Information on Ancestry From Genetic Markers

Carrie Lynn Pfaff,[1] Jill Barnholtz-Sloan,[2] Jennifer K. Wagner,[1] and Jeffrey C. Long[1]*

[1]*Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan*
[2]*Department of Internal Medicine, Wayne State University School of Medicine, Detroit, Michigan*

It is possible to estimate the proportionate contributions of ancestral populations to admixed individuals or populations using genetic markers, but different loci and alleles vary considerably in the amount of information that they provide. Conventionally, the allele frequency difference between parental populations (δ) has been used as the criterion to select informative markers. However, it is unclear how to use δ for multiallelic loci, or populations formed by the mixture of more than two groups. Moreover, several other factors, including the actual ancestral proportions and the relative genetic diversities of the parental populations, affect the information provided by genetic markers. We demonstrate here that using δ as the sole criterion for marker selection is inadequate, and we propose, instead, to use Fisher's information, which is the inverse of the variance of the estimated ancestral contributions. This measure is superior because it is directly related to the precision of ancestry estimates. Although δ is related to Fisher's information, the relationship is neither linear nor simple, and the information can vary widely for markers with identical δs. Fortunately, Fisher's information is easily computed and formally extends to the situation of multiple alleles and/or parental populations. We examined the distribution of information for SNP and microsatellite loci available in the public domain for a variety of model admixed populations. The information, on average, is higher for microsatellite loci, but exceptional SNPs exceed the best microsatellites. Despite the large number of genetic markers that have been identified for admixture analysis, it appears that information for estimating admixture proportions is limited, and estimates will typically have wide confidence intervals. © 2004 Wiley-Liss, Inc.

*Correspondence to: Jeffrey C. Long, Ph.D., Department of Human Genetics, University of Michigan Medical School, 4909 Buhl, Ann Arbor, MI 48109. E-mail: longjc@umich.edu

Received 1 July 2003; Accepted 1 December 2003
Published online 19 February 2004 in WileyInterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.10319

INTRODUCTION

Several applications in the study of human genetics require precise estimates of the proportionate contributions of ancestry from two or more parental populations to an admixed population or individual. It is possible to make these estimates of ancestry using genetic markers. In principle, all loci are affected by admixture in the same way, and each locus ideally reflects the same ancestral contributions. However, it is well-known that genetic loci differ in the amount of information that they provide, and the precision of estimates can vary widely depending on which loci are used for the estimation. An optimal marker for estimating ancestral proportions would have different alleles fixed in each of the parental populations. Unfortunately, optimal loci appear to be very rare in the human genome. In the absence of such loci, markers that demonstrate a large difference in allele frequency between the parental populations (δ) have been preferred for ancestry estimation [Glass and Li, 1953; Reed, 1969; Cavalli-Sforza and Bodmer, 1971; Adams and Ward, 1973; Dean et al., 1994; Shriver et al., 1997; Parra et al., 1998, 2001; Smith et al., 2001; Collins-Schramm et al., 2002]. Large numbers of microsatellite, SNP, and insertion/deletion polymorphisms have been identified as useful for admixture estimation based on this criterion [Shriver et al., 1997; Parra et al., 1998; Smith et al., 2001; Collins-Schramm et al., 2002]. Although this strategy is sensible, several under-appreciated difficulties arise in its application. For example, the minimum allele frequency difference acceptable for markers used in ancestry estimation is subjective. In fact, the cutoff for acceptable markers has steadily decreased over time, from δ=0.5 [Shriver et al., 1997] to δ=0.4 [Parra et al., 1998] to δ=0.3 [Collins-Schramm et al., 2002]. Additionally, it is problematical to apply the δ criterion to loci with more than two alleles. While a composite δ (δ_c) was recently defined as one half the sum of the absolute allele frequency differences at a locus [Shriver et al., 1997; Smith et al., 2001], its appeal is mostly heuristic, and its rigorous statistical properties are unknown. A further difficulty arises when three or more populations have contributed to the founding of the admixed population. δ only applies to a pair of populations, and it is unclear how multiple δs can be combined to provide a single criterion by which the usefulness of a marker can be assessed for admixed populations formed by three or more parental populations.

Here, we derive the Fisherian information (I_m) for estimating ancestral contributions from genetic markers. The information is the inverse of the variance of the maximum likelihood estimator of ancestral contributions. Therefore, the information has a direct relationship to the precision of the estimate. Although δ is an important contributor to the information, when δ <1.0, other factors are also important. These factors include the allele frequencies in the parental populations (*p*) irrespective of δ, and the respective genetic contribution of each parental population to the admixed population (*m*). These three factors (δ, *p*, and *m*) contribute to the total information in a complex manner, and it is difficult to disentangle their individual effects. Fortunately, I_m is easy to compute and apply. The Fisherian information can also be used to develop a marker-selection strategy, which can be applied to the estimation of ancestral contributions when more than two parental populations have contributed to the admixed population.

METHODS

ADMIXTURE MODELS, LIKELIHOOD ESTIMATION, AND INFORMATION

We begin by considering a population that was formed by admixture between two genetically distinct ancestral populations. Assuming that evolutionary pressures other than admixture have been insignificant, the frequency of an allele *k* at an arbitrary locus in the admixed population, p_{Ak}, will be a linear combination of allele frequencies in the ancestral populations, p_{jk} ,

$$p_{Ak} = m_1 p_{1k} + m_2 p_{2k} \tag{1}$$

where m_j is the proportionate contribution of the *jth* ancestral population. Since $m_1 + m_2 = 1.0$, Eq. 1 can be written using only one of the ancestral contributions,

$$p_{Ak} = p_{2k} + m_1 \delta_k \tag{2}$$

where $\delta_k = p_{1k} - p_{2k}$ is the usual measure of allele frequency difference. Eqs. 1 and 2 apply to all alleles at all loci.

The genetic contributions of the ancestral populations can be estimated by maximum likelihood from a sample of genotypes from the admixed population. Assuming codominance and random mating in the admixed population, the log likelihood for the ancestral contributions at the *gth* locus is proportional to

$$\begin{aligned} \ln L_g &\propto \sum_k n_{gk} \ln(p_{gAk}) \\ &= \sum_k n_{gk} \ln\left(p_{g2} + m_1 \delta_l\right) \end{aligned} \tag{3}$$

where $n_{gk} = 2n_{gkk} + \sum_{k \neq l} n_{gkl}$ and is the count of the allele A_k, n_{gkk} is the count of A_kA_k homozygotes, and n_{gkl} is the number of A_kA_l heterozygotes. The summation in Eq. 3 is taken over all alleles at all loci. Notice that $\sum_k n_{gk} = 2N$ is twice the number of individuals in the sample, and barring missing data, *2N* is the same for all loci. The log likelihood across multiple loci is the sum of the log likelihoods for the individual loci, $\ln L = \sum_g \ln L_g$. Estimates of individual admixture are obtained by treating each individual as a sample of size one [Chakraborty et al., 1986] because the same likelihood applies to samples of any size.

Differentiating Eq. 3 with respect to m_1 gives

$$\frac{\partial \ln L}{\partial m_1} = \sum_g \sum_k n_{gk} \frac{\delta_{gk}}{p_{gAk}}. \tag{4}$$

The maximum likelihood estimate of the ancestral contribution, $\hat{m}_1$, is obtained by setting Eq. 4 equal to zero and solving for m_1. This usually requires a numerical procedure, but the problem is not particularly challenging. The expected Fisherian information is obtained from the expected negative second derivative of the likelihood

$$\mathrm{E}[I(m_1)] = -\mathrm{E}\left[\frac{\partial^2 \ln L}{\partial m_1^2}\right] = 2N \sum_g \sum_k \frac{\delta_{gk}^2}{\hat{p}_{gAk}} \tag{5}$$

where $\hat{p}_{gAk} = p_{g2k} + \hat{m}_1 \delta_{gk}$ is the expected frequency of the *k*th allele in the admixed population or individual. Note that information for ancestry is additive over loci. The expected variance of the admixture estimate is $V(m_1) = 1/E[I(m_1)]$ and $V(m_2)=V(m_1)$ because of the relationship $m_2=1-m_1$.

The model can be expanded to accommodate any number of parental populations. For brevity and concreteness, we consider the case in which three parental populations have contributed to the admixed population, such that the frequency of the kth allele in the admixed population is

$$\begin{aligned} p_{gAk} &= m_1 p_{g1k} + m_2 p_{g2k} + m_3 p_{g3k} \\ &= p_{g3k} + m_1 \delta_{g1k} + m_2 \delta_{g2k} \end{aligned} \tag{6}$$

where the ancestral contributions sum to 1.0, and the δ coefficients are defined $\delta_{g1k} = p_{g1k} - p_{g3k}$ and $\delta_{g2k} = p_{g2k} - p_{g3k}$. As before, Eq. 6 applies to all alleles at all loci. The log-likelihood function remains as in Eq. 3,

$$\begin{aligned} \ln L &= \sum_g \sum_k n_{gk} \ln(p_{gAk}) \\ &= \sum_g \sum_k n_{gk} \ln(p_{g3k} + m_1 \delta_{g1k} + m_2 \delta_{g2k}) \end{aligned} \tag{7}$$

except that the allele frequency component is expanded as in Eq. 6. It is worth noting that the constraint $\sum_j m_j = 1.0$ ensures that the outcome of analysis is unaffected by the way parental populations are numbered, or which population is subtracted from the others. Maximum likelihood estimates for the ancestral contributions are obtained from the log likelihood function by setting the partial derivatives,

$$\frac{\partial \ln L}{\partial m_j} = \sum_g \sum_k n_{gk} \frac{\delta_{gjk}}{p_{gAk}} \tag{8}$$

equal to zero, and solving simultaneously.

The information with respect to the admixture parameters is presented in matrix form. By extension of Eq. 5, the information matrix has as its elements the negative expected second partial derivatives of the log likelihood function

$$I(m_i, m_j) = -E\left[\frac{\partial^2 \ln L}{\partial m_i \partial m_j}\right] = 2N \sum_g \sum_k \frac{\delta_{gik}\delta_{gjk}}{\hat{p}_{gAk}} \tag{9}$$

The element in the ith row and ith column is the expected information for the estimate of the parameter m_i and the element in the ith row and jth column is the expected shared information for the estimates of the parameters m_j and m_j. The inverse of the information matrix provides the expected variance-covariance matrix of estimated admixture proportions. As before, for multiple loci, the likelihoods and information are summed over all loci. The inverse of the information matrix provides the expected variances of the estimates along the diagonal and covariances on the off-diagonals. The maximum likelihood estimate of $\hat{m}_3 = 1 - \hat{m}_1 - \hat{m}_2$ and the associated variance is obtained from $V(m_3, m_3) = V(m_1, m_1) + V(m_2, m_2) - 2 \cdot V(m_1, m_2)$ [Edwards, 1992].

Analysis of multiple parental populations raises a new issue in choosing marker loci for the estimation of ancestry. Namely, the information supplied by a marker locus for the contribution of one ancestral population is not necessarily independent of the information that the marker supplies for the contribution of another ancestral population. The covariance between parameter estimates measures the degree of redundancy. For populations formed by more than two parental populations, the most informative marker loci are those that simultaneously add to the diagonal elements of the information matrix without greatly increasing the off-diagonal elements.

MODEL POPULATIONS AND GENETIC DATA

In order to investigate how genetic markers contribute information on ancestry, we constructed a set of model populations that were designed to tease apart the intricacies of how parental population allele frequencies, δs, and the actual admixture proportions contribute information. Three of the model populations were based on actual African and European allele frequencies, but differed with respect to their proportionate contributions (Table I). Four more model populations were formed from African, European, and Native American allele frequencies. The ancestral proportions for these model populations were selected from the literature, and represent the variety of ancestry found in contemporary admixed populations. The specific ancestral proportions used are shown in Table II.

The genetic loci and allele frequencies we used to represent the ancestral source populations were taken from 2,492 SNPs from Gabriel et al. [2002] for which European and Sub-Saharan African frequencies were available (http://www-genome.wi.mit.edu/mpg/hapmap/hapstruc.html), 377 autosomal microsatellite loci from the HGDP-CEPH Human Genome Diversity Cell Line Panel with frequencies for European, African, and Native American populations [Weber and Broman, 2001; Cann et al., 2002; Rosenberg et al., 2002] (available

TABLE I. Optimal marker selection I

	Ancestral contributions		Information[a]		
Model population (i)	African (m_{AF})	European (m_E)	S_1	S_2	S_3
1	0.1	0.9	**75.8**	65.6	24.6
2	0.5	0.5	23.4	**25.3**	16.5
3	0.9	0.1	39.4	40.5	**52.3**

[a] S_i is set of 10 most informative markers for population i.

TABLE II. Optimal marker selection II

		Ancestral contributions			Determinant of information matrix			
Model population	Reference population	African (m_{AF})	European (m_E)	Native. American (m_{NA})	S_4	S_5	S_6	S_7
4	African American (SC)[a]	0.866	0.118	0.016	**6,712**	3,470	2,476	2,735
5	Afro-Uruguayan[b]	0.47	0.38	0.15	377	538	461	388
6	Mexican American[c]	0.03	0.68	0.29	741	1,693	**2,013**	1,062
7	Arhuaco (Columbia)[d]	0.217	0.003	0.78	4,838	6,530	6,288	**10,889**

[a] Population frequencies from Parra et al. [2001].
[b] Population frequencies from Sans et al. [2002].
[c] Population frequencies from Long et al. [1991].
[d] Population frequencies from Yunis et al. [1994].

at http://research.marshfieldclinic.org/genetics/Freq/FreqInfo.htm), and 39 ancestry informative SNP markers selected specifically for admixture analysis [Shriver et al., 2003].

MARKER SELECTION STRATEGY

For each model population (P_i), we selected the set (S_i) of 10 loci that gave the most information for the population. We chose to use sets of 10 markers for convenience, but the following selection processes can be applied to sets of any size. For model populations formed from two parental populations, we computed the information for each marker locus, and then pooled the information from the 10 most informative loci. Selecting the optimal set of loci when there are three (or more) parental populations is complicated by the fact that the information about ancestral proportions is partially confounded. The determinant of the information matrix

$$\det\left[I(m_i, m_j)\right] = I(m_1, m_1)I(m_2, m_2) - I(m_1, m_2)^2 \qquad (10)$$

is used to get an overall assessment of the information. A large value for the determinant meets the requirement that the diagonal elements of *I(m_i, m_j)* are large relative to the off-diagonal elements. For each of the four model trihybrid populations in Table II, we selected a panel of 10 markers by using a D-optimization search algorithm. Briefly, the algorithm works by selecting the pair of loci, from all possible pairs, for which the determinant of the pooled information matrix was greatest. The next locus added to the set was that which, by its inclusion, added the most to the determinant of the information matrix of the set of three. Marker loci were thus individually added until the set contained 10 loci.

A computer program to compute the expected information for a set of genetic markers, and to implement the D-optimization strategy, is available from the corresponding author.

RESULTS

Table I provides the results for the three model populations formed by mixtures of two parental groups: Africans and Europeans. Each genetic marker set (S_1–S_3) consists of the 10 most informative loci for a particular model population (P_1–P_3). For each model population, the marker set optimized to it is substantially more informative than the marker sets optimized to the other model populations. For model population 1, S_1 provides a 1.2-fold increase in information over S_2 and a 3.1-fold increase over S_3. For model population 2, S_2 slightly outperforms S_1 but provides a 1.5-fold increase over S_3. For model population 3, S_3 provides a 1.7-fold increase over S_1 and a 1.3-fold increase over S_2. Note that the performance of S_1 for model population 1 exceeds the performance of S_3 for model population 3, which exceeds the performance of S_2 for model population 2. These differences in performance underscore the importance of the admixture proportions in determining the informativeness of genetic markers. Note that the least amount of information for ancestry estimation is obtained for model population 2, which was formed by equal contributions from Africans and Europeans. Surprisingly, the best set of 10 markers for model population 1, which was constructed to have 10% African ancestry and 90% European ancestry, was more informative than the 10 best markers for model population 3, which was constructed to have 90% African ancestry and 10% European ancestry. This illustrates the fact that, in addition to the actual mixtures of ancestry, the allele frequencies in the ancestral populations are important contributors to the information.

Table II provides the results for the four model populations formed by mixtures of three parental groups: Africans, Europeans, and Native Americans. For admixed populations formed from three parental populations, the information on genetic ancestry is completely specified by a two by two information matrix. As explained above, the determinant of an information matrix provides a summary measure of its content, and the determinant provides a useful criterion in selecting optimal markers. Each genetic marker set (S_4-S_7) consists of the 10 most informative microsatellite loci for the respective model population (P_4-P_7). The basic trends observed for populations formed by mixing two parental groups are observed again for populations formed by mixing three parental groups. First, the most information on ancestry for a three-way mixed population is obtained from the marker set that was optimized for that model population, and a marker set optimized for a different model population can perform quite poorly. Second, the least information is obtained for model population 5, which has substantial contributions from all three parental sources. This leads to the general conclusion that the more even the mix of ancestry, the less information is available to estimate ancestral contributions. Third, parental population allele frequencies contribute importantly to information on ancestry. The optimal situation occurs where the major fraction of ancestry is derived from the less heterozygous parental population, and the minor fraction of ancestry is derived from the more heterozygous parental population. For instance, S_7, the set of 10 markers optimized for P_7, was very highly informative for P_7, a population that was composed of mostly Native American ancestry with a minor component of African ancestry.

Some loci were included in more than one optimal set. Not surprisingly, the Duffy locus, which has a null allele that is nearly fixed in Sub-Saharan Africans and nearly absent in non-Africans, was included in the optimal marker set for each of the three model populations. Two microsatellites, D2S1400 and D7S1808, were included in all four sets in Table II. The informativeness of D2S1400 is due to substantial differences in allele frequencies across all regions (Fig. 2A). Sub-Saharan Africans are highly polymorphic and segregate several alleles with a large number of repeat units that are absent in non-Africans. These alleles would firmly indicate African admixture into Europeans or Native Americans. Only one allele (111 bp) is common in Native Americans, and so this locus would be useful for documenting gene flow from non-Native Americans into Native Americans. However, the 111-bp allele is less informative for tracing low levels of Native American gene flow into Europeans or Sub-Saharan Africans because both of these populations already possess it at polymorphic frequencies. Although D7S1808 appears in all four optimal sets, it is substantially less informative than D2S1400. The allele frequency profiles (Fig. 2B) demonstrate that D7S1808 has little ability to distinguish between Sub-Saharan African and European ancestry. However, D7S1808 is similar to D2S1400 in having one allele (252 bp) that is very common in Native Americans. On the whole, the locus is very useful for tracing gene flow into Native Americans from non-Native Americans.

For the marker sets examined here, a few exceptional SNPs demonstrated very high information values, but microsatellites, on average, contained more information on ancestry than the SNPs. For example, the mean information for microsatellite loci is 1.27 ± 0.05 ($m_E = 0.9$), compared to 0.44 ± 0.02 ($m_E = 0.9$) for SNPs (Fig. 3).

DISCUSSION

The formulas that we provide for the information associated with an estimated admixture proportion depend critically on the population genetic model defined in Eqs. 1 and 6. Although this model is widely used for admixture estimation in both populations [Elston, 1971; Chakraborty, 1986] and individuals [Chakraborty et al., 1986; Williams et al., 2000], we recognize that it is an oversimplification because it ignores processes such as genetic drift and natural selection in the admixed population, and it optimistically treats the parental population allele frequencies as known constants. Using a model that allows for genetic drift in the hybrid population, Long [1991] derived a variance formula for weighted least squares estimates of admixture proportions. The inverse of this formula reduces algebraically to Eq. 9 divided by a constant that is determined by the extent of genetic drift. In other words, genetic drift reduces the amount of information on ancestry provided by a locus, but it will not change the relative rankings of loci. Allowing for

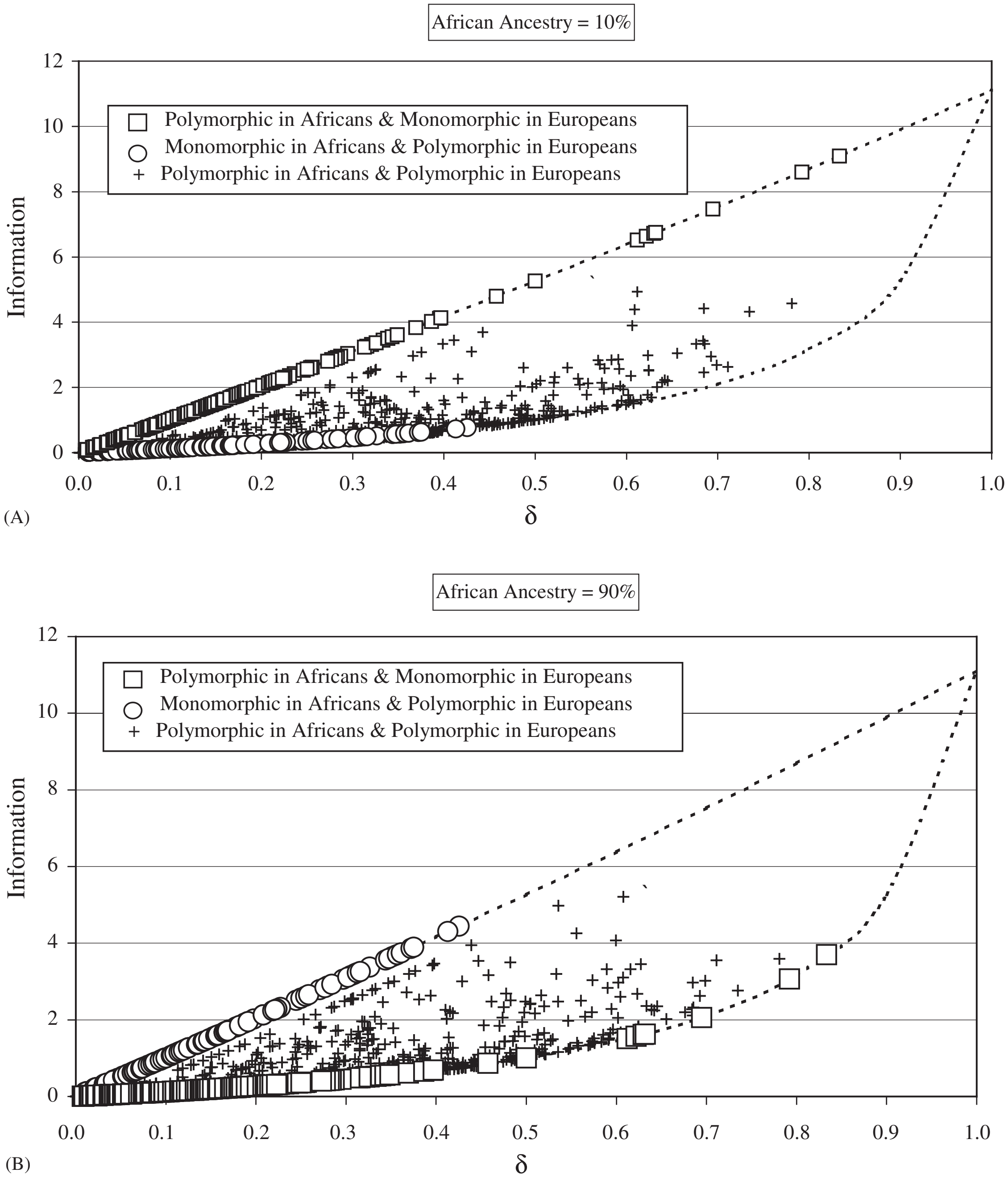

Fig. 1. Relationship between marker information and δ for SNPs (A, B) and composite δ (δ_c) for microsatellites (C, D). The symbols (squares, circles, crosses) in A and B indicate patterns of monomorphism in the African and European source populations. The microsatellite loci depicted by circles in C and D are polymorphic in all populations. The dotted lines in A, B, C, and D indicate the upper and lower bounds for information on ancestry at different levels of δ (or δ_c).

uncertainty in parental population allele frequencies will undoubtedly further diminish the information for ancestry provided by a locus. It is important to note that there are two sources of this uncertainty: 1) estimating allele frequencies from samples drawn from putative source populations, and 2) incorrect identification of parental source populations. Estimation of allele frequencies from samples is the less serious issue, but investigators should realize that estimates of frequencies for rare alleles may be quite imprecise, and that larger sample sizes may be necessary with microsatellite loci. Incorrect identification of parental source populations is the more serious issue because the

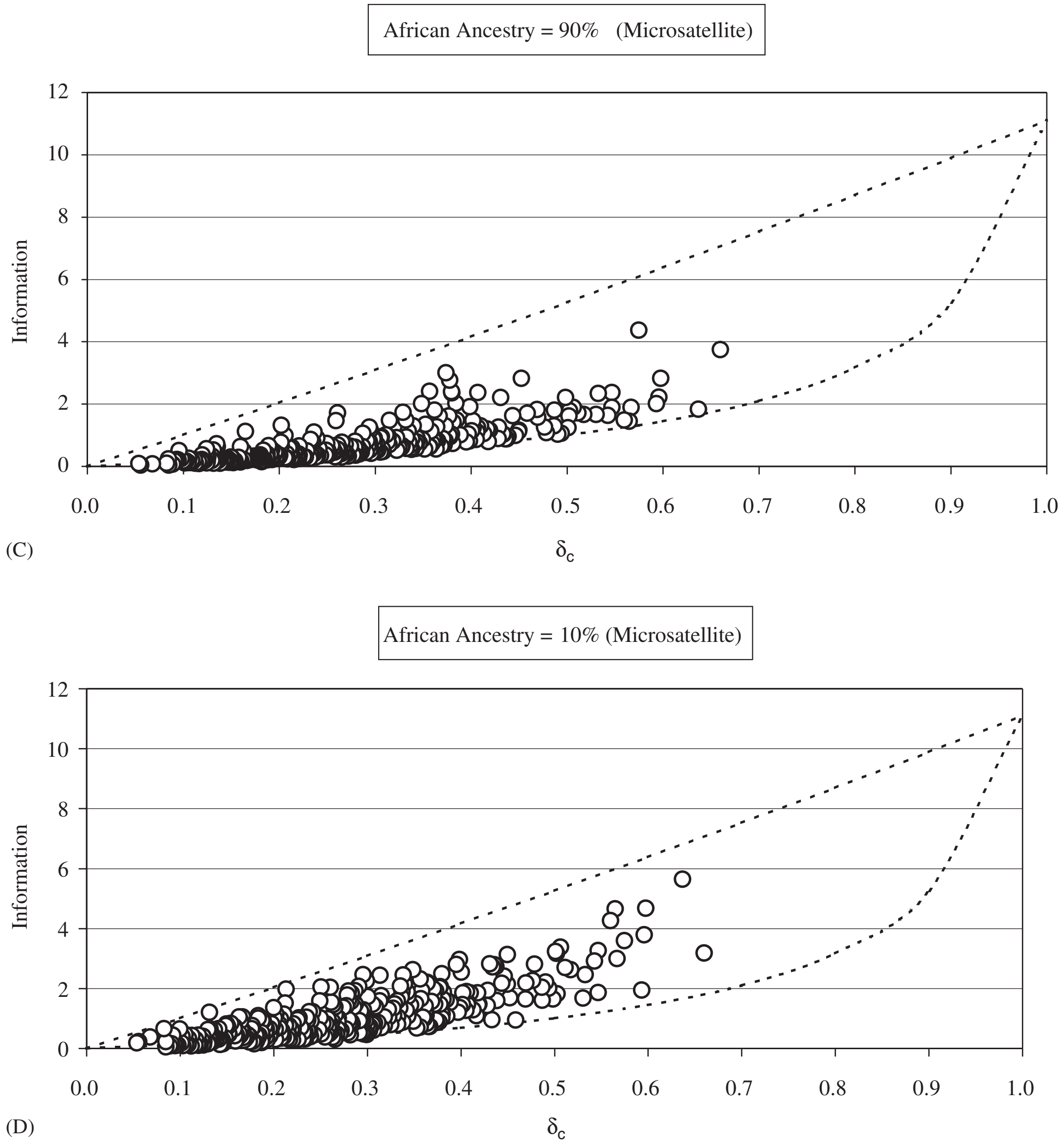

Fig. 1. (*Continued*)

ancestral sources of admixed populations might be an amalgam that requires complex sampling, as in the case of African Americans [Adams and Ward, 1971], or simply no longer exist, as in the case of Mexican Americans [Long et al., 1991]. Nevertheless, we do not expect that a more realistic population genetic model that allows for error in parental population allele frequencies will change the basic insights gleaned by thorough analysis of the model employed here.

It is well-known that an optimal marker for estimating ancestry is one that has a fixed, unique allele in each parental population (i.e., δ=1.0). Duffy, which has a δ value that approaches one for African and European populations, exemplifies this type of marker. Duffy has very high information on ancestry for populations formed from a mixture between African and non-African populations, and its usefulness as an admixture marker is well-known [Reed, 1973; Shriver et al., 1997, 2003; Parra et al., 1998, 2001; Lautenberger et al., 2000; McKeigue et al., 2000; Pfaff et al., 2001]. However, Reed [1973] showed that in order to estimate individual ancestry from two parental populations, 18 ideal markers (i.e., δ=1.0) are needed to obtain an estimate with a 95% confidence interval of 0.20 when m_1=0.1, and 72 ideal loci are required to decrease the confidence

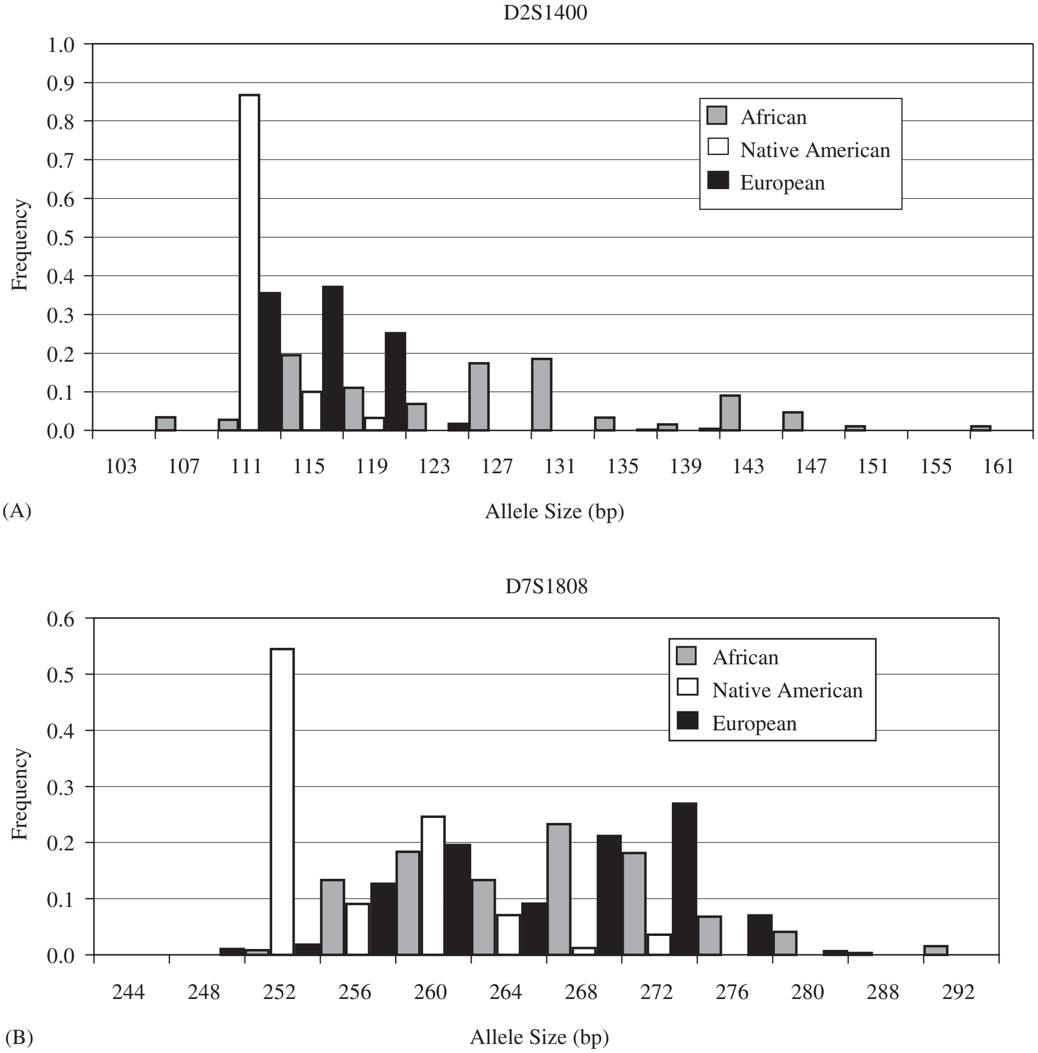

Fig. 2. Allelic frequencies for markers D2S1400 (A) and D7S1808 (B). Data are from Gabriel et al. [2002] and Shriver et al. [2003].

interval to 0.10. The number of ideal markers in the genome is unlikely to meet this requirement. Although the degree to which the markers used here are genomically representative is largely unknown, it does appear that ideal ancestry markers (δ=1.0) are relatively rare.

Thus we are challenged to select the most informative markers from the remaining pool of suboptimal markers. On average, microsatellite markers, which are more likely to have alleles present in one parental population that are not found in the other parental population(s), contain more ancestry information than SNPs. Additionally, a single microsatellite can be informative for more than two parental populations, while a single SNP can only provide information on ancestry from two parental populations. In any case, as shown in Figure 3, markers with reasonably high information on ancestry (both SNPs and microsatellites) are currently available in the public domain.

The imprecise nature of δ as an indicator of information when δ <1.0 (shown in Fig. 1) has thus far been ignored. As shown above, the information on ancestry depends on the interaction between parental population allele frequencies (irrespective of δ) and the admixture proportion (*m*). For admixed populations formed from two parental populations, the markers with the highest information for any given δ are those for which the parental population that contributes the minor proportion of genes to the admixed population has an allele that is not present in the other parental population (Fig. 1A,B). Conversely, the markers with the lowest information for a given δ are those for which the parental

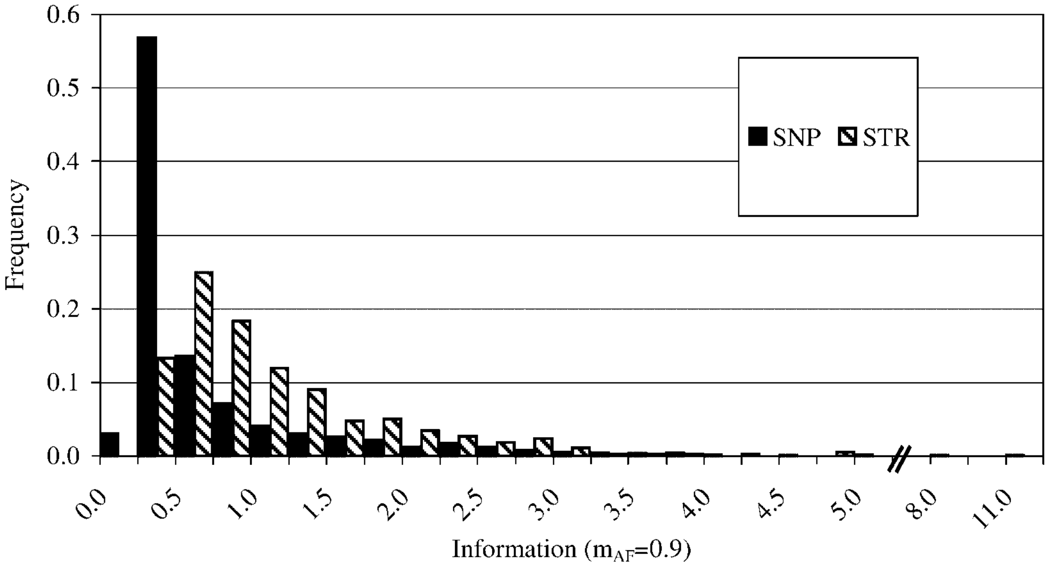

Fig. 3. Distribution of information on ancestry for SNP and STR loci when m_{AF}=0.9. SNP data are shown in solid bars, and STR data are shown in hatched bars.

population that contributes the majority of genes to the admixed population has an allele that is not present in the parental population that contributes the minor proportion of genes. The lowest information is found when each parental population has contributed an equal proportion of genes to the admixed population. The combined effects of allele frequencies, δ, and admixture proportions explain some of the more complicated findings presented in Results. For example, marker set 1 (optimized to detect African ancestry) performs better than marker set 3 (optimized to detect European ancestry) for model population 2 (composed of an equal mix of African and European ancestry). This likely owes to the fact that African populations harbor more unique polymorphic alleles than do non-African populations [Gabriel et al., 2002; Stephens et al., 2001].

McKeigue [1998] recognized these problems and suggested using Wright's statistic F_{ST} [Wright, 1969] computed for the parental populations as a criterion for selecting markers for ancestry estimation. For an admixed population formed by an equal mixture of two parental populations, F_{ST} is equal to the expected information from a single, randomly drawn allele. However, when the parental populations have made unequal contributions to the admixed population, this equality does not hold. An additional advantage of using information as a criterion for selecting markers is that any number of parental populations can be accommodated, and it does not require that the parental populations have made proportionately equal contributions.

The usefulness of a given marker for ancestry estimation can depend largely on the characteristics of the admixed population and the parental populations that contributed to it. Thus, selecting a panel of markers specifically for application in an admixed population of interest can maximize the information and minimize the standard error. For example, the information for model populations P_1–P_3 obtained from the set of 39 markers [Shriver et al., 2003] selected specifically for admixture estimation is 56.5, 37.5, and 66.0, respectively, which correspond to standard errors of individual admixture estimates equal to 0.09, 0.12, and 0.09. The information for sets S_1, S_2, and S_3 in P_1, P_2, and P_3, respectively (shown in Table I), correspond to standard errors of 0.08, 0.15, and 0.13, respectively. In other words, for P_1, S_1 provides a more precise estimate of individual ancestry with only 10 markers than does the generalized admixture set using 39 markers. For P_2, expanding S_2 to contain the best 16 markers (chosen from among those available in all three data sources) increases the information to 37.2 and decreases the standard error to 0.12, making it comparable to the generalized panel of 39 markers. Similarly, expanding S_3 to contain the best 14 markers decreases its standard error to 0.09, thereby making its ancestry estimates comparable in precision to those of the generalized set. Therefore, the amount of genotyping necessary to achieve a given level of precision can be greatly reduced by assembling customized sets of markers.

It is important to point out that the standard errors presented here, in the best-case scenario of S_1 in model population P_1, correspond to 95% confidence intervals of greater than 0.3. In order to obtain ancestry estimates with more desirable precision, marker panels will need to be much larger. For example, selecting the best set of 100 markers for P_1 increases the information to 404, and decreases the standard error to 0.04, which corresponds to a 95% confidence interval of 0.16. Doubling the number of markers in the set to include the best 200 markers only increases the information to 632, slightly decreasing the 95% confidence interval to 0.12. The return in information for the increase in the number of markers in the panel is necessarily diminishing, since the highly informative markers are added to the panel early in the selection process.

Customizing marker sets for admixed populations formed by three (or more) parental populations is also possible, but this requires consideration regarding the study objective. As discussed, one way to select the maximally informative set of markers is to use the determinant of the information matrix. This approach will identify the panel of markers that has the most independent information on ancestry. In some cases, however, the objective may be to estimate, with as much precision as possible, the proportionate contribution of one of the ancestral populations, at the expense of precision in the estimates of the other parental populations. In this case, selecting markers based on the determinant of the information matrix may be less desirable than taking the inverse of the information matrix and using the resulting variance/covariance matrix to maximize the precision of the estimate of interest.

In any case, we advocate the use of Fisher's information over δ or F_{ST} as the selection criterion for markers to be used in ancestry estimation. Unlike δ and F_{ST}, information is directly related to the precision of the estimate. Additionally, we suggest maximizing information by selecting marker panels with regard to the specific admixture characteristics of the populations. When appropriate to the research question, it may also be possible to increase the available information by selecting as study populations those admixed populations that have advantageous admixture proportions with respect to information.

REFERENCES

Adams J, Ward RH. 1973. Admixture studies and detection of selection. Science 180:1137–1143.

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

Cavalli-Sforza LL, Bodmer WF. 1971. The genetics of human populations. San Fransisco: Freeman.

Chakraborty R. 1986. Gene admixture in human populations: models and predictions. Yrbk Phys Anthropol 29:1–43.

Chakraborty R, Ferrell RE, Stern MP, Haffner SM, Hazuda HP, Rosenthal M. 1986. Relationship of prevalence of non-insulin-dependent diabetes mellitus to Amerindian admixture in the Mexican Americans of San Antonio, Texas. Genet Epidemiol 3:435–454.

Collins-Schramm HE, Phillips CM, Operario DJ, Lee JS, Weber JL, Hanson RL, Knowler WC, Cooper R, Li H, Seldin MF. 2002. Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. Am J Hum Genet 70:737–750.

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

Edwards AWF. 1992. Likelihood. Baltimore: Johns Hopkins University Press.

Elston RC. 1971. The estimation of admixture in racial hybrids. Ann Hum Genet 35:9–17.

Glass B, Li CC. 1953. The dynamics of racial intermixture—an analysis based on the American Negro. Am J Hum Genet 5: 1–19.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Lautenberger JA, Stephens JC, O'Brien SJ, Smith MW. 2000. Significant admixture linkage disequilibrium across 30 cM around the FY locus in African Americans. Am J Hum Genet 66:969–978.

Long JC. 1991. The genetic structure of admixed populations. Genetics 127:417–428.

Long JC, Williams RC, McAuley JE, Medis R, Partel R, Tregellas WM, South SF, Rea AE, McCormick SB, Iwaniec U. 1991. Genetic variation in Arizona Mexican Americans: estimation and interpretation of admixture proportions. Am J Phys Anthropol 84:141–157.

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.

McKeigue P, Carpenter J, Parra E, Shriver M. 2000. Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach using Markov chain simulation: application to African-American populations. Ann Hum Genet 64:171–186.

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD. 1998. Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63:1839–1851.

Parra EJ, Kittles RA, Argyropoulos G, Pfaff CL, Hiester K, Bonilla C, Sylvester N, Parrish-Gause D, Garvey WT, Jin L, McKeigue PM, Kamboh MI, Ferrell RE, Pollitzer WS, Shriver MD. 2001. Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. Am J Phys Anthropol 114:18–29.

Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, Hutchinson RG, Ferrell RE, Boerwinkle E, Shriver MD. 2001. Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet 68:198–207.

Reed TE. 1969. Caucasian genes in American Negroes. Science 165:762–768.

Reed TE. 1973. Number of gene loci required for accurate estimation of ancestral population proportions in individual human hybrids. Nature 244:575–576.

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

Sans M, Weimer TA, Franco MH, Salzano FM, Bentancor N, Alvarez I, Bianchi NO, Chakraborty R. 2002. Unequal contributions of male and female gene pools from parental populations in the African descendants of the city of Melo, Uruguay. Am J Phys Anthropol 118:33–44.

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 JP, 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.

Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N, Baron A, Jackson T, Argyropoulos G, Jin L, Hoggart CJ, McKeigue PM, Kittles RA. 2003. Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 112:387–399.

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han J-H, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shaner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruano G, Vovis GF. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489–493.

Weber JL, Broman KW. 2001. Genotyping for human whole-genome scans: past, present, and future. Adv Genet 42: 77–96.

Williams RC, Long JC, Hanson RL, Sievers ML, Knowler WC. 2000. Individual estimates of European genetic admixture associated with lower body-mass index, plasma glucose, and prevalence of type 2 diabetes in Pima Indians. Am J Hum Genet 66:527–538.

Wright S. 1969. Evolution and the genetics of populations. Volume 2. Chicago: University of Chicago Press.

Yunis JJ, Ossa H, Salazar M, Delgado MB, Deulofeut R, de la Hoz A, Bing DH, Ramos O, Yunis EJ. 1994. Major histocompatibility complex class II alleles and haplotypes and blood groups of four Amerindian tribes of northern Colombia. Hum Immunol 41:248–258.