JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 12, Number 9, 2005

Pp. 1183–1201

Algorithms for Selecting Informative Marker Panels for Population Assignment

NOAH A. ROSENBERG

ABSTRACT

Given a set of potential source populations, genotypes of an individual of unknown origin at a collection of markers can be used to predict the correct source population of the individual. For improved efficiency, informative markers can be chosen from a larger set of markers to maximize the accuracy of this prediction. However, selecting the loci that are individually most informative does not necessarily produce the optimal panel. Here, using genotypes from eight species—carp, cat, chicken, dog, fly, grayling, human, and maize—this univariate accumulation procedure is compared to new multivariate "greedy" and "maximin" algorithms for choosing marker panels. The procedures generally suggest similar panels, although the greedy method often recommends inclusion of loci that are not chosen by the other algorithms. In seven of the eight species, when applied to five or more markers, all methods achieve at least 94% assignment accuracy on simulated individuals, with one species—dog—producing this level of accuracy with only three markers, and the eighth species—human—requiring ∼13–16 markers. The new algorithms produce substantial improvements over use of randomly selected markers; where differences among the methods are noticeable, the greedy algorithm leads to slightly higher probabilities of correct assignment. Although none of the approaches necessarily chooses the panel with optimal performance, the algorithms all likely select panels with performance near enough to the maximum that they all are suitable for practical use.

Key words: ancestry inference, informativeness, microsatellites, population structure.

1. INTRODUCTION

SITUATIONS OFTEN ARISE IN WHICH THE SOURCE POPULATION or populations for genetic material from individuals of unknown origin must be determined (Anderson and Thompson, 2002; Davies *et al.*, 1999; Guinand *et al.*, 2002; Hansen *et al.*, 2001; Lowe *et al.*, 2001; Manel *et al.*, 2005; Waser and Strobeck, 1998; Ziv and Burchard, 2003). In a typical scenario, allele frequencies at a set of loci are given for several predefined groups, and using their genotypes at these loci, unknown individuals are each assigned to a single source population (Banks and Eichert, 2000; Baudouin *et al.*, 2004; Buchanan *et al.*, 1994; Paetkau *et al.*, 1995; Primmer *et al.*, 2000; Pritchard *et al.*, 2000; Rosenberg *et al.*, 2003).

Department of Human Genetics, Bioinformatics Program, and the Life Sciences Institute, University of Michigan, Ann Arbor, MI.

In an increasing number of species, the number of markers for which allele frequencies are available exceeds that required for accurate assignments. Thus, to perform assignment procedures efficiently, the panel of loci for genotyping of unknowns can be chosen from a larger collection of markers to contain as much information about ancestry as possible. Two of the questions that arise in the selection of an efficient panel are:

1. Given a collection of L loci and a desired number of markers $M < L$ to genotype, which markers should constitute a panel of size M?
2. How should the number of markers to genotype, M, be determined?

To answer question 2, for each number of markers from 1 to L, a measure of the "performance" of marker panels (either random panels or those selected using answers to question 1 can be evaluated, and M can be chosen as the smallest number for which the performance exceeds a specified threshold (Bamshad *et al.*, 2003; Banks *et al.*, 2003; Bernatchez and Duchesne, 2000; Campbell *et al.*, 2003; Cornuet *et al.*, 1999; Edwards, 2003; Manel *et al.*, 2002; Risch *et al.*, 2002; Rosenberg *et al.*, 2001, 2003; Turakulov and Easteal, 2003). In this analysis, one of several possible procedures for measuring performance can be used.

Question 1 poses greater difficulties. A simple answer suggests evaluation of an information-content statistic for each marker, followed by assembly of a panel consisting of the M most "informative" markers, or of any M markers individually more informative than a specified threshold (Collins-Schramm *et al.*, 2002; Dean *et al.*, 1994; Manel *et al.*, 2002; Rosenberg *et al.*, 2001, 2003; Shriver *et al.*, 1997). These approaches produce higher performance than use of random markers (Rosenberg *et al.*, 2001, 2003). However, they need not lead to the set with maximal performance (Pfaff *et al.*, 2004; Rosenberg *et al.*, 2003): in Fig. 1, Locus 1 is most informative, but the most informative *pair* of loci is {Locus 2, Locus 3}. In fact, in Fig. 1, the two loci that are most informative individually comprise the *least* informative pair.

The explanation for why selecting markers that are most informative individually need not lead to an optimal panel lies in the fact that the ability of a marker to assign an individual correctly depends on the source population of the individual. A panel of markers that are generally useful for all source populations may be less efficient than a panel of markers that are generally poor but in which for each m, the mth marker is extremely informative for the mth source population. Consider $K \geq 3$ populations and a set X of K loci, numbered 1 through K, in which each locus has K alleles. For locus $m \in X$, the mth allele has

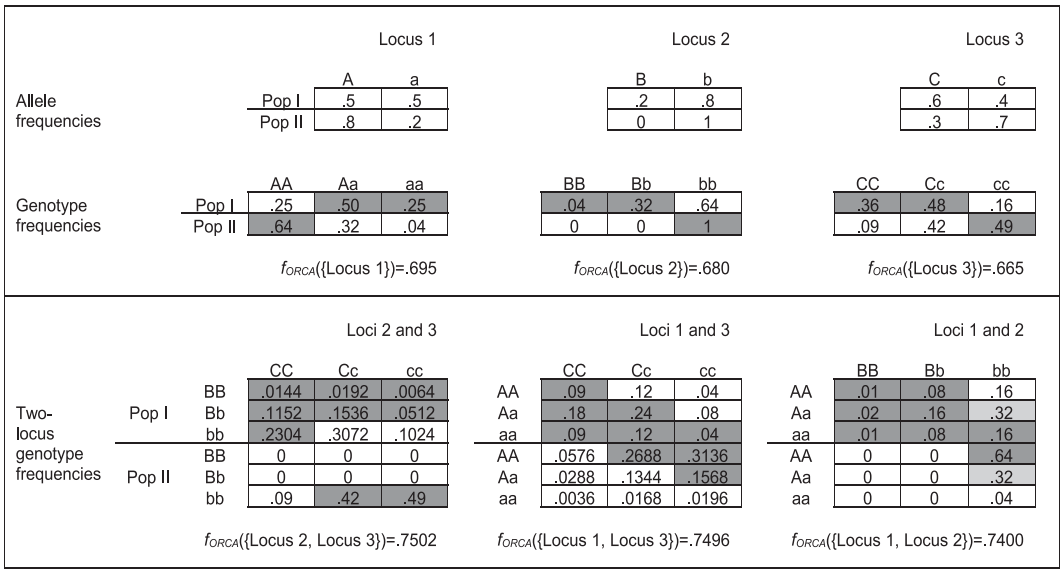

FIG. 1. A set of three loci—with each locus statistically independent of the others in each of two populations—for which the pair most informative about ancestry does not consist of the two most informative loci; f_{ORCA} (Section 2.2) gives the probability that a multilocus genotype is assigned correctly. The frequency of each genotype in the population in which the genotype is most frequent is shaded (lightly, in case of a tie). For the set containing all three loci, it can be shown that $f_{ORCA}(\{\text{Locus 1, Locus 2, Locus 3}\}) = 0.7928$.

frequency 1 in the mth population and frequency 0 in all other populations; all other alleles of locus m have frequency $1/(K-1)$ in all except the mth population. Using the f_{ORCA} measure of performance (Section 2.2), which gives the probability of correct assignment if multilocus genotypes are assigned to the population from which they are most likely to have originated, for one of these loci, the probability that an individual is assigned to the correct source population is $2/K$. However, using all K loci, this probability is 1, because every possible multilocus genotype for the K loci is observed in only one population.

Now consider a second set Y of K loci, each of which also has K alleles. For each locus in Y, the mth allele has frequency $1-(K-1)\varepsilon$ in the mth population and frequency ε in all other populations, where $0 < \varepsilon < (K-2)/[K(K-1)]$. Using f_{ORCA}, the probability of correct assignment for such a locus is $1-(K-1)\varepsilon$, which is larger than $2/K$. Thus, any locus in Y gives a higher correct assignment probability than any locus in X. However, because every genotype is found in every population, no multilocus genotype at the loci in Y can be assigned with certainty to a particular source population. Consequently, the probability of correct assignment for the set Y is less than 1, and the panel of generally informative markers is less useful than the panel of markers that are each informative for only one source population.

To account for the fact that the performance of a set of markers need not be expressible solely in terms of performances of individual markers, I consider multivariate algorithms for selecting efficient panels of size M from among L loci. In analytical and simulation-based forms, these algorithms, as well as sequential accumulation of individually informative markers, are applied to data from various species. Using simulated individuals, the performances of the algorithms in population assignment are then compared.

2. ALGORITHMS FOR SELECTING MARKER PANELS

2.1. Algorithms based on a given "performance function"

For a finite set S_L containing L loci, denote the set of all its subsets by $\mathcal{P}(S_L)$. Let ϕ denote the empty set.

Definition. *Consider a function $f : \mathcal{P}(S_L) \to \mathbb{R}$. Function f is a* performance function *for S_L if $f(T_1) \le f(T_2)$ for any finite $T_1, T_2 \subset S_L$ with $T_1 \subset T_2$.*

Informally, a performance function measures "performance" of a collection of markers in population assignment; higher values indicate better performance, so that subsets of a set of markers have equal or poorer performance than the set itself. Specific examples are discussed in Section 2.2; we will see later that if f is the function that measures the probability of correctly assigning individuals when each multilocus genotype for a set of loci is assigned to the population in which it is most likely to occur, then f is a performance function.

Question 1 from Section 1 can now be rephrased: given a set S_L of L loci, a performance function f and a positive integer $M < L$, identify the subset of S_L that maximizes f over all subsets of S_L with cardinality M. Several methods can be used to choose a set S_M to serve as a candidate for this optimal set.

In each of the following approaches, ties for the choice of set are broken randomly. The first method is to evaluate f for all possible candidate sets.

Method 1 (exhaustive evaluation). *For $M \ge 1$, define*

$$S_M = \arg \max_{\substack{T \in \mathcal{P}(S_L) \\ |T| = M}} f(T).$$

Two computational difficulties arise in application of Method 1. First, for sufficiently large M, evaluation of $f(T)$ may be impractical when $|T| = M$. Second, for sufficiently large L and M, even if it were possible to evaluate $f(T)$ when $|T| = M$, the number ${}_LC_M$ of subsets of S_L with cardinality M is very large and the subsets cannot be exhaustively tested.

If L or M is large enough that Method 1 is not feasible, an algorithm that is less computationally intensive but that produces only an approximately optimal set can be used. The three such algorithms that

follow each sequentially accumulate loci to marker panels. Thus, Methods 2–4 are all greedy algorithms, in that each constructs S_M from S_{M-1} together with the "best" remaining locus by some criterion. For convenience, however, Methods 2, 3, and 4 are labeled "univariate," "greedy," and "maximin," respectively.

The simplest of the three computationally feasible algorithms is the procedure discussed in Section 1, which proposes evaluation of loci individually, and which defines S_M as the set containing the M loci that have the highest individual values.

Method 2 (univariate accumulation). *Define* $S_0 = \phi$, *and for* $M \geq 1$, *define*

$$S_M = S_{M-1} \cup \left\{ \arg \max_{v \in S_L \setminus S_{M-1}} f(\{v\}) \right\}.$$

This algorithm is convenient, but as discussed in Section 1, if f depends on interactions among contributions of individual markers, the procedure might fail to choose the set with maximal performance.

To incorporate multivariate dependence of f while reducing the computational burden of Method 1, a procedure can be used that chooses the next marker in the panel conditional on the information obtained from those markers that have already been included.

Method 3 (greedy accumulation). *Define* $S_0 = \phi$, *and for* $M \geq 1$, *define*

$$S_M = S_{M-1} \cup \left\{ \arg \max_{v \in S_L \setminus S_{M-1}} f(\{v\} \cup S_{M-1}) \right\}.$$

To choose the Mth marker, this algorithm evaluates each of the remaining markers together with the $M-1$ markers that have already been chosen and selects the marker that gives the highest value of f. Method 3 is more computationally feasible than Method 1 in that for each M, only $L - M + 1$ rather than ${}_LC_M$ sets must be tested. However, like Method 1, Method 3 is not practical if M is sufficiently large. This procedure is also not guaranteed to locate the set with maximal performance (Fig. 1).

The final algorithm takes into account multivariate dependence and has greater computational feasibility than Methods 1 and 3, but also does not necessarily maximize performance.

Method 4 (maximin accumulation). *Choose* $r \geq 2$ *small enough that for* $M \leq r$, S_M *is obtained by Method 1. For* $M > r$, *define*

$$S_M = S_{M-1} \cup \left\{ \arg \max_{v \in S_L \setminus S_{M-1}} \left[\min_{\substack{T \in \mathcal{P}(S_{M-1}) \\ |T| = r-1}} f(\{v\} \cup T) \right] \right\}$$

Note that this algorithm has two parts: for small M, exhaustive evaluation is performed. For larger M, the method accumulates loci that contribute new information, conditional on the information from markers that have already been selected: it chooses the Mth marker from among the remaining markers as the one with the maximal value of the minimum f, where the minimum is taken across all sets in which the other $M - 1$ markers are among those that have already been selected.

Other "hybrid" algorithms are possible. In increasing order of ability to locate the set with maximal performance, but also in increasing order of difficulty of computation, the methods are ordered 2, 4, 3, 1. Thus, for a given set of loci, as M is increased, Method 1 can be used until ${}_LC_M$ becomes too large for exhaustive evaluation of all subsets of cardinality M. Method 3 can then be used to add new loci to the existing set until M becomes too large for evaluation of *any* sets with cardinality M. Method 4 can then be used until ${}_LC_r$ is too large for evaluation of all subsets of cardinality r, reducing r to 2 as the computational burden increases. Finally, if no other options are available, Method 2 is likely to be feasible in any realistic scenario. The specific choice of the performance function f affects the values of L and M at which the various algorithms become impractical.

2.2. Performance functions

To evaluate the potential of a set of loci to provide information about ancestry, I consider an analytical approach and a closely related simulation approach. The simulation procedure and modifications of it are frequently used to assess performance of sets of loci (Banks and Eichert, 2000; Banks *et al.*, 2003; Buchanan *et al.*, 1994; Campbell *et al.*, 2003; Paetkau *et al.*, 1995, 2004; Waser and Strobeck, 1998); the analytical approach uses the formula that underlies the simulation procedure (Rosenberg *et al.*, 2003).

Consider a set S_M containing loci $m = 1, 2, \ldots, M$, with locus m having alleles $j = 1, 2, \ldots, N^{(m)}$. Consider populations $i = 1, 2, \ldots, K$, with the relative frequency of allele j of locus m in population i equaling $p_{ij}^{(m)}$. Suppose that at each locus, in each population, the two alleles of a diploid individual are independent: that is, for each i, j, h, and m, an individual in population i has genotype jh at locus m with probability $(2 - \delta_{jh}) p_{ij}^{(m)} p_{ih}^{(m)}$, where jh is the same genotype as hj and δ_{jh} is 1 if $j = h$ and 0 otherwise. Suppose also that within each population, genotypes are independent across loci, and that for each i, individuals of unknown origin have prior probability $1/K$ of having derived from population i. If we consider decision rules where each possible multilocus diploid genotype has a specified probability of being assigned to each of the potential source populations, the rule that produces the *optimal rate of correct assignment* (*ORCA*) simply assigns an individual to the population from which its genotype is most likely to have originated (Rosenberg *et al.*, 2003). The probability that an individual is assigned to its correct population of origin is

$$f_{ORCA}(S_M) = \sum_{j_1^{(1)}=1}^{N^{(1)}} \sum_{j_2^{(1)}=j_1^{(1)}}^{N^{(1)}} \sum_{j_1^{(2)}=1}^{N^{(2)}} \sum_{j_2^{(2)}=j_1^{(2)}}^{N^{(2)}} \cdots \sum_{j_1^{(M)}=1}^{N^{(M)}} \sum_{j_2^{(M)}=j_1^{(M)}}^{N^{(M)}} \max_{i \in \{1,2,\ldots,K\}} \left[\frac{1}{K} \prod_{m=1}^{M} (2 - \delta_{j_1^{(m)} j_2^{(m)}}) p_{ij_1^{(m)}}^{(m)} p_{ij_2^{(m)}}^{(m)} \right]. \quad (1)$$

For the empty set, $f_{ORCA}(\phi) = 1/K$. It can be shown that f_{ORCA} is indeed a performance function (Theorem 2 in the appendix).

Conveniently, because of its relationship to assignment by most likely source population, $f_{ORCA}(S_M)$ can be approximated using the following simulation.

1. From a uniform prior on $\{1, 2, \ldots, K\}$, simulate the source population, q, of an individual.
2. Independently for each locus $m \in S_M$, simulate two independent alleles, $j_1^{(m)}$ and $j_2^{(m)}$, from the allele frequency distribution of population q.
3. Compute

$$\gamma = \arg \max_{i \in \{1,2,\ldots,K\}} \left[\frac{1}{K} \prod_{m=1}^{M} (2 - \delta_{j_1^{(m)} j_2^{(m)}}) p_{ij_1^{(m)}}^{(m)} p_{ij_2^{(m)}}^{(m)} \right].$$

   In case of a tie in the value of the product for two or more values of i, randomly assign one of these i to equal γ. If $\gamma = q$, the individual is assigned correctly.
4. Repeat steps 1–3 many times, computing the fraction of simulated individuals that are correctly assigned. The result is $\tilde{f}_{ORCA}(S_M)$.

For the empty set, $\tilde{f}_{ORCA}(\phi) = 1/K$. An advantage of evaluating the less precise $\tilde{f}_{ORCA}$ rather than f_{ORCA} is that the simulation can be performed quickly for large values of M, whereas if the number of terms summed in Equation (1), or $\prod_{m=1}^{M} (N^{(m)} + 1) N^{(m)}/2$, is large, then Equation (1) cannot realistically be evaluated. In a strict sense, $\tilde{f}_{ORCA}$ is only approximately a performance function (Corollary 6 in the appendix), as stochasticity makes it possible for a set of loci to have a lower value of $\tilde{f}_{ORCA}$ than one of its proper subsets; however, because of its close relationship to f_{ORCA}, $\tilde{f}_{ORCA}$ is treated here as a performance function.

3. DATA

Methods 2–4 and f_{ORCA} and $\tilde{f}_{ORCA}$ are applied to selection of marker panels using data from eight species (Table 1). The datasets each consist of unphased individual multilocus diploid genotypes for autosomal microsatellite loci (Goldstein and Schlötterer, 1999) spread throughout the genomes of their respective species. They span a wide range in number of markers and populations, as well as in levels of genetic diversity within populations and of genetic divergence across populations.

4. IMPLEMENTATION

4.1. Computation of f_{ORCA} and $\tilde{f}_{ORCA}$

Allele frequencies at a locus were estimated from the data using the ratios of the numbers of observed copies of alleles to the total number of observations for the locus. In each dataset, for each locus and population, individuals were assumed to have two independent alleles. This assumption of Hardy–Weinberg proportions holds for most locus-population pairs, although the fraction of pairs at which it is violated is large in some populations (Irion *et al.* [2003] for example). A substitute for this assumption is replacement of the product of allele frequencies, $(2-\delta_{j_1^{(m)}j_2^{(m)}})p^{(m)}_{ij_1^{(m)}}p^{(m)}_{ij_2^{(m)}}$, in Equation 1 and in Step 3 of the simulation procedure, with the genotype frequency $p^{(m)}_{i(j_1^{(m)}j_2^{(m)})}$, and simulation from the genotype frequency distribution in Step 2 rather than from the allele frequency distribution. However, the large number of possible genotypes compared to typical per-population sample sizes makes it more difficult to obtain accurate estimates of genotype frequencies than of allele frequencies. When sample sizes are too small for this approach to be feasible, genotype frequencies estimated from allele frequencies and a single parameter measuring the deviation from Hardy–Weinberg proportions—the inbreeding coefficient (Ayres and Balding, 1998)—could potentially be used. For simplicity, however, Hardy–Weinberg proportions were assumed here. Additionally, because markers were generally widely spaced across the genomes of the various species, in each population, genotypes at different loci were assumed to be independent.

Because the allele frequencies were estimated from samples that were in general small compared to the numbers of alleles at loci, similarly to previous implementations (Banks and Eichert, 2000; Campbell *et al.*, 2003; Paetkau *et al.*, 2004; Waser and Strobeck, 1998), a slight alteration was made to the computation of $\tilde{f}_{ORCA}$: $1/(Z+1)$ was substituted in place of allele frequencies of 0 in Step 3, where Z is the largest number of alleles genotyped at any locus in any population. This substitution reflects the fact that even if its sample frequency is 0, an allele may be present in a population at nonzero frequency. However, because the simulations were performed assuming that the sample frequencies equal the true allele frequencies (Step 2), the use of $1/(Z+1)$ in place of a true frequency of 0 systematically decreases $\tilde{f}_{ORCA}$ compared to f_{ORCA}; note that this change has little effect if most alleles are found in most populations, so that allele frequencies of 0 are rare. The corresponding substitution of 0 with $1/(Z+1)$ was not made in computation of f_{ORCA}, as this substitution can only *increase* the value of f_{ORCA} and therefore is anticonservative.

4.2. Selection of marker panels

Because of the sizeable number of alleles at the microsatellite loci in the data, for sets of approximately four or more loci, the number of possible multilocus genotypes was quite large and evaluation of f_{ORCA} for $M \geq 4$ proceeded very slowly. Thus, in Method 4, $r = 2$ was chosen, as the use of even $r = 3$ was impractical for the datasets with the largest numbers of loci and alleles. In each of the datasets, the number of loci was sufficiently small that f_{ORCA} and $\tilde{f}_{ORCA}$ could be evaluated relatively rapidly for all ${}_LC_2$ pairs of loci.

Of the eight possible combinations of methods (1, 2, 3, and 4) and functions (f_{ORCA} and $\tilde{f}_{ORCA}$), five were practical to implement on the datasets for all possible values of M and L: panels were obtained using both f_{ORCA} and $\tilde{f}_{ORCA}$ with Methods 2 and 4 and using $\tilde{f}_{ORCA}$ with Method 3. In case two or more loci tied in their values according to any criterion, one of these loci was selected randomly to be the next locus accumulated to the chosen set.

TABLE 1. DATASETS[a]

Dataset	*Carp*	*Cat*	*Chicken*	*Dog*	*Fly*	*Grayling*	*Human*	*Maize*
Source	David *et al.* (2005)	Beaumont *et al.* (2001)	Rosenberg *et al.* (2001)	Irion *et al.* (2003)	Kauer *et al.* (2003)	Koskinen *et al.* (2002)	Rosenberg *et al.* (2002)	Matsuoka *et al.* (2002)
Number of individuals	45	276	600	1338	94	623	1056	193
Number of populations	9	2	20	28	4	18	7	9
Number of loci	11[b]	9	27	100	77[c]	14[b]	377	99
Percent missing data	3.2	3.0	0.8	14.0	23.6[d]	0.8	3.8	4.5
Gene diversity[e] (mean across populations)	0.549	0.739	0.473	0.619	0.556	0.436	0.739	0.778
Gene diversity[e] (mean across loci)	0.779	0.761	0.677	0.793	0.554	0.707	0.771	0.830
Mean number of alleles per locus	10.2	12.4	12.1	16.0	7.4	17.0	12.4	27.3
Mean percent of all alleles present in a given population	30.0	86.2	30.6	38.1	61.2	22.6	69.1	36.9
Percent of alleles found only in one population (private alleles)	42.0	27.8	31.0	20.6	27.1	40.3	17.0	32.0
Mean frequency of private alleles in their population of occurrence	0.275	0.027	0.066	0.047	0.040	0.126	0.010	0.034
Largest number of alleles sampled at any one locus in any one population (Z)	10	398	60	112	60	96	482	98
Genetic divergence across populations[f]	0.315	0.065	0.313	0.228	0.038	0.398	0.059	0.068

[a]The population groups in the various species are as follows: carp—9 wild, edible, and ornamental breeds; cat—2 groups (Scottish wildcats, house cats); chicken—20 breeds, mostly from Europe; dog—28 domestic breeds from seven breed categories; fly—4 populations, from Harare (Zimbabwe), Katowice (Poland), Naples and Rome (Italy), and Sengwa (Zimbabwe); grayling—18 European populations, mostly from Scandinavia; human—7 regional groups (Africa, Europe, Middle East, Central/South Asia, East Asia, Oceania, America); maize—9 regional groups (Andes, South America excluding Andes, Guatemala and Southern Mexico, Caribbean, Lowland Western and Northern Mexico, Highland Mexico, Eastern and Central USA, Southwestern USA, Northern Mexico).

[b]Locus *13-14* in the David *et al.* (2005) data and loci *BFRO4*, *BFRO7*, and *Cocl23* in the Koskinen *et al.* (2002) data were omitted because no genotypes were available in one or more of the populations.

[c]74 X-chromosomal loci from the Kauer *et al.* (2003) data were omitted.

[d]The large amount of missing data results from recoding one of the two alleles in homozygous inbred individuals as missing (Kauer *et al.*, 2003).

[e]An unbiased estimator of gene diversity, or expected heterozygosity, was used (Nei, 1987, Equation 8.31).

[f]Divergence was measured with the F_{st} statistic (Weir, 1996, Equation 5.3), using the GDA program (Lewis and Zaykin, 2001); larger values (in [0, 1]) indicate greater divergence.

TABLE 2. ROBUSTNESS OF SETS OF SELECTED MARKERS THAT WERE OBTAINED USING SIMULATIONS WITH $\tilde{f}_{ORCA}$[a]

	Kendall coefficient of concordance of rankings			*Mean across loci of standard deviation of rank across replicates*		
Species	*Method 2 (univariate)*	*Method 3 (greedy)*	*Method 4 (maximin)*	*Method 2 (univariate)*	*Method 3 (greedy)*	*Method 4 (maximin)*
Carp	0.977	0.740	0.972	0.39	1.46	0.49
Cat	0.923	0.808	0.910	0.66	1.18	0.70
Chicken	0.984	0.445	0.961	0.90	5.71	1.38
Dog	0.966	0.183	0.981	5.12	26.74	3.85
Fly	0.979	0.208	0.978	3.09	20.29	3.17
Grayling	0.976	0.622	0.973	0.55	2.31	0.63
Human	0.940	0.125	0.980	26.09	105.63	15.21
Maize	0.983	0.177	0.988	3.55	26.85	2.98

[a]For Method 4, if the marker ranked 1 was not included in the top-ranked pair, the two loci in the top-ranked pair were assigned rank 1.5 and the top-ranked marker was assigned rank based on the later stage at which it re-entered the list. This scenario was generally unusual, occurring for none of the 10 replicates in chicken, grayling, and maize, 1 of 10 in carp, cat, and dog, and 3 of 10 in fly, but 8 of 10 in human.

In application of $\tilde{f}_{ORCA}$, it is necessary to simulate enough individuals that robust rankings are obtained. Thus, for each dataset, 10 replicates were performed for each of Methods 2, 3, and 4, using 1,000 individuals to evaluate $\tilde{f}_{ORCA}$ for each proposed panel. For each replicate, each locus was associated with a number in $\{1, 2, \ldots, L\}$, indicating the step at which the locus was accumulated to the set of selected markers (for example, in Fig. 1, using Method 3, Locus 1 is added at the first step, Locus 3 at the second step, and Locus 2 at the third step). The Kendall coefficient of concordance (Gibbons, 1985, p. 250) of the ten marker "rankings" obtained in this manner was then computed. Also, the mean across loci of the standard deviation of locus "ranks" across replicates was calculated. Except for those based on simulations with the greedy algorithm, rankings in independent replicates were highly concordant, and loci varied little in rank across replicates (Table 2). For the greedy algorithm, after enough markers for nearly perfect assignment have been accumulated, additional markers are selected essentially randomly, because none of the markers contribute to an increase in performance. Thus, less concordance of marker sets is to be expected if the number of markers is sufficient for highly accurate assignment. Even though the marker panels in replicate simulations differed in composition, however, these panels had very similar performance. For all datasets and each possible number of markers M, the values of $\tilde{f}_{ORCA}$ for the 10 panels suggested by the greedy algorithm were nearly always within 0.04 of each other; only occasionally was the range of the 10 values larger than 0.01.

Thus, use of 1,000 simulated individuals to compute $\tilde{f}_{ORCA}$ was assumed to be sufficient for selection of marker panels; to be conservative, in all $\tilde{f}_{ORCA}$ computations other than those that underlie Table 2 (and those applied to random sets of markers—Section 4.3), 10,000 individuals were simulated. In larger datasets for which the simulation time with this number of individuals is prohibitive, fewer individuals could potentially be used, with a consequent decline in robustness of the rankings obtained.

In addition to the variability that results from the stochasticity of simulation, sampling provides a separate source of variability for rankings. However, in a previous analysis (Rosenberg *et al.*, 2003), using a performance function similar to f_{ORCA} with datasets of comparable complexity to those in Table 1, values of the performance function and the associated rankings based on Method 2 showed little variation across datasets in which bootstrap resamples of individuals were taken. Thus, although it might be nontrivial for the smallest of the datasets in Table 1, the impact of sampling variation on marker rankings was not investigated here.

4.3. Evaluation of performance

After marker panels were chosen using each of the five approaches, simulations were used to evaluate $\tilde{f}_{ORCA}$ on the panels selected with Methods 2 and 4. For each set of loci, 10,000 individuals were simulated. For panels obtained with Method 3, this evaluation of performance was based on the same simulations used to select the markers.

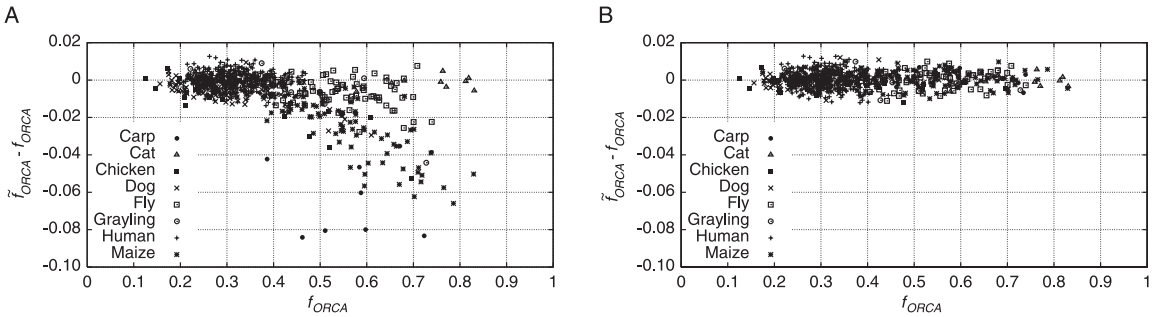

FIG. 2. Relationship of the difference between simulated and analytically obtained probabilities of correct assignment to the analytical probability. (**A**) Simulations performed with replacement of allele frequencies of 0 by $1/(Z+1)$. (**B**) Simulations performed retaining allele frequencies of 0. Both graphs were generated using the same simulated individuals. In (A), locus 3-4 in carp lies below the graph at (0.734, −0.151).

The performances of the five approaches were compared to those of random sets of markers. Each marker was associated with a random number. For each number of loci M, $\tilde{f}_{ORCA}$ was evaluated (using 1,000 simulated individuals) for the set containing the M markers with the M highest random numbers. This procedure was repeated for 100 random orderings of the markers.

5. RESULTS

As described in Section 2.2, $\tilde{f}_{ORCA}$ in principle estimates by simulation the same quantity computed analytically by f_{ORCA}. Similar values of $\tilde{f}_{ORCA}$ and f_{ORCA} for individual markers were observed in all data sets, with $\tilde{f}_{ORCA} < f_{ORCA}$ more often than $f_{ORCA} < \tilde{f}_{ORCA}$ (Fig. 2A). The generally smaller values of $\tilde{f}_{ORCA}$ compared to f_{ORCA} result from the substitution of $1/(Z+1)$ for 0 in Step 3 of the computation of $\tilde{f}_{ORCA}$. This interpretation is supported by the fact that when the substitution is not made, the simulated and analytically obtained values are nearly equal (Fig. 2B). Note that for cat, in which most alleles have nonzero frequencies in both populations (Table 1), the substitution has little impact on the simulation.

Although $\tilde{f}_{ORCA}$ was sometimes ∼10% smaller than f_{ORCA} (Fig. 2A), when the same algorithm was applied to selection of panels—Method 2 or 4—locus ranks when $\tilde{f}_{ORCA}$ was used as the performance function were nearly identical to those obtained with f_{ORCA} (Tables 3 and 4). With the same algorithm applied, correlation coefficients of marker rankings based on the analytically computed f_{ORCA} and the simulated $\tilde{f}_{ORCA}$ were in most datasets larger than 0.99 (Table 3).

TABLE 3. SPEARMAN COEFFICIENTS OF RANK CORRELATION BETWEEN PAIRS OF RANKINGS OF MARKERS

Pair of rankings		*Spearman coefficient*							
(Method, function)	*(Method, function)*	*Carp*	*Cat*	*Chicken*	*Dog*	*Fly*	*Grayling*	*Human*	*Maize*
(2, f_{ORCA})	(2, $\tilde{f}_{ORCA}$)	0.936	0.983	0.995	0.997	0.997	0.996	0.996	0.994
(4, f_{ORCA})	(4, $\tilde{f}_{ORCA}$)	0.945	0.983	0.996	0.993	0.994	1.000	0.996	0.993
(2, f_{ORCA})	(3, $\tilde{f}_{ORCA}$)	0.864	0.933	0.488	0.161	0.410	0.780	0.229	0.216
(2, f_{ORCA})	(4, f_{ORCA})	0.927	0.983	0.991	0.985	0.981	0.956	0.992	0.988
(2, f_{ORCA})	(4, $\tilde{f}_{ORCA}$)	0.982	1.000	0.987	0.984	0.979	0.956	0.991	0.985
(2, $\tilde{f}_{ORCA}$)	(3, $\tilde{f}_{ORCA}$)	0.809	0.900	0.511	0.154	0.402	0.793	0.227	0.211
(2, $\tilde{f}_{ORCA}$)	(4, f_{ORCA})	0.955	0.950	0.991	0.982	0.979	0.943	0.989	0.982
(2, $\tilde{f}_{ORCA}$)	(4, $\tilde{f}_{ORCA}$)	0.945	0.983	0.991	0.982	0.979	0.943	0.989	0.983
(3, $\tilde{f}_{ORCA}$)	(4, f_{ORCA})	0.818	0.917	0.510	0.183	0.454	0.833	0.228	0.230
(3, $\tilde{f}_{ORCA}$)	(4, $\tilde{f}_{ORCA}$)	0.827	0.933	0.515	0.183	0.448	0.833	0.233	0.236

TABLE 4. RANKS OF MARKERS SELECTED USING EACH OF FIVE PROCEDURES[a]

		Locus rank (carp)							*Locus rank (cat)*				
Locus	f	$(2, f)$	$(2, \tilde{f})$	$(3, \tilde{f})$	$(4, f)$	$(4, \tilde{f})$	*Locus*	f	$(2, f)$	$(2, \tilde{f})$	$(3, \tilde{f})$	$(4, f)$	$(4, \tilde{f})$
21-22	0.739	1	1	1	1	1	Fca96	0.831	1	1	1	1	1
3-4	0.734	2	4	2	3	3	Fca90	0.819	2	2	2	2	2
51-52	0.723	3	2	3	2	2	Fca23	0.812	3	3	3	3	3
85-86	0.671	4	3	5	5	4	Fca8	0.771	4	5	4	4	4
69-70	0.598	5	7	6	6	5	Fca45	0.763	5	4	6	6	5
111-112	0.588	6	6	9	7	6	Fca43	0.759	6	6	7	5	6
57-58	0.584	7	5	7	4	7	Fca77	0.680	7	7	5	7	7
29-30	0.511	8	8	4	8	8	Fca126	0.653	8	8	9	8	8
17-18	0.462	9	9	8	10	10	Fca35	0.578	9	9	8	9	9
55-56	0.387	10	10	11	9	9							
89-90	0.337	11	11	10	11	11							

		Locus rank (chicken)							*Locus rank (dog)*				
Locus	f	$(2, f)$	$(2, \tilde{f})$	$(3, \tilde{f})$	$(4, f)$	$(4, \tilde{f})$	*Locus*	f	$(2, f)$	$(2, \tilde{f})$	$(3, \tilde{f})$	$(4, f)$	$(4, \tilde{f})$
LEI228	0.696	1	1	1	1	1	FH2165	0.610	1	1	1	1	1
LEI234	0.609	2	2	2	2	3	FH2233	0.553	2	2	2	3	3
LEI192	0.521	3	4	6	4	5	FH2199	0.520	3	3	39[b]	4	5
LEI194	0.517	4	3	3	3	2	FH2200	0.518	4	4	3	2	2
LEI94	0.477	5	5	9	6	6	FH2202	0.465	5	5	72[b]	5	6
MCW34	0.437	6	6	5	5	4	CFMSAT	0.454	6	6	6	6	4
MCW206	0.424	7	7	4	7	7	FH2138	0.425	7	9	53[b]	12	15
MCW183	0.398	8	8	20	8	8	FH2289	0.425	8	8	77[b]	11	11
MCW295	0.327	9	9	7	11	11	FH2247	0.424	9	7	5	10	12
ADL268	0.313	10	10	15	10	10	PEZ18	0.416	10	10	8	8	8
MCW81	0.305	11	11	10	9	9	FH2313	0.401	11	11	9	9	10
MCW103	0.125	27	27	8	27	27	C05.771	0.391	12	12	4	7	7
							C10.404	0.362	18	18	23[b]	15	9
							PEZ03	0.323	27	30	7	24	24
							AHT136	0.209	84	84	10	79	83

Locus	f	Locus rank (fly) (2, f)	(2, $\tilde{f}$)	(3, $\tilde{f}$)	(4, f)	(4, $\tilde{f}$)
2189/2190	0.739	1	1	1	1	1
2117/2118	0.708	2	2	6	3	3
1908/1909	0.701	3	5	5	7	5
2185/2186	0.700	4	3	3	4	4
2137/2138	0.680	5	4	2	2	2
2151/2152	0.679	6	9	22	8	9
1761/1762	0.676	7	6	10	6	6
2127/2128	0.664	8	7	27	9	8
2141/2142	0.660	9	8	15	15	15
2115/2116	0.659	10	12	19	11	10
2147/2148	0.655	11	14	25	10	12
2139/2140	0.652	12	10	4	5	7
2109/2110	0.597	22	25	9	14	14
2145/2146	0.568	31	30	7	23	21
1932/1933	0.566	32	32	8	24	23

Locus	f	Locus rank (grayling) (2, f)	(2, $\tilde{f}$)	(3, $\tilde{f}$)	(4, f)	(4, $\tilde{f}$)
One2	0.728	1	1	1	1	1
BFRO13	0.594	2	2	2	2	2
BFRO18	0.509	3	3	4	3	3
BFRO12	0.500	4	4	8	6	6
Ogo2	0.500	5	5	5	5	5
BFRO10	0.453	6	6	7	8	8
BFRO11	0.438	7	7	3	4	4
BFRO15	0.428	8	9	13	7	7
BFRO5	0.424	9	8	10	10	10
MST73	0.374	10	10	6	9	9
MST85	0.242	13	13	9	13	13

Locus	f	Locus rank (human) (2, f)	(2, $\tilde{f}$)	(3, $\tilde{f}$)	(4, f)	(4, $\tilde{f}$)
D21S2055	0.468	1	1	15	6	6
D2S1356	0.467	2	3	2	1.5	1.5
D22S683	0.463	3	2	12	5	5
D1S1589	0.461	4	4	1	9	8
D9S1871	0.456	5	6	3	4	4
D8S560	0.452	6	5	5	1.5	1.5
D14S1007	0.444	7	10	4	3	3
D16S3401	0.441	8	8	116[b]	11	10
D7S2477	0.440	9	7	29	14	11
F13A1-D6S	0.437	10	11	9	7	9
D20S851	0.434	11	9	187[b]	10	16
D2S2986	0.423	15	15	6	17	14
D9S1779	0.409	19	23	7	8	7
D11S2000	0.403	22	32	10	33	19
D5S1501	0.402	24	28	8	40	24

Locus	f	Locus rank (maize) (2, f)	(2, $\tilde{f}$)	(3, $\tilde{f}$)	(4, f)	(4, $\tilde{f}$)
BNGL244	0.829	1	1	1	1	1
BNGL619	0.786	2	2	2	2	2
MC1046	0.765	3	3	26[b]	3	3
MC1191	0.738	4	4	5	4	5
MC1523	0.719	5	7	23[b]	7	7
MC1940	0.716	6	10	8	6	6
DUP28	0.711	7	9	53[b]	8	13
MC1662	0.710	8	11	62[b]	10	10
MC1890	0.702	9	13	7	5	11
MC1194	0.701	10	5	46[b]	9	9
MC1371	0.694	11	8	12	12	8
BNGL105	0.687	12	12	4	15	4
MC1288	0.681	13	6	3	11	12
MC1740	0.658	17	16	6	13	14
MC1329	0.509	50	53	9	54	52
MC1931	0.455	66	63	10	55	54

[a] Only those markers that appear among the 10 top-ranked for one or more of the five procedures are shown. For the maximin algorithms, the situation in which the top-ranked pair of loci did not include the top-ranked locus was treated in the same way as in Table 2. Symbols f and $\tilde{f}$ refer to f_{ORCA} and $\tilde{f}_{ORCA}$, respectively.

[b] In the non-human datasets, these loci were added in the greedy algorithm at a stage when the previous combination of loci already produced perfect assignment ($\tilde{f}_{ORCA} = 1$), so that the exact rank of these loci is less important than the fact that they were not among the first loci to be accumulated to the panel (for the human dataset with the greedy algorithm, performance with the first 72 markers exceeded 0.998, and to reduce computation time, the remaining 305 markers were randomly ordered with ranks between 73 and 377).

For pairs of rankings that used the same marker selection algorithm but different performance functions, correlation coefficients were generally larger than for pairs that used different marker selection algorithms and the same performance function (Table 3). However, Methods 2 and 4 produced highly correlated rankings and lists with similar composition, regardless of whether f_{ORCA} or $\tilde{f}_{ORCA}$ was used as the performance function (Tables 3 and 4). Note also that for chicken, a previous application of a univariate procedure based on a heterozygosity performance function (Rosenberg *et al.*, 2001) produced the same choice of the seven best-performing markers as Method 4 with f_{ORCA}.

Partly because of the fact that after enough markers for nearly perfect assignment have been selected, the greedy algorithm chooses new markers in an essentially random manner, lists of high-performing markers suggested by Method 3 were not very closely related to those obtained using the other algorithms (Tables 3 and 4). Especially for the larger datasets—dog, fly, human, and maize—the lists contained markers that were not included in panels suggested using the other algorithms. Simultaneously, many markers that were obtained using other algorithms did not appear among the lists suggested by the greedy method.

When $\tilde{f}_{ORCA}$ was evaluated for panels recommended by the selection algorithm/performance function combinations, performance was substantially higher than that of random panels (Fig. 3). Other than in the human dataset, in which performance differed across combinations for many choices of the number of loci, all five combinations had nearly identical performance for most numbers of loci. In the human data, as the number of loci ranged from 2 to 28, the greedy algorithm with $\tilde{f}_{ORCA}$ averaged 0.013 higher performance than the univariate algorithm with f_{ORCA}, and 0.015 higher than the univariate algorithm with $\tilde{f}_{ORCA}$. Over this range, the combinations involving the univariate algorithm were also slightly outperformed by those involving the maximin algorithm. When performance differences were noticeable in the other datasets—for example, in the situations when it exceeded 0.015 (carp with 2 loci, chicken with 2, 4, 5, and 6 loci, fly with 2, 3, and 4 loci, and grayling with 2 and 4 loci)—as was true in humans, the greedy algorithm with $\tilde{f}_{ORCA}$ generally outperformed the other approaches. For each algorithm, performance function, and dataset, performance appeared to converge as the number of loci increased.

6. DISCUSSION

Several combinations of marker selection algorithms and performance functions appropriate for choosing a panel for use in ancestry inference have been suggested. As a consequence of the fact that $\tilde{f}_{ORCA}$ has expected value equal to f_{ORCA} (Rosenberg *et al.*, 2003), the analytical and simulated performance functions produced nearly identical panels. The panels obtained by straightforward selection of the most informative individual markers, although this procedure does not take into account interactions among markers, had nearly identical composition and performance to those obtained by the maximin procedure, which in the cases studied, makes use of bivariate interactions. The greedy procedure, although its recommended markers differed from those of the other procedures, generally did not produce substantially different performance.

The similarity in performance of the various procedures suggests that although counterexamples do exist, performance of a set of markers can *almost* be decomposed into univariate contributions of individual loci, with only a small contribution of bivariate and higher-order interactions. The greedy method is perhaps appropriate when slightly higher performance is desired. However, when simplicity, robustness, and ease of computation are needed, performance changes little when the univariate or maximin procedure is used in its place. Although none of the three algorithms—univariate, greedy, or maximin—is guaranteed to identify the panel of maximal performance, each likely selects panels that have performance sufficiently close to the optimum that any of the algorithms is suitable for use with data.

The dataset in which the greedy procedure did produce a consistent (though slight) increase in performance—the human data—was both the one with the largest number of markers and the one in which the number of markers required for assignment was largest. These two aspects of the dataset are likely to be partly responsible for the improved performance of the greedy algorithm, as the careful selection of a marker panel has the greatest potential impact when the number of possible choices is particularly large and when the assignment problem is sufficiently difficult to allow different panels to vary substantially in their performance. Further investigation of the influence of various dataset characteristics on assignment success will help to determine the generality of this claim.

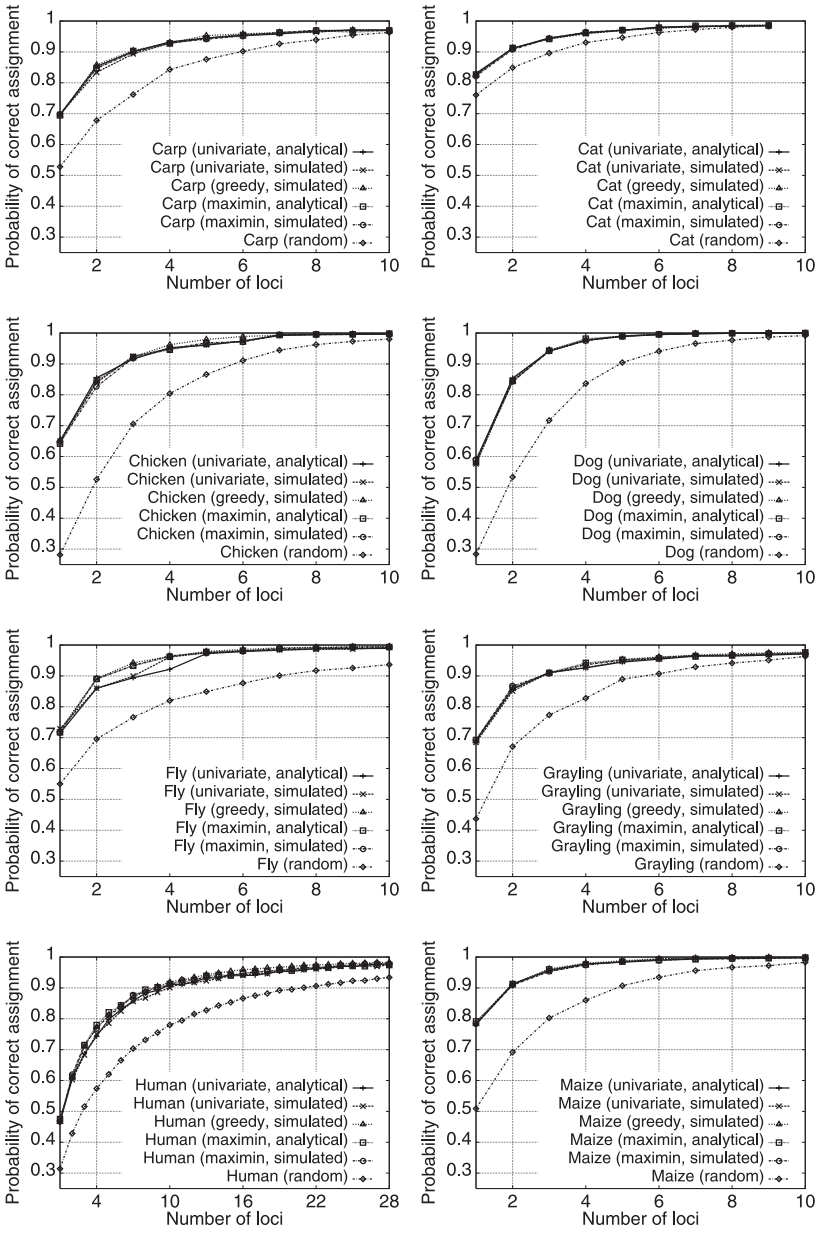

FIG. 3. Probability of correct assignment as a function of number of markers for five methods of selecting marker sets. The median probability of correct assignment based on 100 random orderings of the markers is also shown. For zero markers (not shown), the correct assignment probability is the reciprocal of the number of populations in the dataset. Note that the x-axis is scaled differently for the plots with the human dataset.

APPENDIX

It has sometimes been observed for certain ancestry inference procedures that accuracy of inference does not necessarily increase as markers are accumulated (Alaska Department of Fish and Game, 2000). This appendix investigates the relationship of f_{ORCA} and $\tilde{f}_{ORCA}$ to the number of loci, demonstrating that f_{ORCA} does have the property that accumulating loci increases performance, and that $\tilde{f}_{ORCA}$ "almost" has this property. Thus, when evaluating the performance of $\tilde{f}_{ORCA}$ in assignment of individuals, although exceptions can occur, incorporating additional loci generally increases performance (as was observed in Fig. 3).

Lemmas 1 and 3 give bounds for f_{ORCA}. Lemma 1 motivates the choice $f_{ORCA}(\phi) = 1/K$, so that nonempty sets of loci produce correct assignment probabilities at least as large as those obtained with no loci. Theorem 2 proves that f_{ORCA} is a performance function: if allele frequencies are known, the probability of correct assignment for the procedure that assigns individuals to their most likely source populations increases as additional loci are considered. Theorem 4 shows that the values of f_{ORCA} for a nested sequence of sets of loci converge to a constant, providing an explanation for the apparent convergence of performance in Fig. 3. This constant need not equal 1—for example, consider an infinite set of loci in which for each locus l, each allele has frequency $1/N^{(l)}$ in every population. For this set of loci, f_{ORCA} equals $1/K$.

Corollary 6 of Theorem 5 explains the assertion that $\tilde{f}_{ORCA}$ is "almost" a performance function. It shows that if enough simulated individuals are used in the evaluation of $\tilde{f}_{ORCA}$, with high probability, accumulation of additional loci either increases performance, does not affect it, or decreases it by a small amount. Corollary 8 of Theorem 7 gives a similar result in case f_{ORCA} is computed using sample allele frequencies rather than true frequencies. If large enough samples are used in evaluation of this estimate, $\hat{f}_{ORCA}$, accumulating loci is likely to either increase performance, not affect it, or decrease it by a small amount.

Finally, Corollary 10 of Theorem 9 shows that if $\tilde{\hat{f}}_{ORCA}$—performance based on simulations that employ estimated allele frequencies—is used to evaluate assignments, then both the number of simulations and the sample sizes can be made large enough so that with high probability, accumulating additional loci either increases performance, does not affect it, or decreases it by a small amount. This result demonstrates that even under realistic conditions—in which simulations rather than the analytical formula are used and allele frequencies are estimated rather than known—the genotyping of additional loci is likely to increase the probability of correct assignment.

We now introduce additional notation before proving the theorems. Consider a vector $\mathbf{Q}$ of nonnegative numbers $q_1, q_2, \ldots, q_K$ with $\sum_{i=1}^{K} q_i = 1$. The choice of $\mathbf{Q}$ corresponds to a prior probability distribution for the source population of an individual; the purpose of introducing a general prior is to allow more general assignment rules. With the prior distribution $\mathbf{Q}$, the probability of correct assignment if individuals are assigned to their most likely source populations, denoted $f_{\mathbf{Q}}$, is obtained by replacing $1/K$ with q_i in Equation (1). The following form of this generalized quantity is more convenient for the proofs than is that of Equation (1) (though it is less convenient for evaluation due to its increased number of terms):

$$f_{\mathbf{Q}}(S_M) = \sum_{j_1^{(1)}=1}^{N^{(1)}} \sum_{j_2^{(1)}=1}^{N^{(1)}} \sum_{j_1^{(2)}=1}^{N^{(2)}} \sum_{j_2^{(2)}=1}^{N^{(2)}} \cdots \sum_{j_1^{(M)}=1}^{N^{(M)}} \sum_{j_2^{(M)}=1}^{N^{(M)}} \max_{i \in \{1,2,\ldots,K\}} \left[q_i \prod_{m=1}^{M} p^{(m)}_{ij_1^{(m)}} p^{(m)}_{ij_2^{(m)}} \right]. \tag{2}$$

The function f_{ORCA} (Equation (1)) is the special case of $f_{\mathbf{Q}}$ in which $q_1 = q_2 = \ldots = q_K = 1/K$.

Define $\tilde{f}_{\mathbf{Q}}$ by the simulation procedure in Section 2.2, replacing Step 1 with simulation of q from the prior $\mathbf{Q}$. For a set of loci T, let $\tilde{f}_{\mathbf{Q},\alpha}(T)$ be the (random) value of $\tilde{f}_{\mathbf{Q}}(T)$ obtained from α simulated individuals. Let $\hat{f}_{\mathbf{Q},(n_1,n_2,\ldots,n_K)}(T)$ be the (random) value of $f_{\mathbf{Q}}(T)$ obtained using allele frequency estimates from a sample with $n_i \geq 1$ observations in population i and abbreviate $(n, n, \ldots, n)$ by $\mathbf{n}$. Let $\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T)$ be the (random) value of $\tilde{f}_{\mathbf{Q}}(T)$ obtained using α simulated individuals based on allele frequency estimates from a sample with size vector $\mathbf{n}$. For the empty set, define $\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(\phi) = \hat{f}_{\mathbf{Q},\mathbf{n}}(\phi) = \tilde{f}_{\mathbf{Q},\alpha}(\phi) = f_{\mathbf{Q}}(\phi) = \max_{i \in \{1,2,\ldots,K\}} q_i$.

Until now, we have viewed f_{ORCA} and its extensions as real-valued functions on sets of sets. For a given set with M loci, it is convenient to also view them as functions on the set of possible allele frequencies for the loci. In this framework, these functions have domain $\Delta^K_{N^{(1)}} \times \Delta^K_{N^{(2)}} \ldots \times \Delta^K_{N^{(M)}}$, where $\Delta_N = \{(p_1, p_2, \ldots, p_N | p_j \geq 0, \sum_{i=1}^{N} p_j = 1\}$ is the set of possible allele frequencies for a single population at a locus with N alleles. Henceforth, allele frequencies of 0 are retained in all computations.

Let $W_M = \{(j_1^{(1)}, j_2^{(1)}, \ldots, j_1^{(M)}, j_2^{(M)}) \mid j_1^{(m)}, j_2^{(m)} \in \{1, 2, \ldots, N^{(m)}\}$ for each $m\}$. Consider a denumerable set of loci S_L, and without loss of generality, label the loci $1, 2, 3, \ldots$. Let R_M denote the subset of S_L that contains loci $1, 2, \ldots, M$ (R_0 is the empty set). Note that in the main text, S_L is assumed to be finite, in order to guarantee that maxima exist for functions on $\mathcal{P}(S_L)$. Theorem 4 below, however,

specifically assumes infinite S_L. In case S_L is finite, this result applies to infinite sets obtained by appending loci to S_L so that for each positive integer l, at locus $L+l$, each allele has frequency $1/N^{(L+l)}$ in every population.

Lemma 1. *For any $v \in S_L$, $f_{\mathbf{Q}}(\{v\}) \geq \max_{i\in\{1,2,\ldots,K\}} q_i$.*

Proof. It suffices to show that for each $\mathbf{Q}$, $f_{\mathbf{Q}}(R_1) \geq \max_{i\in\{1,2,\ldots,K\}} q_i$. Using Equation (2), for any $k \in \{1, 2, \ldots, K\}$,

$$f_{\mathbf{Q}}(R_1) = \sum_{j_1=1}^{N^{(1)}} \sum_{j_2=1}^{N^{(1)}} \max_{i\in\{1,2,\ldots,K\}} \left[q_i\, p_{ij_1}^{(1)} p_{ij_2}^{(1)} \right] \geq \sum_{j_1=1}^{N^{(1)}} \sum_{j_2=1}^{N^{(1)}} q_k\, p_{kj_1}^{(1)} p_{kj_2}^{(1)}.$$

Using the fact that for any m

$$\sum_{j_1=1}^{N^{(m)}} \sum_{j_2=1}^{N^{(m)}} p_{kj_1}^{(m)} p_{kj_2}^{(m)} = 1, \tag{3}$$

it follows that $f_{\mathbf{Q}}(R_1) \geq q_k$. Because this inequality holds for each k, $f_{\mathbf{Q}}(R_1) \geq \max_{i\in\{1,2,\ldots,K\}} q_i$. ■

Theorem 2. *Function $f_{\mathbf{Q}}$ is a performance function for S_L.*

Proof. It suffices to show that for any $M \geq 1$, $f_{\mathbf{Q}}(R_{M-1}) \leq f_{\mathbf{Q}}(R_M)$. For $M = 1$, the result follows from Lemma 1. Otherwise, writing out $f_{\mathbf{Q}}(R_{M-1})$ and $f_{\mathbf{Q}}(R_M)$ (Equation (2)), it suffices to show that for any $\mathbf{j} \in W_{M-1}$,

$$\max_{i\in\{1,2,\ldots,K\}} \left[q_i \prod_{m=1}^{M-1} p_{ij_1^{(m)}}^{(m)} p_{ij_2^{(m)}}^{(m)} \right] \leq \sum_{j_1^{(M)}=1}^{N^{(M)}} \sum_{j_2^{(M)}=1}^{N^{(M)}} \max_{i\in\{1,2,\ldots,K\}} \left[q_i \prod_{m=1}^{M} p_{ij_1^{(m)}}^{(m)} p_{ij_2^{(m)}}^{(m)} \right]. \tag{4}$$

For $\mathbf{j} \in W_{M-1}$, abbreviate $a_i = q_i \prod_{m=1}^{M-1} p_{ij_1^{(m)}}^{(m)} p_{ij_2^{(m)}}^{(m)}$. For $j_1^{(M)}, j_2^{(M)} \in \{1, 2, \ldots, N^{(M)}\}$ and $k \in \{1, 2, \ldots, K\}$,

$$a_k\, p_{kj_1^{(M)}}^{(M)} p_{kj_2^{(M)}}^{(M)} \leq \max_{i\in\{1,2,\ldots,K\}} \left[a_i\, p_{ij_1^{(M)}}^{(M)} p_{ij_2^{(M)}}^{(M)} \right].$$

Summing this inequality over all possible $j_1^{(M)}, j_2^{(M)}$ and using Equation (3), we obtain

$$a_k \leq \sum_{j_1^{(M)}=1}^{N^{(M)}} \sum_{j_2^{(M)}=1}^{N^{(M)}} \max_{i\in\{1,2,\ldots,K\}} \left[a_i\, p_{ij_1^{(M)}}^{(M)} p_{ij_2^{(M)}}^{(M)} \right].$$

Because this inequality holds for each k, we can take the maximum of the left hand side to obtain (4). ■

Lemma 3. *For any $T \subset S_L$, $f_{\mathbf{Q}}(T) \leq 1$.*

Proof. By definition of $f_{\mathbf{Q}}$, the result holds for the empty set. Otherwise, it suffices to show that the result holds for every R_M, $M \geq 1$. Choose M, and for each $k \in \{1, 2, \ldots, K\}$, let

$$W_{M,k} = \left\{ \mathbf{j} \in W_M \mid \arg \max_{i\in\{1,2,\ldots,K\}} \left[q_i \prod_{m=1}^{M} p_{ij_1^{(m)}}^{(m)} p_{ij_2^{(m)}}^{(m)} \right] = k \right\}.$$

For a given **j**, if $i_1, i_2, \ldots, i_C$ tie for the maximum, place **j** in $W_{M,\min_{c\in\{1,2,\ldots,C\}} i_c}$. Rearranging Equation (2),

$$f_{\mathbf{Q}}(R_M) = \sum_{i=1}^{K} q_i \sum_{\mathbf{j}\in W_{M,i}} \prod_{m=1}^{M} p^{(m)}_{ij_1^{(m)}} p^{(m)}_{ij_2^{(m)}}$$

$$\leq \sum_{i=1}^{K} q_i \sum_{\mathbf{j}\in W_M} \prod_{m=1}^{M} p^{(m)}_{ij_1^{(m)}} p^{(m)}_{ij_2^{(m)}}$$

$$= \sum_{i=1}^{K} q_i \prod_{m=1}^{M} \sum_{j_1^{(m)}=1}^{N^{(m)}} \sum_{j_2^{(m)}=1}^{N^{(m)}} p^{(m)}_{ij_1^{(m)}} p^{(m)}_{ij_2^{(m)}}.$$

Applying Equation (3), it follows that $f_{\mathbf{Q}}(R_M) \leq \sum_{i=1}^{K} q_i = 1$. ■

Theorem 4. *If $m_1 \leq m_2$ implies $T_{m_1} \subset T_{m_2} \subset S_L$ for all m_1, m_2, then $\{f_{\mathbf{Q}}(T_m)\}_{m=1}^{\infty}$ converges to a number in* $[\max_{i\in\{1,2,\ldots,K\}} q_i, 1]$.

Proof. As a consequence of Theorem 2, the sequence is monotonically nondecreasing. As a consequence of Lemmas 1 and 3, it is bounded below by $\max_{i\in\{1,2,\ldots,K\}} q_i$ and above by 1. Using the fact that monotonic bounded sequences of real numbers converge (Rudin, 1976, Theorem 3.14), the result follows. ■

Theorem 5. *Consider $T \subset S_L$. As $\alpha \to \infty$, $\tilde{f}_{\mathbf{Q},\alpha}(T)$ converges almost surely in and probability to $\tilde{f}_{\mathbf{Q}}(T)$.*

Proof. The almost sure convergence is a consequence of the strong law of large numbers (Serfling, 1980, Theorem 1.8B), using the fact that for any T, $\mathbb{E}[\tilde{f}_{\mathbf{Q},1}(T)] = f_{\mathbf{Q}}(T)$ (see Section 2.2). Convergence in probability then follows (Serfling, 1980, Theorem 1.3.1). ■

Corollary 6. *Consider $T \subset S_L$, sets $T_1, T_2 \subset T$ with $T_1 \subset T_2$, and $\epsilon_1, \epsilon_2 > 0$. There exists α^* such that if $\alpha \geq \alpha^*$, then $\mathbb{P}[\tilde{f}_{\mathbf{Q},\alpha}(T_2) > \tilde{f}_{\mathbf{Q},\alpha}(T_1) - \epsilon_1] > 1 - \epsilon_2$.*

Proof. Applying Theorem 5 and the definition of convergence in probability, there exists α^* such that for $\alpha \geq \alpha^*$, both $\mathbb{P}[|\tilde{f}_{\mathbf{Q},\alpha}(T_1) - f_{\mathbf{Q}}(T_1)| < \epsilon_1/2] > 1-\epsilon_2/2$ and $\mathbb{P}[|\tilde{f}_{\mathbf{Q},\alpha}(T_2) - f_{\mathbf{Q}}(T_2)| < \epsilon_1/2] > 1-\epsilon_2/2$. Then $\mathbb{P}[\tilde{f}_{\mathbf{Q},\alpha}(T_1) < f_{\mathbf{Q}}(T_1) + \epsilon_1/2] > 1 - \epsilon_2/2$ and $\mathbb{P}[f_{\mathbf{Q}}(T_2) < \tilde{f}_{\mathbf{Q},\alpha}(T_2) + \epsilon_1/2] > 1 - \epsilon_2/2$, from which $\mathbb{P}[(\tilde{f}_{\mathbf{Q},\alpha}(T_1) < f_{\mathbf{Q}}(T_1) + \epsilon_1/2) \cap (f_{\mathbf{Q}}(T_2) < \tilde{f}_{\mathbf{Q},\alpha}(T_2) + \epsilon_1/2)] > 1 - \epsilon_2$. The intersection in this expression has probability less than or equal to that of $\tilde{f}_{\mathbf{Q},\alpha}(T_1) + f_{\mathbf{Q}}(T_2) < f_{\mathbf{Q}}(T_1) + \tilde{f}_{\mathbf{Q},\alpha}(T_2) + \epsilon_1$, which, using Theorem 2 to obtain $f_{\mathbf{Q}}(T_1) \leq f_{\mathbf{Q}}(T_2)$, has probability less than or equal to that of $\tilde{f}_{\mathbf{Q},\alpha}(T_2) > \tilde{f}_{\mathbf{Q},\alpha}(T_1) - \epsilon_1$. ■

Theorem 7. *Consider $T \subset S_L$. As $n \to \infty$, $\hat{f}_{\mathbf{Q},\mathbf{n}}(T)$ converges almost surely and in probability to $\hat{f}_{\mathbf{Q}}(T)$.*

Proof. If $T = \phi$ the result is trivial. Otherwise, by the strong law of large numbers (Serfling, 1980, Theorem 1.8B), for each i, m, and $j^{(m)}$, as $n \to \infty$, the sample frequency $\hat{p}^{(m)}_{ij^{(m)},n}$ estimated from sample size n converges almost surely to the true frequency $p^{(m)}_{ij^{(m)}}$. Because each component sample frequency converges a.s. to the appropriate true frequency, the sample frequency vector converges a.s. to the true frequency vector (Serfling, 1980, Problem 1.P.2b). As a composition of sums, products, and maxima, $f_{\mathbf{Q}}$ is continuous on $\Delta^K_{N^{(1)}} \times \Delta^K_{N^{(2)}} \ldots \times \Delta^K_{N^{(M)}}$, and it follows that $\hat{f}_{\mathbf{Q},\mathbf{n}}(T)$ converges a.s. to $f_{\mathbf{Q}}(T)$ (Serfling, 1980, Theorem 1.7i). Convergence in probability follows (Serfling, 1980, Theorem 1.3.1). ■

Corollary 8. *Consider $T \subset S_L$, sets $T_1, T_2 \subset T$ with $T_1 \subset T_2$, and $\epsilon_1, \epsilon_2 > 0$. There exists n^* such that if $n \geq n^*$, then $\mathbb{P}[\hat{f}_{\mathbf{Q},\mathbf{n}}(T_2) > \hat{f}_{\mathbf{Q},\mathbf{n}}(T_1) - \epsilon_1] > 1 - \epsilon_2$.*

Proof. Using Theorem 7, $\hat{f}_{\mathbf{Q},\mathbf{n}}(T_2)$ converges in probability to $f_{\mathbf{Q}}(T_2)$ and $\hat{f}_{\mathbf{Q},\mathbf{n}}(T_1)$ converges in probability to $f_{\mathbf{Q}}(T_1)$ (trivially if $T_1 = \phi$). The remainder of the proof follows the same argument as in the proof of Corollary 6, using $\hat{f}_{\mathbf{Q},\mathbf{n}}$ in place of $\tilde{f}_{\mathbf{Q},\alpha}$ and n, n^* in place of α, α^*. ■

Theorem 9. *Consider $T \subset S_L$. As $\alpha, n \to \infty$, $\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T)$ converges in probability to $f_{\mathbf{Q}}(T)$.*

Proof. Let $\epsilon_1, \epsilon_2 > 0$ and $D_{\alpha,\mathbf{n}} = \mathbb{P}[|\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T) - f_{\mathbf{Q}}(T)| > \epsilon_1]$. Because $|\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T) - f_{\mathbf{Q}}(T)| \leq |\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T) - \hat{f}_{\mathbf{Q},\mathbf{n}}(T)| + |\hat{f}_{\mathbf{Q},\mathbf{n}}(T) - f_{\mathbf{Q}}(T)|$, it follows that $D_{\alpha,\mathbf{n}} \leq \mathbb{P}[|\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T) - \hat{f}_{\mathbf{Q},\mathbf{n}}(T)| > \epsilon_1/2] + \mathbb{P}[|\hat{f}_{\mathbf{Q},\mathbf{n}}(T) - f_{\mathbf{Q}}(T)| > \epsilon_1/2]$. By definition of convergence in probability, it suffices to show that there exist (α^*, n^*) such that for any $\alpha \geq \alpha^*$ and any $n \geq n^*$, $D_{\alpha,\mathbf{n}} < \epsilon_2$. By Theorem 7, using the definition of convergence in probability, there exists n^* such that for $n \geq n^*$, $\mathbb{P}[|\hat{f}_{\mathbf{Q},\mathbf{n}}(T) - f_{\mathbf{Q}}(T)| > \epsilon_1/2] < \epsilon_2/2$. Applying Chebyshev's inequality (Durrett, 1996, p. 15) and using $\mathbb{E}[\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T)] = \hat{f}_{\mathbf{Q},\mathbf{n}}$ (see Section 2.2),

$$\begin{aligned}\mathbb{P}[\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T) - \hat{f}_{\mathbf{Q},\mathbf{n}}(T)| > \epsilon_1/2] &\leq \mathrm{Var}[\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T)]/[(\epsilon_1/2)^2] \\ &= 4\mathrm{Var}[\tilde{\hat{f}}_{\mathbf{Q},1,\mathbf{n}}(T)]/(\alpha\epsilon_1^2) \\ &< 4/(\alpha\epsilon_1^2),\end{aligned}$$

where the last step follows from the fact that the variance of a random variable on [0,1] is less than 1. The bound $4/(\alpha\epsilon_1^2)$ applies for any n. Choosing $\alpha^* > 8/(\epsilon_1^2\epsilon_2)$, $\mathbb{P}[\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T) - \hat{f}_{\mathbf{Q},\mathbf{n}}(T)| > \epsilon_1/2] < \epsilon_2/2$. ■

Corollary 10. *Consider $T \subset S_L$, sets $T_1, T_2 \subset T$ with $T_1 \subset T_2$, and $\epsilon_1, \epsilon_2 > 0$. There exist α^* and n^* such that if $\alpha \geq \alpha^*$ and $n \geq n^*$, then $\mathbb{P}[\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T_2) > \tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T_1) - \epsilon_1] > 1 - \epsilon_2$.*

Proof. By Theorem 9, there exist (α^*, n^*) so that for $\alpha \geq \alpha^*$ and $n \geq n^*$, both $\mathbb{P}[|\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T_1) - f_{\mathbf{Q},\mathbf{n}}(T_1)| < \epsilon_1/2] > 1 - \epsilon_2/2$, and $\mathbb{P}[|\tilde{\hat{f}}_{\mathbf{Q},\alpha,\mathbf{n}}(T_2) - f_{\mathbf{Q},\mathbf{n}}(T_2)| < \epsilon_1/2] > 1 - \epsilon_2/2$. The argument in the proof of Corollary 6 then applies. ■

ACKNOWLEDGMENTS

This research was supported in part by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences. I thank S. Kalinowski for helpful conversations, an anonymous reviewer for comments on the manuscript, V. Plagnol for a very careful reading of the Appendix and for correcting an earlier error in the proof of Theorem 9, and D. Irion, C. Schlötterer, M. Koskinen, and Y. Vigouroux for assistance with the dog, fly, grayling, and maize datasets, respectively.

REFERENCES

Alaska Department of Fish and Game. 2000. *SPAM Version 3.2: User's Guide*, Alaska Department of Fish and Game, Anchorage.

Anderson, E.C., and Thompson, E.A. 2002. A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* 160, 1217–1229.

Ayres, K.L., and Balding, D.J. 1998. Measuring departures from Hardy–Weinberg: A Markov chain Monte Carlo method for estimating the inbreeding coefficient. *Heredity* 80, 769–777.

Bamshad, M.J., Wooding, S., Watkins, W.S., Ostler, C.T., Batzer, M.A., and Jorde, L.B. 2003. Human population genetic structure and inference of group membership. *Am. J. Human Genet.* 72, 578–589.

Banks, M.A., and Eichert, W. 2000. WHICHRUN (version 3.2): A computer program for population assignment of individuals based on multilocus genotype data. *J. Hered.* 91, 87–89.

Banks, M.A., Eichert, W., and Olsen, J.B. 2003. Which genetic loci have greater population assignment power? *Bioinformatics* 19, 1436–1438.

Baudouin, L., Piry, S., and Cornuet, J.M. 2004. Analytical Bayesian approach for assigning individuals to populations. *J. Hered.* 95, 217–224.

Beaumont, M., Barratt, E.M., Gottelli, D., Kitchener, A.C., Daniels, M.J., Pritchard, J.K., and Bruford, M.W. 2001. Genetic diversity and introgression in the Scottish wildcat. *Mol. Ecol.* 10, 319–336.

Bernatchez, L., and Duchesne, P. 2000. Individual-based genotype analysis in studies of parentage and population assignment: How many loci, how many alleles? *Can. J. Fish. Aquat. Sci.* 57, 1–12.

Buchanan, F.C., Adams, L.J., Littlejohn, R.P., Maddox, J.F., and Crawford, A.M. 1994. Determination of evolutionary relationships among sheep breeds using microsatellites. *Genomics* 22, 397–403.

Campbell, D., Duchesne, P., and Bernatchez, L. 2003. AFLP utility for population assignment studies: Analytical investigation and empirical comparison with microsatellites. *Mol. Ecol.* 12, 1979–1991.

Collins-Schramm, H.E., Phillips, C.M., Operario, D.J., Lee, J.S., Weber, J.L., Hanson, R.L., Knowler, W.C., Cooper, R., Li, H., and Seldin, M.F. 2002. Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. *Am. J. Human Genet.* 70, 737–750.

Cornuet, J.-M., Piry, S., Luikart, G., Estoup, A., and Solignac, M. 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* 153, 1989–2000.

David, L., Rosenberg, N.A., Lavi, U., Feldman, M.W., and Hillel, J. 2005. Genetic diversity and population structure inferred from the partially duplicated genome of *Cyprinus carpio* L. Submitted.

Davies, N., Villablanca, F.X., and Roderick, G.K. 1999. Determining the source of individuals: Multilocus genotyping in nonequilibrium population genetics. *Trends Ecol. Evol.* 14, 17–21.

Dean, M., Stephens, J.C., Winkler, C., Lomb, D.A., Ramsburg, M., Boaze, R., Stewart, C., Charbonneau, L., Goldman, D., Albaugh, B.J., Goedert, J.J., Beasley, R.P., Hwang, L.-Y., Buchbinder, S., Weedon, M., Johnson, P.A., Eichelberger, M., and O'Brien, S.J. 1994. Polymorphic admixture typing in human ethnic populations. *Am. J. Human Genet.* 55, 788–808.

Durrett, R. 1996. *Probability: Theory and Examples*, 2nd ed., Duxbury, Belmont, CA.

Edwards, A.W.F. 2003. Human genetic diversity: Lewontin's fallacy. *BioEssays* 25, 798–801.

Gibbons, J.D. 1985. *Nonparametric Statistical Inference*, 2nd ed., Marcel Dekker, New York.

Goldstein, D.B., and Schlötterer, C., eds. 1999. *Microsatellites: Evolution and Applications*, Oxford University Press, Oxford, UK.

Guinand, B., Topchy, A., Page, K.S., Burnham-Curtis, M.K., Punch, W.F., and Scribner, K.T. 2002. Comparisons of likelihood and machine learning methods of individual classification. *J. Hered.* 93, 260–269.

Hansen, M.M., Kenchington, E., and Nielsen, E.E. 2001. Assigning individual fish to populations using microsatellite DNA markers. *Fish Fish.* 2, 93–112.

Irion, D.N., Schaffer, A.L., Famula, T.R., Eggleston, M.L., Hughes, S.S., and Pedersen, N.C. 2003. Analysis of genetic variation in 28 dog breed populations with 100 microsatellite markers. *J. Hered.* 94, 81–87.

Kauer, M., Dieringer, D., and Schlötterer, C. 2003. Nonneutral admixture of immigrant genotypes in African *Drosophila melanogaster* populations from Zimbabwe. *Mol. Biol. Evol.* 20, 1329–1337.

Koskinen, M.T., Nilsson, J., Veselov, A., Potutkin, A.G., Ranta, E., and Primmer, C.R. 2002. Microsatellite data resolve phylogeographic patterns in European grayling, *Thymallus thymallus*, Salmonidae. *Heredity* 88, 391–401.

Lewis, P.O., and Zaykin, D. 2001. GDA (Genetic Data Analysis): Computer program for the analysis of allelic data (version 1.0 d16c). *hydrodidyon.eeb.uconn.edu/people/plewis*

Lowe, A.L., Urquhart, A., Foreman, L.A., and Evett, I.W. 2001. Inferring ethnic origin by means of an STR profile. *Forensic Sci. Int.* 119, 17–22.

Manel, S., Berthier, P., and Luikart, G. 2002. Detecting wildlife poaching: Identifying the origin of individuals with Bayesian assignment tests and multilocus genotypes. *Conserv. Biol.* 16, 650–659.

Manel, S., Gaggiotti, O.E., and Waples, R.S. 2005. Assignment methods: Matching biological questions with appropriate techniques. *Trends Ecol. Evol.* 20, 136–142.

Matsuoka, Y., Vigouroux, Y., Goodman, M.M., Sanchez, G.J., Buckler, E., and Doebley, J. 2002. A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci. USA* 99, 6080–6084.

Nei, M. 1987. *Molecular Evolutionary Genetics*, Columbia University Press, New York.

Paetkau, D., Calvert, W., Stirling, I., and Strobeck, C. 1995. Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol.* 4, 347–354.

Paetkau, D., Slade, R., Burden, M., and Estoup, A. 2004. Genetic assignment methods for the direct, real-time estimation of migration rate: A simulation-based exploration of accuracy and power. *Mol. Ecol.* 13, 55–65.

Pfaff, C.L., Barnholtz-Sloan, J., Wagner, J.K., and Long, J.C. 2004. Information on ancestry from genetic markers. *Genet. Epidemiol.* 26, 305–315.

Primmer, C.R., Koskinen, M.T., and Piironen, J. 2000. The one that did not get away: Individual assignment using microsatellite data detects a case of fishing competition fraud. *Proc. R. Soc. Lond. B* 267, 1699–1704.

Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Risch, N., Burchard, E., Ziv, E., and Tang, H. 2002. Categorization of humans in biomedical research: Genes, race and disease. *Genome Biol.* 3, comment2007.

Rosenberg, N.A., Burke, T., Elo, K., Feldman, M.W., Freidlin, P.J., Groenen, M.A.M., Hillel, J., Mäki-Tanila, A., Tixier-Boichard, M., Vignal, A., Wimmers, K., and Weigend, S. 2001. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* 159, 699–713.

Rosenberg, N.A., Li, L., Ward, R., and Pritchard, J.K. 2003. Informativeness of genetic markers for inference of ancestry. *Am. J. Human Genet.* 73, 1402–1422.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. 2002. Genetic structure of human populations. *Science* 298, 2381–2385.

Rudin, W. 1976. *Principles of Mathematical Analysis*, 3rd ed., McGraw-Hill, New York.

Serfling, R.J. 1980. *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Shriver, M.D., Smith, M.W., Jin, L., Marcini, A., Akey, J.M., Deka, R., and Ferrell, R.E. 1997. Ethnic-affiliation estimation by use of population-specific DNA markers. *Am. J. Human Genet.* 60, 957–964.

Turakulov, R., and Easteal, S. 2003. Number of SNPS loci needed to detect population structure. *Human Hered.* 55, 37–45.

Waser, P.M., and Strobeck, C. 1998. Genetic signatures of interpopulation dispersal. *Trends Ecol. Evol.* 13, 43–44.

Weir, B.S. 1996. *Genetic Data Analysis II*, Sinauer, Sunderland, MA.

Ziv, E., and Burchard, E.G. 2003. Human population structure and genetic association studies. *Pharmacogenomics* 4, 431–441.

Address correspondence to:
Noah A. Rosenberg
Department of Human Genetics,
Bioinformatics Program, and the Life Sciences Institute
University of Michigan
2017 Palmer Commons
100 Washtenaw Ave., Box 2218
Ann Arbor, MI 48109-2218

E-mail: rnoah@umich.edu