Am. J. Hum. Genet. 73:1402–1422, 2003

Informativeness of Genetic Markers for Inference of Ancestry*

Noah A. Rosenberg,[1] Lei M. Li,[1] Ryk Ward,[2] and Jonathan K. Pritchard[3]

[1]Program in Molecular and Computational Biology, University of Southern California, Los Angeles; [2]Institute of Biological Anthropology, University of Oxford, Oxford, United Kingdom; and [3]Department of Human Genetics, University of Chicago, Chicago

Inference of individual ancestry is useful in various applications, such as admixture mapping and structured-association mapping. Using information-theoretic principles, we introduce a general measure, the informativeness for assignment (I_n), applicable to any number of potential source populations, for determining the amount of information that multiallelic markers provide about individual ancestry. In a worldwide human microsatellite data set, we identify markers of highest informativeness for inference of regional ancestry and for inference of population ancestry within regions; these markers, which are listed in online-only tables in our article, can be useful both in testing for and in controlling the influence of ancestry on case-control genetic association studies. Markers that are informative in one collection of source populations are generally informative in others. Informativeness of random dinucleotides, the most informative class of microsatellites, is five to eight times that of random single-nucleotide polymorphisms (SNPs), but 2%–12% of SNPs have higher informativeness than the median for dinucleotides. Our results can aid in decisions about the type, quantity, and specific choice of markers for use in studies of ancestry.

Introduction

Inference of individual ancestry from genetic markers is helpful in diverse situations, including admixture and association mapping, forensics, prediction of medical risks, wildlife management, and studies of dispersal, gene flow, and evolutionary history (Shriver et al. 1997; Davies et al. 1999; Primmer et al. 2000; Manel et al. 2002; Bamshad et al. 2003; Campbell et al. 2003; Ziv and Burchard 2003). Statistical methods for ancestry inference use multilocus genotypes and population allele frequencies, either specified in advance or estimated during the inference process, to assign populations of origin to individuals (Smouse et al. 1982; Paetkau et al. 1995; Rannala and Mountain 1997; Cornuet et al. 1999; Pritchard et al. 2000; Guinand et al. 2002).

Because use of highly informative markers can reduce the amount of genotyping required for ancestry inference, it is desirable to measure the extent to which specific markers contribute to this inference. Several approaches have previously been used for measuring these locus contributions (table 1). However, despite their various features in specific scenarios, all of these measures are either difficult to compute, not designed specifically for estimating marker information content, or not applicable to sets with many potential source populations.

Here, using information-theoretic and decision-theoretic approaches, we introduce new criteria: the *informativeness for assignment*, the *optimal rate of correct assignment*, and the *informativeness for ancestry coefficients*. The choice of statistic for use in identifying markers for ancestry inference depends on the inference algorithm that is being used (table 1). The new statistics, as convenient and statistically motivated general measures applicable to any number of alleles and populations, may be useful both in admixed and in multisource human groups, such as those that have formed in the Western Hemisphere by the intermixing of Africans, Native Americans, and Europeans. We first define the statistics, consider their relationships with δ and F_{st} (two criteria that are often used to measure marker information content), and study the number of markers needed for inference. We demonstrate that the criteria are highly correlated and proceed using only the informativeness for assignment, or, simply, the *informativeness*. Informativeness rankings of loci in human microsatellite data are found to be robust, and use of markers of highest informativeness is observed to reduce the number of markers needed for inference of population structure. We consider the relationships of informativeness values in different subsets of the human population, and the relative informativeness of microsatellites and SNPs. Tables A, B, C, D, and E (online only) provide lists of informativeness ranks for various sets of source populations.

Theory

Consider populations $i = 1,2,\ldots,K$ and loci $l = 1,2,\ldots,L$, with $K \geq 2$ and $L \geq 1$. Locus l has alleles

Received May 29, 2003; accepted for publication October 2, 2003; electronically published November 20, 2003.

Address for correspondence and reprints: Dr. Noah A. Rosenberg, Program in Molecular and Computational Biology, University of Southern California, 1042 West 36th Place, DRB 289, Los Angeles, CA 90089. E-mail: noahr@usc.edu

* This article is dedicated to the memory of Ryk Ward.

0002-9297/2003/7306-0017$15.00

Table 1

Measures of Marker Information Content

Criterion	Description	Features	Limitations
Absolute allele frequency difference (δ)	$\lvert p_{11} - p_{21} \rvert$	Is related to amount of linkage disequilibrium in an admixture model (Chakraborty and Weiss 1988); is related to probability of correct assignment in a multilocus no-admixture model (Risch et al. 2002); is related to Fisher information curvature criterion for $K = 2$ (eq. [18]); is related to *ORCA* for $K = 2$ (eq. [11])	Requires that only two populations be possible sources; does not take into account all available information about allele frequencies (Stephens et al. 1999; Campbell et al. 2003); statistical features do not apply to the Shriver et al. (1997) multiallelic extension of δ
F_{st}	Excess in probability of identity of alleles from the same population compared with randomly chosen alleles (Excoffier 2001, for example)	Is related, for biallelic markers, to the quotient of expected posterior and prior variance of ancestry in a population equally admixed from two sources[a] (McKeigue 1998; Molokhia et al. 2003)	Performs only slightly better than random markers (Rosenberg et al. 2001)
Expected heterozygosity[b]	$1 - \sum_{j=1}^{N} p_j^2$ (bias correction can be applied in estimation from data)	Performs better than random markers (Rosenberg et al. (2001) and fig. 3)	Measures the amount of variation but not the differences across populations
Number of alleles[b]	N	Performs better than random markers (Rosenberg et al. 2001)	Measures the amount of variation but not the differences across populations; is useful only for multiallelic markers that have variation in number of alleles
Fisher information curvature criterion	Reciprocal of the largest eigenvalue of the information matrix for maximum-likelihood estimation of ancestry coefficients (Gomulkiewicz et al. 1990; Millar 1991)	Enables predictions about approximate variances of ancestry estimates (see "Number of Markers" subsection of the "Theory" section); information matrix is additive across loci that are independent within populations	Depends on unknown ancestry coefficients and requires computation for many possible parameter values; largest eigenvalue gives an upper bound that might not be generally applicable across the parameter space
Pairwise Kullback-Leibler divergence[c]	$\sum_{j=1}^{N} p_{1j} \log \frac{p_{1j}}{p_{2j}} + p_{2j} \log \frac{p_{2j}}{p_{1j}}$ (Brenner 1998; Smith et al. 2001; Anderson and Thompson 2002)	Provides a natural measure, for $K = 2$, of average potential for assignment of an allele to one population compared with the other; has a natural multilocus extension; enables measurement of contributions of specific alleles	Requires that only two populations be possible sources; has upwardly biased estimates in small samples
Informativeness for assignment (I_n)	Equation (4)	Provides a natural measure of potential for assignment of an allele to one population compared with the "average" population; has a natural multilocus extension; enables measurement of contributions of specific alleles or populations; performs better than random or highly heterozygous markers (fig. 3)	Has upwardly biased estimates in small samples
Informativeness for ancestry coefficients (I_a)	Equation (14)	Provides a natural measure of potential for assignment of an allele to a point on the set of all possible ancestry coefficient vectors; has a natural multilocus extension; enables measurement of contributions of specific alleles	Has upwardly biased estimates in small samples; is difficult to compute in samples with populations of equal sample size
Optimal rate of correct assignment (ORCA)	Equation (10)	Gives the probability of correct assignment of an allele using the decision rule with lowest risk; has a natural multilocus extension (eq. [12]); enables measurement of contributions of specific alleles	Has upwardly biased estimates in small samples

NOTE.—Notation is defined in the "Theory" section. All criteria apply to multiallelic loci in any number of populations, except where specified.

[a] A multiallelic statistic related to this ratio was suggested by Molokhia et al. (2003).

[b] Can also be calculated by using average values of the statistic across populations rather than by using values for the whole collection of populations.

[c] A similar statistic based on genotype frequencies was suggested by Shriver et al. (1997). Some authors multiply by a factor of 1/2 in the formula for this statistic.

$j = 1,2,\ldots,N^{(l)}$. The relative frequency for allele j of locus l in population i is $p_{ij}^{(l)}$; this quantity represents a parametric rather than a sample frequency. The (parametric) average frequency of allele j at locus l is defined as

$$p_j^{(l)} = \sum_{i=1}^{K} \frac{p_{ij}^{(l)}}{K} \ . \qquad (1)$$

We use "log" to denote the natural logarithm, with $0 \log 0 = 0$.

Informativeness for Assignment: the No-Admixture Model

In the no-admixture model, individuals are each assumed to originate from one of K populations. Suppose we are given a random individual, whose (random) population assignment is Q, with $Q \in \{1,2,\ldots,K\}$. The probability that the individual belongs to population i is $\mathbb{P}(Q = i)$; we assume that each population has the same initial probability of being the source of an unknown individual, so that $\mathbb{P}(Q = i) = 1/K$ for all i. The (random) genotype of one of the individual's two alleles at locus l is $J^{(l)}$.

Our aim is to measure the amount of "information" gained about Q from knowledge of $J^{(l)}$ and to compare this quantity across different values of l. This measurement can be performed in a natural way, through use of an information-theoretic framework. If the value of $J^{(l)}$ is unknown, there is uncertainty, or entropy, regarding the value of the random variable Q. Once the value of $J^{(l)}$ is known, the entropy of Q decreases. The reduction in uncertainty about Q due to knowledge of $J^{(l)}$ is the *mutual information*, $I_n(Q;J^{(l)}) = H_n(Q) - H_n(Q|J^{(l)})$, where $H_n(Q)$ is the initial entropy of Q, and $H_n(Q|J^{(l)})$ is the conditional entropy of Q given knowledge of $J^{(l)}$ (the subscript "n" refers to the *no-admixture* model). Using standard definitions (Cover and Thomas 1991, chapter 2), and leaving off superscripts for convenience, we have

$$\begin{aligned} H_n(Q) &= -\sum_{i=1}^{K} \mathbb{P}(Q = i) \log \mathbb{P}(Q = i) \\ &= \log K \ , \end{aligned} \qquad (2)$$

$$\begin{aligned} H_n(Q|J) &= -\sum_{j=1}^{N}\sum_{i=1}^{K} \mathbb{P}(Q = i, J = j) \log \mathbb{P}(Q = i|J = j) \\ &= -\sum_{j=1}^{N}\sum_{i=1}^{K} \mathbb{P}(J = j|Q = i)\mathbb{P}(Q = i) \\ &\quad \times \log \frac{\mathbb{P}(J = j|Q = i)\mathbb{P}(Q = i)}{\mathbb{P}(J = j)} \\ &= -\sum_{j=1}^{N}\sum_{i=1}^{K} \frac{p_{ij}}{K} \log \frac{p_{ij}}{p_j K} \ , \end{aligned} \qquad (3)$$

and

$$I_n(Q;J) = \sum_{j=1}^{N} \left(-p_j \log p_j + \sum_{i=1}^{K} \frac{p_{ij}}{K} \log p_{ij} \right) . \qquad (4)$$

We refer to $I_n(Q;J)$ as the *informativeness for assignment*. For a given set of populations, the minimal I_n of 0 occurs when all alleles have equal frequencies in all populations. The maximal value, $\log K$, occurs when $N \geq K$ and no allele is found in more than one population.

This measure of the amount of "information" about the ancestry Q contained in the genotype J also arises from a likelihood approach. The quantity $\sum_{j=1}^{N}\sum_{i=1}^{K} (p_{ij}/K) \log p_{ij}$ can be viewed as the expected log-likelihood associated with drawing an allele randomly from the set of populations $\{1,2,\ldots,K\}$. The term $\sum_{j=1}^{N} p_j \log p_j$ is the expected log-likelihood associated with drawing an allele from a hypothetical "average" population whose allele frequencies equal the mean across the K populations. Thus, equation (4) gives the expected logarithm of the likelihood ratio whose numerator is the likelihood that an allele is assigned to one of the populations and whose denominator is the likelihood that an allele is assigned to the "average" population. When the minimum description length principle (Barron et al. 1998) is used, up to a constant, I_n also equals the expected reduction, upon observation of J, in the length of the optimal coding of the random variable Q. It gives the average (taken across populations) Kullback-Leibler information (Kullback 1959, p. 6) for distinguishing population-specific allele frequency distributions from the distribution for the "average population." For $K = 2$, I_n is similar to a previously proposed statistic based on Kullback-Leibler information (table 1).

The expression for $I_n(Q;J)$ also has a close correspondence with the G statistic obtained from a contingency table, each of whose N columns of K elements gives the relative frequencies of an allele in the K populations. This G statistic is given by (Sokal and Rohlf 1995, p. 737)

$$\begin{aligned} G_n(Q;J) &= 2\sum_{j=1}^{N}\sum_{i=1}^{K} p_{ij} \log p_{ij} \\ &\quad - \sum_{j=1}^{N} \left(\sum_{i=1}^{K} p_{ij} \right) \log \left(\sum_{i=1}^{K} p_{ij} \right) \\ &\quad - \sum_{i=1}^{K} \left(\sum_{j=1}^{N} p_{ij} \right) \log \left(\sum_{j=1}^{N} p_{ij} \right) \\ &\quad + \left(\sum_{j=1}^{N}\sum_{i=1}^{K} p_{ij} \right) \log \left(\sum_{j=1}^{N}\sum_{i=1}^{K} p_{ij} \right) . \end{aligned}$$

Using equation (1) and $\sum_{j=1}^{N} p_{ij} = 1$, we can simplify to

$$G_n(Q;J) = 2KI_n(Q;J) .$$

Thus, for any number of populations and alleles, identifying loci of high informativeness for assignment is equivalent to identifying loci with large values of G_n.

Note that $I_n(Q;J)$ is a sum over alleles. The contribution of allele j to informativeness for assignment is

$$I_n(Q;J = j) = -p_j \log p_j + \sum_{i=1}^{K} \frac{p_{ij}}{K} \log p_{ij} . \qquad (5)$$

For fixed p_j, the minimal allelic informativeness of 0 occurs when $p_{ij} = p_j$ for all i. For fixed $p_j \leq 1/K$, the maximum, $p_j \log K$, occurs when $p_{ij} = Kp_j$ for exactly one value of i and $p_{ij} = 0$ for all other values.

Similarly, it is possible to write $I_n(Q;J)$ as a sum over populations, with the contribution of population i to the informativeness for assignment equaling

$$I_n(Q = i;J) = \sum_{j=1}^{N} \frac{p_{ij}}{K} \log \frac{p_{ij}}{p_j} . \qquad (6)$$

Equations (5) and (6) enable calculation of the specific contribution of an allele or population to I_n. These computations are useful, since alleles at a locus might differ in their importance for assignment of individuals to populations. Populations might also differ in their degree of difference from the "average" population, so that assignment to some populations is easier than assignment to others. Henceforth, we use "informativeness" only in relation to the I_n statistic (and the I_a statistic, to be defined later), although we use "informative" and "information" more generally, to describe "ability to infer ancestry." Note that, if the prior assignment of individuals is not uniformly distributed, general priors, $\mathbb{P}(Q = i) = q_i$, can be accommodated by replacing $\mathbb{P}(Q = i)$ with q_i in the derivations of equations (2) and (3) and $1/K$ with q_i in equations (1), (4), (5), and (6).

I_n can also be extended for use in assignment to populations of multilocus diploid genotypes rather than of single alleles. For convenience, we treat diploid genotypes as being ordered so that (j_1,j_2) differs from (j_2,j_1) (an unordered genotype is assigned randomly to one of the possible ordered genotypes). If we assume both Hardy-Weinberg proportions and independence of loci within populations, equation (4) applies with $\prod_{l=1}^{L} p_{ij_1^{(l)}} p_{ij_2^{(l)}}$ in place of p_{ij}, $\sum_{i=1}^{K} (1/K) \prod_{l=1}^{L} p_{ij_1^{(l)}} p_{ij_2^{(l)}}$ in place of p_j, and with the sum taken over all $\prod_{l=1}^{L} [N^{(l)}]^2$ possible multilocus genotypes, $\{(j_1^{(1)},j_2^{(1)}),(j_1^{(2)},j_2^{(2)}),\ldots,(j_1^{(L)},j_2^{(L)})\}$. Although this sum is difficult to evaluate if the number of possible multilocus genotypes is large, it can, in principle, predict the informativeness of multilocus sets; note that informativeness is not additive over loci, since loci that are independent within populations may still contribute to ancestry inference in a correlated manner.

Relationship of I_n to δ and F_{st}

For the simplest case in which informativeness is of interest—namely, for $K = N = 2$—it is possible to relate I_n to δ and to F_{st}. In the mathematical development, we let δ equal the signed difference between the frequencies of allele 1 in two populations, $p_{11} - p_{21}$, and, without loss of generality, we assume that $p_{11} \geq p_{21}$, so that $\delta \in [0,1]$; when applied to data, it is implicit that δ refers to the absolute difference, $|p_{11} - p_{21}|$. Denoting $\sigma = p_{11} + p_{21}$, we must have $\sigma \in [\delta,2 - \delta]$. Simplifying equation (4) in terms of δ and σ, we obtain (fig. 1*A*)

$$I_n(Q;J) = -\frac{1}{2} \log [\sigma^{\sigma}(2 - \sigma)^{2-\sigma}] + \frac{1}{4} \log [(\sigma + \delta)^{\sigma+\delta}(2 - \sigma - \delta)^{2-\sigma-\delta} \times (\sigma - \delta)^{\sigma-\delta}(2 - \sigma + \delta)^{2-\sigma+\delta}] . \qquad (7)$$

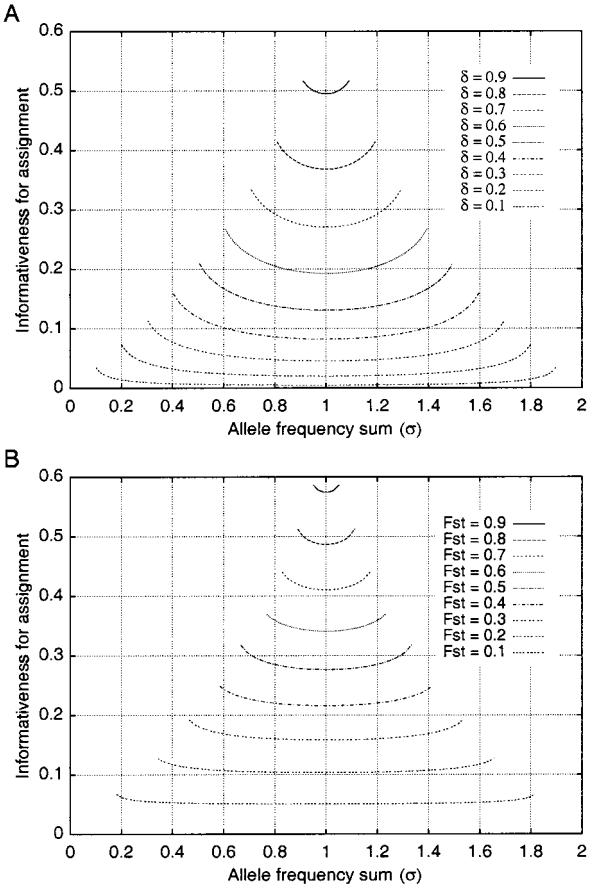

Figure 1 Relationship of informativeness for assignment (I_n) to δ (*A*), and F_{st} (*B*). The plots are based on two alleles and two source populations, and they use equations (7) and (8).

As it should be, I_n is invariant with respect to a transposition of alleles ($\delta \to -\delta$ and $\sigma \to 2 - \sigma$), of populations ($\delta \to -\delta$), or of both alleles and populations ($\sigma \to 2 - \sigma$).

For $K = 2$ and $N = 2$, F_{st} for a locus (henceforth used interchangeably with F), can be written as (modified from p. 167 of Weir 1996)

$$F = \frac{\delta^2}{\sigma(2 - \sigma)} \ . \tag{8}$$

Solving equation (8) for δ and using equation (7), we can express I_n in terms of F and σ (fig. 1*B*).

For a fixed value of δ or F, a biallelic marker is best able to infer ancestry if one of its alleles is absent in one of the populations: if the random genotype J equals this allele, there is no uncertainty about the origin of an individual. In other words, for a fixed δ (Campbell et al. 2003) or F, the markers with the greatest ability to infer ancestry have a value of σ near one of the extremes. The statistic I_n captures this aspect of ancestry inference ability (fig. 1; table 2): for a fixed δ, informativeness declines from its maximum at $\sigma = \delta$ to its minimum at $\sigma = 1$ and then climbs to a second maximum at $\sigma = 2 - \delta$; for a fixed F, the minimal informativeness is at $\sigma = 1$, and the maxima are at $\sigma = 2F/(1 + F)$ and at $\sigma = 2/(1 + F)$.

A comparison of figure 1*A* and 1*B* demonstrates that informativeness varies less across values of σ for a fixed F than for a fixed δ; thus, I_n is more closely related to F than to δ. By considering the difference between the maximal and minimal informativeness over values of σ (table 2), it can be shown that the value of F predicts the value of I_n to within 0.0417, whereas δ predicts I_n only to within 0.0849. The mean difference between the upper and lower bounds on I_n, given the value of F, is 0.0282; the corresponding mean difference between the upper and lower bounds on I_n, given δ, is more than twice as large, equaling 0.0569 (fig. 2*A* and 2*B*).

An additional consequence of equation (8) and the requirement that $\sigma \in [\delta, 2 - \delta]$ is that δ can be used to predict F fairly accurately, and vice versa (fig. 2*C*). This is useful in cases in which only one of these two measures has been reported. Given δ and allowing σ to vary over $[\delta, 2 - \delta]$, F ranges from a minimum of δ^2, when $\sigma = 1$, to a maximum of $\delta/(2 - \delta)$, when $\sigma = \delta$ or $\sigma = 2 - \delta$. Thus, either δ^2 or $\delta/(2 - \delta)$ can be regarded as a substitute for F; for any values of δ and σ, it can be shown that $\delta/(2 - \delta) - 0.0902 \leq F \leq \delta^2 + 0.0902$. The maximal discrepancy between the two approximations, 0.0902, is attained when $\delta \approx 0.3820$ (table 2). The mean of the two bounds, or $(\delta + 2\delta^2 - \delta^3)/(4 - 2\delta)$, is always within 0.0451 of F. The accuracy of such simple approximations as $F \approx \delta^2$ and $F \approx \delta/(2 - \delta)$ is perhaps somewhat surprising.

Predictions of δ from F are slightly less accurate than the reverse predictions. Given F, as σ ranges over $[2F/(1 + F), 2/(1 + F)]$, δ ranges from a minimum of $2F/(1 + F)$ at $\sigma = 1$ to a maximum of $\sqrt{F}$ at $\sigma = 2F/(1 + F)$ or $\sigma = 2/(1 + F)$. The maximal width of this range of δ, given F, is 0.1349, and it is attained at $F \approx 0.0874$.

Optimal Rate of Correct Assignment

If we use the no-admixture model, another way to measure marker information content is to pursue a decision-theoretic approach. We adopt an assignment rule in which observing that one of the two alleles of a random individual is j leads to assignment to population i with probability d_{ij}, where $\sum_{i=1}^{K} d_{ij} = 1$ for each j. We also choose a cost function $c_{ii'}$, which gives the penalty for assignment to population i' when the correct population is i. Under the assumption of a uniform prior on the population of origin, or $\mathbb{P}(Q = i) = 1/K$ for all i, the aim is to choose an assignment rule, or a set of values of d_{ij}, that minimizes the expected value of the loss (Weiss 1961, p. 69), or

$$r(d) = \sum_{i=1}^{K} \sum_{j=1}^{N} \left(\frac{p_{ij}}{K} \sum_{i'=1}^{K} d_{i'j} c_{ii'} \right) . \tag{9}$$

If $c_{ii'}$ is taken to equal 0 when $i = i'$ and 1 otherwise, minimizing equation (9) is equivalent to maximizing the probability of correct assignment, or $\sum_{j=1}^{N} \sum_{i=1}^{K} (p_{ij}/K) d_{ij}$. The *optimal rate of correct assignment* (ORCA) is the probability of correct assignment when the optimal rule is used. To determine this rule, note that, for allele j, the maximum of the linear function $\sum_{i=1}^{K} (p_{ij}/K) d_{ij}$ over the set $\{d_{1j}, d_{2j}, \ldots, d_{Kj} : d_{ij} \geq 0, \sum_{i=1}^{K} d_{ij} = 1\}$ must occur at one of the vertices of the set. Consequently, this maximum occurs when each allele is always assigned to the population in which it is most frequent, and it equals $\max_{i \in \{1,2,\ldots,K\}} p_{ij}/K$. Adding across alleles, we obtain

$$\text{ORCA} = \sum_{j=1}^{N} \max_{i \in \{1,2,\ldots,K\}} \frac{p_{ij}}{K} \ . \tag{10}$$

Similarly to I_n, the minimal value of ORCA, $1/K$, occurs when all alleles have equal frequencies in all populations, and the maximal value, 1, occurs when $N \geq K$ and no allele is found in more than one population. Also similarly to I_n, general prior assignment probabilities, $\mathbb{P}(Q = i) = q_i$, can be accommodated by replacing $1/K$ with q_i in equations (9) and (10).

Table 2

Relationship of informativeness for assignment (I_n), δ, and F_{st}

Quantity (A)	Expression in Terms of σ and Other Quantity (B)	Allowable Values of σ in Terms of A	Allowable Values of A in Terms of B	Maximal Width of the Range of A Given the Value of B	Minimal Informativeness for Assignment in Terms of σ and A	Maximal Informativeness for Assignment in Terms of σ and A	Maximal Width of the Range of Informativeness Given the Value of A
δ	$\sqrt{\sigma(2-\sigma)F}$	$[\delta, 2-\delta]$	$\left[\frac{2F}{1+F}, \sqrt{F}\right]$	.1349 at $F \approx .0874$	$\frac{1}{2}\log\left[(1+\delta)^{1+\delta}(1-\delta)^{1-\delta}\right]$ at $\sigma = 1$	$\frac{1}{2}\log\left[\frac{4(1-\delta)^{1-\delta}}{(2-\delta)^{2-\delta}}\right]$ at $\sigma = \delta$ or $\sigma = 2-\delta$	$\log\left(\frac{4\sqrt{6}}{9}\right) \approx .0849$ at $\delta = 1/2$
F	$\frac{\delta^2}{\sigma(2-\sigma)}$	$\left[\frac{2F}{1+F}, \frac{2}{1+F}\right]$	$\left[\delta^2, \frac{\delta}{2-\delta}\right]$	$5\phi - 8 \approx .0902$ at $\delta = 2-\phi \approx .3820$	$\frac{1}{2}\log\left[(1+\sqrt{F})^{1+\sqrt{F}}(1-\sqrt{F})^{1-\sqrt{F}}\right]$ at $\sigma = 1$	$\frac{1}{2}\log\left[4^{F/(1+F)}(1-F)^{(1-F)/(1+F)}(1+F)\right]$ at $\sigma = \frac{2F}{1+F}$ or $\sigma = \frac{2}{1+F}$	.0417 at $F \approx .4772$

NOTE.—In the upper row of the table, $A = \delta$ and $B = F$. In the lower row, $A = F$ and $B = \delta$. $\phi = (1+\sqrt{5})/2$.

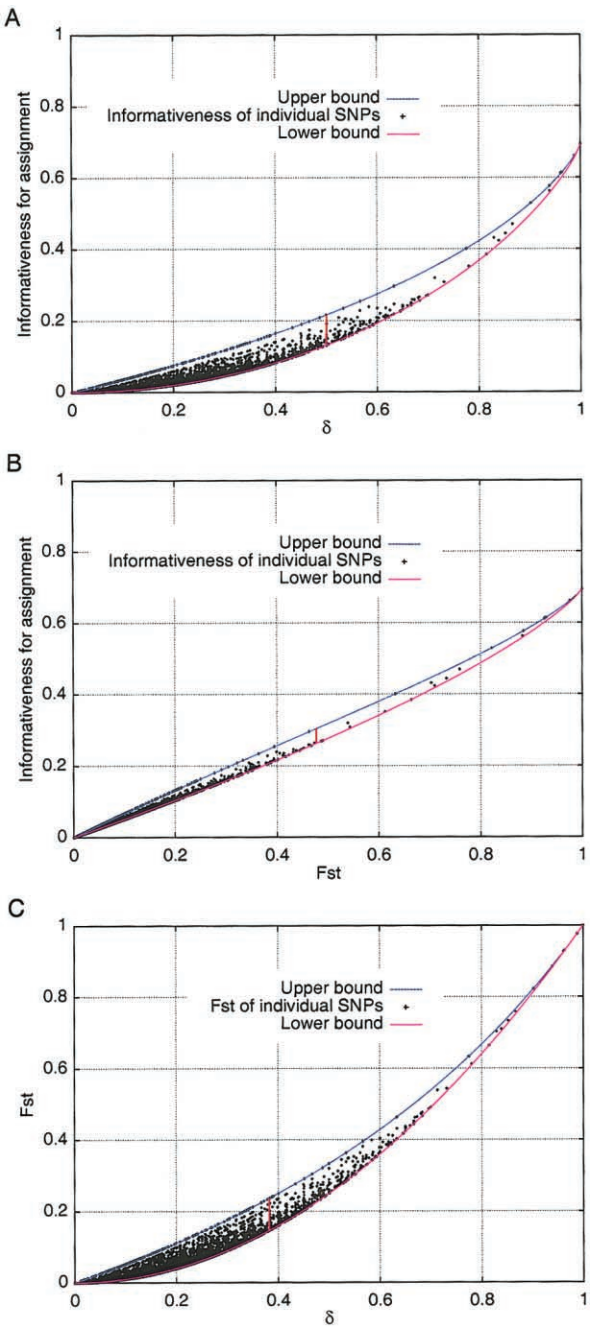

Figure 2 Informativeness for assignment (I_n), δ, and F_{st} for 8,714 SNPs, based on allele frequency estimates in African Americans and European Americans. *A*, I_n vs. δ. *B*, I_n vs. F_{st}. *C*, F_{st} vs. δ. Upper and lower bounds for the dependent variable, given the independent variable, are taken from table 2. A red vertical line marks the point of greatest difference between the upper and lower curves. Mean differences between upper and lower curves are $3/4 - \log 2 \approx 0.0569$ (*A*), $[16 \log 2 + 2 - 6(\log 2)^2 - \pi^2]/12 \approx 0.0282$ (*B*), and $2 \log 2 - 4/3 \approx 0.0530$ (*C*). Spearman rank correlation coefficients between the variables are 0.921, 0.998, and 0.943 in A, B, and C, respectively.

Note that, for $K = N = 2$ with $p_{11} \geqslant p_{21}$, we obtain $(p_{11} + p_{22})/2$ for ORCA, or

$$\text{ORCA} = \frac{1 + \delta}{2} . \tag{11}$$

Similarly to I_n, ORCA can be extended for evaluation of sets of many loci. Because the maximal correct assignment probability when (Hardy-Weinberg) diploid genotypes rather than individual alleles are assigned to populations equals $\sum_{j_1=1}^{N} \sum_{j_2=1}^{N} \max_{i \in \{1,2,\ldots,K\}} (1/K) p_{ij_1} p_{ij_2}$, when multilocus diploid genotypes (at loci that are independent within populations) are assigned, the corresponding probability is

$$\text{ORCA} = \sum_{j_1^{(1)}=1}^{N^{(1)}} \sum_{j_2^{(1)}=1}^{N^{(1)}} \sum_{j_1^{(2)}=1}^{N^{(2)}} \sum_{j_2^{(2)}=1}^{N^{(2)}} \ldots \sum_{j_1^{(L)}=1}^{N^{(L)}} \sum_{j_2^{(L)}=1}^{N^{(L)}} \max_{i \in \{1,2,\ldots,K\}} \frac{1}{K} \prod_{l=1}^{L} p_{ij_1^{(l)}}^{(l)} p_{ij_2^{(l)}}^{(l)} . \tag{12}$$

Equation (12) can, in principle, predict the probabilities of correct assignment of sets of one or more loci in procedures (Buchanan et al. 1994; Paetkau et al. 1995; Banks et al. 2003) that assign multilocus genotypes to their most likely source populations.

Informativeness for Ancestry Coefficients: the Admixture Model

We have introduced two new measures that can facilitate assignment of individuals to populations. Often, however, a goal of ancestry inference is to estimate "ancestry coefficients" for an individual whose ancestry is from two or more populations (Rannala and Mountain 1997; Pritchard et al. 2000; Anderson and Thompson 2002). Such an individual has a vector of K ancestry coefficients that sum to 1, where the coefficient for population i gives the fraction of the individual's genome that derives from population i. Ancestry is now a random vector $\mathbf{Q}$ rather than a discrete random variable. The mutual information $I_a(\mathbf{Q};J)$ that quantifies the amount of information about $\mathbf{Q}$ provided by knowledge of J is of interest. With $\mathbf{Q} = (Q_1, Q_2, \ldots, Q_K)$, where Q_i is the (random) ancestry coefficient for the ith population and $\sum_{i=1}^{K} Q_i = 1$, we have

$$\Pr(J = j | \mathbf{Q} = \mathbf{q}) = \sum_{i=1}^{K} p_{ij} q_i . \tag{13}$$

Similarly to the no-admixture case, any assumptions could be made for the initial probability distribution of $\mathbf{Q}$. That is, any distribution defined on the set $\{\mathbf{Q}: Q_i \geqslant 0, \sum_{i=1}^{K} Q_i = 1\}$ is suitable. For simplicity, we assume that this distribution is uniform: all collections of ancestry coefficients that sum to 1 are a priori equally likely.

Using the definition of mutual information for continuous random variables (Cover and Thomas 1991, chapter 9), $I_a(\mathbf{Q};J)$ equals (see the appendix)

$$I_a(\mathbf{Q};J) = \sum_{j=1}^{N}\left[p_j\left(-\log p_j + 1 - \frac{|S_{K+1}^{(2)}|}{K!}\right) + \sum_{i=1}^{K}\left(\frac{p_{ij}^K \log p_{ij}}{K\prod_{\substack{i'=1\\ i'\neq i}}^{K}(p_{ij}-p_{i'j})}\right)\right], \quad (14)$$

where $S_{K+1}^{(2)}$ is a Stirling number of the first kind (Abramowitz and Stegun 1965, p. 833) (the first six values of $|S_{K+1}^{(2)}|$, starting at $K = 2$, are 3, 11, 50, 274, 1,764, and 13,068). The quantity I_a is termed the "*informativeness for ancestry coefficients.*" I_a has a similar multilocus extension to that of I_n, and, for $K = N = 2$, relationships of I_a to δ and F are qualitatively similar to the corresponding relationships of I_n to δ and F (not shown). For example, for fixed δ, the maxima of I_a over σ occur at $\sigma = \delta$ and $\sigma = 2 - \delta$, and the minimum is at $\sigma = 1$.

Number of Markers

The statistic I_a suggests a way to prioritize markers for use in inference of ancestry coefficients, but it does not have a simple relationship with the number of markers needed for this inference. However, using a maximum-likelihood approach, it is possible to approximate this number of markers. Equation (13) gives the likelihood of ancestry coefficients $(q_1,q_2,\ldots,q_{K-1})$ in a haploid one-locus model, with $q_K = 1 - q_1 - \ldots - q_{K-1}$. The expected Fisher information matrix, $\mathbf{U}$, for the likelihood function has dimensions $(K-1) \times (K-1)$ and, for each i and i', the (i,i')th element equals (Millar 1991, eq. [A.3])

$$\mathbf{U}_{ii'}(q_1,q_2,\ldots,q_{K-1}) = \sum_{j=1}^{N}\frac{(p_{ij}-p_{Kj})(p_{i'j}-p_{Kj})}{p_{Kj} + \sum_{m=1}^{K-1}(p_{mj}-p_{Kj})q_m}. \quad (15)$$

Multiplying equation (15) by 2, we can obtain the corresponding value for (Hardy-Weinberg) diploids. When standard maximum-likelihood theory is used (Elandt-Johnson 1971), the variance-covariance matrix of the ancestry coefficient maximum-likelihood estimates is approximated by $\mathbf{U}^{-1}$. For $K \geq 3$, a straightforward transformation of $\mathbf{U}$ can enable inclusion of q_K in the matrix (Millar 1991); for $K = 2$, the maximum-likelihood estimates $\hat{q}_1$ and $\hat{q}_2$ have equal variances.

Using equation (15), in the (diploid) case of $K = 2$, we obtain

$$\mathrm{Var}(\hat{q}_1) = \left[\sum_{j=1}^{N}\frac{2(p_{1j}-p_{2j})^2}{p_{2j} + q_1(p_{1j}-p_{2j})}\right]^{-1}. \quad (16)$$

For biallelic markers ($N = 2$), equation (16) reduces to

$$\mathrm{Var}(\hat{q}_1) = \frac{(p_{21} + q_1\delta)(1 - p_{21} - q_1\delta)}{2\delta^2}. \quad (17)$$

If we consider all possible values of q_1 and assume that $p_{11} \geq p_{21}$, the largest value of equation (17) occurs at $q_1 = (1 - 2p_{21})/(2\delta)$, producing an upper bound for the approximate variance equal to

$$\mathrm{Var}(\hat{q}_1) = \frac{1}{8\delta^2}. \quad (18)$$

Because the information matrix for a set of loci that are independent within populations is the sum of the matrices for the individual loci, the number of independent markers, all with the same value δ, that are required to achieve $\mathrm{Var}(\hat{q}_1) = V$, is

$$L = \frac{1}{8\delta^2 V}. \quad (19)$$

Using equation (19), 35 biallelic markers with $\delta = 0.6$ are necessary for achieving an SD of 0.1, in agreement with a previous suggestion of ~40 such markers (Hoggart et al. 2003).

The number of independent biallelic markers required for accurate estimation of ancestry coefficients in the two-population admixture model (table 3) is considerably larger than the number required for assignment in corresponding no-admixture models (Risch et al. 2002; Campbell et al. 2003). However, our computations assume that estimation of ancestry coefficients occurs by maximum likelihood; other estimation procedures or use of dependencies between markers might reduce the number of markers needed. In addition, since it is based on an upper bound for the variance and not on the general

Table 3

Number of Biallelic Markers Required for Achieving a Specified Standard Deviation in Ancestry Coefficient Estimates for a Two-Population Model (eq. [19])

	No. of Markers Required for an SD of			
δ	.2	.1	.05	.01
.9	4	16	62	1,544
.8	5	20	79	1,954
.7	7	26	103	2,552
.6	9	35	139	3,473
.5	13	50	200	5,000
.4	20	79	313	7,813
.3	35	139	556	13,889
.2	79	313	1,250	31,250
.1	313	1,250	5,000	125,000

Table 4

Data Sets and Spearman Rank Correlation Coefficients of I_n, ORCA, and I_a

Data Set	Description of Groups	Correlation Coefficient		
		I_n and *ORCA*	I_n and I_a	ORCA and I_a
World-52	52 populations representing seven regions	.920	...	...
World-5	5 regional groups (Africa, Eurasia, East Asia, Oceania, America)	.956	.994	.956
World-7	7 regional groups (Africa, Europe, Middle East, Central/South Asia, East Asia, Oceania, America)	.953	.990	.947
Africa	6 populations (Bantu [Kenya], Mandenka, Yoruba, San, Mbuti Pygmy, Biaka Pygmy)	.945	.986	.937
Europe	8 populations (Orcadian, Adygei, Russian, Basque, French, Italian, Sardinian, Tuscan)	.878	.973	.867
Middle East	4 populations (Mozabite, Bedouin, Druze, Palestinian)	.873	.994	.891
Central/South Asia	9 populations (Balochi, Brahui, Makrani, Sindhi, Pathan, Burusho, Hazara, Uygur, Kalash)	.883	...	...
East Asia	18 populations (Han, Han [N. China], Dai, Daur, Hezhen, Lahu, Miao, Oroqen, She, Tujia, Tu, Xibo, Yi, Mongola, Naxi, Cambodian, Japanese, Yakut)	.915	...	...
Oceania	2 populations (Melanesian, Papuan)	.921	.998	.940
America	5 populations (Karitiana, Surui, Colombian, Maya, Pima)	.934	.989	.945

expression, equation (19) might further overestimate the number of markers needed. Note that, because only the upper bound and not the general expression is directly related to δ, markers with high values of I_n and I_a rather than of δ might often produce smaller variances at values of q_1 relevant to individuals under consideration.

Estimation of Informativeness

I_n, ORCA, and I_a have been defined parametrically, as inherent properties of a marker together with a set of populations. In practice, however, estimates made from data rather than parametric allele frequencies must be used. For a given locus, let the number of copies of allele j observed at the locus in population i equal n_{ij}, and let the total number of observations in population i equal n_i. A simple estimator of informativeness statistics is the count estimate, in which p_{ij} is estimated by n_{ij}/n_i, and the estimated $\hat{p}_{ij}$ values are inserted in place of the parametric values.

This estimator can produce biased estimates; consider, for example, two samples taken from the same population. Since allele counts in the two samples are likely to differ by chance, markers will have positive estimated I_n for distinguishing the two samples and estimated ORCA >1/2 when parametric informativeness equals 0 and parametric ORCA is 1/2. This bias is not of major concern when the goal is to compare informativeness estimates for different loci through use of the same sample (Brenner 1998), since a systematic bias affects all loci in a similar manner. In addition, in comparisons of locus informativeness across different samples, sample-specific biases should affect all loci similarly, and the relationship between informativeness estimates of a locus in two samples is preserved even if the estimates are biased.

I_n, ORCA, and I_a have been defined using frequencies of alleles rather than of diploid genotypes. If alleles within individuals are not independent, so that within-population genotype frequencies do not correspond to Hardy-Weinberg proportions, the definitions can be applied treating J as a random diploid genotype, and the count estimates of genotype frequencies can be used in estimation. We do not consider this issue further.

Data

We consider various subsets (table 4) of a data set of 377 microsatellite markers—45 dinucleotides, 58 trinucleotides, and 274 tetranucleotides—genotyped in 1,056 individuals from 52 human populations (Cann et al. 2002; Rosenberg et al. 2002; Zhivotovsky et al. 2003; Human Diversity Panel Genotypes Web site; Human STRP Screening Sets Web site). Names of regions and regional affiliations of populations are the same as in the article by Rosenberg et al. (2002). At least 50 of the 377 markers are among those reported in table 1 of the article by Collins-Schramm et al. (2002), and at least 212 of them are included in table 1 of the article by Smith et al. (2001).

To compare informativeness for microsatellites and SNPs, we consider SNPs that have been studied in African Americans, European Americans, and East Asians and that were found to have few enough errors for use in analysis of population divergence (Akey et al. 2002; Joshua Akey's Homepage). From the Akey et al. (2002) data, we exclude several types of SNPs: (1) SNPs ge-

notyped by the Whitehead Institute, whose European American sample differed from that used by the other genotyping centers; (2) SNPs with unknown sample sizes or with sample size <40 (20 individuals) in at least one of the three groups; (3) SNPs with unknown, nonunique, or nonautosomal map positions; and (4) SNPs whose frequencies were obtained by DNA pooling or for which one or more of the reported allele frequencies could not be expressed as a rounded quotient of an integer and the reported sample size. The 8,714 SNPs we use were all genotyped by Celera, Motorola, or Orchid.

Statistical Properties of Informativeness

In this section, we demonstrate that the proposed informativeness statistics are, indeed, useful measures. First, we show that the statistics I_n, ORCA, and I_a produce similar estimates, so that we can proceed using only one statistic, the *informativeness for assignment*. Second, we demonstrate that the I_n statistic is robust, in that rankings of locus informativeness do not vary greatly across resamples of the data. Third, we show that the I_n statistic does indeed measure ability to infer ancestry, in that population structure inference using markers of high informativeness requires fewer markers than inference using markers of low informativeness.

Relationship between I_n, ORCA, and I_a

For each of the data sets in table 4, we computed I_n, ORCA, and I_a for each of the loci, using the allele count estimates with equations (4), (10), and (14). For each data set, Spearman rank correlation coefficients (Gibbons 1985, p. 226) of locus I_n and ORCA, I_n and I_a, and ORCA and I_a values were computed. Loci for which two or more populations had an identical allele frequency estimate were not used in the latter two calculations, to avoid obtaining denominators of 0 in the computation of I_a. For the World-52, Central/South Asia, and East Asia data sets, in which many populations had the same sample sizes and therefore had numerous opportunities to produce equal allele frequency estimates in two or more populations, there were many such loci, and the correlation coefficients involving I_a were not computed.

Rankings of loci by I_n, ORCA, and I_a were all highly correlated, with the largest correlations observed between I_n and I_a (table 4). Thus, for convenience, in the remainder of this article we restrict attention to estimates of I_n, or, simply, the *informativeness,* and we assume that all three measures have similar properties.

Locus Informativeness Rankings

For each of the 10 data sets (table 4), markers were ranked from highest to lowest estimated informativeness (table A [online only]). To assess the robustness of these rankings, or the extent to which they are affected by the particular choice of individuals included in the data, we performed bootstrap replicates.

Individuals were resampled with replacement within groups, holding group sample sizes fixed. For each replicate, informativeness was estimated for each locus, and loci were ranked by estimated informativeness. In some replicates, for at least one group and one locus, the resample included only individuals who did not have genotypes at the locus. This situation arose only for data sets in which some groups had small sample sizes ($\leqslant$10). In these data sets, it was possible for a resample to consist solely of copies of a few individuals. Thus, if these few individuals had no data at a locus, the resample also had no data. For each data set, excluding these replicates, which were discarded, 1,000 resamples were performed. For data sets in which all groups had larger sample sizes (>10), it was not necessary to discard any replicates.

To assess the variability of I_n values across bootstrap replicates, for each locus, we computed the ratio of the SD of the bootstrap values of I_n to the value estimated from the data, and we averaged this quantity across loci. Three statistics were used to compare the locus informativeness rankings that were estimated from the data and those that were obtained in the bootstrap replicates: (1) the mean across replicates of $R_{\mathbf{V}_d,\mathbf{V}_b}$, where $\mathbf{V}_d$ denotes the vector of informativeness ranks based on the data, $\mathbf{V}_b$ denotes the vector of ranks based on the bth bootstrap replicate, and R denotes the Spearman rank correlation coefficient; (2) the Kendall coefficient of concordance of the 1,000 bootstrap replicates (Gibbons 1985, p. 250); (3) the mean across loci of the mean across replicates of the absolute deviation between the rank of a locus in the data and its rank in the replicate. This third statistic was also computed using only the 50 loci of highest estimated informativeness.

Although informativeness fluctuated noticeably across replicates for individual loci, rankings in different replicates were highly concordant with each other and were highly correlated with the rankings based on the estimates from the data (table 5). Similar patterns of correlation across bootstrap replicates were observed for all of the data sets. The World-52, World-5, and World-7 data sets, which contained the most data, produced the most robust informativeness ranks and values; the least robust were found for Oceania, the smallest data set.

The fluctuation of ranks of individual loci across replicates indicates that exact ranks of loci (such as in tables A, B, C, D, and E [online only]) should be regarded with

Table 5

Robustness of Informativeness Statistics

Data Set	No. of Replicates Discarded	Mean Ratio of Bootstrap SD to Estimated I_n Value	Mean ± SD Spearman Rank Correlation Coefficient	Kendall Coefficient of Concordance of Ranks	Mean ± SD Absolute Deviation of Ranks	Mean ± SD Absolute Deviation of Ranks (50 Loci of Highest I_n)
World-52	1,460	.064	.977 ± .002	.974	17.53 ± 8.27	7.38 ± 4.38
World-5	0	.119	.978 ± .003	.959	17.24 ± 5.60	8.57 ± 4.42
World-7	0	.106	.980 ± .003	.964	16.44 ± 5.30	8.19 ± 4.09
Africa	27	.215	.909 ± .018	.852	34.75 ± 12.57	16.13 ± 9.06
Europe	514	.274	.845 ± .021	.770	45.39 ± 17.40	17.81 ± 10.92
Middle East	0	.317	.839 ± .020	.734	46.35 ± 15.18	22.51 ± 11.47
Central/South Asia	108	.223	.880 ± .017	.818	39.71 ± 15.32	16.80 ± 12.27
East Asia	392	.170	.919 ± .009	.883	32.47 ± 13.95	13.75 ± 10.13
Oceania	0	.683	.836 ± .034	.723	46.45 ± 15.72	21.44 ± 9.74
America	0	.180	.935 ± .012	.883	29.04 ± 10.81	13.56 ± 6.28

NOTE.—See the "Locus Informativeness Rankings" subsection of the "Statistical Properties of Informativeness" section for a description of the quantities in the table.

caution. However, fluctuations were small enough that the markers of highest informativeness usually had low ranks in bootstrap replicates (rightmost column of table 5). Thus, confidence can be placed in general statements such as a locus being "among the most informative markers" in a data set.

Performance of Markers of High Informativeness in Ancestry Inference

One way to test the utility of I_n as a measurement of the ability of a marker to infer ancestry is to check whether the population structure inferred using the markers of highest I_n more closely approximates the population structure inferred using all of the markers than does the population structure inferred using the markers of lowest I_n. Using the computer program *structure* (Pritchard et al. 2000; available from the Pritchard Lab Web site), with five clusters and the full data of 1,056 individuals, we previously found that the genetically inferred population structure corresponded fairly closely to the five regions in the World-5 data set (Rosenberg et al. 2002). Thus, if I_n indeed measures ability to infer ancestry, informativeness of a locus in the World-5 data set should correlate well with the contribution of the locus to population structure inference using five clusters.

We therefore ran *structure* with all 1,056 individuals in the data, using the markers of highest informativeness for the World-5 data set. For various choices of the number of markers, M, five *structure* runs were performed with the M markers of highest I_n, and five runs were performed with the M most heterozygous markers (table S4 of Rosenberg et al. [2002]). Expected heterozygosity was used for comparison, because, among several statistics studied in a previous analysis (Rosenberg et al. 2001), it produced the greatest reduction in the number of markers needed for inference. One run was performed with each of 20 random sets of M markers; for each value of M, random sets were chosen independently of the sets that were selected for the other values of M. Five runs were performed with the M markers of lowest I_n. All *structure* runs used five clusters, and, as in the study by Rosenberg et al. (2002), they employed the admixture model for individual ancestry (Pritchard et al. 2000), the F model for allele frequency correlations (Falush et al. 2003), and a burn-in period of length 20,000 followed by 10,000 iterations.

The similarity coefficient C (Rosenberg et al. 2002) was used to compare runs with subsets of the markers against 10 runs that employed all 377 markers and were performed by Rosenberg et al. (2002). As in that study, the normalization required in the computation of C was based on the runs that used all of the markers. For each value of M, $M < 377$, each of the 10 runs that used all 377 markers was compared with each of the five runs that used the M markers of highest informativeness, for a total of 50 comparisons. For $M = 377$, the 90 pairwise comparisons of the 10 full-data runs were performed. For each M, the first quartile, median, and third quartile of the distribution of the 50 values were obtained (90 values for $M = 377$). Comparisons to the full-data runs were made in an analogous manner, using the runs based on the least informative, most heterozygous, and random markers. For the random markers and $M < 377$, the similarity coefficient distribution was based on 200 comparisons.

Figure 3 indicates that, in general, fewer loci chosen according to the highest informativeness were required than random loci for inferring a population structure similar to that obtained with all the loci. This pattern was observed especially for small and intermediate values of M; although similarity coefficients at these M often varied considerably across runs, runs based on the markers of highest I_n generally produced greater similarity coefficients than those based on random or highly heterozygous markers. For larger values of M, the difference in similarity coefficients across criteria was less pronounced, partly because the sets of markers chosen by different criteria had greater overlap than for small values of M. However, runs that used the markers of lowest I_n produced similarity coefficients that were considerably smaller than those obtained by the other sets of markers. Many more of the markers of lowest I_n than of those of highest I_n were required to obtain inferred population structures that were visually similar to that inferred using the full data (fig. 4). Thus, high informativeness is a useful indicator of the ability of a marker to infer ancestry; more dramatically, low informativeness suggests that a locus is not of great utility for inference of ancestry.

Comparison of Rankings across Data Sets

For pairs of data sets, we computed correlation coefficients of locus informativeness (table 6). Most pairs of rankings had correlations of at least 0.2. Markers that had high informativeness for inference of regional ancestry tended to be informative for inference within several regions. One exception was that informativeness in the America data set was not correlated with informativeness in the World-5 and World-7 data sets.

The highest correlations for pairs of regions occurred for regions that were geographically proximate, such as Central/South Asia and East Asia. All correlations for pairs of regions, among those that included two of Africa, Europe, Middle East, Central/South Asia, and East Asia, were larger than correlations that involved Oceania or America. The smallest correlation for a pair of regions was between informativeness in Africa and in-

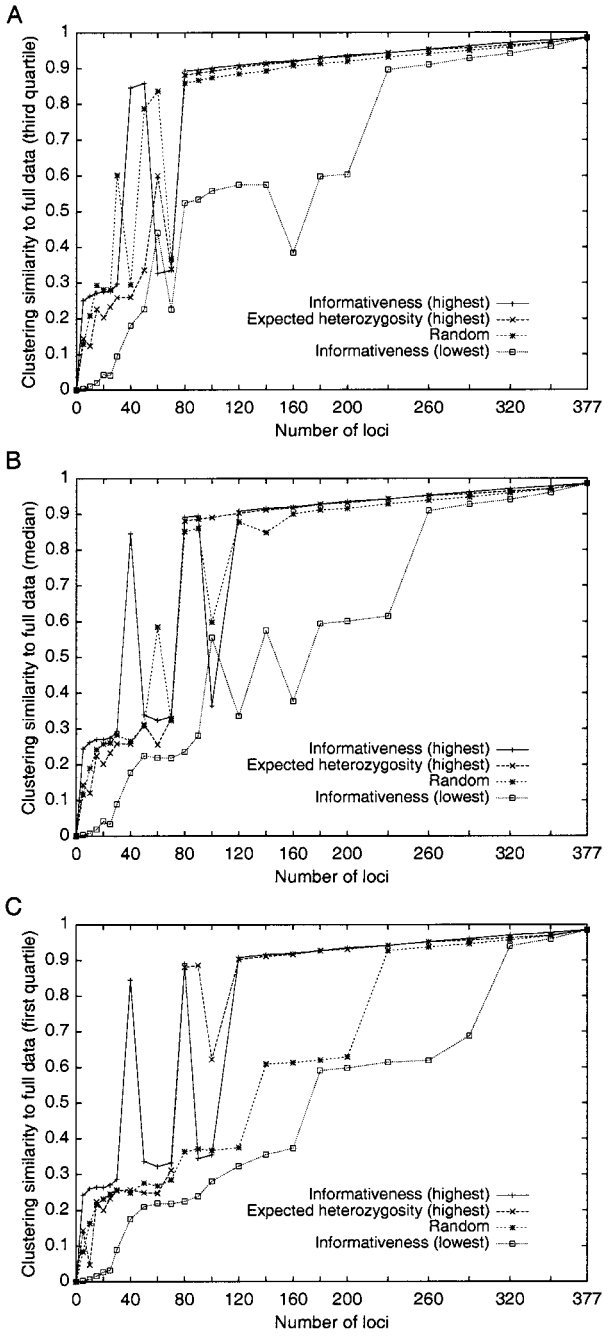

Figure 3 Similarity coefficients for runs based on reduced sets of markers and runs based on the full data. Sets of markers were chosen with each of four methods: highest informativeness, highest expected heterozygosity, random, and lowest informativeness.

formativeness in America. Larger absolute levels of informativeness in Africa, America, and Oceania (fig. 5) are consistent with the greater observed differentiation among populations in these regions (Rosenberg et al. 2002).

Most loci were ranked poorly in at least one data set (table A [online only]). D11S2000, D16S3401, D16S422, D21S2055, and D3S2427 were the only markers to rank among the 75 most informative in all data sets; note that D21S2055 was one of three loci identified by Zhivotovsky et al. (2003) as unusually variable. D13S285 and D7S1804 were highly informative in all seven regional data sets (rank $\leq$75) but were less informative in at least one of the three worldwide data sets (rank >75). Conversely, D14S1007, D1S235, D22S683, D2S1356, D8S560, D9S1779, D9S1871, NA-D18S-2, and NA-D5S-1 were highly informative in the worldwide data sets (rank $\leq$25) but were less informative in one or more of the regional data sets (rank >75).

Microsatellites and SNPs

Dinucleotide loci, which show the most variation among the markers in these data (Zhivotovsky et al. 2003), were generally more informative than tetranucleotide loci (table 7), consistent with the generally greater differentiation of dinucleotides across human populations (Ruiz Linares 1999; Rosenberg et al. 2003). Dinucleotides were usually also more informative than trinucleotides, but, in many cases, trinucleotides and tetranucleotides had similar levels of informativeness. However, for the worldwide data sets and for Africa, tetranucleotides were by far the least informative class of microsatellite. For example, although 73% of the loci were tetranucleotides, in the World-7 data set, the 25 loci of highest informativeness included only 7 tetranucleotides. Of the 100 loci of lowest informativeness in this data set, 97 were tetranucleotides.

To compare informativeness of microsatellites and SNPs, we determined the informativeness of microsatellites for assignment with three source populations: Africans, Europeans, and East Asians. For these groups, we also determined informativeness for each pairwise combination of source populations (tables B and C [online only]). Similarly, we estimated informativeness of SNPs among African Americans, European Americans, and East Asians. Because the individuals and populations in the microsatellite and SNP data sets were not the same, our comparison of microsatellite and SNP informativeness can only be regarded as approximate. Inclusion of some extremely isolated populations in the microsatellite data but not in the SNP data might exaggerate the relative informativeness of microsatellites. However, this effect might be counteracted by a SNP ascertainment procedure that produced greater divergence across populations than is characteristic of randomly chosen SNP markers (J. Akey, personal communication); the microsatellite data likely show little or

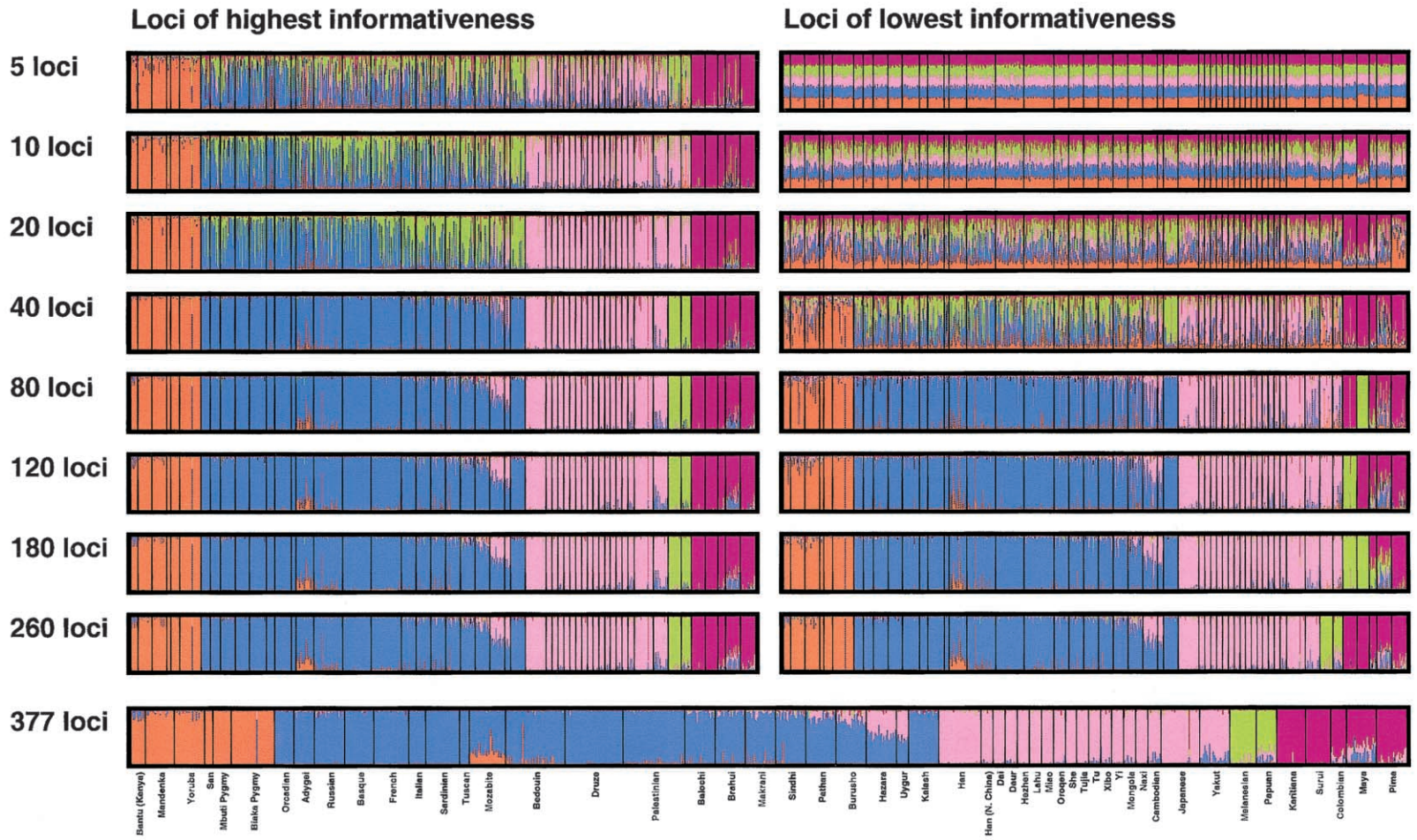

Figure 4 Inferred population structure with five clusters, based on markers of highest and lowest informativeness and plotted using *distruct* (available from Noah Rosenberg's Homepage). Each individual is represented by a thin vertical line, which is partitioned into five colored segments that represent the individual's estimated ancestry coefficients in the five clusters. Black lines separate individuals of different populations, which are labeled below the figure. The left-right order of individuals is the same in all plots. The bottom plot is the same as is shown in figure 1 of Rosenberg et al. (2002); each of the other graphs is based on the highest-likelihood run among five runs with the relevant set of loci.

Table 6

Pearson (above the Diagonal) and Spearman (below the Diagonal) Correlation Coefficients of Locus Informativeness Values for Pairs of Data Sets

	World-52	World-5	World-7	Africa	Europe	Middle East	Central/South Asia	East Asia	Oceania	America
World-52	...	.835	.867	.766	.640	.635	.782	.728	.433	.319
World-5	.860	...	.986	.586	.356	.411	.494	.402	.272	−.007
World-7	.887	.987	...	.605	.380	.431	.526	.422	.278	.024
Africa	.754	.608	.621	...	.505	.521	.598	.521	.317	.180
Europe	.626	.391	.415	.490	...	.500	.600	.607	.310	.387
Middle East	.579	.419	.434	.480	.481	...	.651	.590	.324	.338
Central/South Asia	.746	.508	.536	.542	.582	.570	...	.733	.368	.374
East Asia	.685	.397	.420	.455	.565	.492	.670	...	.402	.444
Oceania	.379	.242	.246	.276	.316	.298	.306	.370	...	.232
America	.286	−.006	.023	.151	.376	.297	.338	.413	.231	...

no such effect (Rosenberg et al. 2002). The small-sample upward bias in informativeness might also impact relative informativeness estimates.

For each set of source populations, randomly chosen microsatellites had greater informativeness than random SNPs (fig. 6). The ratios of median dinucleotide informativeness to median SNP informativeness were 7.8 (Africans vs. Europeans), 6.8 (Africans vs. East Asians), 5.1 (Europeans vs. East Asians), and 5.3 (Africans vs. Europeans vs. East Asians). The ratios of means were 4.3, 3.7, 2.8, and 3.8, respectively, and the 50th percentile of dinucleotide informativeness corresponded to the 96th, 95th, 88th, and 98th percentiles of SNP informativeness.

One threshold proposed for declaring a SNP to be highly informative is $\delta = 0.5$ (Shriver et al. 1997), a value exceeded by 1.9%, 4.6%, and 2.7% of the SNPs (among those polymorphic in the relevant pair of populations) for African Americans and European Americans, African Americans and East Asians, and European Americans and East Asians, respectively. The value $\delta = 0.5$ corresponds to $F_{st} \in [0.250,0.333]$ and $I_n \in [0.131,0.216]$ (table 2); for corresponding comparisons, averaging across the three classes of loci, 26.0%, 42.0%, and 12.4% of microsatellites exceed the lower bound of I_n, and 5.9%, 10.2%, and 1.5% exceed the upper bound.

Discussion

In this article, we have introduced new statistics, I_n, ORCA, and I_a, for measuring the information provided by loci about ancestry. I_n, which is highly correlated with ORCA and I_a (table 4), is robust, in that it gives similar results in bootstrap replicates (table 5). The statistic is effective for inference of ancestry, in that population structure is more easily inferred using markers that have high values of I_n than using those that have low values (figs. 3 and 4). Although it is closely related to δ in the case of biallelic markers in two source populations (figs. 1 and 2; table 2), unlike δ, I_n captures the dependence of information content on the position of allele frequencies in the unit interval.

Use of markers of highest informativeness is desirable for reduction of genotyping effort in such situations as forensics (Shriver et al. 1997; Lowe et al. 2001), admixture mapping (Dean et al. 1994; McKeigue 1998), and structured-association mapping (Pritchard and Donnelly 2001; Hoggart et al. 2003). In these scenarios, it is desirable to maximize information about individual ancestry at minimal cost. For the case of admixture mapping, the additional constraint that loci must be located in candidate regions of the genome applies; unlike other ancestry inference scenarios, admixture mapping makes use not of the ancestry of an individual as a whole but of particular parts of an individual genome. Thus, ideal marker sets for admixture mapping must have representation in regions of interest as well as high informativeness.

Highly informative markers are also useful in testing for population stratification in case-control genetic association studies (Pritchard and Rosenberg 1999), although the test does not use individual ancestry estimates. The goal is to determine whether cases and controls differ in ancestry to such an extent that an excess number of random markers will, by chance, be associated with disease status. Because they have the greatest potential to differentiate among ancestry groups, the most informative markers offer the greatest power to reject the null hypothesis of no genomewide allele-frequency differences between cases and controls; thus, their use offers a cautious approach in dealing with population stratification. If allele-frequency differences are detected, these markers are ideal for structured-association methods that employ individual ancestry estimates to avoid identifying the associations that result from ancestry differences rather than from true association with disease status (Pritchard and Donnelly 2001). The number of these markers

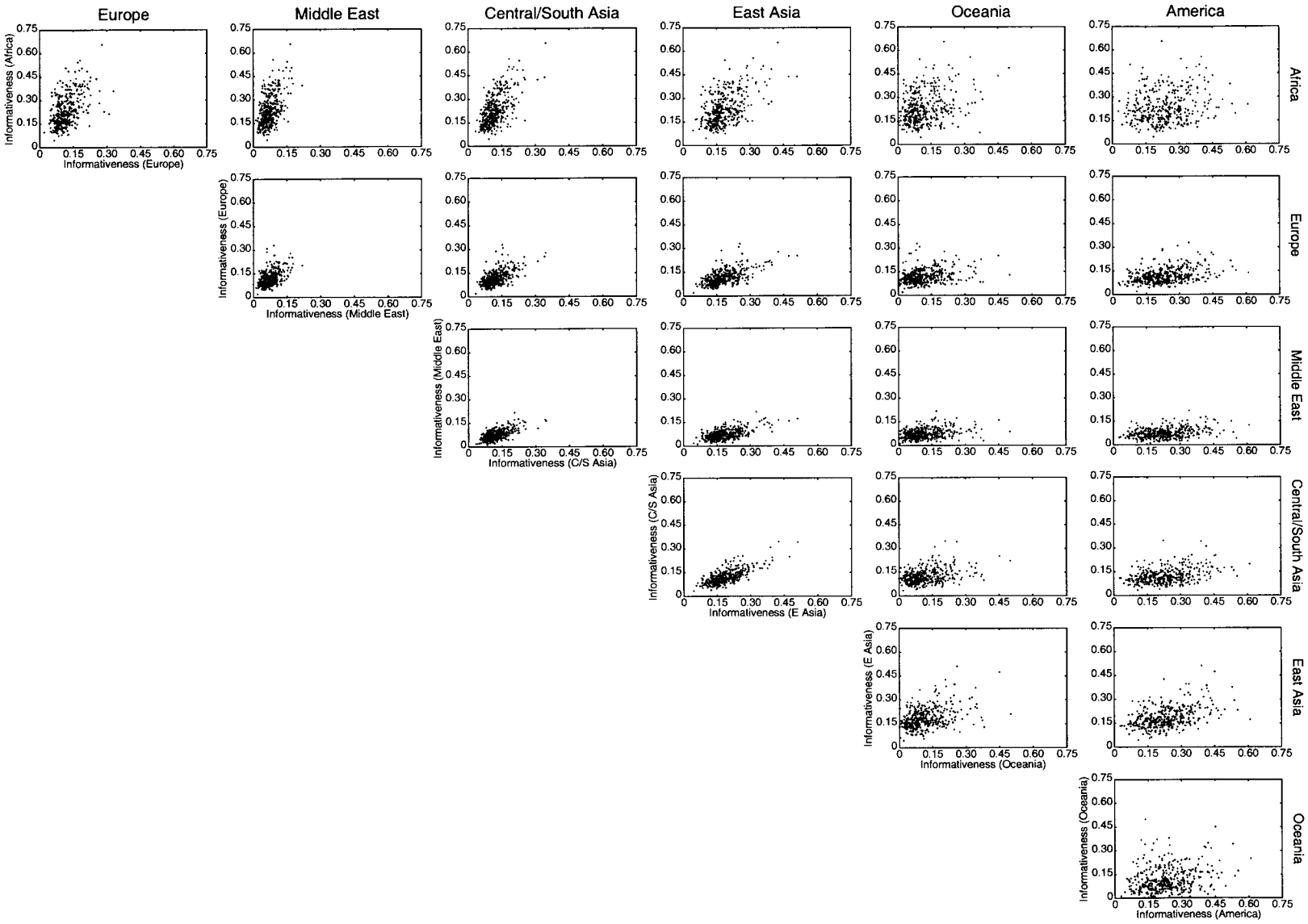

Figure 5 Correlations of informativeness for pairs of regional data sets

Table 7

Informativeness and Repeat Size

	Mean Informativeness Rank			*P* Value (Two-Tailed Mann-Whitney Test)		
Data Set	Dinucleotides	Trinucleotides	Tetranucleotides	Dinucleotides vs. Trinucleotides	Dinucleotides vs. Tetranucleotides	Trinucleotides vs. Tetranucleotides
World-52	94	151	213	<.001	<.001	<.001
World-5	86	119	221	.009	<.001	<.001
World-7	85	119	221	.011	<.001	<.001
Africa	107	161	209	<.001	<.001	.001
Europe	136	201	195	<.001	.001	.790
Middle East	140	189	197	.011	.002	.564
Central/South Asia	124	205	196	<.001	<.001	.603
East Asia	138	206	194	<.001	.002	.466
Oceania	168	189	192	.314	.176	.789
America	185	232	181	.048	.865	<.001

needed for desired precision in estimated ancestry coefficients can be approximated using the maximum-likelihood model.

Consequently, the panels in tables A, B, C, D, and E (online only) can provide a resource for tests of population stratification. For example, in European Americans, the test might use markers that are most informative for distinguishing among various types of European ancestry (table A [online only]); in Hispanic Americans, it might employ markers that are most informative for distinguishing European from Native American ancestry (table B [online only]) or for distinguishing European, Native American, and African ancestry (table C [online only]). Note that panels in tables A–E utilize groups and classifications that might not be identical to those needed in applications: for example, if ancestry inference in African Americans is of interest, the African and European groups in our data do not fully represent the groups from which African Americans have descended. However, we observed that informativeness in one region was often highly correlated with informativeness in another region (table 6; fig. 5). Thus, while the most informative markers in a data set need not be the most informative for use with a different collection of groups, this imperfect panel of markers is likely to be considerably more informative than a random panel. The observed pattern, in which informativeness correlations were highest for neighboring geographic regions, is likely to be a consequence of the correlation of allele frequencies that results from shared ancestry (Ramachandran et al., in press). Populations from neighboring regions typically share ancestors more recently, so that their allele frequencies are more strongly correlated.

Two exceptions to the general pattern of correlation across data sets were Oceania and America, in which informativeness was not very highly correlated with informativeness in other regions. The small correlations likely indicate that many of the markers that are extremely variable in other regions by chance must not have been highly variable in founder groups of Oceania and the Americas. That informativeness patterns across di-, tri-, and tetranucleotides were different in Oceania and America from those of the other data sets suggests that bottlenecks were strong enough to obscure the typical patterns of variation for these three classes of markers.

Thus, to identify a panel of markers that are generally useful for inference of regional ancestry and for population ancestry inference within regions (Hoggart et al. 2003), it is most difficult to find markers that are informative both within continental Eastern Hemisphere regions and within Oceania and the Americas. We have identified a small number of generally informative markers; many more loci will need to be screened if markers that are informative in every region are to be found. Alternatively, a general panel might be assembled by collecting markers useful for inference between specific pairs of groups. Such a procedure may be advantageous, because, unlike sequential accumulation of generally informative markers, it avoids duplication of effort by accounting for the possibility that markers of high informativeness can provide information about ancestry in different ways. A systematic procedure to identify maximally informative sets or loci that are conditionally optimal, given the markers that have already been chosen, might use multilocus I_n, multilocus ORCA (eq. [12]), or a decision tree (Guinand et al. 2002).

Although random microsatellites are considerably more informative than random SNPs for distinguishing among pairs of populations, and highly informative loci constitute a greater fraction of microsatellites than of SNPs, the right-hand tail of the distribution of SNP informativeness crosses that of microsatellites (fig. 6), suggesting that, if enough SNPs are screened, a set with informativeness comparable to that of the set of the most informative microsatellites can be found. This observation may be less applicable to the problem of distinguishing among K source populations for $K > 2$. For a locus with N alleles, if $N \geqslant K$, the informativeness of the locus can potentially be as large as $\log K$, whereas,

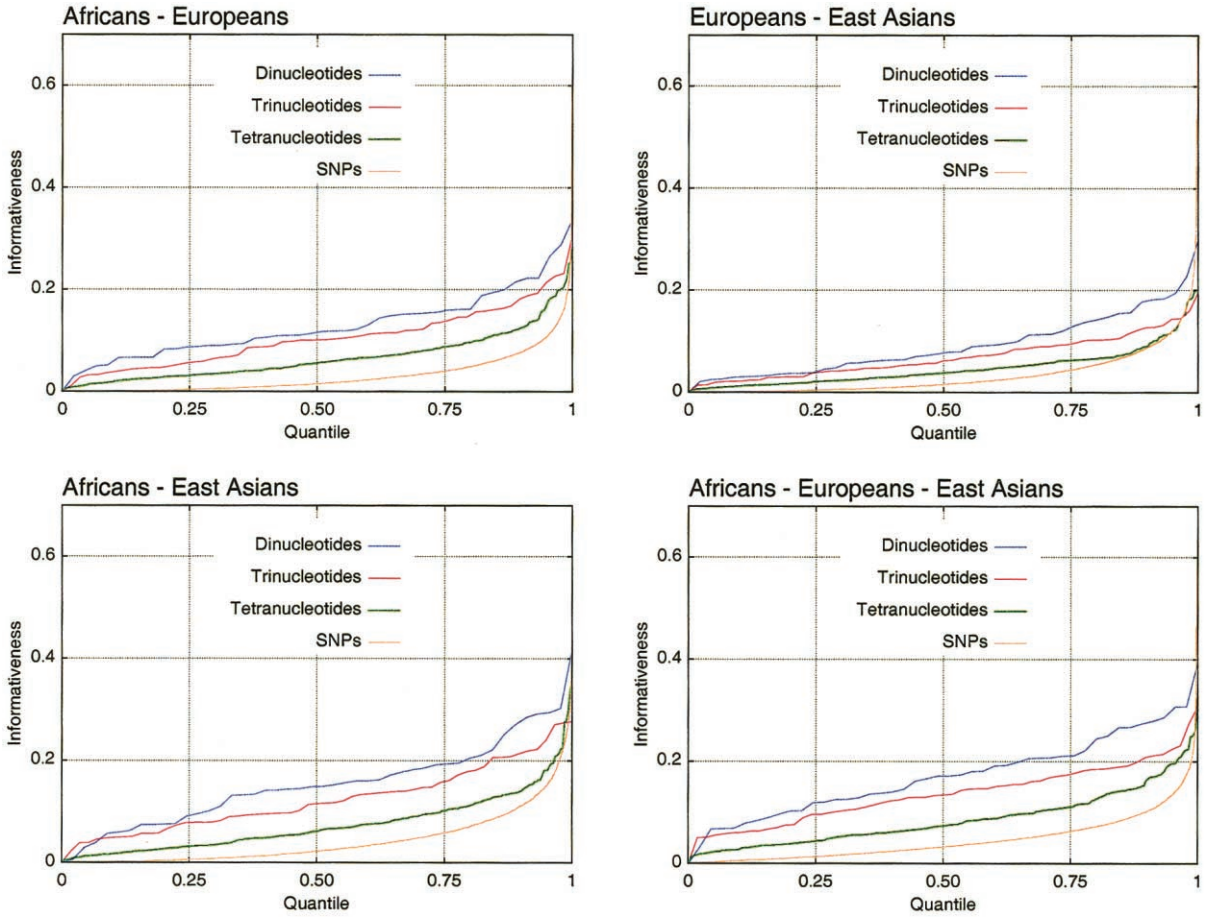

Figure 6 Informativeness quantiles for microsatellites and SNPs. For each set of populations, curves follow the same relative order over most of the domain (from top to bottom: dinucleotides, trinucleotides, tetranucleotides, and SNPs). SNPs were genotyped in African Americans and European Americans rather than in Africans and Europeans.

if $N = 2$, the maximal informativeness is no larger than $\log 2$, regardless of the value of K. Because microsatellites in the data of Rosenberg et al. (2002) have an average of 12.4 alleles, for relatively large values of K, microsatellites have greater potential for higher information content than SNPs, most of which are biallelic. Thus, for large K, the relative performance of microsatellites compared with SNPs will likely be greater than is seen in figure 6 for sets of two and three source populations.

Thus, for inferring ancestry among groups such as African Americans, European Americans, and East Asians, for which genomewide SNP allele frequencies have already been obtained (Akey et al. 2002), use of the most informative known SNPs is likely to be most efficient. However, because informativeness for distinguishing among populations such as different Native American groups is not correlated with informativeness for distinguishing among major regional groups (table 6), SNPs chosen by their informativeness in other scenarios will likely be considerably less useful in these populations than randomly chosen microsatellites. Until the most informative SNPs are identified for a set of populations of interest, use of microsatellites, especially dinucleotides, may lead to greater statistical efficiency in inference of ancestry. Of course, technical problems associated with dinucleotides (Ghebranious et al. 2003) might outweigh the efficiency that derives from their use, and factors such as laboratory fixed costs and difficulties in multiplexing might make the application of less informative markers more economical. The decision about which markers to use for inference of ancestry in any particular context should incorporate a combination of economic, technical, and statistical concerns.

Acknowledgments

N.A.R. was supported by an NSF Postdoctoral Fellowship in Biological Informatics. We thank J. Akey for assistance with the SNP data set, P. Calabrese and D. Conti for discussions, and E. Ziv and two reviewers for thoughtful comments on the manuscript.

Appendix

Informativeness for Ancestry Coefficients

Define $\Delta_{K-1} = \{(Q_1, \ldots, Q_{K-1}) : Q_i \geq 0, \sum_{i=1}^{K-1} Q_i \leq 1\}$ and suppose $Q_K = 1 - \sum_{i=1}^{K-1}$ on Δ_{K-1}. The probability density function for $\mathbf{Q} = (Q_1, Q_2, \ldots, Q_K)$, which we denote by $f_Q(\mathbf{q})$, can be regarded as a function defined on Δ_{K-1}. An elementary calculation shows that $\int_{\Delta_{K-1}} dQ_1 \ldots dQ_{K-1} = 1/(K-1)!$, and, therefore,

$$f_Q(\mathbf{q}) = (K-1)! \ . \tag{A1}$$

Similarly to the discrete no-admixture model, we can apply the continuous-variable analogue of mutual information (Cover and Thomas 1991, chapter 9) to define informativeness in the admixture model, or $I_a(\mathbf{Q}; J)$ (the subscript "a" refers to the *admixture model*). The informativeness for ancestry coefficients is the difference of entropy $H_a(\mathbf{Q})$ and conditional entropy $H_a(\mathbf{Q}|J)$. By use of equation (A1) and the definition, the entropy is

$$H_a(\mathbf{Q}) = -\int_{\Delta_{K-1}} f_Q(\mathbf{q}) \log f_Q(\mathbf{q}) dQ_1 \ldots dQ_{K-1} = -\log(K-1)! \ . \tag{A2}$$

The conditional entropy of $\mathbf{Q}$ given J is given by

$$H_a(\mathbf{Q}|J) = -\sum_{j=1}^{N} \int_{\Delta_{K-1}} \mathbb{P}(J = j|\mathbf{Q} = \mathbf{q}) f_Q(\mathbf{q}) \log \frac{\mathbb{P}(J = j|\mathbf{Q} = \mathbf{q}) f_Q(\mathbf{q})}{\mathbb{P}(J = j)} dQ_1 \ldots dQ_{K-1} \ . \tag{A3}$$

By use of equations (13) and (A1), the following integral can be evaluated:

$$\Pr(J = j) = \int_{\Delta_{K-1}} \mathbb{P}(J = j|\mathbf{Q} = \mathbf{q}) f_Q(\mathbf{q}) dQ_1 \ldots dQ_{K-1} = p_j \ . \tag{A4}$$

Setting $a_{ij} = p_{ij}(K-1)!/p_j$ and substituting equations (13), (A1), and (A4) into (A3), we have

$$H_a(\mathbf{Q}|J) = -\sum_{j=1}^{N} p_j \int_{\Delta_{K-1}} \left(\sum_{i=1}^{K} a_{ij} q_i\right) \log\left(\sum_{i=1}^{K} a_{ij} q_i\right) dQ_1 \ldots dQ_{K-1} \ . \tag{A5}$$

If we assume that, for all j, if $i \neq i'$, then $p_{ij} \neq p_{i'j}$, by applying the result of Rosenberg and Stong (2003) with $K-1$ in place of k to the function $f(x) = x^K \log x/K!$, it can be shown that the integral in equation (A5) evaluates to

$$\sum_{i=1}^{K} \left[\frac{a_{ij}^K \log a_{ij}}{K! \prod_{\substack{i'=1 \\ i' \neq i}}^{K} (a_{ij} - a_{i'j})}\right] - \left(\sum_{i=1}^{K} a_{ij}\right)\left(\frac{|S_{K+1}^{(2)}| - K!}{(K!)^2}\right), \tag{A6}$$

where $S_{K+1}^{(2)} = (-1)^{K+1} K!(1 + 2^{-1} + 3^{-1} + \ldots + K^{-1})$ is a Stirling number of the first kind. Finally, inserting equation (A6) into (A5) and simplifying gives

$$H_a(\mathbf{Q}|J) = -\sum_{j=1}^{N} \left[\sum_{i=1}^{K} \left(\frac{p_{ij}^K \log[p_{ij}(K-1)!/p_j]}{K \prod_{\substack{i'=1 \\ i' \neq i}}^{K} (p_{ij} - p_{i'j})}\right) + p_j\left(1 - \frac{|S_{K+1}^{(2)}|}{K!}\right)\right] . \tag{A7}$$

The expression for the mutual information (eq. [14]) is obtained by use of $I_a(\mathbf{Q}; J) = H_a(\mathbf{Q}) - H_a(\mathbf{Q}|J)$ with equations (A2) and (A7). Note that the definition of I_a is sensible only if no two populations share the same frequency for

any allele. It is appropriate to assume that parametric allele frequencies are unequal in different populations; however, when allele frequencies are estimated from samples of small and equal size, this assumption will often not be met.

Electronic-Database Information

The URLs for data presented herein are as follows:

Human Diversity Panel Genotypes, Center for Medical Genetics, http://research.marshfieldclinic.org/genetics/Freq/FreqInfo.htm (for microsatellite genotypes)
Human STRP Screening Sets, Center for Medical Genetics, http://research.marshfieldclinic.org/genetics/sets/combo.html (for Marshfield panel 10)
Joshua Akey's Homepage, http://cgi.uc.edu/~jakey/ (for SNP allele frequencies)
Noah Rosenberg's Homepage, http://www.cmb.usc.edu/~noahr/distruct.html (for *distruct* software)
Pritchard Lab, http://pritch.bsd.uchicago.edu/ (for *structure* software)

References

Abramowitz MA, Stegun IA (1965) Handbook of mathematical functions. Dover, New York

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. Genetics 160:1217–1229

Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB (2003) Human population genetic structure and inference of group membership. Am J Hum Genet 72:578–589

Banks MA, Eichert W, Olsen JB (2003) Which genetic loci have greater population assignment power? Bioinformatics 19:1436–1438

Barron A, Rissanen J, Yu B (1998) The minimum description length principle in coding and modeling. IEEE Trans Inform Theory 44:2743–2760

Brenner CH (1998) Difficulties in the estimation of ethnic affiliation. Am J Hum Genet 62:1558–1560

Buchanan FC, Adams LJ, Littlejohn RP, Maddox JF, Crawford AM (1994). Determination of evolutionary relationships among sheep breeds using microsatellites. Genomics 22: 397–403

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

Cann HM, de Toma C, Cazes L, Legrand M-F, Morel V, Piouffre L, Bodmer J, et al (2002) A human genome diversity cell line panel. Science 296:261–262

Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci USA 85:9119–9123

Collins-Schramm HE, Phillips CM, Operario DJ, Lee JS, Weber JL, Hanson RL, Knowler WC, et al (2002) Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. Am J Hum Genet 70:737–750

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

Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York

Davies N, Villablanca FX, Roderick GK (1999) Determining the source of individuals: multilocus genotyping in nonequilibrium population genetics. Trends Ecol Evol 14:17–21

Dean M, Stephens JC, Winkler C, Lomb DA, Ramsburg M, Boaze R, Stewart C, et al (1994) Polymorphic admixture typing in human ethnic populations. Am J Hum Genet 55: 788–808

Elandt-Johnson RC (1971) Probability models and statistical methods in genetics. Wiley, New York

Excoffier L (2001) Analysis of population subdivision. In: Balding DJ, Bishop M, Cannings C (eds) Handbook of statistical genetics. John Wiley & Sons, Chichester, UK, pp 271–307

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Ghebranious N, Vaske D, Yu A, Zhao C, Marth G, Weber JL (2003) STRP screening sets for the human genome at 5 cM density. BMC Genomics 4:6

Gibbons JD (1985) Nonparametric statistical inference. 2nd ed. Marcel Dekker, New York

Gomulkiewicz R, Brodziak JKT, Mangel M (1990) Ranking loci for genetic stock identification by curvature methods. Can J Fish Aquat Sci 47:611–619

Guinand B, Topchy A, Page KS, Burnham-Curtis MK, Punch WF, Scribner KT (2002) Comparisons of likelihood and machine learning methods of individual classification. J Hered 93:260–269

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM (2003) Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72:1492–1504

Kullback S (1959) Information theory and statistics. Dover, Mineola, New York

Lowe AL, Urquhart A, Foreman LA, Evett IW (2001) Inferring ethnic origin by means of an STR profile. Forensic Sci Int 119:17–22

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

McKeigue PM (1998) Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in

admixed populations, by conditioning on parental admixture. Am J Hum Genet 63:241–251

Millar RB (1991) Selecting loci for genetic stock identification using maximum likelihood, and the connection with curvature methods. Can J Fish Aquat Sci 48:2173–2179

Molokhia M, Hoggart C, Patrick AL, Shriver M, Parra E, Ye J, Silman AJ, et al (2003) Relation of risk of systemic lupus erythematosus to west African admixture in a Caribbean population. Hum Genet 112:310–318

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

Primmer CR, Koskinen MT, 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 JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. Theor Popul Biol 60:227–237

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Ramachandran S, Rosenberg NA, Zhivotovsky LA, Feldman MW. Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. Hum Genomics (in press)

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. Proc Natl Acad Sci USA 94:9197–9201

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

Rosenberg N, Stong R (2003) Problem 11039. Amer Math Monthly 110:743

Rosenberg NA, Burke T, Elo K, Feldman MW, Freidlin PJ, Groenen MAM, Hillel J, et al (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics 159:699–713

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

——— (2003) Response to comment on "Genetic structure of human populations." Science 300:1877

Ruiz Linares A (1999) Microsatellites and the reconstruction of the history of human populations. In: Goldstein DB, Schlötterer C (eds) Microsatellites: evolution and applications. Oxford University Press, Oxford, pp 183–197

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 60:957–964

Smith MW, Lautenberger JA, Shin HD, Chretien J-P, Shrestha S, Gilbert DA, O'Brien SJ (2001) Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. Am J Hum Genet 69:1080–1094

Smouse PE, Spielman RS, Park MH (1982) Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. Am Nat 119:445–463

Sokal RR, Rohlf FJ (1995) Biometry. 3rd ed. Freeman, New York

Stephens JC, Smith MW, Shin HD, O'Brien SJ (1999) Tracking linkage disequilibrium in admixed populations with MALD using microsatellite loci. In: Goldstein DB, Schlötterer C (eds) Microsatellites: evolution and applications. Oxford University Press, Oxford, pp 211–224

Weir BS (1996) Genetic data analysis II. Sinauer, Sunderland, MA

Weiss L (1961) Statistical decision theory. McGraw-Hill, New York

Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet 72:1171–1186

Ziv E, Burchard EG (2003) Human population structure and genetic association studies. Pharmacogenomics 4:431–441