sibship inference in a one-generation sample (for example, Painter, 1997; Smith *et al.*, 2001; Thomas and Hill, 2000, 2002; Wang, 2004; Butler *et al.*, 2004; Konovalov *et al.*, 2004; Berger-Wolf *et al.*, 2007; Almudevar and Anderson, 2012), and for both sibship and parentage assignments in a two-generation sample (for example Emery *et al.*, 2001; Wang and Santure, 2009; Wang, 2012). These methods also vary vastly in how marker information is used statistically in the inference procedure. Some depend on the exclusion rules derived from Mendelian segregation law (for example, Butler *et al.*, 2004; Konovalov *et al.*, 2004; Berger-Wolf *et al.*, 2007), some rely on the likelihood (for example, Painter, 1997; Thomas and Hill, 2000; Emery *et al.*, 2001; Wang, 2004) or its surrogate such as the pairwise likelihood score (for example, Smith *et al.*, 2001; Wang, 2012) of the entire sample of individuals, while others screen feasible sibship assignments by exclusions and then choose the best assignment based on a score that is part of a likelihood function (for example, Almudevar and Anderson, 2012).

In general, exclusion based methods are simpler and computationally more efficient than likelihood based methods. However, they are powerless for a sample containing numerous small sibships and do not apply to lowly polymorphic markers such as single-nucleotide polymorphisms and restriction fragment length polymorphisms. They also have difficulties in dealing with genotyping errors of data, in estimating multiple relationships, such as full and half sibships and parentage (Wang, 2012), and in utilizing known relationships (for example, maternity) and prior information (for example, sibship size distribution). Likelihood methods are

Institute of Zoology, Zoological Society of London, London, UK
Correspondence: Dr J Wang, Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, UK.
E-mail: jinliang.wang@ioz.ac.uk
Received 14 August 2012; revised 18 March 2013; accepted 26 March 2013; published online 24 April 2013

flexible, robust, accurate and apply to various markers (for example, dominant and codominant markers with two or more alleles, with or without genotyping errors). However, they are computationally demanding, especially in the case of polygamy of both sexes and markers suffering from genotyping errors (Wang and Santure, 2009; Wang, 2012).

All methods in reconstructing pedigrees from marker data are statistical and thus could suffer from both type I (falsely related) and type II (falsely unrelated) errors, where the error types are defined with unrelated as the null hypothetical relationship. In principle, exclusion based methods should have few type II errors, if the markers follow Mendelian segregation law and have no typing errors and mutations. The exclusion rules are inherently sufficient for exclusion, but insufficient for nonexclusion in sibship or parentage analyses. Although a set of individuals can be confidently excluded from a full sibship if their genotypes conform to any of the exclusion rules (for example, displaying more than four alleles or more than two types of homozygotes) at any locus, they cannot be assigned to a full sibship with confidence. For example, m individuals with the same genotype AA and n individuals with the same genotype BB at a diploid codominant locus are always compatible with, and thus nonexcludable from, a full sibship, because the $m+n$ genotypes can be generated by a pair of parents with the same genotype AB. However, the likelihood that these $m+n$ individuals are full siblings relative to the likelihood that they come from two distinctive full-sib families (one with m individuals of genotype AA, and the other with n individuals of genotype BB) diminishes rapidly with an increasing value of $m+n$. For another example, any two unrelated individuals will never be excluded from a full-sib family, and any three or more unrelated individuals will never be excluded from a single full-sib family for a marker with two alleles. While unrelated individuals identified by exclusion methods are highly likely to be true in the absence of mutations and mistypings, related individuals (for example, siblings) inferred by exclusion methods may actually be unrelated. In contrast, likelihood methods may suffer from both type I and type II errors. It thus seems as if likelihood methods are less accurate than exclusion methods, but in fact the opposite is true. Any robust and accurate statistical inference framework must have intrinsically a delicate balance between the occurrences of type I and II errors. If a statistical framework is inherently more prone to one type of errors than to the other, then its power or/and application scope must be quite limited. Indeed exclusion based methods for sibship reconstruction perform well only in the special case of a sample containing few small sibships and containing highly polymorphic codominant markers without genotyping errors and mutations.

Despite the popularity of likelihood methods in pedigree reconstruction, surprisingly little work has been done to investigate their frequencies of type I (for exclusion methods, see Almudevar and Field, 1999; Wang, 2007) and II errors, and the factors affecting these frequencies. Such work helps in improving the likelihood methodology and aids in optimizing the experimental design of a practical pedigree analysis. In a likelihood sibship analysis, unrelated individuals may be inferred as siblings (type I error, sibship fusion) and siblings may be inferred as unrelated (type II error, sibship split). The latter was reported by several simulation studies (for example, Thomas and Hill, 2000; Butler *et al.*, 2004), but was regarded as unimportant when marker information is sufficient (Wang, 2004; Wang and Santure, 2009). Recently, Almudevar and Anderson (2012) indicated that the splitting of a large sibship is an inherent property of the maximum likelihood methods, and derived a formula to predict the amount of marker information required to prevent sibship splitting.

In this study, I will formally investigate the frequencies of type I and II errors of likelihood methods in sibship reconstruction by both analytical treatment and simulations. I will then propose an improvement on the likelihood methods to reduce sibship splitting by scaling down the number of inferred siblings with an identical genotype at a locus used in likelihood calculations. Simulations are then conducted to compare the accuracy of the original and improved likelihood methods for reconstructing the sibship of a large sample of individuals coming from many small full-sib families, a few very large full-sib families and a mixture of small and large full-sib families. The methods were also applied to the analysis of a salmon data set. I showed that my scaling scheme reduces effectively the splitting of large sibships, reduces type II errors with little increase in type I errors, and improves the overall accuracy of sibship assignments except when marker information is very scarce and family sizes are uniformly small.

FREQUENCY OF TYPE I ERRORS

Multiple small sibships, such as singletons, might be spuriously fused by likelihood methods because they have by chance similar genotypes compatible with a single sibship. The rate of these errors increases with a decrease in marker information, and an increase in the proportion of small sibships in a sample. A fusion may involve true sibships of various sizes, but occurs most often between two singletons. Therefore I will focus on calculating the frequency of fusing two unrelated individuals into a single full-sib family to indicate the tendency of type I errors. Such an error occurs when the two individuals has a higher likelihood as full siblings than the likelihood as unrelated.

Let's consider a marker with n codominant alleles, A_i for $i=1, 2, \ldots, n$, in a population under Hardy–Weinberg equilibrium. There are seven types of pairs of genotypes (ignoring order) as listed in Table 1. The probability of observing each pair under the null (H0: unrelated, UR) and alternative (H1: full-sibs, FS) hypothesis are the likelihood of UR, L_0, and the likelihood of FS, L_1, respectively. Given the frequency p_i for A_i in the population, L_0 and L_1 can be derived for each pair of unrelated genotypes drawn at random from the population (Table 1). Fusion occurs when $L_1 > L_0$, which is always true for pairs of identical genotypes $\{A_iA_i, A_iA_i\}$ and $\{A_iA_j, A_iA_j\}$. It is also true for pairs of similar genotypes sharing one or more alleles, which are $\{A_iA_i, A_iA_j\}$ and $\{A_iA_j, A_iA_k\}$ (where $j \neq i$ and $k \neq i, j$) when the shared allele A_i has a small frequency of $p_i < 1/3$ and $p_i < 1/6$, respectively. Denoting the sets of alleles with frequencies smaller than

Table 1 Likelihoods for pairs of individuals

Dyad	L_0	L_1	$(L_1 > L_0)$?
$\{ii,ii\}$	p_i^4	$\frac{1}{4}p_i^2(1+p_i)^2$	Yes
$\{ii,ij\}$	$2p_i^3p_j$	$\frac{1}{2}p_i^2(1+p_i)p_j$	Yes, $p_i < \frac{1}{3}$
$\{ii,jj\}$	$p_i^2p_j^2$	$\frac{1}{2}p_i^2p_j^2$	No
$\{ii,jk\}$	$2p_i^2p_jp_k$	$\frac{1}{2}p_i^2p_jp_k$	No
$\{ij,ij\}$	$4p_i^2p_j^2$	$\frac{1}{2}p_ip_j(1+p_i+p_j+2p_ip_j)$	Yes
$\{ij,ik\}$	$4p_i^2p_jp_k$	$\frac{1}{2}p_i(1+2p_i)p_jp_k$	Yes, $p_i < \frac{1}{6}$
$\{ij,kl\}$	$4p_ip_jp_kp_l$	$p_ip_jp_kp_l$	No

The first column lists allele indexes in a genotype dyad, where $j \neq i$, $k \neq \{i,j\}$ and $l \neq \{i,j,k\}$. The second and third columns give the likelihood that the two individuals are unrelated (L_0) and full siblings (L_1), respectively. The last column indicates whether $L_1 > L_0$ or not, and the conditions for $L_1 > L_0$ if any.

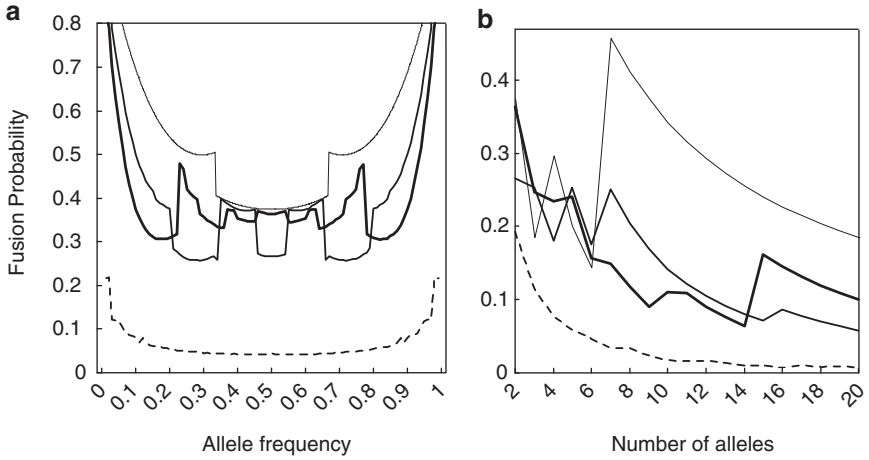

Figure 1 Fusion probability as a function of the allele frequencies at 1 (thin line), 2 (thick line), 3 (very thick line) and 40 (broken line) biallelic loci (**a**), and as a function of the number of equifrequent alleles at 1 (thin line), 2 (thick line), 3 (very thick line) and 10 (broken line) loci (**b**).

1/3 and 1/6 by Ψ and Ω, respectively, I obtain the fusion probability by summing the frequencies of dyads that have $L_1 > L_0$,

$$Q_1 = \sum_{i=1}^{n} p_i^4 + \sum_{\substack{i=1 \\ i\in\Psi}}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} 4p_i^3 p_j + \sum_{i=1}^{n} \sum_{j=i+1}^{n} 4p_i^2 p_j^2 + \sum_{\substack{i=1 \\ i\in\Omega}}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \sum_{\substack{k=j+1 \\ k\neq i}}^{n} 8p_i^2 p_j p_k \quad (1)$$

In the case of equifrequent ($p_i = 1/n$) alleles, (1) reduces to $Q_1 = (2n-1)/n^3$, $Q_1 = (6n-5)/n^3$ and $Q_1 = (4n^2-6n+3)/n^3$ when $n = 2$–3, $n = 4$–6, and $n > 6$, respectively.

It is difficult to derive a simple expression of Q_1 for multiple loci, because the number of genotype combinations increases rapidly with an increasing number of loci (L). Simulations are used instead to obtain Q_1 when $L > 1$. The results are shown in Figure 1a for $L = 1$–3 and $L = 40$, each locus having two alleles of variable frequencies. For the case of $L = 1$, both simulated and analytical (calculated from (1)) Q_1 values were obtained, which are almost identical and thus only the analytical results were presented in Figure 1a for clarity. The first to notice is that the relationship between allele frequency (P) and Q_1 is nonlinear and complicated, especially when P is intermediate. This is because both the number of types and the frequencies of genotype pairs having $L_1 > L_0$ vary with P. At $P = 1/3$, genotype pair $\{A_iA_i, A_iA_j\}$ has exactly $L_1 = L_0$ with a frequency of 0.099. The abrupt changes in Q_1 at $P = 1/3$ and 2/3 are caused by this mating type. The curve of Q_1 as a function of allele frequency becomes smooth only when L becomes large.

Similarly, for the case of equifrequent alleles, Q_1 does not decline smoothly with an increasing number of alleles at a locus (Figure 1b). A four allele and seven allele locus has a higher fusion probability than a three allele and six allele locus, respectively. With an increasing number of loci, however, the curve becomes smoother, Q_1 becomes smaller, and loci with more alleles always have smaller Q_1 values. The abrupt increases in Q_1 at $n = 4$ and 7 for a single locus disappear when $L \gg 1$. When $L = 10$, for example, the Q_1 values at $n = 3$, 4, 6 and 7 are 0.12, 0.08, 0.05 and 0.03, respectively.

It should be emphasized that, strictly speaking, the above results apply to pairwise approaches to full-sib inference, where each pair of individuals are analyzed in isolation. The Q_1 value is usually higher than that of a joint sibship analysis, in which all sampled individuals are considered for sibship assignments jointly, as implemented in Colony (Wang, 2004). This is understandable from a simple example. Suppose three unrelated individuals X, Y and Z have genotypes $\{A_iA_j\}$, $\{A_iA_i\}$ and $\{A_jA_j\}$, respectively at a locus with $n = 4$ equifrequent alleles. With the pairwise approach, {(X, Y)} and {(X, Z)} will be assigned full sibship, and {(Y), (Z)} will be assigned non-sibship according to Table 1. This results in two errors out of three pairwise relationships. In a joint sibship analysis, the five possible sibship configurations are {(X, Y, Z)}, {(X, Y), (Z)}, {(X, Z), (Y)}, {(Y, Z), (X)}, {(X), (Y), (Z)}, with log-likelihood values −7.62, −7.40, −7.40, −9.01, −7.62. The maximum likelihood configuration is {(X, Y), (Z)} or {(X, Z), (Y)}, each with a single error out of three pairwise relationships. Joint sibship analysis uses marker information more efficiently and as a result has a higher accuracy. It also avoids conflict inferences typical of the pairwise approach.

FREQUENCY OF TYPE II ERRORS

A large sibship consisting of many siblings might be spuriously split by the likelihood method when marker information in support of the sibship is insufficient. The split happens when the likelihood of the sibship is smaller than the product of the likelihoods of two or more split sibships. For example, mating $A_iA_i \times A_iA_j$ may produce a sibship consisting of two subsets of offspring. One subset has m_1 genotypes A_iA_i, and the other has m_2 genotypes A_iA_j. When marker information is insufficient (that is, both A_i and A_j are not rare) in support for a large ($m_1 + m_2 \gg 1$) sibship, it might be split into two subsets, each as a reconstructed sibship to result in a larger overall likelihood.

There is no doubt that likelihood methods for sibship reconstruction have the risk of splitting large sibships. The questions are how often these type II errors occur, how severe type II errors are relative to type I errors, and what factors affect the rate of type II errors. Large sibship splitting was noticed in both pairwise likelihood (for example, Butler *et al.*, 2004) and joint likelihood (for example, Thomas and Hill, 2000) methods of sibship reconstruction. Almudevar and Anderson (2012) made some simple analysis of the splitting of large sibships, and pointed out that it is an inherent problem of the likelihood method. They derived a criterion $p^{4L} \gg 0.5^m$ (where L is the number of loci, m is the true sibship

size and P is the representative allele frequency) to predict sibship splitting by likelihood methods.

It is difficult to derive an exact yet simple expression for the split probability even for a single equifrequent-allele locus. Part of the difficulty comes from the many possible offspring genotype combinations that must be considered for a large sibship produced by one or two heterozygous parents. When both parents are homozygotes (see Table 1), then all offspring will be of the same genotype, and split does not occur regardless of m and allele frequencies. To gain some insight into the sibship split problem, let us consider a sibship consisting of m offspring produced by mating $A_iA_i \times A_jA_k$ at a locus with n equifrequent alleles ($P=1/n$). According to Mendelian segregation law, the mating produces two possible offspring genotypes, A_iA_j and A_iA_k, with the same probability of ½. Therefore, the number of A_iA_j genotypes follows a binomial distribution, $m_1 \sim B(m, 0.5)$. When $m_1=0$ or m with a probability of 0.5^{m-1}, split does not occur. When $0<m_1<m$, the sibship splits into a sibship of m_1 genotypes A_iA_j and m_2 genotypes A_iA_k, if the split likelihood,

$$L_2 = \frac{1}{n^8}\prod_{i=1}^{2}\left(2+2^{3-2m_i}n(n-2)+2^{2-m_i}(2n-1)\right)$$

is larger than the nonsplit likelihood,

$$L_1 = \frac{2^{2-2m}(2^m+2n)}{n^4}$$

Both L_1 and L_2 can be obtained from the general sibship likelihood function (for example, Wang, 2004; Equation (2) below) or more conveniently from one of the polynomial functions for 14 feasible full-sib genotype configurations (for example, Painter, 1997). Figure 2a plots the ratio L_1/L_2 as a function of m, for different values of m_1 at a locus with $n=10$ equifrequent alleles. As can be seen, split is guaranteed to occur when m is large ($\geqslant 12$); otherwise, it occurs if m_2 is much larger or smaller than m_1.

The split probability for this type of mating can be obtained by summing over the frequencies of all possible offspring genotype combinations (with constraints m_1, $m_2=0$–m and $m_1+m_2=m$) that have $L_1<L_2$ weighted by their occurrence probabilities. The resultant expression is complicated and unenlightening, but some numerical values calculated from it for a locus with $n=5$, 10, 20 and 40 equifrequent alleles are shown in Figure 2b. At the same m value, split probability decreases very rapidly with an increase in n or a decrease in allele frequency (remember herein $P=1/n$), as found by Almudevar and Anderson (2012). At the same allele frequency (or value of n), split probability increases rapidly with an increasing sibship size, m. My derivation gives numerical results similar to Almudevar and Anderson's criterion $p^{4L} \gg 0.5^m$. For $P=1/n$ with $n=5$, 10, 20 and 40, for example, both methods predict that split occurs with a probability of 1 for a sibship with $m>9$, 13, 17, 21 offspring (Figure 2b), respectively. However, my method can also calculate the n and m values that result in a split probability smaller than 1.

Neither the above analysis nor Almudevar and Anderson's criterion is complete because, out of the seven possible matings listed in Table 1, only one is considered. Different mating types will produce sibships highly variable in the tendency of splitting up in marker based reconstruction. Sibship from the two mating types in which both parents are homozygotes (Table 1) will never be split, with a probability of $1/n^2$ for a locus with n equifrequent alleles. However, as a simple yet good approximation, the criterion is useful in delineating the relationship among P, L and m for determining large sibship splitting in the likelihood methods.

AN IMPROVED LIKELIHOOD METHOD

The above analyses on sibship fusion and splitting probabilities show that both type I and type II errors are 'inherent' properties of the likelihood methods. How often these errors occur depends on many factors, the most important being marker information content and the actual sibship size distribution. When most individuals have no siblings in a sample, sibship fusion is more problematic than sibship splitting. When the sample is dominated by a few very large sibships, splitting and type II errors are more severe than fusion and type I errors. However, irrespective of the actual sibship size distributions, both types of errors should decrease rapidly with an increase in marker information.

Both the above sibship splitting analysis and Almudevar and Anderson's criterion suggest that sibship splitting probability increases rapidly with an increase in actual sibship size. Therefore, I propose to reduce sibship splitting by scaling down the number of siblings displaying the same genotype in likelihood calculations. The original likelihood function for a pure full sibship with unknown parental genotypes at a single locus of n alleles is

$$\begin{aligned} L(\mathrm{FS} \mid g, m) &= \Pr(g, m \mid \mathrm{FS}) \\ &= \sum_{w=1}^{n} p_w \sum_{x=1}^{n} p_x \sum_{y=1}^{n} p_y \sum_{z=1}^{n} p_z \prod_{i=1}^{d} \Pr(g_i \mid w, x; y, z)^{m_i} \end{aligned} \qquad (2)$$

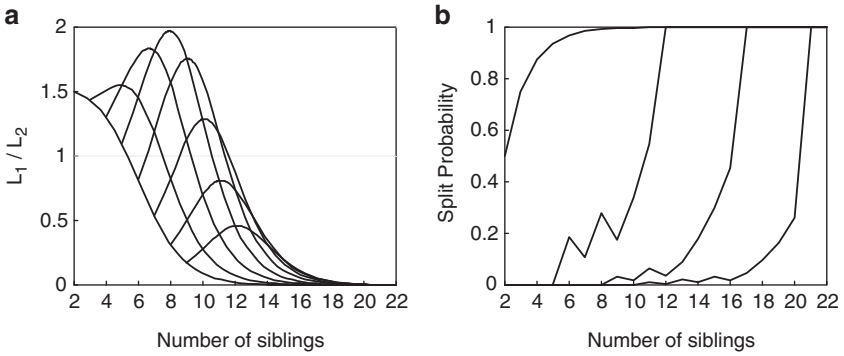

Figure 2 Likelihood ratio (L_1/L_2, left) and split probability (right) as a function of the number of siblings (m) produced by mating $A_iA_i \times A_jA_k$ at a locus. (**a**) assumes a locus with 10 equifrequent alleles, where curve i (=1–8) counting from the left (that is, starting at $m=i+1$) refers to $m_1=i$. (**b**) assumes a locus with different numbers (n) of equifrequent alleles, where curve i counting from the left refers to $n=5$, 10, 20, and 40, respectively.

where $g=(g_1, g_2, \ldots, g_d)$ are the d distinctive genotypes and $m=(m_1, m_2, \ldots, m_d)$ are the counts of the d distinctive genotypes observed among the inferred siblings, w and x index the alleles in one parent and y and z index the alleles in the other parent. The probability of an offspring genotype g_i given parental alleles $\{w, x\}$ and $\{y, z\}$, $\Pr(g_i|w, x; y, z)$, can be derived from Mendelian segregation law, with the possibility of accommodating genotyping errors (Wang, 2004).

A sibship splits when m_i is large and marker information in support of the sibship is insufficient, as shown above. A solution to the problem is to scale down m_i to reduce sibship splitting and thus the frequencies of type II errors. However, scaling down m_i could potentially promote the fusion of small sibships and increase the frequency of type I errors. Other factors that affect both type I and II errors and the scaling effects are the number of alleles (n) and the rate of genotyping errors (e, the sum of allelic dropout rate and false allele rate) at a locus. Having considered the known effects of n, e, and m_i on sibship fusion and splitting, and after extensive experimentation on simulated and empirical data, I arrived at the scaling scheme

$$M_i=\left(m_i^q+1-\delta_{1,q}\right)\left(1-\delta_{1,m_i}\right)+\delta_{1,m_i}, \quad (3)$$

where $q=(0.5+1/n)(1-2e)+2e$ with $e\leqslant 0.5$, the Kronecker delta $\delta_{1,m_i}=1$ and 0 when $m=1$ and $m_i\neq 1$, respectively, and $\delta_{1,q}=1$ and 0 when $q=1$ and $q\neq 1$, respectively. Note that q is calculated from n, which is the number of alleles at a locus observed in a sample of individuals, and e, which is the mistyping rate estimated and supplied by a researcher.

With this scaling scheme, m_i at a locus with more alleles observed in a sample of individuals (thus a lower average allele frequency and a higher power for excluding false sibship in general) and with a lower mistyping rate has a smaller q value and is more downscaled. The minimum q value is 0.5 when n is big and e is small, which leads to $M_i=\sqrt{m_i}+1$ for $m_i>1$. The maximum q value is 1 when $n=2$ or $e=0.5$, which leads to $M_i\equiv m_i$ (that is, no scaling). The reason that m_i at a less informative locus (small n or high e, or both) is less downscaled is that the rate of type I errors is expected to be high for such a locus, and any downscaling would make this problem worse. This *ad hoc* scaling scheme is arrived at by balancing the frequencies of both type I and type II errors, considering the locus specific properties (e and n), which may affect these errors. The performance of the scaling scheme was checked by numerous simulations under various parameter combinations, some of which being shown below. In all simulations, the scheme was shown to effectively reduce the splitting of large sibships (if present) with little increase in the fusion of small sibships (if present).

To understand how the scaling works to reduce sibship splitting, let us consider a numerical example. Suppose a set of $L=10$ microsatellites, each having 10 equifrequent alleles ($p=0.1$), are used to reconstruct the sibship of $m=200$ full siblings contained in a sample of individuals. This large sibship is highly likely to be split into two or more sibships by the original likelihood method, because the splitting criterion $p^{4L}\gg 0.5^m$ is met, where $p^{4L}=10^{-40}$ and $0.5^m=6.2\times 10^{-61}$. By applying the scaling scheme, we have $q=0.6$, $M=m^q+1=25$, $0.5^M=3\times 10^{-8}$ and the splitting criterion $p^{4L}\gg 0.5^M$ is not met (that is, $p^{4L}\ll 0.5^M$). As a result, this large sibship will be reconstructed without splitting by the improved likelihood method.

With the scaling scheme, the likelihood of a sibship is still calculated by (2), but m_i should be replaced by M_i. Similarly, the likelihood function for a more complicated pedigree (for example, containing half-sib, full-sib and parent–offspring relationships) (Wang and Santure, 2009) still applies, except m_i is replaced by M_i.

SIMULATIONS

Simulations were conducted to compare the rates of type I and II errors committed by the likelihood methods with and without the scaling. Because the relative importance of type I and II errors depends strongly on the distribution of actual sibship sizes in a sample, I simulated samples with three very different family size distributions.

I simulated a large sample containing many full-sib families of highly variable sizes to challenge the methods. The sample contains nine sets of full sibships, with set i ($i=1$–9) having 2^{9-i} sibships, and each sibship having 2^{i-1} offspring. The sample has thus a total number of 2304 offspring distributed in 511 sibships, and contains 64 256 full-sib dyads and 2 588 800 non-sib dyads. This kind of data are highly challenging, because the presence of the many very small sibships (for example, 256 sibships with each having just one offspring and 128 sibships with each having only two offspring) means a high potential of type I errors, and the presence of a few very large sibships (for example, the largest sibship has 256 offspring) means also a high chance of type II errors. The exclusion based sibship assignment methods, for example, are guaranteed to fuse the 256 singletons into 128 or fewer sibships, no matter how many markers are used.

As two extremes, samples containing uniformly large or small sibships are also possible in practice. Simulations were also conducted to compare the accuracy of the likelihood method with and without scaling in these two extreme cases. I simulated a sample containing 600 singletons, and a sample containing four large families, each family consisting of 150 full siblings. For the first sample, sibship splitting or type II errors are impossible, but sibship fusion or type I errors are expected to be severe when marker information is scarce. For the second sample, sibship fusion or type I errors should be rare, but sibship splitting or type II errors are expected to be frequent except when marker information is sufficiently high or the scaling scheme is applied.

Simulations considered different numbers of loci, different numbers of alleles per locus and different allele frequency distributions. The allele frequency distribution at an n-allele locus was assumed to be triangular, equal or highly skewed, with allele i ($=1, 2, \ldots, n$) having a frequency of $i/(n(n+1)/2)$, $1/n$ and $2^{i-1}/(2^n-1)$, respectively. Note that, in the case of a locus having a large number of alleles in a highly skewed frequency distribution, a large proportion of the alleles will have very low frequencies in the population and will not be observed in a sample of individuals except the sample is exceptionally large and sibship sizes are small. For a given full-sib family, parental genotypes were generated independently across loci and parents, assuming Hardy–Weinberg equilibrium and linkage equilibrium. Given the parental genotypes, each offspring multilocus genotype was then generated, following Mendelian segregation law. The offspring genotype data were then subjected to sibship analyses, using the default parameter settings in Colony program.

For each parameter combination, 50 replicate data sets were simulated and analyzed. Accuracy was measured by the frequencies that true full-sib dyads and non-sib dyads are correctly identified, denoted by P_{FS} and P_{NS}, respectively. The total accuracy was measured by the frequency that a dyad of any relationship is correctly identified, denoted by P_{TS}. These frequencies were calculated for each replicate data set, and then averaged across replicates. The rates of type I and II errors and the rate of any type of errors are thus calculated by $1-P_{NS}$, $1-P_{FS}$ and $1-P_{TS}$, respectively.

SIMULATION RESULTS

Simulation results showed that compared with no scaling, scaling of m_i slightly increased type I errors (note the very small scales of y axis for

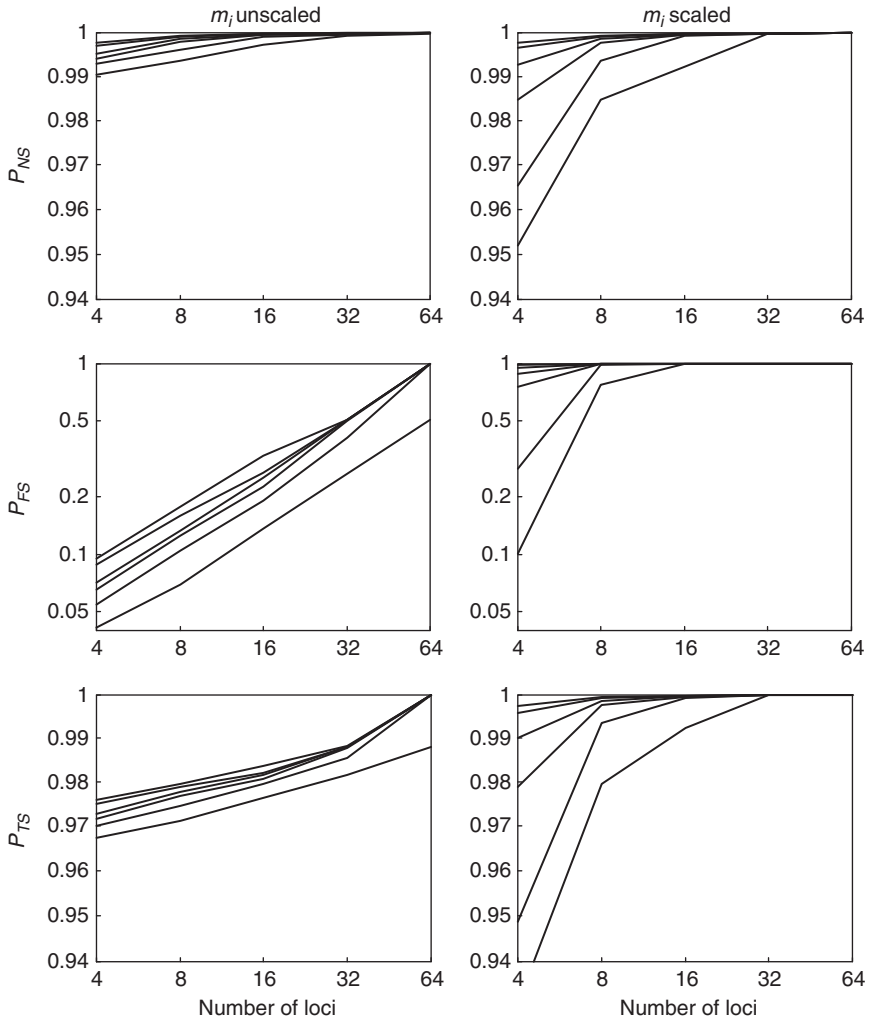

Figure 3 Sibship assignment accuracy as a function of the number of loci for a sample containing sibships of highly variable sizes. The left and right columns show P_{NS} (upper), P_{FS} (middle) and P_{TS} (lower) when m_i is not scaled and is scaled, respectively. In each graph, the lines counting from bottom up refer to 3, 4, 5, 6, 8 and 10 alleles per locus in a triangular frequency distribution, respectively. Note both axes are in log scale.

P_{NS}), especially when a sample contains many small sibships (Figures 3 and 5) and marker information is scarce (that is, few loci and few alleles per locus or a highly skewed allele frequency distribution). The scaling scheme increased type I errors by about 5% in the extreme case of all sampled individuals being in singletons and only four loci, each having six alleles in a highly skewed frequency distribution, being used in the estimation (Figure 5). However, it should be noticed that this level of marker information is unrealistically too low nowadays in practical sibship analyses. With 8 or more loci, little increase in type I errors is incurred by the scaling scheme.

In contrast, the scaling scheme reduced dramatically type II errors (note the large log scales of y axis for P_{FS}) whenever a sample contains a large sibship (Figures 3 and 4). To prevent the largest sibship containing 256 offspring from splitting (that is, $P_{FS}=1$), 8 and 64 loci, each having $n=4$–10 alleles, would be required when the scaling scheme is and is not applied, respectively. While 8 loci are on the lower limit, 64 loci are much above the higher limit of the range of typical empirical data sets in most relationship analyses.

Considering both sibship fusion and split errors, the overall accuracy is improved by the scaling scheme, except when sibships are uniformly small so that sibship splitting is impossible (Figure 5) and when marker information is too scarce (that is, when 4 loci, each having 3–4 alleles in a triangular distribution or six alleles in a highly skewed distribution, are used). Using a modest 8 loci, each having eight alleles in a triangular distribution, the likelihood method with scaling yields an overall accuracy of $P_{TS}=0.999$ for the very challenging data set containing sibships of highly variable sizes (Figure 3). In contrast, the likelihood method without scaling requires 64 loci to attain the same accuracy.

ANALYSIS OF AN EMPIRICAL DATA SET

A salmon data set (Herbinger *et al.*, 1999) was also analyzed comparatively by the likelihood methods with and without the scaling. It comprises 759 fish in 12 full-sib families whose sizes are 8, 10, 31, 51, 54, 59, 64, 69, 75, 91, 107 and 140, respectively. Each individual is genotyped at 4 microsatellite loci, which have 8, 10, 11 and 14 observed alleles in the sample. This data set was analyzed repeatedly in testing previous methods (for example, Smith *et al.*, 2001; Butler *et al.*, 2004; Berger-Wolf *et al.*, 2007; Wang, 2012; Almudevar and Anderson, 2012), and is ideal for demonstrating the power of exclusion based methods because it has large family sizes,

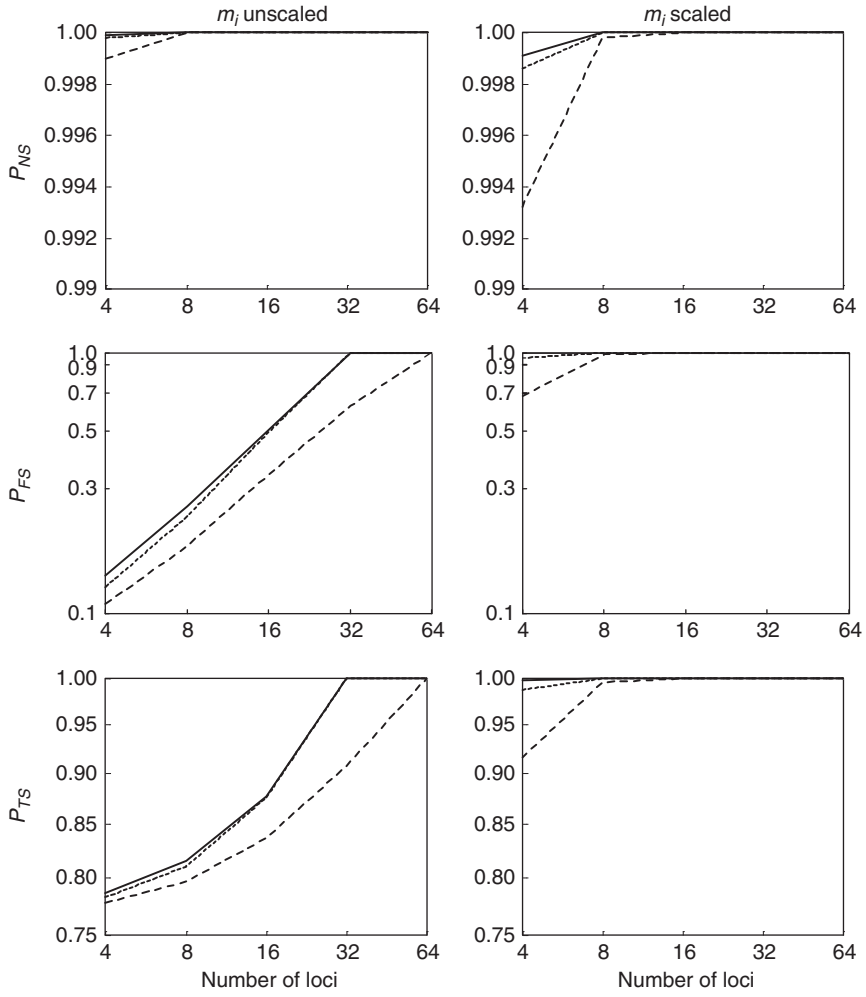

Figure 4 Effect of scaling on the sibship assignment accuracy for a sample containing uniformly large sibships. A sample contains four sibship, each having 150 full siblings. The left and right columns show P_{NS} (upper), P_{FS} (middle) and P_{TS} (lower) as a function of the number of loci (each having six alleles), when m_i is not scaled and is scaled, respectively. In each graph, the continuous, short dashed and long dashed lines refer to equal, triangular and highly skewed allele frequency distribution at a locus, respectively. Note both axes are in log scale.

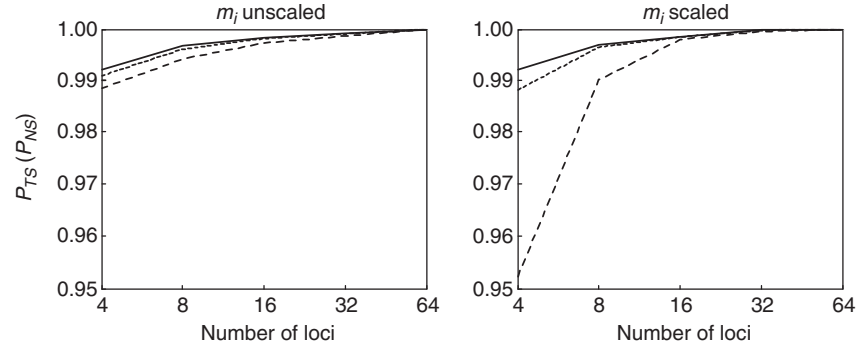

Figure 5 Effect of scaling on the sibship assignment accuracy for samples containing uniformly small sibships. A sample contains 600 sibships, each having one offspring. The left and right columns show P_{NS} ($= P_{TS}$ because all offspring are unrelated) as a function of the number of loci (each having six alleles), when m_i is not scaled and is scaled, respectively. In each graph, the continuous, short dashed and long dashed lines refer to equal, triangular and highly skewed allele frequency distribution at a locus, respectively. Note both axes are in log scale.

contains no missing data and no apparent genotyping errors, and has highly polymorphic codominant markers.

To check the convergence of the likelihood methods with and without scaling, three replicate runs were initiated with different random number seeds. The best log-likelihood values and numbers of inferred full-sib dyads as a function of the number of iterates in the simulated annealing algorithm employed in searching for the maximum likelihood configuration were shown in Figure 6.

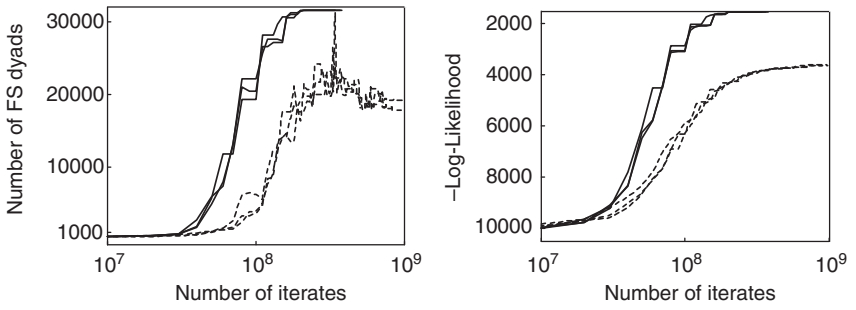

Figure 6 The inferred numbers of sib dyads (left) and log-likelihood values (right) as a function of the number of iterates (*x* axis) in simulated annealing for the salmon data set. The results for three replicate runs were plotted in continuous lines when m_i is scaled and in broken lines when m_i is not scaled. For clarity, only the results from iterate 10^7 onwards are shown.

First, the likelihood maximization algorithm converges reliably for this data set when m_i is scaled. The three replicate runs render the same best configuration, which is identical to the known true configuration with the same maximum log-likelihood value of −1537.18. In contrast, the three replicate runs yield different best configurations with similar likelihood values when m_i is not scaled. The final best configurations of the three replicate runs without scaling have partitioned the 759 fish into 21, 20 and 20 sibships, with a log-likelihood value of −3618.36, −3627.87 and −3625.81, respectively. All of these configurations are much better (in terms of likelihood) than the true configuration, which has a log-likelihood value of −3841.31. One of the replicate runs actually recovered the true configuration at iterates of about 5×10^9, but the true configuration was quickly abandoned and replaced by better configurations with higher likelihood values. It is not surprising that the algorithm does not converge reliably for this data set when m_i is not scaled, because the marker information is scarce and sibship sizes are large, such that there exist numerous configurations with the same or very similar likelihood values (see Figures 1 and 2).

Second, the scaling down of m_i causes an increase in likelihood values. However, absolute likelihood values are usually meaningless, because they are determined by many factors, including the scaling and the amount of data. More data (markers) always lead to a lower likelihood. The more interesting and useful property of a likelihood method is the relative likelihood values, which are used in likelihood ratio tests for model selection and parameter estimation.

Third, all three replicates with scaling render the same maximal likelihood configuration, which is identical to the known pedigree of the sample. In contrast, the three replicates without scaling yield best configurations that have 20–21 sibships. A close examination of the best configurations showed that, while 8 of the 12 true sibships were correctly and consistently reconstructed, 4 true sibships were split differently among the three replicates. The largest sibship with 140 siblings were split into 3 clusters in all 3 replicates.

Fourth, the replicates with scaling run much fewer iterates ($\sim 3.7 \times 10^8$ vs $\sim 9.8 \times 10^8$) and took much less time ($\sim$35 vs $\sim$76 minutes) to finish than the replicates without scaling. This is again due to the numerous configurations with the same or similar likelihood when m_i is not scaled, which cause difficulties for the algorithm to climb to the global maximum likelihood configuration.

Sibship splitting is expected to be frequent when the likelihood method without scaling m_i is applied to a sample containing large sibships but little marker information, like the salmon data set. In such a situation, the simulated annealing algorithm has difficulty of convergence, as reflected by the large variation between replicate runs in likelihood values (see Figure 6). The difficulty comes from the existence of many tied configurations with very similar or identical likelihood values, which form many deep valleys in the likelihood landscape for the algorithm to traverse. The previous implementation of the simulated annealing algorithm in Colony program did not allow for extensive searches, to save computational time, for the maximum likelihood configurations. The algorithm is now improved, making adaptive searches depending on the difficulty of the data. A quick search is implemented for an easy data set, which has a lot of marker information and thus a simple likelihood landscape, while an extensive search is executed for a difficult data set, which has little marker information and thus a potentially complicated likelihood landscape with many peaks of the same or similar height.

To understand the impact of scaling on sibship splitting, I consider in detail the numerical example provided by the largest sibship of the salmon data set that has 140 siblings. This sibship can be split into two sibships, one having a single genotype and the other having the

Table 2 Log likelihoods of split against nonsplit sibships for the largest sibship (140 siblings) in the salmon data set, calculated with and without scaling

Splitting		*Scaling (no split $L_0 = -356.1$)*			*No scaling (no split $L_0 = -703.3$)*		
Locus	*Sibship sizes*	L_1	L_2	L_t	L_1	L_2	L_t
1	78, 62	−210.3	−181.0	−391.3	−349.9	−281.8	−631.7
2	34, 106	−111.4	−288.2	−399.6	−142.0	−538.4	−680.4
2	40, 100	−124.4	−276.9	−401.3	−165.3	−509.3	−674.6
2	38, 102	−119.7	−280.5	−400.2	−157.0	−519.0	−676.0
2	28, 112	−101.7	−299.3	−401.0	−122.6	−567.5	−690.1
3	33, 107	−104.9	−290.2	−395.1	−138.7	−543.2	−681.9
3	34, 106	−105.5	−288.3	−393.8	−142.3	−537.7	−680.0
3	34, 106	−108.2	−288.6	−396.8	−144.4	−538.4	−682.8
3	39, 101	−117.0	−279.2	−396.3	−162.3	−514.1	−676.5
4	47, 93	−133.3	−264.0	−397.3	−186.5	−475.3	−661.8
4	39, 101	−121.5	−278.7	−400.2	−160.5	−514.1	−674.6
4	25, 115	−98.1	−304.9	−403.0	−114.0	−582.0	−696.0
4	29, 111	−104.0	−297.3	−401.2	−126.1	−562.6	−688.8

The actual sibship of 140 offspring was split into two, the first containing individuals sharing a single genotype and the second containing the rest of the 140 individuals, at each of the four loci. The sizes of the first and second split sibships are listed in column 2, the log-likelihood values of the first and second split sibships are listed in columns 3 and 4 with scaling, and columns 6 and 7 without scaling. The total log likelihoods of the two split sibships, L_t, are listed in columns 5 and 8, with and without scaling respectively. The corresponding log likelihoods of the actual nonsplit sibship, L_0, are listed in the column heads.

rest of the genotypes at a locus, in a total of 13 possible ways (Table 2). Of course, many more partitions that have potentially a high likelihood are possible when the sibship is allowed to be split into three or more sibships, and when splitting is on other criteria of genotype combinations. Using the allele frequencies calculated from the entire data set, I calculated the log likelihood of the actual sibship of the 140 offspring (L_0) and the log likelihoods of each split sibship (L_1 and L_2) and their sum (L_t), with and without scaling m_i (Table 2). As can be seen, any sibship splitting always results in a big drop ($L_t < L_0$) and a big increase ($L_t > L_0$) in log likelihood when scaling is applied and not applied, respectively. These likelihood changes due to splitting mean that the likelihood methods will reconstruct this sibship completely without splitting when the scaling scheme is applied, but will split it into two or more sibshsips when the scaling scheme is not applied.

DISCUSSION

Sibship reconstruction from genotype data is a difficult statistical inference problem. It is much more challenging than the now widely applied parentage analyses for a number of reasons. First, the problem of inferring sibship is inherently more difficult than inferring parentage. While a parent–offspring dyad must always share at least one allele identical in state (IIS) at each locus, a full-sib dyad may share 2, 1 or 0 allele IIS at a locus, and these sharing patterns vary greatly among loci due to Mendelian segregation and depending on allele frequencies. This much more variation in allele sharing patterns makes it particularly difficult and error prone to identify siblings from other candidate relationships. This is especially so when closely competing relationships, such as half sibship, are also included in the candidate relationships. Second, a sibship analysis has many more configurations to consider than a parentage analysis. Although an offspring has either 0, 1, or 2 parents included in the candidate parents, it may have 0 to $N-1$ siblings, where N is the sample size. In fact, the number of possible sibship configurations is a Bell number, which increases faster than exponential with N. Enumerating all possible sibship configurations is feasible only when N is very small, say $N<10$. Otherwise, one has to resort to a Monte Carlo searching algorithm, such as simulated annealing to construct and consider just a small fraction of the configurations, which have relatively high likelihood values.

Because of its difficulty and complexity, sibship inference has potentially a high risk of statistical errors of both types. Non-siblings might be mistaken as siblings (type I errors), because they happen to have similar genotypes congruent with a sibship. This problem deteriorates rapidly with a decrease in marker information, as shown in Figure 1, an increase in sample size and an increase in the frequency of small sibships. Exclusion based methods are especially prone to type I errors. Any two unrelated diploid individuals, for example, are always compatible with, and thus not excludable from, a full sibship. The advantages of the likelihood methods are that they control this type of errors more effectively by employing Mendelian segregation law quantitatively rather than just qualitatively (Wang, 2012), and they have the option to use a prior to further reduce errors. A prior can be constructed and integrated into the likelihood framework to reduce type I errors (Almudevar and Anderson, 2012), as is implemented in the program Colony. However, it is difficult to conceive a good prior that favours both small and large sibships to minimize both type I and type II errors in all situations. This is not a problem in situations, in which sibships are known to be uniformly either small (such as in species of low fecundity) or large (such as in some experimental systems involving highly fecund species), but is problematic for a sample with both small and large sibships, as considered in my simulations. Irrespective of the actual sibship size distributions, however, an increase in marker information always improves sibship assignments by reducing both type I and II errors.

True siblings might be mistaken as non-siblings (type II errors), because their genotypes, albeit compatible with a full sibship, are dissimilar enough to justify a split of the sibship based on likelihood values. This problem was analyzed recently by Almudevar and Anderson (2012). They showed that likelihood methods tend to split large sibships, and derived a criterion to predict the occurrences of splitting. My analytical and simulation results confirm their conclusion in general. As shown in this study, scaling down the number of siblings displaying the same single locus genotype used in likelihood calculations can effectively reduce type II errors without causing a substantial increase in type I errors. The overall accuracy considering both types of errors is improved by this scaling, except when a sample contains many very small sibships or marker information is extremely (and unrealistically) scarce.

It is worth noting that the scaling scheme is an *ad hoc* rule obtained by considering the known effects on type I and II errors of factors, such as sibship size, marker polymorphism and mistyping rate, and by extensive experimentations using simulations. In developing this scaling scheme, many more simulations considering many more parameter combinations (for example, mistyping rate) were conducted than those presented in this paper. All these simulations yield the same conclusion presented in this study. However, the scaling scheme is by no means the best scheme applicable optimally to all situations, and there may be room for improvement. It is also worth mentioning that the scaling scheme reduces, not eliminates, the splitting of large sibships. In theory, the improved likelihood method still has the risk of splitting a sibship if it is sufficiently large relative to the amount of marker information. However, I would argue that, with the typical marker information now used in sibship analyses, no sibship should be split by the improved likelihood method except when the actual sibship is exceptionally large. Using a set of 10 markers, each having 10 equifrequent alleles, for example, the improved likelihood should not split a sibship as large as 2000 siblings. In most practical situations, the more worrying problem is sibship fusion (type I errors) rather than sibship splitting (type II errors).

In this study, I focused on the simple case of full sibship analysis. The results and conclusions are, however, applicable to the more complicated cases involving half sibship and parentage assignments (Wang and Santure, 2009) as is verified by simulations not shown in this paper. Full sibship assignments are an integral component of the more complicated analyses, and any improvement in full sibship assignments helps in the inference of other relationships. For example, when a large full sibship is split, parentage assignments for this sibship cannot be completely correct. On the other hand, the presence and assignment of parentage help to reduce sibship splitting. The scaling scheme can therefore improve indirectly the inference of other relationships (half sibship and parentage) by reducing the splitting of large full sibships (data not shown).

My simulations demonstrate that the likelihood methods are powerful, robust and accurate in reconstructing sibships. This is true even for a large sample (2304 individuals) containing many small (singletons) and a few very large (256 siblings) sibships, which were reconstructed with few errors using a realistic number of 8–16 markers (Figure 3) when the scaling scheme is applied. A real data set may be much more complicated than my simulations. It may contains, for example, genotyping errors and half sibships and some

background relationships (for example, cousins), which are assumed absent by the current sibship analysis (likelihood or exclusion based) methods. These are, however, challenges common to all sibship reconstruction methods, and necessitate more marker data than indicated by my simulations to achieve satisfactorily accurate results.

The scaling scheme described in this study has now been implemented in the program Colony, downloadable from http://www.zsl.org/science/research-projects/software/colony,1154,AR.html.

DATA ARCHIVING

There were no data to deposit.

CONFLICT OF INTEREST

The author declares no conflict of interest.

ACKNOWLEDGEMENTS

We thank the editor and three anonymous referees for helpful comments on earlier versions of this manuscript.

Almudevar A, Anderson EC (2012). A new version of PRT software for sibling groups reconstruction with comments regarding several issues in the sibling reconstruction problem. *Mol Ecol Res* **12**: 164–178.

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

Berger-Wolf TY, Sheikh SI, DasGupta B, Ashley MV, Caballero IC, Chaovalitwongse W *et al.* (2007). Reconstructing sibling relationships in wild populations. *Bioinformatics* **23**: I49–I56.

Blouin MS (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *TREE* **18**: 503–511.

Butler K, Field C, Herbinger CM, Smith BR (2004). Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. *Mol Ecol* **13**: 1589–1600.

Emery AM, Wilson IJ, Craig S, Boyle PR, Noble LR (2001). Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid. *Mol Ecol* **10**: 1265–1278.

Epstein MP, Duren WL, Boehnke M (2000). Improved inference of relationship for pairs of individuals. *Am J Hum Genet* **67**: 1219–1231.

Herbinger CM, O'Reilly PT, Doyle RW, Wright JM, O'Flynn F (1999). Early growth performance of Atlantic salmon full-sib families reared in single tanks versus in mixed family tanks. *Aquaculture* **173**: 105–116.

Jones OR, Wang J (2010). Molecular marker-based pedigrees for animal conservation biologists. *Anim Conserv* **13**: 26–34.

Konovalov D, Manning AC, Henshaw MT (2004). KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Mol Ecol Notes* **4**: 779–782.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Mol Ecol* **7**: 639–655.

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

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

Pemberton JM (2008). Wild pedigrees: the way forward. *Proc R Soc London Ser B* **275**: 613–621.

Sieberts SK, Wijsman EM, Thompson EA (2002). Relationship inference from trios of individuals, in the presence of typing error. *Am J Hum Genet* **70**: 170–180.

Smith BR, Herbinger CM, Merry HR (2001). Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics* **158**: 1329–1338.

Thomas SC, Hill WG (2000). Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics* **155**: 1961–1972.

Thomas SC, Hill WG (2002). Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genet Res* **79**: 227–234.

Wang J (2004). Sibship reconstruction from genetic data with typing errors. *Genetics* **166**: 1963–1979.

Wang J (2007). Parentage and sibship exclusions: higher statistical power with more family members. *Heredity* **99**: 205–217.

Wang J (2012). Computationally efficient sibship and parentage assignment from multilocus marker data. *Genetics* **191**: 183–194.

Wang J, Santure AW (2009). Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics* **181**: 1579–1594.