Theor Appl Genet (2006) 112: 472–482
DOI 10.1007/s00122-005-0148-3

ORIGINAL PAPER

Hongmei Zhang · Hal Stern

Assessment of ancestry probabilities in the presence of genotyping errors

Received: 1 July 2005 / Accepted: 21 October 2005 / Published online: 24 November 2005

Abstract This paper extends an approach for estimating the ancestry probability, the probability that an inbred line is an ancestor of a given hybrid, to account for genotyping errors. The effect of such errors on ancestry probability estimates is evaluated through simulation. The simulation study shows that if misclassification is ignored, then ancestry probabilities may be slightly overestimated. The sensitivity of ancestry probability calculations to the assumed genotyping error rate is also assessed.

Introduction

Establishing parentage

Establishing parentage is important in many biological settings. It secures legal relationships in human beings and helps to protect intellectual property in plant varieties.

A number of methods exist for establishing parentage; some of these are reviewed below. The starting point for the work reported here is the approach of Berry et al. (2002) in identifying ancestors of hybrid lines from among a collection of inbred lines using simple sequence repeats (SSR) marker profiles. The Berry et al. approach calculates the posterior probability of an inbred line being an ancestor of a hybrid, given the SSR profiles for the hybrid and a pool of possible ancestors. As it is typical for such methods the calculations do not explicitly account for errors generated during genotyping, though Berry et al. (2002) provide simulation results indicating their method is robust to such errors. We modify the approach to explicitly account for genotyping errors and then study the performance of the two approaches.

Literature review

Biotechnology has created the opportunity to use genetic material to identify ancestors and to determine pairwise relationships. Much work in this area relies on exclusion analysis wherein putative parents are excluded from consideration if the genotypes of the offspring are not consistent with the genotypes of the candidates. Examples of the application of this exclusion approach include Alderson et al. (1999) using SSR markers to determine parentage in brown head cowbirds; Ellstrand (1984) identifying multiple paternities within the fruits of the wild radish based on known multilocus genotypes of the maternal parent; and Chakraborty et al. (1988) deriving an analytical expression for the probability of a male being excluded from paternity assuming the mother of the offspring is known.

Other work in ancestry assessment is developed through the use of probability models for genetic data. Berry et al. (2002) developed such a probability model and then calculated the likelihood for an inbred line being the true ancestor of a given hybrid. That model serves as our starting point so we return to it in the following section. Marshall et al. (1998) identify paternity in a sample of red deer based on likelihood ratio comparisons under a similar model. The likelihood ratio indicates how much more likely the observed genetic data is for an alleged father than for an arbitrary male.

One difficulty is that many of the suggested methods do not account for the possibility of genotyping errors. Such errors could be caused by laboratory mishaps or other errors during the data collection process. Ewen et al. (2000) describe different types of errors in genetic analyses and indicate that such errors can cause serious

Communicated by A. E. Melchinger

H. Zhang (✉)
Department of Mathematics and Statistics,
The University of West Florida, Pensacola, FL 32514, USA
E-mail: hzhang@uwf.edu

H. Stern
Department of Statistics, The University of California,
Irvine, CA 92697, USA

problems for biological inference. One type of error is called relationship misspecification. This refers to the possibility that believed relationships among organisms are incorrect. Such mistakes can be caused by adoption, alternate paternity, or mishandling errors. A second type of error is genotyping error, where the recorded allele score for an offspring is not the same as the allele that would be found in the ancestor. This can occur as a result of mutation or misscoring. Other types of errors include misspecification of allele frequencies and of marker–map distances. Genotyping errors are among the most frequently occurring errors according to Ewen et al. (2000). We focus on genotyping errors because they occur often and because such errors will directly impact our ancestry probability calculation.

There are earlier studies of parentage identification that do allow for genotyping errors. For example, Broman (1998), based on the work of Boehnke and Cox (1997), identifies pairwise relationships among humans by evaluating the likelihood of the observed genotypes for two individuals on all the available loci given a putative relationship between the two individuals; the relationship which maximized the likelihood is the "estimated" relationship. The Broman model allows for constant known error rate along the genome. Marshall et al. (1998) draw paternity inference in red deer based on likelihood ratio comparisons in the presence of genotyping errors. They provide an approach to estimate the genotyping error rate based on the observed number of parent–offspring mismatches. Using an approach similar to that used by Marshall et al. (1998) but with different probability models, San Cristobal and Chevalet (1997) propose another likelihood ratio approach that accommodates genotyping errors. In their work the unknown error rate is estimated using the maximum likelihood approach. The San Cristobal and Chevalet (1997) approach is able to identify parents for various population schemes encountered in animal and plant breeding.

Methods

Review of Berry et al. (2002) approach

Given genetic marker information for a hybrid whose parentage is unclear or unknown and many inbreds that are possible ancestors of the hybrid, Berry et al. (2002) calculate the probability that each inbred is an ancestor of the hybrid under an assumed probability model for the genetic data.

Let i and j denote two possible inbred ancestors from among the available set of ancestor candidates. Berry et al. (2002) calculate the probability that inbreds i and j are the closest ancestors of the hybrid, given the marker information for the hybrid and the inbred candidates. Denote this ancestry probability as $\Pr(\{i, j\}|X)$, where X denotes the collection of genetic information for the hybrid. The notation does not explicitly mention the inbred marker profiles but all calculations are conditional on this information as well. In the application to maize considered below X represents SSR genetic marker profiles. At marker site or locus m the observed SSR marker profile is two alleles from among the n_m alleles found in the population at this locus.

The ancestry probability $\Pr(\{i, j\}|X)$ is calculated using Bayes' rule. Let $\Pr(\{i, j\})$ denote the prior probability assigned to the event that inbreds i and j are the closest ancestors. It is common to assume $\Pr(\{i, j\})$ is the same for all pairs. The likelihood or data distribution for the hybrid's SSR results, given that i and j are the closest ancestors, is $\Pr(X|\{i, j\})$. The set of all possible ancestor pairs from among the K ancestor candidates is $W(K) = \{\{u, v\}, u, v = 1,..., K, u \neq v\}$.

Using Bayes' rule, we have

$$\begin{aligned}\Pr(\{i,j\}|X) &= \frac{\Pr(X|\{i,j\})\Pr(\{i,j\})}{\sum_{(u,v)\in W(K)} \Pr(X|\{u,v\})\Pr(\{u,v\})} \\ &= \frac{\Pr(X|\{i,j\})}{\sum_{(u,v)\in W(K)} \Pr(X|\{u,v\})},\end{aligned} \tag{1}$$

where the last equation follows from the assumption that $\Pr(\{u, v\})$ is constant over ancestor pairs. The probability that a particular inbred candidate i is one of the closest ancestors of the hybrid is just the sum of $\Pr(\{i, v\}|X)$ over all inbreds v with $i \neq v$, i.e.

$$\Pr(i|X) = \sum_{v, v\neq i} \Pr(\{i, v\}|X). \tag{2}$$

One noteworthy point is that the sum of the ancestry probabilities over all pairs is one. This means that probability is assigned to some candidates even if there are no ancestors of the hybrid in the pool of inbred lines. We return to this point in the simulations later.

Calculating $Pr(X|\{u, v\})$

A key element of the Berry et al. approach is the calculation of $\Pr(X|\{u, v\})$, the likelihood of observing data X given a particular ancestor pair. As a first step consider the alleles at a single marker locus. Let X_m denote the set of two unordered allele values of the hybrid at locus m for $m = 1,..., M$, and define $\Pr(X_m|\{u, v\})$ as the probability of observing X_m in the hybrid offspring at locus m, given inbreds $\{u, v\}$ are the closest ancestors.

To calculate this single locus probability, given that $\{u, v\}$ are the closest ancestors of the hybrid, we must determine whether each ancestor passes or does not pass an allele to the hybrid. This implies four exclusive possibilities regarding transmission of genetic information from u and v to the hybrid line. Let Y denote the event that an allele is passed from an ancestor to the hybrid, and N denote the event that an allele is not passed from an ancestor to the hybrid. Then the four possibilities regarding the transmission of genetic information from the two ancestors are YY, YN, NY, and NN, where the

first letter refers to ancestor u and the second refers to ancestor v. Let p denote the probability that one of the ancestor's inbred alleles came through in the hybrid given that the inbred is one of the two closest ancestors. Following Berry et al. (2002), we take p to be the same for each inbred line. The value of p is discussed below. Assuming independence of the inheritance process from the two ancestors the probabilities for the four cases are $P(YY|\{u, v\}) = p^2$, $P(YN|\{u, v\}) = p(1-p)$, $P(NY|\{u, v\}) = (1-p)p$, and $P(NN|\{u, v\}) = (1-p)^2$.

If an inbred's allele on locus m does not come through, then the allele the hybrid has on this locus would be an allele from another inbred ancestral line that is not one of the two putative ancestors under consideration; a result of laboratory error, or the result of mutation. In this case, it is assumed that the allele is chosen randomly from among the available alleles on locus m, with each known allele having probability $1/n_m$ (recall that n_m is the total number of alleles known to exist at the SSR locus m). Note that other assumptions are possible here; the random alleles could be assigned nonuniform probabilities based on the population distribution.

We calculate $\Pr(X_m|\{u, v\})$ by the law of total probability,

$$\begin{aligned}\Pr(X_m|\{u,v\}) =& p^2 \Pr(X_m|YY,\{u,v\}) + p(1-p)\Pr \\ & (Xm|YN,\{u,v\}) \\ & + (1-p)p\Pr(X_m|NY,\{u,v\}) + \\ & (1-p)^2 \Pr(X_m|NN,\{u,v\}). \end{aligned} \quad (3)$$

The four component probabilities are determined according to the laws of genetics and our assumptions. The calculation of the four components depends on a number of factors: the zygosity of the hybrid, i.e. whether the hybrid is homozygous or heterozygous, the zygosity of the inbreds, and whether there are any missing alleles. An example of this type of calculation follows. Berry et al. (2002) provide further discussion regarding this calculation.

Example Suppose that at locus m we observe hybrid alleles $X_m = (3, 4)$, inbred u has alleles (3, 3), and inbred v has alleles (6, 6). First, note that $\Pr(X_m|YY, \{u, v\}) = 0$ because if v's allele were passed on, then the hybrid would have inherited allele 6. The same argument yields $\Pr(X_m|NY, \{u, v\}) = 0$. If u's allele is passed correctly and v's allele is not passed correctly then $\Pr(X_m|YN, \{u, v\}) = 1/nm$. The remaining case covers the scenario where X_m are two randomly selected alleles, $\Pr(X_m|NN, \{u, v\}) = 2/(n_m)^2$, where the factor of two addresses the fact that the hybrid has distinct (and unordered) alleles.

The value of p depends on the proximity of the ancestor (i.e., parent or grandparent or other). If the ancestor is a parent, then $p = 1$ (though in practice we ought to allow for mutations and/or genotyping errors). If the ancestor is a grandparent, then $p = 0.5$ because a descendent is equally likely to inherit from that ancestor or the other grandparent. Berry et al. (2002) use $p = 0.5$ in their calculations because they found this value to provide robust inferences. If the parents are present (which means $p = 1$), then the lower value protects against mutations and genotyping errors. If the parents are not present and the closest ancestors are in fact more remote than grandparents, then $p = 0.5$ still identifies them because they are the closest match to the hybrid's alleles.

After $\Pr(X_m|\{u, v\})$, $m = 1,..., M$, is calculated, Berry et al. (2002) calculate the joint probability for all loci, $\Pr(X|\{u, v\})$, as the product over the M markers

$$\Pr(X|\{u,v\}) = \prod_m \Pr(X_m|\{u,v\}). \quad (4)$$

The use of the product implicitly assumes independence and ignores genetic linkage. In this sense it is equivalent to using a composite likelihood (Lindsay 1988) rather than an exact multilocus likelihood. The use of exact likelihoods would require a multipoint calculation like that described in Ott (1991) for human genetic data. The data for such a calculation are not available for the maize data in our application. Composite likelihoods have been used in some genetic contexts (see e.g., Devlin et al. 1996; Rannala and Slatkin 2000; Garner and Slatkin 2002) to reduce computational complexity; they are useful approximations but can involve a loss of precision. However, Berry et al. (2002) note that the markers in their application are spread over the ten maize chromosomes and that they obtain similar results for a range of numbers of marker loci. These results suggest that the composite likelihood approximation is adequate for their application. The use of the composite likelihood approximation is also supported by our simulation results, discussed in a later section. In the work reported here we follow Berry et al. (2002) and use the composite likelihood defined by Eq. 4 so that we can focus exclusively on the effect of modifying their approach to accommodate genotyping errors.

Once the joint probabilities for all markers are calculated using the product rule (Eq. 4), then the probability that inbreds i and j are the closest ancestors; $\Pr(\{i, j\}|X)$, is computed using Bayes' rule (Eq. 1) and the probability that a particular inbred candidate is one of the closest ancestors of the hybrid is calculated using Eq. 2.

Robustness of the Berry et al. approach

Berry et al. (2002) did not explicitly allow for genotyping errors in their probability calculations. To determine the importance of genotyping errors and to examine the effect of missing data, Berry et al. (2002) artificially modified their original data (a set of maize hybrids and a collection of inbred lines) by eliminating specific proportions of alleles that had been scored (this gives missing data) and/or by misclassifying (misscoring) other alleles. For example, in order to evaluate the robustness of the approach with respect to misscored alleles, they simulated misscored data at 2% of the loci, 5% of the loci, 10% of the loci, and 25% of the loci for all hybrids and all inbreds. They

examine the effect of these error levels by comparing the number of correctly identified ancestors with and without these extra genotyping errors, and conclude that the method works well as long as the fraction of missing or misscored data is below 15%.

Incorporating genotyping errors

In this section we describe a method that incorporates misscoring or genotyping errors into the ancestry probability calculation (Eq. 1). Our approach calculates $\Pr(X_m|\{u, v\})$ while accommodating genotyping errors only on hybrids. This is certainly optimistic because one expects that errors occur for the alleles of both inbred ancestor candidates and hybrid offsprings. There are, however, technical reasons to expect much lower error rates in inbred lines. One source of error in hybrids (which are generally heterozygous) is allelic competition in DNA annealing and amplification; allelic competition may result in one allele masking the presence of another. This source of error does not typically occur for inbred lines (which are generally homozygous). The methods presented here are developed assuming a known error rate π. This reflects the assumption that there will often be information outside of a particular dataset about the error rate associated with a genotyping technology. Of course, this is not always the case; Zhang and Stern (2003) describe some approaches to estimating π.

Calculating $Pr(X_m|\{u, v\})$ with genotyping errors for hybrids

As indicated in the previous section the key calculation is determining the probability of observing alleles X_m on marker m, given u and v are the closest ancestors. In order to contrast the calculation when genotyping errors are accommodated and the earlier calculation, we use the notation $\Pr(X_m\ |\{u, v\}, \pi)$ for the probability of observing X_m, given that $\{u, v\}$ are the closest ancestors and the error rate is π. As needed subsequently, we also use the notations $P(\{i, j\}|X, \pi)$ and $\Pr(X|\{u, v\}, \pi)$ in place of $P(\{i, j\}|X)$ and $\Pr(X|\{u, v\})$.

Let (a_{tm}, b_{tm}) denote the true allele pair at locus m and $X_m = (a_{om}, b_{om})$ the observed allele pair at locus m for $m = 1,\ldots, M$. Define A_m as the set containing all possible true allele pairs on locus m of a hybrid. When allowing for possible hybrid genotyping errors, we have

$$\begin{aligned}\Pr(X|\{u,v\},\pi) &= \sum_{Am} \Pr((a_{om},b_{om}),(a_{tm},b_{tm})|\{u,v\},\pi)\\ &= \sum_{Am} [\Pr((a_{om},b_{om})|(a_{tm},b_{tm}),\{u,v\},\pi)\\ &\quad \Pr((a_{om},b_{om})|\{u,v\},\pi)]\\ &= \sum_{Am} [\Pr((a_{om},b_{om})|(a_{tm},b_{tm})\pi)\\ &\quad \Pr((a_{tm},b_{tm})|\{u,v\})], \end{aligned} \tag{5}$$

where the final line reflects simplifications based on the following observations: (1) given the true alleles the probability distribution of the observed alleles depends only on the error rates (and not on which pair of ancestors we are talking about); and (2) inheritance of the true alleles does not depend on errors in genotyping technologies. The final term $\Pr((a_{tm}, b_{tm})|\{u, v\})$ is the probability that the hybrid has true alleles (a_{tm}, b_{tm}) on locus m, given $\{u, v\}$ are the closest ancestors. This quantity is precisely the one used by Berry et al. (2002) and described in the previous section. The first term on the final line $\Pr((a_{om}, b_{om})|(a_{tm}, b_{tm}), \pi)$ describes the error process; it measures the probability of observing alleles (a_{om}, b_{om}), given that the true alleles on locus m are (a_{tm}, b_{tm}).

Calculating $Pr((a_{om}, b_{om})|(a_{tm}, b_{tm}), \pi)$

The new aspect of the calculation is the term describing the error process. This term is now described in some detail. We assume a constant genotyping error rate π on every locus. As before nm denotes the number of available alleles on locus m. If a true allele is not correctly identified, then we assume the observed allele is chosen randomly according to a uniform distribution over the remaining available alleles. The value of $\Pr((a_{om}, b_{om})|(a_{tm}, b_{tm}), \pi)$ depends on whether the observed alleles match the true alleles and also on the zygosity of each pair. The details for the different situations are listed below with one example provided for each case.

1. Observed alleles are the same as the true alleles, i.e. $(a_{om}, b_{om}) = (a_{tm}, b_{tm})$. Two cases are considered.

- The homozygous case, i.e. $a_{om} = b_{om}$, $a_{tm} = b_{tm}$. Because the observed alleles and the true alleles are both homozygous on locus m, the labeling of a and b does not matter. The probability of the two hybrid alleles on locus m being correctly identified would be the probability of no errors, $\Pr((a_{om}, b_{om})|(a_{tm}, b_{tm}), \pi) = (1-\pi)^2$.
- The heterozygous case, i.e. $a_{om} \neq b_{om}$, $a_{tm} \neq b_{tm}$. For a specific example, consider the case $(a_{om}, b_{om}) = (a_{tm}, b_{tm}) = (8, 9)$. When the observed alleles and the true alleles are heterozygous, the labeling of the a and b alleles matters. The observed alleles will be equal to the true alleles if there are no errors OR if both alleles are identified with error and the match occurs due to chance. The probability of this event is the sum of the probabilities (1) that both hybrid alleles are correctly identified, e.g. $(8 \rightarrow 8, 9 \rightarrow 9)$, and (2) that each allele is "incorrectly" identified as the other, e.g. $(8 \rightarrow 9,\ 9 \rightarrow 8)$. This yields $\Pr((a_{om}, b_{om})|(a_{tm}, b_{tm}), \pi) = (1-\pi)^2 + \pi^2/(n_m-1)^2$.

2. One of the two observed alleles is the same as one of the two true alleles. Several different situations arise

depending on whether the observed alleles or the true alleles on locus m are homozygous or heterozygous

- The observed alleles are homozygous, but the true alleles are not. In this case, $a_{om} = b_{om}$, $a_{tm} \neq b_{tm}$. Consider the case $(a_{om}, b_{om}) = (8, 8)$, $(a_{tm}, b_{tm}) = (8, 9)$. In this case one of the true alleles was measured correctly (8 measured as 8) and the other is in error (9 converted to 8), thus $\Pr((a_{om}, b_{om})|(a_{tm}, b_{tm}), \pi) = (1-\pi)\pi/(n_m-1)$.
- The true alleles are homozygous, but the observed alleles are not. In this case, $a_{om} \neq b_{om}$, $a_{tm} = b_{tm}$. Consider the case $(a_{om}, b_{om}) = (9, 8)$, $(a_{tm}, b_{tm}) = (8, 8)$. This is similar to the above but there are now two ways to choose the true allele which is measured correctly, $\Pr((a_{om}, b_{om})|(a_{tm}, b_{tm}), \pi) = 2(1-\pi)\pi/(n_m-1)$.
- Both of the observed alleles and the true alleles are heterozygous. In this case, $a_{om} \neq b_{om}$, $a_{tm} \neq b_{tm}$. Consider the case $(a_{om}, b_{om}) = (8, 9)$, $(a_{tm}, b_{tm}) = (8, 6)$. There are two ways to generate the observed data pattern: either the 8 is recorded accurately (with no error) and the 6 is not, OR both alleles are measured/recorded in error. The probability sums the likelihood of these two possibilities, $\Pr((a_{om}, b_{om})|(a_{tm}, b_{tm}), \pi) = (1-\pi)\pi/(n_m-1) + \pi^2/(n_m-1)^2$.

3. Neither of the two observed alleles are the same as the true alleles. In this case, $a_{om} \neq a_{tm}$, $b_{om} \neq b_{tm}$, $a_{om} \neq b_{tm}$, and $b_{om} \neq a_{tm}$. Two situations need to be considered depending on whether the observed hybrid alleles are homozygous or heterozygous. It does not matter whether the true alleles are homozygous or heterozygous because both are evidently measured with error.

- Observed alleles are homozygous, i.e. $a_{om} = b_{om}$. Consider the case $(a_{om}, b_{om}) = (8, 8)$, $(a_{tm}, b_{tm}) = (9, 9)$ or the case $(a_{om}, b_{om}) = (8, 8)$, $(a_{tm}, b_{tm}) = (6, 9)$. In this case $\Pr((a_{om}, b_{om})|(a_{tm}, b_{tm}), \pi) = \pi^2/(n_m-1)^2$ which is the probability that both of the alleles are incorrectly identified.
- Observed alleles are heterozygous, i.e. $a_{om} \neq b_{om}$. The only difference here is that because the observed alleles are heterozygous, we must account for the labeling of the a and b alleles. There are thus two ways in which a given set of true alleles can be recorded in error to yield the observed heterozygous pair, $\Pr((a_{om}, b_{om})|(a_{tm}, b_{tm}), \pi) = 2\pi^2/(n_m-1)^2$.

After $\Pr((a_{om}, b_{om})|(a_{tm}, b_{tm}), \pi)$ is calculated for each marker, and $\Pr((a_{tm}, b_{tm})|\{u, v\})$ is calculated using the Berry et al. method, then we can use Eq. 5 to obtain $\Pr(X_m|\{u, v\}, \pi)$. Finally, we again obtain the likelihood for all the markers, $\Pr(X|\{u, v\}, \pi)$, by using the composite likelihood approximation (i.e., assuming independence),

$$\Pr(X|\{u,v\},\pi) = \prod_m \Pr(X_m|\{u,v\},\pi).$$

The ancestry probabilities follow from Bayes rule,

$$\Pr(\{i,j\}|X,\pi) = \frac{\Pr(X|\{i,j\},\pi)}{\sum_{(u,v)\in W(K)} \Pr(X|(u,v\},\pi)}.$$

The effects of genotyping errors

This section demonstrates the method developed for accommodating genotyping errors using simulated data and compares the resulting inferences to those obtained from the Berry et al. approach. Initially results are provided for the case where data are simulated under the assumption of independent markers. That assumption is consistent with the composite likelihood approach used in our algorithm and by Berry et al. (2002). Toward the end of the section additional simulations are used to assess the impact of linkage on ancestry inferences. Application to real data, from the maize hybrids considered by Berry et al. (2002), is considered in the next section.

Data simulation

The simulated data are generated assuming

1. there are 110 possible markers for each individual
2. nine equally likely alleles identified as {1, 2 ,..., 9} are possible on each marker locus
3. the probability that an ancestor's allele is passed to the hybrid offspring is $p = 0.5$ (i.e., the closest ancestor is a grandparent)
4. inheritance of each marker is independent of the others (i.e., no linkage)
5. the inbred ancestor candidates are homozygous for each marker
6. each ancestor independently passes its alleles to the offspring
7. genotyping errors occur independently with probability π for each of the two hybrid alleles at every locus.

A single data set is generated based on the above assumptions as follows. First, marker profiles for 100 inbred ancestor candidates are generated. Second, two of the inbreds are picked as "true" ancestors. Third, these two ancestors are used to generate a hybrid offspring, introducing genotype errors with error rate π for each allele at every locus. In practical applications the different ancestral and hybrid lines do not have data for each possible marker. To simulate the variability associated with such missing data the hybrid offspring and the two true ancestors have SSR profile information on 100 randomly chosen loci (a possibly different set of 100 loci for each). The number of loci with available marker information is randomly varied for the remaining ancestor

candidates as well (at least 90 loci are always present). An example of the simulated data is provided in Table 1. This format is essentially the same as the format of the SSR marker profiles used by Berry et al. (2002). Note that the hybrid is missing data on marker 3, the first inbred (Inbred1) is missing data on marker 2, and the last inbred (Inbred100) is missing data on marker 1.

Simulation results

We consider two distinct cases in our simulation study. The data are always generated as above; the two cases vary in the makeup of the data that are analyzed. The first scenario covers the case when the true ancestors are among the set of ancestors considered in the ancestry probability calculation (Eq. 1), and the second scenario covers the case where the true ancestors are not present. The latter is included to study the nature of the ancestry probability inferences for the realistic possibility that no ancestor is present.

For each case simulated data is generated for a variety of error rates. Specifically we use values of π ranging from 0.001 to 0.5 (actual values $\pi = 0.001$, 0.01, 0.1, 0.2, 0.3, 0.5). The high values are much higher than that expected in practice but they are useful for illustrating methodology. Posterior ancestry probabilities are obtained for each simulated data set using the "known" value of π used to generate the data set. Sensitivity to the assumed error rate is explored by recalculating ancestry probabilities for a range of error rates that differ from the error rate used to generate the data.

Case 1: true ancestors are present

We first consider the case where SSR marker profile information for the true ancestors is included among the set of ancestor candidates. Fifty different hybrids are simulated from the same two ancestors. Each hybrid is then analyzed separately yielding posterior probabilities for each of the possible ancestors. Analyses are carried out using the Berry et al. approach and the approach proposed here to accommodate genotyping errors. As a summary the posterior probability assigned to each of the two true ancestors is recorded. Table 2 presents summaries of the 100 posterior ancestor probabilities for a range of π values (assuming the same value of π is used to generate and analyze the data). Specifically, the table gives the minimum, first quartile, median, mean, third quartile, and maximum of the empirical distribution of the posterior ancestry probabilities.

Table 1 The simulated data format

Organism code	Marker	Allele1	Allele2
Hybrid1	1	2	3
Hybrid1	2	2	5
Hybrid1	4	3	6
...	...	...	...
Hybrid1	104	1	9
Inbred1	1	7	7
Inbred1	3	4	4
...	...	...	...
Inbred100	2	1	1
Inbred100	3	3	3
...	...	...	...
Inbred100	110	3	3

For $\pi \leq 0.3$, both approaches always correctly identified the true ancestors with high ancestry probability values. Moreover, for $\pi \leq 0.2$ the ancestry probabilities are all 1. For the case with $\pi = 0.3$ the values are all near 1 with the values from the Berry approach slightly larger. These results support the robustness results reported in Berry et al. (2002), the ancestor probability calculation provides accurate inferences about ancestry even with nontrivial genotyping error rate. As the error rate goes even higher to $\pi = 0.5$ (so half of the marker alleles are erroneous) both methods misidentify several offsprings' ancestors. The Berry et al. approach misidentifies 8 out of 100 ancestors and the approach accommodating errors misidentifies 6 out of 100 ancestors. The difference does not seem especially important since error probabilities that high do not seem very realistic.

Figure 1 explores the sensitivity of the inference to the assumed data generating error rate. The median of the 50 probabilities for one of the true ancestors is plotted against the value of π used to calculate the probabilities. There are separate curves for each of the true data-generating error rates. The figure indicates that one need not know the data-generating genotyping error rate precisely when the true ancestors are present among the set of ancestor candidates. The calculated ancestry probability values will be accurate unless an unrealistically large genotyping error rate is assumed (above $\pi = 0.6$).

The results in Table 2 and Fig. 1 are consistent across a range of different simulated data sets (i.e., we repeated the above calculations with different ancestor pairs). When the true ancestors are present in the data, posterior ancestry probabilities for the true ancestors are always quite large, even for error rates much higher than expected in practice and even when the error rate assumed for data analysis is different than the true data-generating error rate. Thus the Berry et al. approach performs quite well in this case. The calculations to accommodate error are arguably more realistic, but hardly seem to matter for the case when the ancestors are present.

Case 2: True ancestors are not present

One limitation of the Bayesian approach is that it implicitly assumes the ancestors are included among the candidates; the posterior ancestor probabilities assigned to all pairs of inbred lines will sum to one (see the formula 1). As one possible use of ancestry calculations concerns legal determinations of intellectual property it is important to understand the performance of the

Table 2 Distribution of estimated ancestry probabilities with true ancestors included in the data

π	Method	Minimum	First quartile	Median	Mean	Third quartile	Maximum
0.001	With Misclas.	1	1	1	1	1	1
	Berry et al.	1	1	1	1	1	1
0.01	With Misclas.	1	1	1	1	1	1
	Berry et al.	1	1	1	1	1	1
0.1	With Misclas.	1	1	1	1	1	1
	Berry et al.	1	1	1	1	1	1
0.2	With Misclas.	1	1	1	1	1	1
	Berry et al.	1	1	1	1	1	1
0.3	With Misclas.	0.9847	0.9999	1	0.9998	1	1
	Berry et al.	0.9934	1	1	0.9999	1	1
0.5	With Misclas.	0.001932	0.9911	0.9995	0.8902	0.99997	1
	Berry et al.	0.00001291	0.9999	1	0.8681	1	1

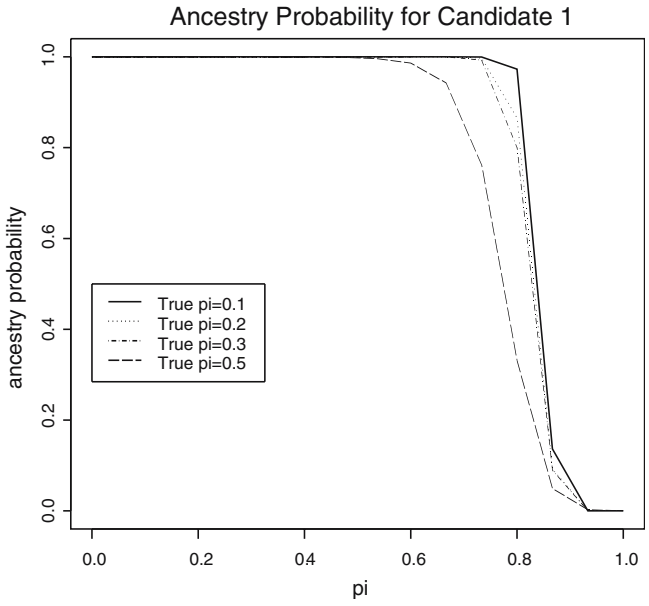

Fig. 1 The relationship between the median ancestry posterior probabilities for one true ancestor and the assumed genotyping error rate (π). Each *curve* corresponds to a different true data-generating genotyping error rate

methods when one or more ancestral line is absent from the candidate set. To study this issue we use the same simulated data as above except that the true ancestors' marker information is removed from the data set prior to analysis.

For each hybrid that is analyzed the two highest probabilities assigned to candidate ancestors are recorded. Table 3 provides a numerical summary of the posterior probabilities of the 100 identified ancestors for a range of data-generating π values (with each analysis carried out assuming the error rate is correctly specified). When the true ancestors are not included among the candidates, the identified "ancestors" are those whose allele scores happen to match the scores of the hybrid offspring by chance, and hence we expect the probability values of being an ancestor should be lower than those reported in Table 2. When the error rate is exceptionally low, like $\pi = 0.001$, the difference between the Berry et al. method and the modified approach is hard to see. However, when the misclassification rate gets higher (e.g., greater than 0.1), there is evidence that the Berry et al. approach assigns greater confidence to the putative (but incorrect) ancestor than the approach that accommodates genotyping errors. The final column in Table 3 gives the proportion of the ancestry probabilities greater than 0.9. Such high ancestry probabilities would likely be interpreted as strong evidence that a correct ancestor had been found; so differences between the two approaches in this column are of practical interest.

The results in Table 3 are for 50 hybrids from the same two simulated ancestors. In order to see the consistency among different simulated data sets, Table 4 provides results concerning the proportion of identified ancestors with probability values higher than 0.9 for a number of different data sets (i.e., simulated from different ancestors). The mean proportion higher than 0.9 and the standard deviation (over 15 simulated data sets with different ancestors) are reported for each π. The general pattern is the same as seen in Table 3.

The results above assume that the same error rate is used to create and analyze the data, i.e., that one knows the correct error rate. We next explore the relationship between the calculated ancestry probabilities and the assumed genotyping error rate π. There are 50 hybrids in the simulated data set and hence 100 posterior ancestry probabilities are obtained from each analysis. Figure 2 gives the median of the 100 ancestry probabilities as a function of π. There are separate graphs for each value of the data-generating error rate. The dotted lines in each graph are empirical 90% posterior intervals for each curve. When we assume that there are no errors ($\pi = 0$ on the horizontal axis) then the ancestry probabilities are identical to those obtained by the Berry et al. approach. The median ancestry probabilities are near or higher than 0.8 when we assume no errors regardless of the magnitude of the true genotyping error rate. Thus as seen earlier two "ancestors" are always identified with relatively high probabilities when we assume no errors—these are

Table 3 Distribution of estimated ancestry probabilities without true ancestors included in the data

π	Method	Minimum	First quartile	Median	Mean	Third quartile	Maximum	Proportion≥0.9
0.001	With Misclas.	0.380	0.788	0.915	0.866	0.994	0.999	0.53
	Berry et al.	0.381	0.789	0.915	0.866	0.994	1	0.53
0.01	With Misclas.	0.401	0.809	0.903	0.860	0.980	0.999	0.51
	Berry et al.	0.402	0.812	0.910	0.862	0.982	1	0.52
0.1	With Misclas.	0.462	0.708	0.924	0.846	0.980	1	0.60
	Berry et al.	0.507	0.738	0.946	0.866	0.988	1	0.65
0.2	With Misclas.	0.372	0.751	0.885	0.835	0.980	1	0.43
	Berry et al.	0.339	0.803	0.950	0.876	0.990	1	0.68
0.3	With Misclas.	0.226	0.460	0.655	0.662	0.903	0.998	0.26
	Berry et al.	0.292	0.609	0.807	0.769	0.982	1	0.39
0.5	With Misclas.	0.178	0.372	0.469	0.552	0.751	1	0.11
	Berry et al.	0.313	0.632	0.826	0.775	0.976	1	0.40

Table 4 Proportion of ancestry probabilities ≥0.90 (with standard deviation in parentheses)

Method	$\pi = 0.001$	$\pi = 0.01$	$\pi = 0.1$	$\pi = 0.3$	$\pi = 0.5$
With Misclas.	0.42 (0.062)	0.43 (0.043)	0.35 (0.043)	0.26 (0.046)	0.09 (0.028)
Berry et al.	0.42 (0.062)	0.44 (0.039)	0.40 (0.031)	0.41 (0.066)	0.41 (0.046)

clearly overestimates of our confidence when the true ancestors are not part of the data set! The posterior ancestry probabilities decrease quickly as the assumed error rate increases until extreme (and unrealistic) error rates are reached. Comparison of Figs. 1 and 2 suggests that incorporating an appropriate genotyping error into the probability calculation may help reveal whether the true ancestors are in the candidate pool. The estimated probabilities assigned to the "most likely ancestors" will be much higher when the true ancestors are in the candidate pool than when they are not. This is especially true for high genotyping error rates.

The effects of linkage

As remarked earlier it is important to assess the sensitivity of ancestry inferences to violations of the independence assumption that is implied by the use of the composite likelihood in our algorithm (and that of Berry et al. 2002). After all it is common to have markers that are linked and for some pairs of markers the linkages can be very close. Here we incorporate linkage in our simulated data and assess the performance of our composite-likelihood-based algorithm for calculating ancestry probabilities.

Data simulation

Our basic approach to simulating marker data remains the same. To address linkage we replace steps 3 and 4 of the original data-generating algorithm (which assumed independent markers inherited with probability 0.5). For the most part we work with pairs of linked markers in which case steps 3 and 4 are replaced by:

3′. If a pair of markers are linked, then the ancestral alleles of the first marker in the pair are passed (not passed) to the hybrid with probability 0.5 as usual. The alleles of the second marker in the pair are also passed (not passed) to the hybrid with probability $p_{\text{link}} > 0.5$, where p_{link} controls the degree of linkage.

4′. If a pair of markers are not linked, then an ancestral allele is passed (not passed) to the hybrid offspring with $p = 0.5$ for each marker and the inheritance is independent of the other member of the pair (note that this is the same as the original algorithm).We vary the number of linked pairs, k (recall that there are 110 simulated markers, hence $k \leq 55$), the degree of linkage, p_{link} ($0.5 < p_{\text{link}} \leq 1$), and the genotyping error rate π.

In addition, as an extreme case, we allow for the possibility of complete linkage of sets of more than two markers. For the complete linkage case if alleles on one marker in the set are inherited (not inherited), then the alleles on all other markers in the set will also be inherited (not inherited).

Simulation results

We initially consider pairs of linked markers. A range of simulated data sets were generated covering all possible combinations of three different values for the number of linked pairs ($k = 11, 22, 55$), three different values of the degree of linkage ($p_{\text{link}} = 0.6, 0.8, 1.0$), and four values of the genotyping error rate ($\pi = 0.1, 0.2, 0.3, 0.5$). For each combination of k, p_{link}, and π, we simulate ten hybrids from different ancestor pairs. Then for each hybrid ancestor probabilities are estimated for all possible ancestors, and the presumed

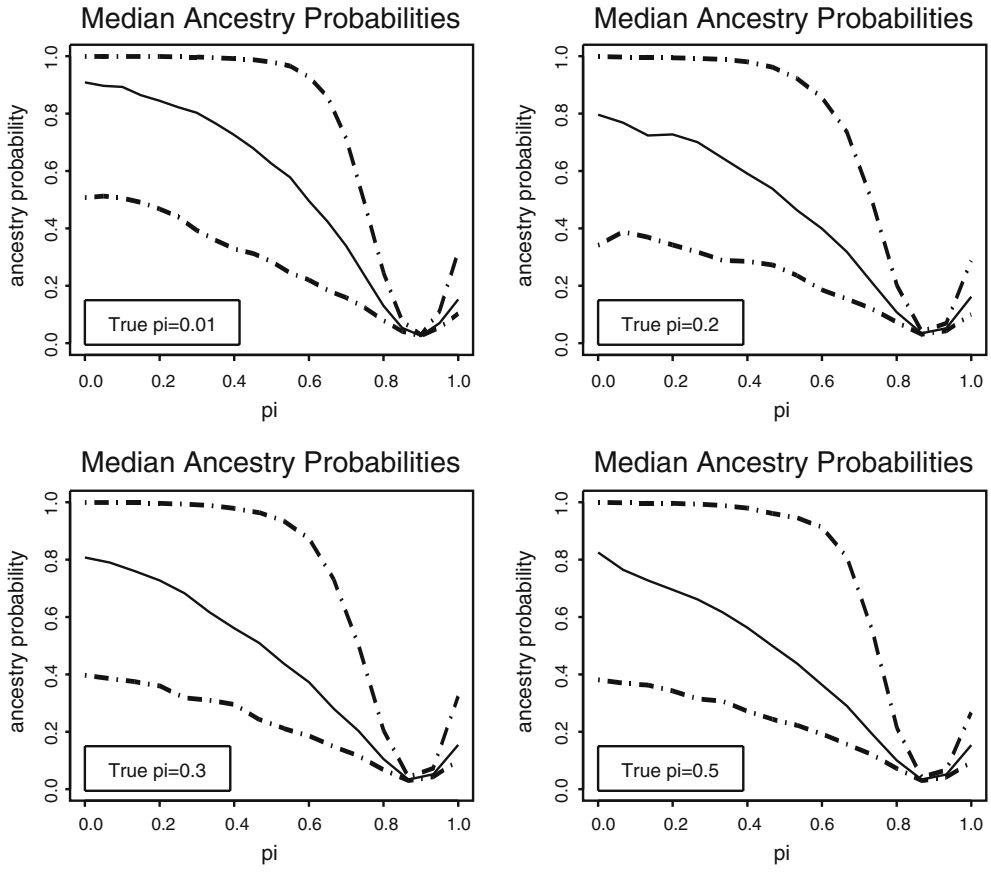

Fig. 2 The relationship between the median ancestry posterior probabilities and the assumed genotyping error rate when true ancestors are not present. Each figure corresponds to a different true data-generating genotyping error rate. The *solid lines* are median ancestry posterior probabilities and the *dashed lines* are empirical 90% intervals

ancestors are taken to be the two with the highest estimated probabilities. It turns out that all ancestors are correctly identified with great confidence by both the Berry et al. approach and our modification except when data are generated using the extremely unrealistic genotyping error rate $\pi = 0.5$. The results suggest a high degree of robustness for ancestry inferences in the presence of pairs of linked markers.

For the complete linkage case we considered sets of 5, 10, or 20 perfectly linked markers. The full set of 110 markers are split into groups of the specified size so that virtually every member is part of a "linkage group". As the size of the set of linked markers increases there is correspondingly less independent information on which to base ancestry inference. Once again we vary the genotyping error rate ($\pi = 0.1$, 0.2, 0.3, 0.5) and simulate ten hybrids from different ancestor pairs for each scenario. Table 5 summarizes the complete linkage results. The proportion of misidentified ancestors (out of 20 true ancestors) is given for each combination of linkage group size and genotyping error rate. Even if the full market profile is comprised of sets of 5 or 10 completely linked markers it is possible to correctly identify all ancestors for moderate and small genotyping error rates. We only make errors if $\pi \geq 0.3$. It is only in the most extreme case, where our 100+ marker profile includes five sets of 20 completely linked markers, that our ability to identify ancestors is compromised.

In all, these results support the use of the composite likelihood approximation for maize ancestry assessment as long as the genotyping error rate is not extreme and as long as there are not a large number of perfectly linked marker sets. The markers in the maize application that motivated Berry et al. (2002) are spread over ten chromosomes so there should not be many closely linked markers.

Real data application

In this section, we revisit the data considered by Berry et al. (2002) using the new approach to accommodating errors. Of course in this case we don't know the true error rate!

The SSR data description

The data contains SSR marker profiles for three hybrids and 118 inbreds on 195 loci. As often happens real data present a number of complications not present in the

Table 5 Proportion of misidentified ancestors with full linkage

Number of markers in full-linkage group	Genotyping error rate (π)			
	0.1	0.2	0.3	0.5
5	0.00	0.00	0.00	0.20
10	0.00	0.00	0.05	0.15
20	0.10	0.10	0.15	0.30

Note: There are 110 markers in total, so for the first case, there are 22 full-linkage groups each with five fully linked markers; for the second case, 11 full-linkage groups each with five fully linked markers; and for the third case, five full-linkage groups each with 20 fully linked markers plus one full-linkage group with ten fully linked markers

simulations. Some inbred lines' marker profiles present more than one type of allele at some loci. The frequency of such events depends on how long the inbreeding process has been carried out for the line in question and whether there has been unintended pollination from other genotypes. This by itself does not present a problem for the algorithms discussed here. However, this also means that some single-cross hybrids show more than two alleles per locus, because their inbred parents are not homozygous. Table 6 shows the format of the SSR data supplied by Stephen Smith, Deanne Wright, and Chongqing Xie (Pioneer Hi-Bred International, Inc.). Line 1 in the table is a typical marker for a hybrid with two distinct alleles assumed to come from the two ancestral inbred lines. The second line shows a marker for which there are three alleles present in the hybrid. This is very likely due to a heterozygous parent (or two). The two inbreds in Table 6 are generally homozygous, but marker 2 for "inbred1" is heterozygous. To take this situation into account, the probability of ancestry might be evaluated as the mean of a number of ancestry probability values, each calculated with a random choice of two of the hybrid's alleles for each locus. Because this is rare in the maize data, we make a single random choice of two alleles per locus and ignore the variation due to the existence of other possibilities.

Results

Table 7 lists the identified ancestors together with the next closest possible relative for each of the three hybrids as calculated using the Berry et al. approach ($\pi = 0$). The identification codes of the inbreds and the inbreds' ancestry probabilities (in parentheses) are shown in the table. Five of the six ancestors are identified as having very high probabilities of being the closest ancestor, the exception being inbred 90s identification as a likely ancestor for the third hybrid.

Table 7 Probability assessment of ancestry from the Berry et al. approach

Hybrid	Ancestor 1	Ancestor 2	Next closest relative
1	100 (0.99979)	16 (0.99934)	15 (0.64373×10^{-3})
2	27 (1.0000)	85 (0.94157)	92 (0.058397)
3	37 (1.0000)	90 (0.75796)	71 (0.23032)

Figure 3 illustrated the effect of allowing for genotyping errors. The figure shows the relationship between the posterior ancestry probabilities for each of the six ancestors identified in Table 7 and the assumed genotyping error rate, π. Each curve in the figure corresponds to a different identified ancestor. Except for inbred 90, all the ancestry probabilities are at least 0.8 even when the value of π is as large as 0.4. From our previous discussion, this suggests that we are likely looking at true ancestors. The most unusual case is inbred 90, which is identified as one of the two ancestors for hybrid 3. The ancestry probability (0.76) for inbred 90 at $\pi = 0$ is not high. This implies that the data may include other candidates that are almost as closely related to the hybrid as inbred 90. The ancestry probability varies as the assumed error rate changes because we may begin to favor another candidate as the closest ancestor.

Inferring the error rate

The analyses presented here calculated ancestry probabilities for a range of assumed error rates. The error model introduced in this paper can be used to infer the error rate. For example, Zhang and Stern (2003) obtain the maximum likelihood estimate for π in the maize data as approximately 0.005, though there is considerable uncertainty given the relatively small amount of

Table 6 Format of the SSR data

Code	Marker	Allele1	Allele2	Allele3	Allele4
Hybrid1	1	8	9		
Hybrid1	2	2	3	1	
...	...	...	...	...	...
Inbred1	1	9	9		
Inbred1	2	4	3		
...	...	...	...	...	...
Inbred11	1	5	5		
Inbred11	2	3	3		
...	...	...	...	...	...

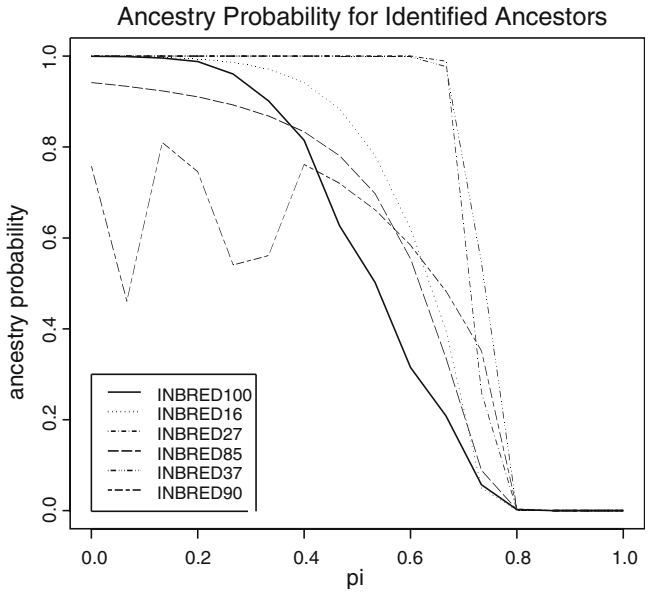

Fig. 3 The relationship between the median ancestry posterior probabilities and the assumed genotyping error rates for the real data application

information contained in the three hybrids. Our collaborators who supplied the maize data indicated that an error rate of 0.05 is more realistic. For error rates as small as these there is little difference between the ancestry probabilities calculated with and without accommodating genotyping errors.

Summary

This paper extends the ancestry probability assessment method of Berry et al. (2002) to account for genotyping errors. For this initial exploration of the issue we assume a known constant error rate across the genome. Simulations are used to study the effect of genotyping errors on ancestry probability calculations and the relationship between the calculated ancestry probabilities and the assumed genotyping error rate π. The simulations indicate that when the error rate is high and the true ancestors are not in the data set, the Berry et al. approach (which does not explicitly address genotyping errors) tends to produce slightly higher ancestry probabilities than the approach that accommodates errors. For low error rates, including those that seem most likely in practice, our simulations support the robustness results of Berry et al. (2002). Their method produces reliable inferences about ancestry.

There are some limitations in our approach. We follow Berry et al. (2002) in using a composite likelihood approximation to the full likelihood of the marker profile. This approximation matches the exact likelihood only if all of the markers segregate independently. The empirical evidence presented here and in Berry et al. (2002) suggests that the composite likelihood approximation is adequate for this application to ancestry determination. It is important that one determines whether such an approximation is adequate before applying the method in other organisms or with different marker profiles. In building up the model to accommodate genotyping errors we only consider genotyping errors for hybrids and assume independent occurrences of errors along the genome. Loosening these assumptions is an area of current research.

References

Alderson GW, Gibbs HL, Sealy SG (1999) Parentage and kinship studies in an obligate brood parasitic bird, the brown-headed cowbird (Molothrus ater), using microsatellite DNA markers. J Hered 90:182–190

Berry AD, Seltzer DJ, Xie C, Wright LD, Smith CS (2002) Assessing probability of ancestry using simple sequence repeat profiles: application to maize hybrids and inbreds. Genetics 161:813–824

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

Broman K, Weber J (1998) Estimation of pairwise relationships in the presence of genotyping errors. Am J Hum Genet 63:1563–1564

Chakraborty R, Meagher TR, Smouse PE (1988) Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. Genetics 118:527–536

Devlin B, Risch N, Roeder K (1996) Disequilibrium mapping: composite likelihood for pairwise disequilibrium. Genomics 36:1–16

Ellstrand NC (1984) Multiple paternity within the fruits of the wild radish, Raphanus sativus. Am Nat 123:819–828

Ewen K, Bahlo M, Treloar S, Levinson D, Mowry N et al (2000) Identification and analysis of error types in high-throughput genotyping. Am J Hum Genet 67:727–736

Garner C, Slatkin M (2002) Likelihood-based disequilibrium mapping for two-marker haplotype data. Theor Popul Biol 61:153–161

Lindsay BG (1988) Composite likelihood methods. Contemp Math 80:221–239

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

Ott J (1991) Analysis of human genetic linkage (revised edition). The Johns Hopkins University Press, Baltimore

Rannala B, Slatkin M (2000) Methods for multipoint disease mapping using linkage disequilibrium. Genet Epidemiol 19(Suppl 1):S71–S77

San Cristobal M, Chevalet C (1997) Error tolerant parent identification from a finite set of individuals. Genet Res 70:53–62

Zhang H, Stern H (2003) Estimating genotyping error rate for maize using SSR profiles. Technical report. Department of Mathematics and Statistics, University of West Florida, Pensacola, FL 32514