DOI: 10.1534/genetics.103.022061

Using Molecular Sizes of Simple Sequence Repeats *vs.* Discrete Binned Data in Assessing Probability of Ancestry: Application to Maize Hybrids

Donald A. Berry,*[,1,2] Deanne Wright,[†,1] Chongqing Xie,[†] Jon D. Seltzer[‡] and J. Stephen C. Smith[†]

**University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, [†]Pioneer Hi-Bred International, Johnston, Iowa 50131 and [‡]Medtronic, Minneapolis, Minnesota 55432*

Manuscript received September 8, 2003
Accepted for publication February 4, 2005

ABSTRACT

Most inferential methods for profiling genotypes based upon the use of DNA fragments use molecular-size data transcribed into discrete bins, which are intervals of DNA fragment sizes. Categorizing into bins is labor intensive with inevitable arbitrariness that may vary between laboratories. We describe and evaluate an algorithm for determining probabilities of parentage based on raw molecular-size data without establishing bins. We determine the standard deviation of DNA fragment size and assess the association of standard deviation with fragment size. We consider a pool of potential ancestors for an index line that is a hybrid with unknown pedigree. We evaluate the identification of inbred parents of maize hybrids with simple sequence repeat data in the form of actual molecular sizes received from two laboratories. We find the standard deviation to be essentially constant over the molecular weight. We compare these results with those of parallel analyses based on these same data that had been transcribed into discrete bins by the respective laboratories. The conclusions were quite similar in the two cases, with excellent performance using either binned or molecular-size data. We demonstrate the algorithm's utility and robustness through simulations of levels of missing and misscored molecular-size data.

THE application of molecular marker technologies to characterize genotypes is of fundamental importance in basic and applied research, including studies to identify the genetic control of complex traits; to determine phylogenies and pedigrees; and to identify animal breeds, plant varieties, or individuals (Narvel *et al.* 2000; Andersson 2001; Cardon and Bell 2001; Barton and Keightley 2002; Dekkers and Hospital 2002; Glazier *et al.* 2002; Cavalli-Sforza and Feldman 2003; Grassi *et al.* 2003; Tang *et al.* 2003; Tommasini *et al.* 2003). Laboratories typically have in-house procedures for converting molecular-size data into bins. Binned data are used for many applications, including parentage analysis and forensic science applications, and were the basis of our previous work (Berry *et al.* 2002, 2003). The comparison of the molecular size of amplified DNA fragments is therefore a fundamental aspect for most applications that use simple sequence repeats (SSRs) to characterize numerous genotypes at many loci (Abe *et al.* 2003; Baek *et al.* 2003; Tang and Knapp 2003; Yu *et al.* 2003), including studies of pedigrees (Chaix *et al.* 2003; Sjakste *et al.* 2003; Vouillamoz *et al.* 2003).

We have introduced algorithms for determining probabilities of ancestry of a hybrid (Berry *et al.* 2002) and an inbred line (Berry *et al.* 2003) based on molecular marker profiles of discrete binned alleles. We assumed no prior knowledge of pedigree. For example, there may be gaps in the pedigree with neither of the parents known. We showed that the algorithms are robust in the presence of missing and/or mistyped data.

Berry (1991) and Berry *et al.* (1992) proposed an alternate method for comparing DNA fragment data that directly compares molecular size instead of relying on binned values. Advantages include avoiding the need to define an arbitrary molecular-weight cutoff between fragments for classifying as "like," having a greater match probability for identical fragment lengths than for fragment lengths that are some distance apart, and establishing match criteria that can be described and used repeatedly in place of informal visual matches that are used in some laboratories. Moreover, utilizing raw molecular-size data can better account for laboratory error and process variability. Finally, and of great practical importance, directly applying molecular-size data eliminates the time-consuming and demanding task of assigning amplified fragments into discrete bins. As laboratories increase their knowledge of a particular marker, they may need to redefine bins; boundaries of bins may be moved and bins may be split or combined. Different systems of binning within the same laboratory and across different laboratories can complicate the science and undermine the legal interpretations of scientific analyses. Using raw

[1] These authors contributed equally to methodological development and to the application of the methods.

[2] *Corresponding author:* Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd., Unit 447, Houston, TX 77030-4009. E-mail: dberry@mdanderson.org

molecular-size data eliminates the need to manage and defend the details of bins and serves to standardize procedures across laboratories.

Despite these advantages and the convenience of raw molecular sizes, their use is not standard. Our literature search found that published analyses of molecular marker data utilizing representatives of the plant kingdom that are derived from gel migration data, including SSRs, use discrete binned data rather than actual molecular sizes. A principal reason for this is the lack of good algorithms that can use measured molecular sizes.

Our objective is to describe and evaluate an algorithm for determining probabilities of ancestry based on raw molecular-size data for an index line and for a pool of potential ancestors. We develop and evaluate the methodology for the case in which the index genotype of unknown pedigree is a hybrid, but the central idea applies as well for inbreds and to other breeding circumstances. The ramifications of our approach to the analysis of SSR data transcend both the type of data and the application that we have investigated (pedigree analysis). A fundamental issue in the analyses of genetic and genomics data is whether two or more observations are alike. An example is the intensity of genetic expression in cDNA microarrays. Measuring the degree of concordance may lead to more powerful conclusions than using yes/no procedures that simply conclude whether there is agreement. Our methods can be adapted to these settings.

METHODS

Algorithm: Our new algorithm for determining probabilities of ancestry of a hybrid line using raw molecular-size SSR data is a variation of the algorithm presented in Berry *et al.* (2002). The latter was developed to determine probabilities of ancestry of a hybrid line based on discrete (binned) SSR alleles. In this article, SSR refers to the actual molecular-size value that is assigned for a genotype at a particular locus. Consider an index hybrid whose parentage is unknown or in dispute. A database containing molecular sizes at a number of loci for this index hybrid along with a set of potential inbred ancestors is available. The objective is to find the (posterior) probability of the closest ancestry for each inbred in the database using this genotypic information.

Consider a pair of possible ancestors, inbred i and inbred j. We calculate the posterior probability that inbreds i and j are in the index's ancestry, repeating this for all pairs of inbreds in the database. Let $P(i, j|\text{SSRs})$ stand for the posterior probability that i and j are ancestors of the index given the molecular-size values for the various SSRs. Let $P(i, j)$ stand for the unconditional (or prior to the data) probability of the same event and let $P(\text{SSRs}\,|i, j)$ be the likelihood for observing the various SSR weights, if in fact i and j are ancestors of the index hybrid. Just as in Berry *et al.* (2002), Bayes' rule relates these various probabilities,

$$P(i, j|\text{SSRs}) = P(\text{SSRs}\,|i, j) \times P(i, j)/\Sigma[P(\text{SSRs}|u, v) \times P(u, v)],$$

where the sum in the denominator is over all pairs of inbreds in the database, indexed by u and v. We use the relatively noninformative (uniform) prior assumption that $P(i, j)$ is the same for all pairs (i, j). Then $P(u, v)$ is a constant, and as a common multiple in the denominator it cancels with $P(i, j)$ in the numerator:

$$P(i, j|\text{SSRs}) = P(\text{SSRs}\,|i, j)/\Sigma P(\text{SSRs}\,|u, v).$$

Considered as a function of the index's SSRs, $P(\text{SSRs}\,|i, j)$ is the probability (more accurately, probability density) of observing these SSRs assuming inbreds i and j are both ancestors. But considered as a function of the pair (i, j), it is the likelihood obtained as a product of likelihoods for observing the index genotype given the potential ancestor (i, j) alleles at each locus, individually.

The single-locus likelihood calculation is fundamentally different when raw molecular-size data are used in place of alleles that have been binned into discrete categories. When using raw molecular-size data, we do not rule out a match between the index and a potential ancestor allele just because they happen to be different. Instead, we assign a likelihood of a match on the basis of the difference between the two measured sizes. We account for laboratory error that occurs in estimating molecular sizes by explicitly considering that an observed molecular size is actually a random deviation from an underlying true molecular size. Although various error distributions are possible, we have found from replicate measurements that a normal distribution fits the actual error distribution well. So the distribution of the difference between a potential ancestor's observed molecular size (a) and the index hybrid's observed molecular size (o), when in fact these molecular sizes are observations of the same allele, can be written as

$$o - a \sim \text{normal}(0, V_{o-a}), \qquad (1)$$

where V_{o-a} is the variance (squared standard deviation) of the observed difference. We consider the matter of estimating V_{o-a} later in this section.

As an example, suppose an offspring's measured molecular size at a particular SSR is $o = 180.2$ bp and a potential ancestor's is $a = 180.6$ bp. Consider three consecutive bins, as shown in Figure 1. The three bins have molecular sizes between 178.5 and 179.5 bp (bin 1), between 179.5 and 180.5 bp (bin 2), and between 180.5 and 181.5 bp (bin 3). Using binning procedures, a and o are in different bins and so they do not match. But they are close to each other, closer even than some sizes in bin 2 that would be called a match for o. In this article, we account for the measurement error involved

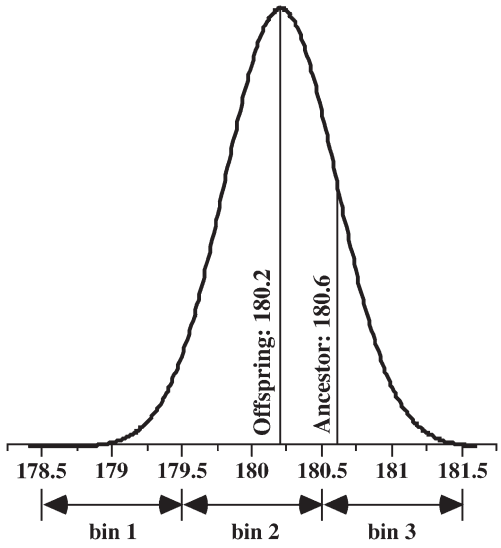

FIGURE 1.—Frequency distribution of a replicate observation when the first observed molecular size is 180.2 bp. The height of the curve is the likelihood of replicate measurements when the first measured molecular size (offspring) is o = 180.2. Values further from 180.2 are less likely. The bins are shown for comparison but play no role in the proposed method.

in determining molecular sizes by assigning a likelihood of a match to a potential ancestor allele depending on the difference between its measured molecular size and that of the offspring allele.

Figure 1 shows the frequency distribution of a replicate observation when the first observed molecular size is 180.2 bp. The frequency distribution of the next molecular size is represented by the height of the curve. Therefore, the frequency distribution in Figure 1 is the likelihood of the ancestor's molecular size at this SSR, as indicated by Equation 1. In Figure 1, there is a positive density at 180.6 bp even though its bin is different from that for the offspring. However, 180.6 is not as probable as are values closer to 180.2 bp. On the other hand, 180.6 is more likely than values between 179.5 and 179.8, even though they are in the same bin as 180.2. Likelihoods decrease to 0 for ancestral molecular sizes sufficiently far from 180.2 bp. However, the decrease is gradual rather than abrupt as it is in the case of binning.

We generalize the algorithm of BERRY *et al.* (2002), using the normal distribution as in Equation 1 to calculate the likelihood of each pair (i, j) of ancestors for the observed molecular sizes of the index hybrid at a particular locus. As in BERRY *et al.* (2002), we allow for alleles to be introduced through additional breeding with i and j prior to the creation of the final hybrid. Thus, since the degree of ancestry of i and j (if any) is unknown, we label the actual probability of i (or j) passing on one of their alleles to the index hybrid to be p. The value $p = 1$ would be appropriate if the closest ancestors in the data pool were parents and there were no mutations or laboratory errors. Allowing for a low rate of mutations and errors we consider $p = 0.99$ instead of 1. However, our primary concern is when the parents are not in the database. Therefore we also consider other values of p, notably $p = 0.50$, which is consistent with the closest ancestors in the database being grandparents.

When inbreds i and j are truly ancestors then there are four possibilities: (I) the alleles of both i and j were passed to the index hybrid, (II) i came through but not j, (III) j came through but not i, and (IV) neither came through. Assuming independence, these have respective probabilities p^2, $p(1 - p)$, $p(1 - p)$, and $(1 - p)^2$. Let (o_1, o_2) represent the offspring alleles at an SSR, so that $P(\text{SSR}\,|i, j) = P(o_1, o_2|i, j)$ stands for the likelihood of observing the offspring alleles at the SSR given that i and j are ancestors. Using the law of total probability this overall likelihood is the (weighted) average over these four cases. For a given value of p,

$$P(o_1, o_2|i, j) = p^2P_1(o_1, o_2|i, j) + p(1 - p)P_2(o_1, o_2|i, j) + p(1 - p)P_3(o_1, o_2|i, j) + (1 - p)^2P_4(o_1, o_2|i, j), \quad (2)$$

where $P_1(o_1, o_2|i, j)$ is the likelihood of the pair (i, j) calculated assuming that case I applies and similarly for cases II–IV. We provide detailed likelihood calculations for case I only, in which the alleles of both i and j are passed to the index hybrid. Let (o_1, o_2) represent the index hybrid (offspring) alleles at the locus under consideration, let (a_{i1}, a_{i2}) represent inbred i's alleles, and let (a_{j1}, a_{j2}) represent inbred j's alleles. This likelihood is computed by considering all possible ways that i and j could have contributed their alleles to the index offspring,

$$P_1(o_1, o_2|i, j) = (1/8)\{f(o_1, a_{i1})f(o_2, a_{j1}) + f(o_1, a_{j1})f(o_2, a_{i1}) + f(o_1, a_{i1})f(o_2, a_{j2}) + f(o_1, a_{j2})f(o_2, a_{i1}) + f(o_1, a_{i2})f(o_2, a_{j1}) + f(o_1, a_{j1})f(o_2, a_{i2}) + f(o_1, a_{i2})f(o_2, a_{j2}) + f(o_1, a_{j2})f(o_2, a_{i2})\}, \quad (3)$$

where f is the likelihood assigned to the corresponding pair of molecular sizes. Equation 3 can also be clarified from a descent graph theory by using meiosis indicators (THOMPSON 1994). Let a and o be the ancestor and index hybrid alleles under consideration. Define

$$f(o, a) = \begin{cases} (1/10)\,\text{normal}(0, 0, V_{o-a}) & \text{if } a \text{ is missing} \\ \text{normal}(a - o, 0, V_{o-a}) & \text{otherwise,} \end{cases} \quad (4)$$

where $\text{normal}(\cdot, 0, V_{o-a})$ is the normal probability density having mean 0 and variance V_{o-a}. When a potential ancestor's allele is missing, the likelihood is one-tenth of the maximum possible likelihood, which corresponds to the exact agreement of the two molecular sizes. In our algorithm for binned alleles (BERRY *et al.* 2002,

2003), the likelihood assigned in this case is $(1/N) \cdot 1$, where N is the number of alleles at a locus and 1 is the maximum possible likelihood; when molecular sizes are used, it is not possible to calculate the number of alleles at a locus, but $N = 10$ is a choice that makes our calculation similar to that of the binned algorithm. Likelihoods corresponding to cases II, III, and IV are computed similarly.

Estimating the variance V_{o-a} of the difference between two separate measurements of a molecular size representing the same allele is critical. This variance is functionally related to V_X, the variance of the measurement of the molecular size of a given allele. We obtained replicate SSR profiles for a number of inbred lines to enable estimating this variance as the variance of the replicate observations. However, some amplified fragments from inbred replicates clearly did not represent the same allele nor were they stutter bands of the same allele. Such discrepancies result from incomplete fixation of SSR loci (even after seven or eight generations of self-pollination) and the consequential amplification of alternate alleles in replicate samples. Therefore, we omitted the most obvious mismatches (>2 bp different), *i.e.*, those that were due to the amplification of alternate alleles prior to calculating the variance. Mismatches that were within the range of 2 bp and that could be caused by discrepancies in amplification and migration from the same allele were included in calculations of variance. We evaluated V_X over a range of molecular sizes (underlying allele values) and markers and found that using a constant was reasonable (see RESULTS and Figure 2).

Table 1 contains example calculations of Equation 3, assuming $p = 0.5$ and $V_{o-a} = 0.06$. The columns labeled hybrid, inbred i, and inbred j give the observed molecular sizes at four distinct loci. From Equation 4, when an allele to be compared with an offspring allele is missing (such as in cases II–IV, where an unknown ancestor other than i or j contributed an allele to the offspring), the likelihood of a match is 0.1629 (one-tenth of the normal density at its mode). The values in column $P_1(o_1, o_2|i, j)$ (case I) are obtained from Equation 3, while columns $P_2(o_1, o_2|i, j)$, $P_3(o_1, o_2|i, j)$, and $P_4(o_1, o_2|i, j)$ (cases II–IV) are obtained via analogous formulas. Column $P(o_1, o_2|i, j)$ gives the probability density of the offspring alleles at the SSR in question, considering cases I–IV. In Table 1, the only heterozygous genotype/SSR is the hybrid at SSR 1. In practice, most loci are homozygous for inbred lines and some are homozygous for hybrids as well.

In Table 1, SSR 1 is an example in which the hybrid alleles could have been observed from those of inbreds i and j since the molecular sizes reported are close to those reported for i and j. As a consequence, the likelihood in case I—which is based on the assumption that both i and j passed an allele to the hybrid—is much greater than that for cases II–IV. Case IV, which is based on the assumption that neither i nor j passed an allele to the offspring, is 0.1629^2 since, in this case, both ancestor alleles are missing and we are using a constant value of V_{o-a}.

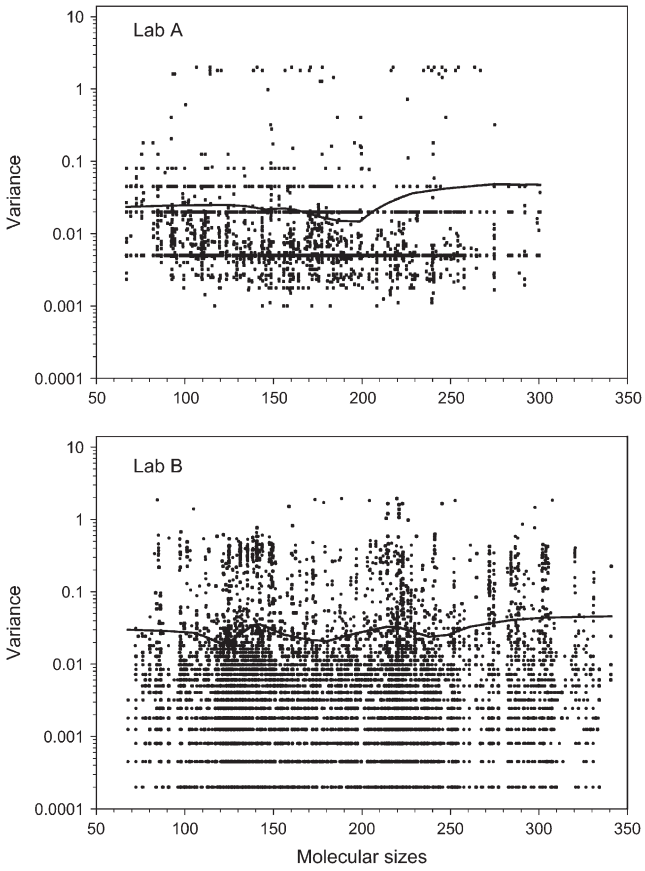

FIGURE 2.—Variance, in semi-log scale, of replicate observations of the same allele across a range of molecular sizes reported by two laboratories. The curves are obtained by using local regression (loess) to fit linear or quadratic functions of the predictors at the centers of neighborhoods. Top, laboratory A data; bottom, laboratory B data.

SSR 2 is an example in which either inbred could be an ancestor of the hybrid or they could both be ancestors. For SSR 3, inbred i could be an ancestor, but it is highly unlikely that inbred j is an ancestor. Finally, SSR 4 is an example where it is unlikely that either inbred contributed an allele to the hybrid, and this is reflected in the very small value for $P(o_1, o_2|i, j)$.

We have shown how to compute the likelihood of observing the index offspring's alleles given a pair of potential ancestors at a particular locus and also the value of $P(\text{SSRs}|i, j)$ by multiplying the various $P(\text{SSR}|i, j)$s. To determine the probability that any particular inbred, say inbred i, is the closest ancestor of the index, sum $P(i, v|\text{SSRs})$ over all inbreds v with $v \neq i$. Call this $P(i|\text{SSRs})$. The maximum of $P(i|\text{SSRs})$ for any inbred i is 1. But since there is one closest ancestor on each side of the family, the sum of $P(i|\text{SSRs})$ over all inbreds i is 2.

A possibility not yet considered is that more than two alleles are observed for an SSR marker run on an

TABLE 1

Example calculations for the likelihood of observing the offspring alleles given that *i* and *j* are ancestors at four distinct loci

SSR	Hybrid (o_1, o_2)	Inbred i (a_{i1}, a_{i2})	Inbred j (a_{j1}, a_{j2})	$P_1(o_1, o_2 \mid i, j)$	$P_2(o_1, o_2 \mid i, j)$	$P_3(o_1, o_2 \mid i, j)$	$P_4(o_1, o_2 \mid i, j)$	$P(o_1, o_2 \mid i, j)$
1	229.18	235.29	229.12	1.2202	0.1257	0.1287	0.0265	0.3753
	235.21	235.29	229.12					
2	108.3	108.57	108.68	0.4337	0.1445	0.0796	0.0265	0.1711
	108.3	108.57	108.68					
3	172.77	172.81	291.32	0	0.2617	0	0.0265	0.0721
	172.77	172.81	291.32					
4	103.45	228.17	222.00	0	0	0	0.0265	0.0066
	103.45	228.17	222.00					

Columns labeled hybrid, inbred *i*, and inbred *j* give the observed molecular-size values at four distinct loci.

individual DNA sample. This can be due to SSR locus duplication, homeology due to alloploidy, more than one individual plant being sampled for DNA extraction, or cross-contamination. We consider all possible pairings of the observed alleles and perform calculations using multiple imputation (Little and Rubin 1987). Namely, we select two of the allelic sizes at random from each line that has more than two and run the algorithm. Then we repeat the process making independent selections of two alleles for each such line and average the results.

As in our earlier work, we assume here that the probability p that i (or j) passed one of its alleles to the index hybrid is the same for both ancestors. This would not be true for some mating designs. It is a simple modification to allow for two different probabilities, say p_i and p_j, for the two ancestors. In this case, Equation 2 would become

$$\begin{aligned} P(o_1, o_2 \mid i, j) = {} & p_i p_j P_1(o_1, o_2 \mid i, j) + p_i(1 - p_j) P_2(o_1, o_2 \mid i, j) \\ & + (1 - p_i) p_j P_3(o_1, o_2 \mid i, j) \\ & + (1 - p_i)(1 - p_j) P_4(o_1, o_2 \mid i, j). \end{aligned}$$

The focus of this article is to determine probabilities of ancestry of a hybrid using molecular-size SSR data. However, by modifying the calculation for the likelihood of observing the index alleles at a locus to account for the self-pollination that occurs in creating an inbred line, we can easily modify this algorithm to assign probabilities of ancestry for inbred lines. For example, if i and j are ancestors and we assume that an allele from each is passed to an intermediate hybrid that is created prior to the self-pollination process to create the final inbred offspring, we have

$$P_1(o \mid i, j) = (1/4)\{f(o, a_{i1}) + f(o, a_{j1}) + f(o, a_{i2}) + f(o, a_{j2})\}.$$

SSR data: We evaluated our new algorithm using molecular-size data from two different laboratories. Laboratory A used 100 SSR loci for 10 hybrids and 52 inbred lines using methods that were essentially identical to those described in Berry *et al.* (2002). Each inbred line was replicated at least twice and several had six-replicated entries. Laboratory B used 195 SSR loci to profile 54 hybrids and 544 inbred lines, as previously described in Berry *et al.* (2002). Fifty-six inbred lines were each replicated twice. SSR loci used by each laboratory provided a sampling of genetic diversity at mapped positions on each chromosome arm of maize. In neither laboratory did staff have knowledge of the identity of the sample genotypes or the pedigree relationships among the genotypes.

For both laboratories, the parental inbred lines of each hybrid were included among a pool of inbred lines that each laboratory profiled; the pool of inbred lines profiled by laboratory B also included all the grandparents of the hybrids along with numerous inbred lines that are very closely related by pedigree, including full-sibs, half-sibs, and inbreds derived from the parents and grandparents of those hybrids. So data from laboratory B provided the greater challenge for determining ancestry. Consequently, we used laboratory B to evaluate robustness of the algorithm with respect to missing data, mistyped data, both missing and mistyped data, and different variances, V_X.

An additional difference in laboratories was that laboratory A used SSR loci that were primarily based on direpeat types while laboratory B used SSRs of di-, tri-, and tetrarepeat types, with the majority being the former two. Thus, bins used by laboratory B tended to be larger (2–3 bases) than those used by laboratory A (1–2 bases). The average variances of molecular sizes were $V_X = 0.0267$ for laboratory A and $V_X = 0.0289$ for laboratory B. Both variances were essentially constant over the range of molecular sizes examined, 70–300 bp (Figure 2). We therefore used $V_X = 0.03$, which gives $V_{o-a} = 0.06$.

RESULTS

For laboratory A, the pedigree of the 10 hybrids was assessed relative to the set of 52 inbreds. Both parents

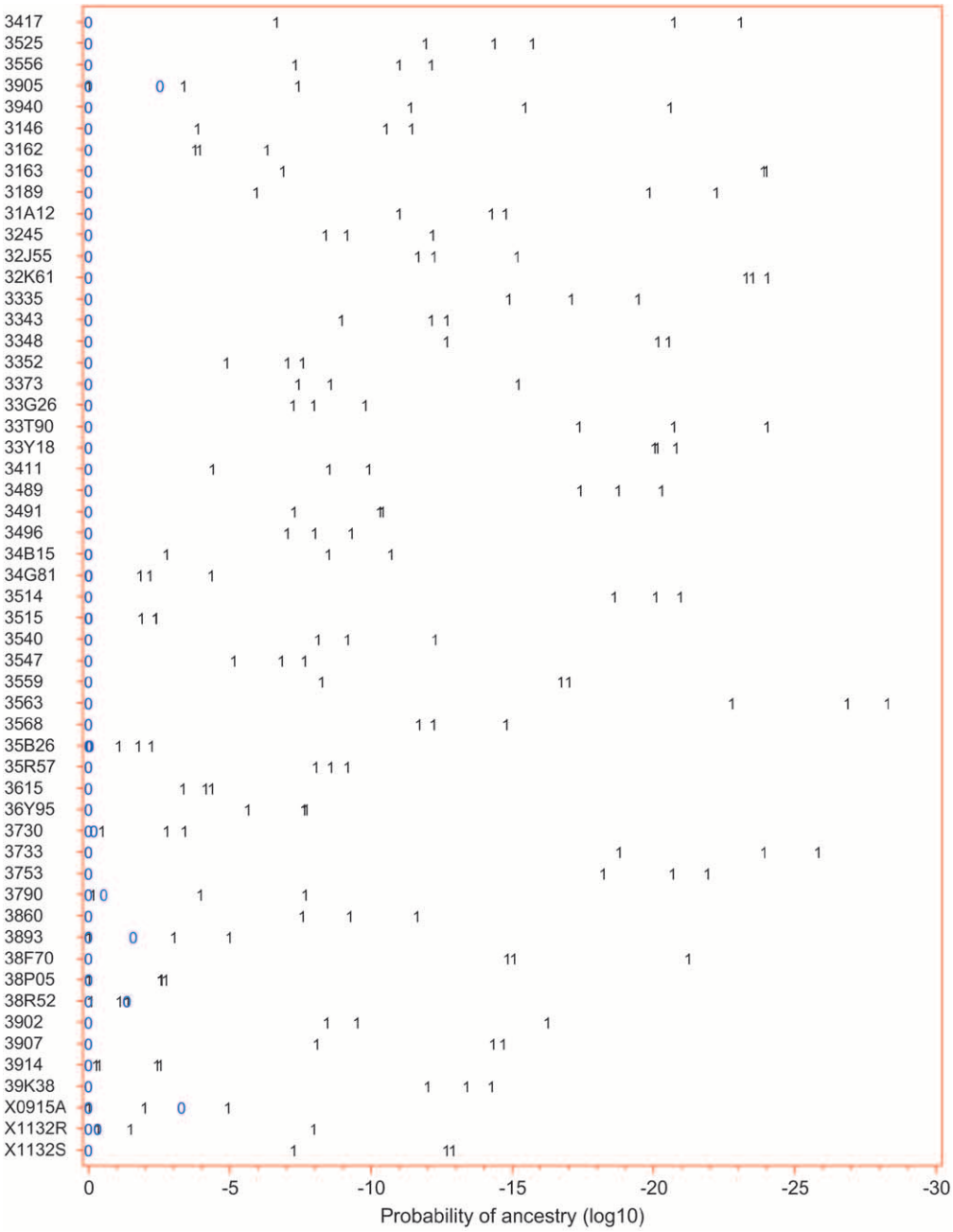

FIGURE 3.—Probability of ancestry, assuming $p = 0.50$, using molecular sizes for the 54 maize hybrids with the set of 544 inbred lines as possible ancestors. Legend: 0 0 0 = parents; 1 1 1 = nonparents.

of all 10 hybrids were correctly identified by our new algorithm, whether p was 0.50 or 0.99. The results for the 54 hybrids profiled by laboratory B, with the associated set of 544 inbred lines as possible ancestors, are presented in Figures 3 ($p = 0.50$) and 4 ($p = 0.99$). They show the top 5 inbreds for each hybrid, provided their probabilities are at least 10^{-30}. For both values of p, all 54 had at least one parent ranked in the top two. For $p = 0.50$, 47 (87%) had both parents with the highest two probabilities. For $p = 0.99$, 52 (96%) had both parents ranking highest. Of the 7 hybrids with one parent incorrectly ranked when $p = 0.50$, another inbred ranked above the true parent in a total of 15 instances. In 9 instances the interloper was an immediate descendant of the true parent, including 6 instances where the true parent was used recurrently in the pedigree of the interloper; in 5 instances the interlopers were full sisters of the true parent; and in one instance it was a grandparent. Of the 2 hybrids that had one parent incorrectly ranked when $p = 0.99$, there were 2 interloping inbreds and both were direct descendants of the respective true parent.

Figure 5 depicts the relationship of parentage probabilities and Malécot's coefficients between the 54 hybrids and the top 10 ranking inbreds (two parents are included in the top 10 ranked inbreds) for each hybrid.

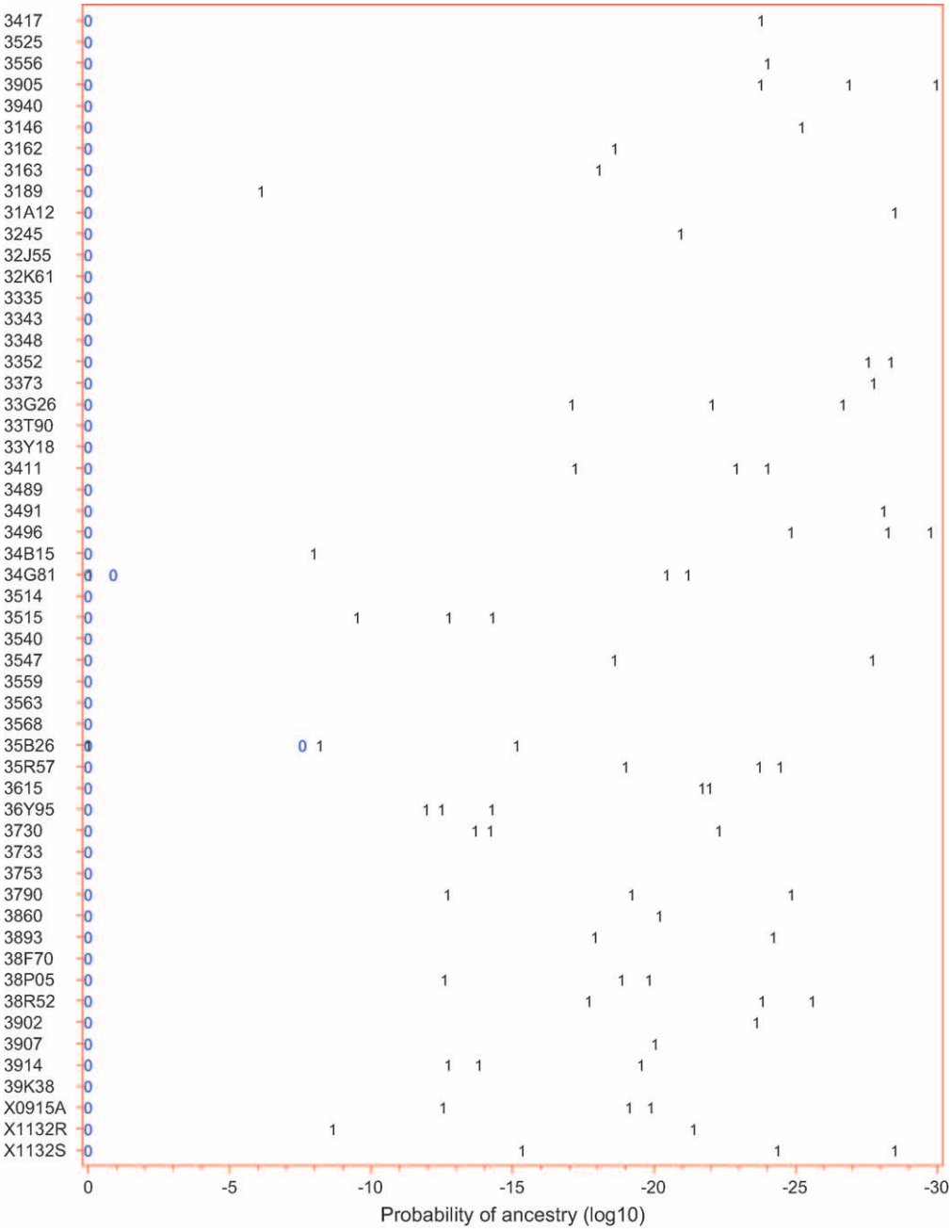

FIGURE 4.—Probability of ancestry, assuming $p = 0.99$, using molecular sizes for the 54 maize hybrids with the set of 544 inbred lines as possible ancestors. 0 0 0, parents; 1 1 1, nonparents.

Several non-parent-inbred lines have Malécot's coefficients similar to those of the parents, but they usually have small probabilities of parentage.

We compared our results with those from the algorithm of BERRY *et al.* (2002), which is based on discrete (binned) allele scores. For laboratory A, both parents of all 10 hybrids were correctly identified, just as for the new algorithm. For laboratory B, the proportions of correct assignments of actual parents were 96% ($p = 0.50$) and 99% ($p = 0.99$), compared with 93% and 98%, respectively, for the new algorithm.

We used data from laboratory B and $p = 0.50$ to examine the ability of the algorithm to identify grandparents of the 54 hybrids after removing the true parents and their direct descendants from the data set. Seven hybrids (13%) identified none of their grandparents, 23 hybrids (43%) identified one grandparent, and 24 hybrids (44%) identified two grandparents, which were ranked into the top two places; two hybrids identified three grandparents that were ranked into the top three places. The algorithm had more difficulty correctly identifying the four grandparents than in identifying parents. This result is not unexpected since the algorithm was designed to identify the most likely members of the pedigree. Moreover, it is customary in plant breeding to develop progeny from crosses of related parents, including making repeated crosses of the same parent. These procedures result in progeny and related lines

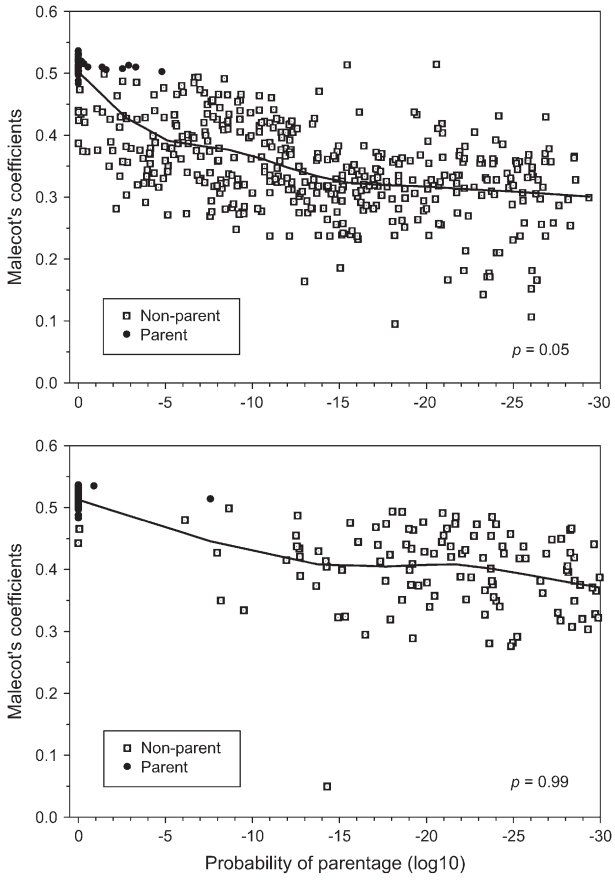

FIGURE 5.—Relationships between Malécot's coefficient and the probability of ancestry, for each of the 54 hybrids along with the top 10 ranking inbreds using our algorithm. Results are shown for those inbreds with probability of ancestry of at least 10^{-30}. The curves are obtained by using local regression (loess) to fit linear or quadratic functions of the predictors at the centers of neighborhoods. Top, results using $p = 0.05$; bottom, results using $p = 0.99$.

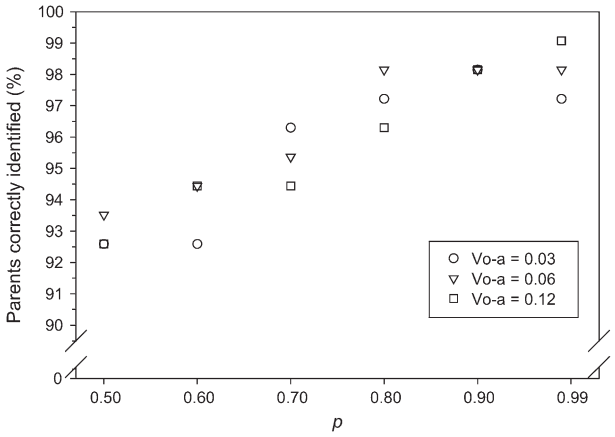

FIGURE 6.—Percentage of parents of the 54 hybrids that were correctly identified across a range of values for p. Results are shown for three different values of the variance parameter, V_{o-a}.

that share a higher degree of relatedness than they share with their grandparents. These types of close relationships would be expected in a pedigree-based breeding program (THOMPSON and MEAGHER 1987).

We evaluated the algorithm's robustness to the value of p, the probability that an ancestor passes its allele to the offspring, by examining the percentage of actual parents correctly identified over a range of values of p for each of the 54 hybrids profiled by laboratory B (Figure 6). True parents were correctly identified more often for larger values of p. We also considered the algorithm's robustness to variance V_{o-a} in Figure 6 by varying it over the range 0.03–0.12. There was no apparent advantage to using a particular value for this variance. Nonetheless, each laboratory, and initially each data set, should include replicated entries to allow a reasonable estimate of variance.

To evaluate robustness of the algorithm to the amount and quality of SSR data, we again used the more extensive data provided by laboratory B. We considered several criteria: (a) the number of loci from which SSR data were considered, (b) the amount of additional missing data (above the natural levels present in the data) for different numbers of loci, (c) the amount of additional mistyped data for different numbers of loci, and (d) the amount of additional missing and mistyped data for different numbers of loci.

To assess the robustness and number of loci needed for pedigree analysis using molecular-size SSR data, we used random subsets of 50 and 100 loci from the 194 loci that were available in the full data set. Figure 7 shows the percentage of parents of the 54 hybrids that were correctly identified in each case along with additional levels (0, 2.5, 5, 10, and 25%) of simulated missing data, misscored data, and combined levels of missing and misscored data. The algorithm identified most parents correctly when data from as few as 100 of the original 194 SSR loci were used with up to 25% additional missing data, 10% additional misscored data, or a combination of 10% additional missing and 10% additional misscored data (~20% combined additional data error). When 50 loci were used, the percentage of correctly identified parents significantly decreased, with only 62–82% of the parents correctly identified for error rates from 0 to 0.10%. For the types of errors we examined, missing data had the least effect on parentage testing, whereas missing plus mistyped data were most likely to lead to erroneously identified parents.

DISCUSSION

The most widely used methods of genotyping using DNA, either simple sequence repeats or variable number tandem repeats, provide measurements of DNA fragments in terms of molecular size. Laboratories can choose to report these raw data or they can transform them into discrete "bin" scores. The latter use is com-

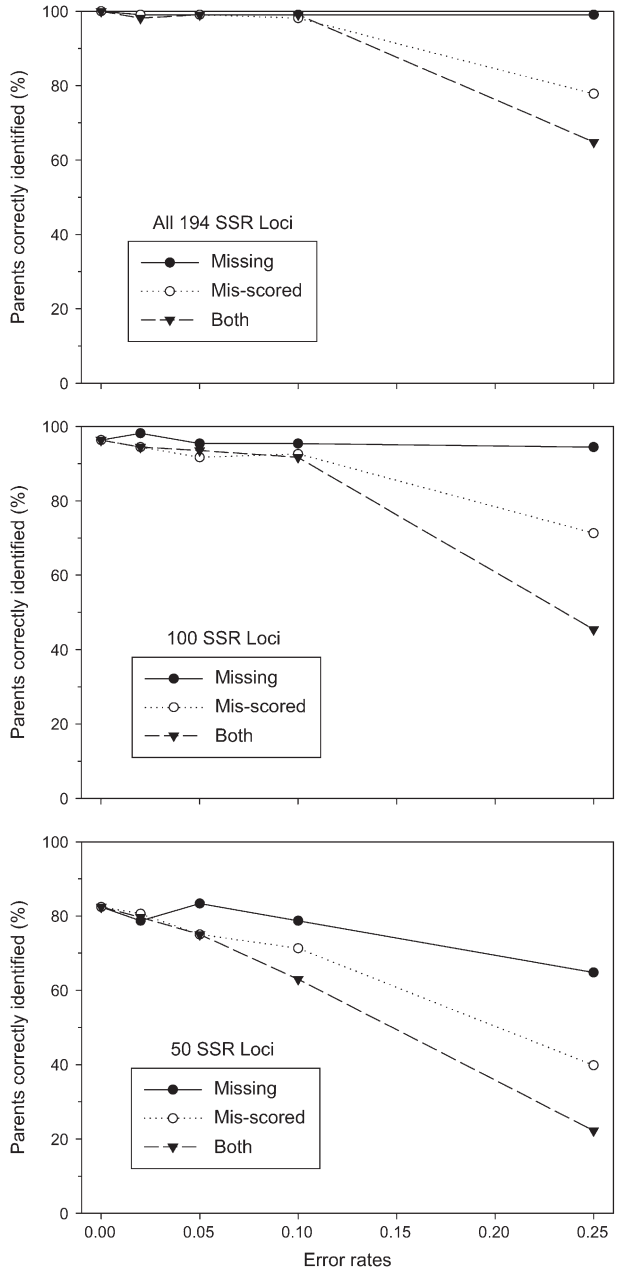

Figure 7.—Percentage of parents of the 54 hybrids that were correctly identified using different levels of simulated error, including missing and misscored data. Results are shown using all 194 SSR loci (top), 100 SSR loci (middle), and 50 SSR loci (bottom).

mon. Berry (1991) proposed a method for comparing raw molecular sizes rather than binned scores. The approach is based on degree of similarity rather than a yes/no decision regarding a match. However, until now there has been a lack of algorithms for comparing molecular sizes. In this article we apply this approach to assessing probability of ancestry.

We evaluated our method of finding the probability of parentage for a hybrid. We utilized two different SSR data sets of maize hybrids and inbred lines. We compared our results using molecular sizes to those from using binned data (Berry *et al.* 2002). Both data sets included a number of hybrids and their inbred parents along with numerous other inbred lines, many of which were very closely related by pedigree and had SSR profiles similar to one or more inbred parents of the hybrid in question. The new algorithm performed well in that it successfully identified the true parents for most of the hybrids. The discrete version of the algorithm actually performed slightly better than the new algorithm for the process employed by laboratory B. This reflects, at least in part, the accuracy of this laboratory's binning process. Creating binned alleles from the raw molecular-size data was labor intensive, requiring at least two full-time personnel. The minor differences between the two methods may not justify the time and resources needed to accurately translate molecular sizes into bin scores. Our results indicate that it is possible to directly apply molecular-size data in the analysis of pedigrees.

The algorithm proposed here requires an estimate of the variance of molecular sizes representing the same allele. We obtained this estimate for each laboratory by using replicated genotypes. Variances based on data from the two laboratories were nearly identical (both $\sim$0.03). This variance did not vary according to fragment size. (This result differs from that for restriction length polymorphism data, where Berry 1991 found the standard deviation to be proportional to band size.) A laboratory implementing pedigree analysis using molecular sizes should include replicates of genotypes to allow for calculating this variance and its possible dependence on molecular size. Our algorithm can readily accommodate any level of variance.

Our new algorithm is robust in that it correctly identified most parents when results from as few as 100 of the original 194 SSR loci were used, with up to 25% additional missing data and an additional 10% misscored data. It was also robust to different estimates of the variance for molecular sizes representing the same allele.

The algorithm presented here is a generalization of the algorithms presented in Berry *et al.* (2002, 2003) for binned allele data. Its results can be compared with our earlier calculations. This is accomplished by modifying the version of f given in Equation 4 as

$$F^*(o, a) = \begin{cases} 1/N & \text{if } a \text{ is missing} \\ 0 & \text{if } o \neq a \\ 1 & \text{if } o = a, \end{cases}$$

where N is the total number of alleles at the SSR in question. This generalization provides a clear connection between the two algorithms, and it greatly reduces the programming required to modify our previous methods.

We evaluated our algorithm for pedigrees of maize hybrids. But it can readily be modified for use in assigning ancestry for inbred lines. In addition, it can be

used for pedigree analyses in other plant and animal species. Moreover, it can be adapted more broadly in the analyses of genomics and other relevant types of data that are naturally continuous. The approach based on quantifying the degree of concordance may allow more powerful inferential methods to replace those based upon overly simplified matching criteria.

LITERATURE CITED

Abe, J., D. H. Xu, Y. Suzuki, A. Kanazawa and Y. Shimamato, 2003 Soybean germplasm pools in Asia revealed by nuclear SSRs. Theor. Appl. Genet. **106:** 445–453.

Andersson, L., 2001 Genetic dissection of phenotypic diversity in farm animals. Nat. Rev. Genet. **2:** 130–138.

Baek, H. J., A. Beharav and E. Nevo, 2003 Ecological-genomic diversity of microsatellites in wild barley, *Hordeum spontaneum*, populations in Jordan. Theor. Appl. Genet. **106:** 397–410.

Barton, N. H., and P. D. Keightley, 2002 Understanding quantitative genetic variation. Nat. Rev. Genet. **3:** 11–21.

Berry, D. A., 1991 Inferences using DNA profiling in forensic identification and paternity cases (with discussion). Stat. Sci. **6:** 175–205.

Berry, D. A., I. W. Evett and R. Pinchin, 1992 Statistical inferences in crime investigations using deoxyribonucleic acid profiling. J. R. Stat. Soc. Ser. C Appl. Stat. **41:** 499–531.

Berry, D. A., J. D. Seltzer, C. Xie, D. L. Wright and J. S. C. Smith, 2002 Assessing probability of ancestry using simple sequence repeat profiles: applications to maize hybrids and inbreds. Genetics **161:** 813–824.

Berry, D. A., J. D. Seltzer, C. Xie, D. L. Wright, E. S. Jones *et al.*, 2003 Assessing probability of ancestry using simple sequence repeat profiles: applications to maize inbred lines and soybean varieties. Genetics **165:** 331–342.

Cardon, L. R., and J. I. Bell, 2001 Association studies for complex diseases. Nat. Rev. Genet. **2:** 91–99.

Cavalli-Sforza, L. L., and M. W. Feldman, 2003 The application of molecular genetic approaches to the study of human evolution. Nat. Genet. **33:** 266–275.

Chaix, G., S. Gerber, V. Razafimaharao, P. Vigneron, D. Verhaegen *et al.*, 2003 Gene flow estimation with microsatellites in a Malagasy seed orchard of *Eucalyptus grandis*. Theor. Appl. Genet. **107:** 705–712.

Dekkers, J. C. M., and F. Hospital, 2002 The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. **3:** 22–32.

Glazier, A. M., J. H. Nadeau and T. J. Aitman, 2002 Finding genes that underlie complex traits. Science **298:** 2345–2349.

Grassi, F., M. Labra, S. Imazio, A. Spada, S. Sgorbati *et al.*, 2003 Evidence of a secondary grapevine domestication centre detected by SSR analysis. Theor. Appl. Genet. **107:** 1315–1320.

Little, R. J. A., and D. B. Rubin, 1987 *Statistical Analysis With Missing Data*. John Wiley & Sons, New York.

Narvel, J. M., W.-C. Chu, W. R. Fehr, P. B. Cregan and R. C. Shoemaker, 2000 Development of multiplex sets of simple sequence repeat DNA markers covering the soybean genome. Mol. Breed. **6:** 175–183.

Sjakste, T. G., I. Rashal and M. S. Roder, 2003 Inheritance of microsatellite alleles in pedigrees of Latvian barley varieties and related European ancestors. Theor. Appl. Genet. **106:** 539–549.

Tang, S., and S. J. Knapp, 2003 Microsatellites uncover extraordinary diversity in native American land races and wild populations of cultivated sunflower. Theor. Appl. Genet. **106:** 990–1003.

Tang, S., V. F. Kishore and S. J. Knapp, 2003 PCR-multiplexes for a genome-wide framework of simple sequence repeat marker loci for cultivated sunflower. Theor. Appl. Genet. **107:** 6–19.

Thompson, E. A., 1994 Monte Carlo likelihood in genetic mapping. Stat. Sci. **9:** 355–366.

Thompson, E. A., and T. R. Meagher, 1987 Parental and sib likelihoods in genealogy reconstruction. Biometrics **43:** 585–600.

Tommasini, L., J. Batley, G. M. Arnold, R. J. Cooke, P. Donini *et al.*, 2003 The development of multiplex simple sequence repeat (SSR) markers to complement distinctness, uniformity and stability testing of rape (*Brassica napus* L.) varieties. Theor. Appl. Genet. **106:** 1091–1101.

Vouillamoz, J., D. Maigre and C. P. Meredith, 2003 Microsatellite analysis of ancient alpine grape cultivars: pedigree reconstruction of *Vitis vinifera* L. 'Cornalin du Valais'. Theor. Appl. Genet. **107:** 448–454.

Yu, S. B., W. J. Xu, C. H. Vijayakumar, J. Ali, B. Y. Fu *et al.*, 2003 Molecular diversity and multilocus organization of the parental lines used in the International Rice Molecular Breeding Program. Theor. Appl. Genet. **108:** 131–140.

Communicating editor: R. W. Doerge