Estimation of Individual Admixture: Analytical and Study Design Considerations

Running title: Inference of Individual Admixture

Hua Tang[1], Jie Peng[2], Pei Wang[2], Neil J. Risch[3]
[1]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109;
[2]Department of Statistics, and [3]Department of Genetics, Stanford University, Stanford, California 94305

Corresponding author: Dr. Hua Tang, Division of Public Health Sciences,

Fred Hutchinson Cancer Research Center, Seattle, Washington 98109.

Email: huatang@fhcrc.org

Fax: (206)667-7004

Telephone (206)667-7041.

Received _____;

ABSTRACT

The genome of an admixed individual represents a mixture of alleles from different ancestries. In the United States, the two largest minority groups, African Americans and Hispanics, are both admixed. An understanding of the admixture proportion at an individual level (individual admixture, or IA) is valuable for both population geneticists and epidemiologists who conduct case-control association studies in these groups. Here we present an extension of a previously described frequentist (maximum likelihood or ML) approach to estimate individual admixture that allows for uncertainty in ancestral allele frequencies. We compare this approach both to prior partial likelihood based methods as well as more recently described Bayesian MCMC methods. Our full ML method demonstrates increased robustness when compared to an existing partial ML approach. Simulations also suggest that this frequentist estimator achieves similar efficiency, measured by the mean squared error criterion, as Bayesian methods but requires just a tiny fraction of the computational time to produce point estimates, allowing for extensive analysis (e.g. simulations) not possible by Bayesian methods. Our simulation results demonstrate that inclusion of ancestral populations or their surrogates in the analysis is required by any method of IA estimation to obtain reasonable results.

keywords: admixture, EM algorithm, maximum likelihood estimate.

INTRODUCTION

An admixed population arises when mating occurs between individuals from reproductively isolated ancestral populations. Historically, anthropologists and population geneticists have focused on estimating admixture composition at a group level. For example, using 12 serum markers, Long (1991) estimated 14 percent European admixture among African-Americans from Claxton, Georgia. Numerous other studies focusing on African Americans and Hispanics have reported similar types of results. Recently, several Bayesian or coalescent-based approaches have been proposed to estimate ancestry proportions in a population, which incorporate into the likelihood function effects of genetic drift and other evolutionary forces (Chikhi *et al.* 2001; Wang 2003).

On the other hand, it has become increasingly recognized that admixture estimation is also important on an individual level (Shriver *et al.* 1997; Ziv and Burchard 2003). A recent article by Falush *et al.* (2003) illustrates some of the important roles that IA can play in answering fundamental scientific questions. One important example is in genetic association studies: controlling the admixture fraction at an individual level is likely to be more efficient than at a group level particularly when one needs to adjust for exogenous risk factors. Furthermore, this task has become far more feasible with the development of high throughput genotyping technologies: as we can now genotype many markers at a low cost, more accurate estimation of IA is now practical.

An early approach to estimating IA can be found in Hanis *et al.* (1986), in which the authors envisioned that IA would become useful in disease association studies as abundant DNA markers became available. However, their approach assumes that ancestral allele frequencies are known *a priori*, and the only parameter to be estimated is IA. In practice, these ancestral frequencies are often estimated based on a small number of individuals, and therefore are subject to large sample errors.

More recently, a Bayesian approach, which uses unlinked genotypes to infer population substructure, was implemented via Markov Chain Monte Carlo (MCMC) methods in a program, `structure` (Pritchard *et al.* 2000). Falush *et al.* (2003) extended the implementation to accommodate linked markers. Compared to the approach of Hanis *et al.* (1986), `structure` has the advantage that the ancestral allele frequencies are inferred using information on both ancestral and admixed individuals. Although the original focus of this approach was to identify discrete clusters roughly corresponding to subpopulations, it can also be applied to an admixture model. Applications of `structure` in population genetic analyses can be found in Rosenberg *et al.* (2002) and Bonilla *et al.* (2004), among many others. Hoggart *et al.* (2003) identified

several weaknesses in `structure` and proposed additional Bayesian MCMC-based solutions that solve some of these problems. However, at least two thorny issues plague `structure` and other MCMC algorithms as a genomic control approach: reliable assessment of convergence and sensitivity of parameter estimates to prior distributions of model parameters.

An additional consideration in the MCMC implementations of the Bayesian approach is that they are extremely computationally intensive, particularly when compared to a frequentist method. In this paper, we describe a new approach to maximum likelihood estimation which extends the prior work of Hanis *et al.* (1986); two different implementations for computing these maximum likelihood estimates (MLE) of IA are given. While the likelihood function we consider resembles that in Hanis *et al.*(1986), we consider both ancestral allele frequency and IA as unknown parameters. To distinguish the two ML methods, the rest of the paper will refer to the approach of Hanis *et al.* (1986) as the partial ML estimator, since the estimated ancestral allele frequencies are plugged into the likelihood function as known parameters. The proposed method will be referred to as the full ML estimator, or simply the MLE. As with `structure`, the full ML estimator allows ancestral allele frequencies to be estimated using both ancestral and admixed individuals. We also construct bootstrap confidence intervals for IA estimates. Although aspects of our implementation bear close resemblance to the Bayesian MCMC algorithms of both Pritchard *et al.* (2000) and Hoggart *et al.* (2003), it is actually based on a philosophy more similar to Hanis *et al.* (1986). We view IA as an unknown but fixed parameter, which can be computed precisely if one can observe the ancestry of each nucleotide in a person's genome. In practice, this quantity cannot be directly computed because most alleles are only partially informative regarding their ancestral origins and because of the finite number of markers genotyped.

We demonstrate that the ML estimator we derive offers a number of significant advantages. First, under realistic circumstances it provides comparable efficiency to the Bayesian approach, yet requires only a tiny fraction of the computation time. Second, under some conditions involving small number of ancestors and markers, the MLE produces less biased estimates than Bayesian approaches. Third, because it is likelihood based, it lends itself easily to questions of study design, for example marker choice. We provide some examples in this regard which support and complement prior findings of Rosenberg *et al.* (2003) and Pfaff *et al.* (2004). Finally, we show that our full maximum likelihood estimator is less biased than the partial ML estimator when the populations used to define the ancestral groups are imprecise.

METHODS

MODEL

In this paper, a person's *individual admixture fraction* (IA) is represented by a vector, $Q_i = (q_{i1}, \ldots, q_{iK})$, in which each coordinate corresponds to the probability that a randomly sampled allele from individual i originates from a specific ancestral population, k. We assume a simple population model, in which a known number of ancestral populations contribute to the admixed group of interest. No pedigree information is available. The genotype data consist of I^0 individuals from the admixed group, as well as I^k subjects from each of the K ancestral populations. For these latter K groups, we would in theory obtain DNA samples from the founding ancestors who contributed to the admixed population. Clearly, this is impractical. Instead, we use as proxies contemporary populations whose ancestors were closely related to the true ancestral populations. These individuals will be referred to as *pseudo-ancestors* in subsequent sections. We assume that all $I = I^0 + I^1 + \ldots + I^K$ individuals are genotyped at the same set of M markers, although in practice each admixed individual and pseudo-ancestor may only be genotyped at a subset of the markers. Our goal is to estimate IA for the admixed individuals, $Q_i = (q_{i1}, \ldots, q_{iK}), (i = 1, \ldots, I^0)$.

Determining the appropriate number of ancestral populations, K, is related to the long-standing problem of estimating the number of clusters in a mixture. However, as we discuss later, in the types of applications we intend it may be impractical to determine the number of ancestral populations from the genotype of admixed individuals alone. Moreover, the notion of ancestral populations is not well defined. For the present study, we focus on gene flow among continentally separated populations. For example, we think of African-Americans as a group with mixed European and African ancestry. Therefore, contemporary European Americans and West Africans are included in the sample, whenever possible, to represent the ancestral populations of African-Americans. The effect of genetic drift is ignored in the present method. One potential approach to model the divergence between the true and the pseudo-ancestral populations is to introduce additional hyper-parameters (Patterson *et al.* 2004). Diagnostic procedures for detecting inappropriate pseudo-ancestors are under development. Finally, we assume that in the ancestral populations, there is neither Hardy-Weinberg disequilibrium (dependency between two alleles) at individual loci nor linkage disequilibrium between loci. In the admixed group, we assume Hardy-Weinberg equilibrium conditioning on the admixture fraction.

Let $G_{ima}(i = 1, \ldots, I; m = 1, \ldots, M; a = 1, 2)$ be an allele at marker m in individual i. Assume there are L_m alleles ($L_m \geq 2$) at marker m. Denote the allele frequency of marker m in the k^{th} ancestral

population as $(p_{m1k}, \ldots, p_{mL_mk})$. In describing our method, we treat the pseudo-ancestors as indigenous. In other words, for an individual from the k^{*th} ancestral group,

$$q_{ik} = \begin{cases} 1 \text{ if } k = k^* \\ 0 \text{ otherwise,} \end{cases}$$

and IA are estimated for $i = 1, \ldots, I^0$ only. When the context is clear, we use G to denote the collection of all genotypes. Similarly, we will use the short-hand notation P and Q to denote the collections of ancestral allele frequencies and admixture fractions, respectively. The log likelihood function can be written as

$$\ell(G|P,Q) = \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{a=1}^{2} \sum_{l=1}^{L_m} \mathbf{1}(G_{ima} = l) \log(r_{iml}). \tag{1}$$

where $\mathbf{1}$ is the indicator function and

$$r_{iml} = \sum_{k=1}^{K} p_{mlk} q_{ik} \tag{2}$$

is the frequency of allele l at marker m given the IA of person i. This likelihood function models the ancestries of all alleles as independent. In the initial generations after admixing, dependency in ancestry can be high even among unlinked markers. Subsequently, correlation in ancestry decays as a result of independent segregation of chromosomes and recombination between linked markers. In practice, the lack of linkage disequilibrium (LD) between markers can often be taken as an indication that the correlation in ancestry between alleles is weak, and thus the likelihood function in (1) holds approximately. However, we did evaluate implications of this assumption of no LD as described below.

The next section describes two algorithms for computing the maximum likelihood estimates (MLE) of P and Q. The first algorithm computes the MLE by a Newton-Raphson method in two stages. It is computationally more efficient than the second algorithm, and it is easy to implement in situations where only a small number of ancestral populations are present (K is small), and where only a few alleles are observed at each marker (L is small). Applicable in a more general setting and numerically more stable, a second implementation uses the EM algorithm (Dempster *et al.* 1977). We also describe a bootstrap method for assessing the uncertainties.

A TWO-STAGE NEWTON-RAPHSON ALGORITHM

This algorithm is motivated by the following observation: given Q, the ancestral allele frequencies of different markers, P, are orthogonal to one another. Likewise, given P, Q are independent vectors and can be estimated following Hanis *et al.* (1986). Thus, the dimensionality of the estimation problem is greatly

reduced if we fix either P or Q. A two-stage algorithm, which iteratively estimates P and Q, is outlined in Algorithm 1.

Algorithm 1: A Two-stage Newton-Raphson Algorithm

Step 0: Randomly assign initial values, $P^{(0)}$ and $Q_i^{(0)}(i = 1, \ldots, I^0)$, subject to the constraints $\sum_{l=1}^{L_m} p_{mlk} = 1$ and $\sum_{k=1}^{K} q_{ik} = 1$.

Step 1: Compute $Q^{(n)}$ by solving the score equations with respect to Q, treating P as known:

$$\frac{\partial(G|Q,P)}{\partial q_{ik}} \|_{P=P^{(n-1)}} = 0 \tag{3}$$

for $i = 1, \ldots, I^0; k = 1, \ldots, K-1$.

Step 2: Compute $P^{(n)}$ by solving the score equations with respect to P, treating Q as known:

$$\frac{\partial(G|Q,P)}{\partial p_{mlk}} \|_{Q=Q^{(n)}} = 0 \tag{4}$$

for $m = 1, \ldots, M$, $l = 1, \ldots, L_m - 1$, and $k = 1, \ldots, K$.

Iterate Steps 1 and 2 until convergence.

In Step 1, the estimation of Q amounts to solving I^0 independent systems of equations, each corresponding to an admixed individual and having $K-1$ unknown variables. The solutions to all systems are constrained in the region: $\Omega_{Qi} = \{(q_{i1}, \ldots, q_{i(K-1)}) : q_{ik} \in [0,1], \sum_{k=1}^{K-1} q_{ik} \leq 1\}$. Similarly, the estimation of P in Step 2 amounts to solving M independent systems of equations, each corresponding to a marker and having $(L_m - 1) \times K$ unknown variables. The solutions to all systems are constrained in the region: $\Omega_{P_{mk}} = \{(p_{m1k}, \ldots, p_{m(L_m-1)K}) : p_{mlk} \in [0,1], \sum_{l=1}^{L_m-1} p_{mlk} \leq 1, k = 1, \ldots, K\}$. Since both Ω_Q and Ω_P are convex sets and the log likelihood function, (1), with respect to Q and P is concave, the Newton-Raphson method is an appropriate and efficient root-searching algorithm. The convergence is reached when each score equation in (3) and (4) is either solved or is proven to have no solution. In the latter case, the MLE is on the boundary.

Although in theory this two-stage Newton-Raphson algorithm applies to arbitrary numbers of K and L, the implementation is more difficult as K or L grow. Additionally, because of the need to compute and to invert the information matrix, the Newton-Raphson algorithm may become numerically unstable. Therefore, we developed an alternative approach that is especially suitable for problems with large values of K or L.

AN EM ALGORITHM

To implement the EM algorithm, we introduce an unobservable variable, $Z_{ima} \in \{1, \ldots, K\}$, which corresponds to the ancestral origin of allele G_{ima}. The augmented log likelihood function of (G, Z) is:

$$l(G, Z|P, Q) = \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{a=1}^{2} \sum_{l=1}^{L_m} \sum_{k=1}^{K} \mathbf{1}(G_{ima} = l,\, Z_{ima} = k) \log(p_{mlk} q_{ik}).$$

If Z were observable, it would be straightforward to compute the MLE of P and Q directly. Specifically, Q is estimated by the empirical distribution of Z, while P is estimated by counting alleles originated from a given population. Treating Z as missing variables, the maximization step of our EM algorithm computes the MLE of P and Q conditional on the current expectation of Z:

$$E_{imak}^{(n)} = E[\mathbf{1}(Z_{ima} = k)|P^{(n)}, Q^{(n)}, G].$$

The variables, E, are updated in the expectation step. The EM algorithm for computing the MLE of P and Q is outlined in Algorithm 2.

Algorithm 2: An EM Algorithm

Step 0: Assign initial values for $E_{imak}^{(0)}$. If an individual is from the k*-th pseudo-ancestral group, let

$$E_{imak}^{(0)} = \begin{cases} 1 \text{ if } k = k^* \\ \\ 0 \text{ otherwise} \end{cases}.$$

For the admixed individuals, the initial values of E are randomly assigned as long as each person's IA vector sums to 1.

Maximization Step. Compute $P^{(n)}$ by:

$$p_{mlk}^{(n)} = \frac{\sum_{i=1}^{I} \sum_{a=1}^{2} \mathbf{1}(G_{ima} = l) E_{imak}^{(n-1)}}{\sum_{i=1}^{I} \sum_{a=1}^{2} E_{imak}^{(n-1)}},$$

and compute $Q^{(n)}$ by:

$$q_{ik}^{(n)} = \frac{\sum_{m=1}^{M} \sum_{a=1}^{2} E_{imak}^{(n-1)}}{2M}.$$

Expectation Step. Expectation of the missing variables is computed by Bayes rule:

$$E_{imak}^{(n)} = E(\mathbf{1}(Z_{ima} = k)|G, P^{(n)}, Q^{(n)}) = \frac{p_{mlk}^{(n)} q_{ik}^{(n)}}{\sum_{k'=1}^{K} p_{mlk'}^{(n)} q_{ik'}^{(n)}}.$$

Iterate Steps 1 and 2 until convergence.

Convergence is declared when, in the successive iteration, the differences in the estimates of Q and P fall below a small threshold. In simulations reported below, we set the threshold to be 10^{-9}; in most cases, convergence is achieved in less than 200 iterations. As expected, the EM algorithm converges to the same solutions as the two-stage Newton-Raphson algorithm. We note that Z is also introduced in Pritchard *et al.* (2000). Rather than computing the expectation of this unobservable variable, `structure` samples from the conditional distribution, $P(Z \mid G, P^{(n)}, Q^{(n)})$.

BOOTSTRAP CONFIDENCE INTERVAL

The sample variance of the IA estimates can be estimated using the inverse of the Fisher information matrix. A confidence region of IA can be constructed on the basis of asymptotic normal theory of the MLE. However, there are at least two reasons for preferring a bootstrap approach. First, with the increase in K, M, and the number of observed alleles at each marker, the dimension of the information matrix grows rapidly. We again face the numerical difficulty of inverting a large matrix. More importantly, the sampling distribution of the MLE of both P and Q are likely to be asymmetric, especially when the point estimate falls close to 0 or 1.

In contrast, a bootstrap approach is simple to implement and encounters no numerical difficulties. Because it is based on the empirical distribution, a bootstrap confidence interval has a second order accuracy, and can correct for bias and skewness. We construct the bias-corrected and accelerated (BCa) interval following Efron and Tibshirani (1993). We observe through simulation that resampling markers alone or individuals alone lead to underestimation of the sampling error; thus each bootstrap sample is obtained by sampling both individuals and markers with replacement. Individuals are resampled to maintain the numbers of individuals in each of the admixed and the ancestral groups. IA of an admixed individual is then estimated using the bootstrap sample. IA of an admixed individual, who is not sampled in a bootstrap sample, is estimated as if the individual is present; his/her genotype does not contribute to the estimation of the ancestral allele frequencies.

SIMULATION MODELS

We use simulated data to evaluate the performance of the proposed likelihood-based approach. Our first three simulations use available empirical data and ignore the effect of genetic drift. Genotype data are simulated under an ideal condition in which the pseudo-ancestral groups have the identical allele

distributions as the true ancestral populations. The last dataset is generated using a Wright-Fisher model under neutrality (Nei 1987). In all simulations, the number of ancestral populations, K, is known without error.

In a recent study, 4833 SNPs were genotyped by the SNP Consortium in a panel that includes 30 African-Americans and 30 European Americans (Matise *et al.* 2003; Clark *et al.* 2003). The absolute allele frequency difference, referred to as the δ-values, between these two samples is shown in Figure 1, and is used to approximate the allele frequency difference between our simulated ancestral populations. On the one hand, the small number of individuals genotyped tends to exaggerate the upper tail of this distribution; on the other hand, since the African-American group contains European admixture, the δ-values we observe should be smaller than that between an indigenous African population and the European American population. Overall, the empirical distribution in Figure 1 may be a reasonable approximation of the genetic differentiation between Africans and Europeans.

SIMULATION 1: SNP DATA. The first simulation aims to compare efficiency of the full MLE to that of the Bayesian estimator implemented in the program, `structure`. The sample consists of 500 admixed individuals, and 250 individuals from each of the two ancestral populations, X and Y. Because there are only two ancestral populations, the IA vector is determined by a scalar representing, say, the contribution of population Y. We model the underlying distribution of IA in the admixed group as a mixture of two distributions: with a probability of 0.2, this fraction is sampled from a uniform distribution, $U[0.1, 0.9]$; with a probability of 0.8, this fraction is sampled from a normal distribution (truncated at 0 and 1), with a mean of 0.15 and a standard deviation of 0.05. Such a distribution is chosen to mimic the estimated IA distribution in African-Americans (unpublished results). In the simulated data, the mean admixture fraction is 0.21. To generate allele frequencies, we assume that 200 SNPs have been pre-selected from the SNP Consortium panel. Further, these SNPs are chosen to have a minimum δ-value of 0.30, and represent the top 15% of all SNPs. For each marker, conditioning on the δ-value, the allele frequencies in the ancestral populations are simulated from a uniform distribution. Conditional on a person's IA, genotypes are simulated assuming independence among all alleles. For comparison, we report IA estimates obtained by program `structure` (Version 2.1), with a run consisting of $10,000$ burn-in iterations followed by $50,000$ further iterations and using default values for all other parameters. The length of the iterations is chosen following the African-American example in Falush *et al.* (2003).

As a measure of the efficiency of an estimator we use the root mean squared error (RMSE) which, in

our setting, can be approximated using the simulated samples as:

$$\widehat{\mathrm{RMSE}} = \left[\frac{1}{I^0} \sum_{i=1}^{I^0} (\widehat{q}_i - q_i) \right]^{\frac{1}{2}},$$

where $\widehat{q}_i$ and q_i are the estimated and true IA, respectively.

SIMULATION 2: MICROSATELLITE DATA. In a second simulation experiment, we compare the informativeness of SNP markers to that of microsatellite (short tandem repeat polymorphism, or STR) markers. To facilitate comparison, the same uniform-normal mixture model used in the previous simulation was adopted to generate the IA of 500 individuals. The pseudo-ancestral groups again consist of 250 individuals from each ancestral population, and all individuals are genotyped at 200 STR markers. The ancestral allele frequencies of the microsatellites are sampled from the empirical distribution of 284 STR markers, estimated using a European group and a sub-Saharan African group from the World Human Diversity Panel (Weber and Browman 2001; Cann *et al.* 2002; Rosenberg *et al.* 2002).

SIMULATION 3: A CASE OF LESS INFORMATION. The next simulation experiment studies the efficiency of the IA estimate when the information in the data is less than the previous two simulations for two reasons: first, only a small number of pseudo-ancestors are available; second, the markers are only moderately informative and are small in number. As with the previous two experiments, the IA of an admixed individual is sampled from the normal-uniform mixture distribution, and the admixed group consists of 500 individuals. The pseudo-ancestors consist of 20 individuals from the ancestral population which contributed more to the admixed group, and 60 individuals from the other ancestral population. Sixty SNP markers are selected to have a δ-value of at least 0.2. The averaged δ-value in the simulated data is 0.33.

SIMULATION 4: EFFECTS OF LD AND GENETIC DRIFT. All simulations above were performed ignoring the effects of genetic drift. One consequence of the genetic drift is that the allele frequencies of pseudo-ancestors may deviate from those among the true ancestors. Such difference could conceivably cause bias in the IA estimates. To examine the robustness property of our IA estimate in the presence of genetic drift, the final simulation experiment we present incorporates a simple evolution model, which roughly corresponds to the intermixture model in Long (1991) and scenario V in Falush *et al.* (2003): a progenitor population gave rise to two reproductively isolated parental populations (denoted as X and Y, respectively) 1010 generations ago. A one-time admixing occurred 10 generations between X and Y before

the present generation, resulting in a mean admixture fraction of 0.2. The effective population sizes of the progenitor population, of X, of Y, and of the admixed group are 7500, 5000, 5000, and 2000, respectively, and each remains constant in time. Genealogies are generated according to coalescent theory (Kingman 1982) and using the program `ms` (Hudson 2002). Each genealogy represents a chromosomal segment of 5cM. Nucleotide sequences are then generated under Kimura's two-parameter model (Nei 1987) using the program `seq-gen` (Rambaut and Grassly 1997). Thus, each genealogy gives rise to a cluster of linked SNPs. In this fashion, we generate 120 unlinked chromosomal segments; each segment harboring between 1-8 SNPs. In total, the dataset contains 421 SNPs; the median distance between neighboring markers on the same segment is 0.99cM. The mean allele frequency difference between populations X and Y (at the time of sampling) is 0.15. A dataset is formed by randomly sampling 500 haplotypes (250 individuals) from populations X and Y and 1000 haplotypes (500 individuals) from the admixed individuals.

Full Likelihood versus Partial Likelihood Estimates. We use a subset of the same simulated data described above (10 generation admixing time) to compare the full ML method with the partial ML of Hanis *et al.*(1986). To do so, we randomly select 100 SNP markers and 50 pseudo-ancestors from each of the two ancestral populations, as well as 500 admixed individuals. To emulate a situation in which the pseudo-ancestors deviate from the true ancestors, 5% of the alleles in each of the pseudo-ancestral individuals are randomly selected and switched from one SNP allele to the other. We estimate IA using both the method of Hanis et al (1986), which computes the ancestral allele frequencies from the pseudo-ancestors only and considers them as fixed values, as well as the full ML method described above. Recall that in this simulation, the true IA for all individuals is near 0.2.

RESULTS

SIMULATION 1: SNP DATA

The results of Simulation 1, which compared the estimation of IA using `structure` and the EM algorithm, are shown in Figure 2(a). In this example, the two methods produce nearly identical estimates (correlation coefficient $r > 0.99$). The RMSE for the estimates using `structure` is 0.052; the corresponding quantity using the EM algorithm is 0.053. Both estimates are nearly unbiased. Figure 2(b) shows the EM estimates along with the 90% BCa intervals for the EM algorithm, using 500 bootstrap samples. The coverage probability is 89.4%; that is, the true IA of 447 out of the 500 admixed individuals fall within the estimated confidence interval. The median length of the BCa interval is 0.17. The 90% credibility intervals produced

by `structure` have similar coverage probability (89%) and median interval length (0.17). We have also analyzed this set of data using program `ADMIXMAP`, which implements the MCMC approach described in Hoggart *et al.* (2003). Estimates obtained from `ADMIXMAP` are highly correlated with both the MLE and the `structure` estimates. To produce the point estimates of IA using the EM algorithm on an intel 2GHz processor required less than 1 minute; bootstrap analysis took 3.5 hr. The analysis using `structure` required 5 hr. Because the primary focus of `ADMIXMAP` is association analysis, the program includes computational components which are unnecessary for the sole purpose of the inference on IA. Thus, the computation time required by `ADMIXMAP` is substantially longer (17 hr), but may not be comparable to that of `structure` or the EM algorithm. The performance of the three approaches is summarized in Table I; clearly this comparsion is somewhat arbitrary, since the computation time depends on the length of the MCMC chains (for `structure` and `ADMIXMAP`) and the number of bootstrap samples (for the EM algorithm). In any event, for producing an IA estimate, MLE required at least 300 times less computational time than the Bayesian approach implemented in `structure`.

Importance of Ancestral Populations A question of concern is the importance of the pseudo-ancestral groups. It has been pointed out that the ancestral populations are unidentifiable unless some of the admixed individuals are nearly "pure," that is, their genomes are derived predominantly from a single ancestral population (Pritchard *et al.* 2000). How many and how close to indigenous these individuals need to be in order to identify the ancestral populations, however, has not been investigated. To answer this question, we ran `structure` using the 500 admixed individuals alone. Among four independent runs of `structure` on the same data, the mean estimated IA was 0.4 three times and 0.07 once. Recall that the true IA has a mean of 0.21. Allowing the two parental populations to have unequal prior admixture proportions (different alpha parameter values in the Dirichlet distribution) does not ameliorate the bias. These results suggest that the MCMC algorithm fails to converge. The estimates, however, are highly correlated with the true value ($r = 0.94$). The EM algorithm estimated a biased mean IA of 0.35, while a run of `ADMIXMAP` reports a mean IA of 0.07. It is worth noting that in the admixed sample, 12% of the 500 individuals have an IA less than 0.1, and 2.2% have an IA above 0.8. Thus, this example demonstrates that a significant number of pseudo-ancestors or individuals with extreme IA values are required for any of the existing approaches to perform well. On the other hand, in situations where the underlying IA distribution is sufficiently broad, our simulations indicate that just a handful of pseudo-ancestors are required to achieve nearly unbiased estimates.

SIMULATION 2: MICROSATELLITE DATA

In simulation 2 we compared relative information content of SNPs versus microsatellites. In this simulation, `structure` and ML again produce essentially identical estimates. The MLE of IA is nearly unbiased, and the coverage probability of the BCa is 89%. The RMSE is 0.075 and the median length of the BCa interval is 0.23, both slightly greater than the corresponding quantities in the SNP simulation (Simulation 1). Therefore, under comparable conditions, pre-selected SNPs with δ-values of at least 0.3 appear as informative as random microsatellite markers. These results are largely consistent with those previously obtained by Rosenberg *et al.*(2002) and Pfaff *et al.*(2004).

SIMULATION 3: A CASE OF LESS INFORMATION

Here we considered the impact of using small number of ancestral individuals and SNPs that are only moderately informative. The MLE and `structure` estimated IA are plotted against true IA in Figure **??**. The simulated IA has a mean of 0.237; our MLE of IA has a mean of 0.251. The RMSE is 0.11. Although the coverage probability is acceptable (87%), the 90% confidence intervals are very wide with a median length of 0.31. Further, assuming perfect knowledge of the true ancestral allele frequencies does not improve the estimates substantially. Thus, the poor result is largely attributed to the small number and lack of informativeness of the SNPs used. In comparison, `structure` estimates appear to have an upward bias, with a mean of 0.37. The RMSE of the `structure` estimate is 0.17. Allowing the two parental populations to have unequal prior admixture proportions (different alpha parameter values in the Dirichlet distribution), the mean estimated IA from `structure` was 0.147, this time indicating a downward bias. Thus, it appears that in at least some circumstances of modest information, MLE provides unbiased IA estimates whereas `structure` does not.

SIMULATION 4: EFFECTS OF LD AND GENETIC DRIFT

In this simulation we examined the impact of genetic drift or misrepresentation of ancestral allele frequencies in the pseudo-ancestors used in the analysis. Figure **??** compares `structure` estimates of IA (assuming the correct linkage map) to the MLE (ignoring linkage). Both estimates appear unbiased and are highly correlated ($r = 0.98$). The RMSE for the EM estimates is 0.071, while that for the `structure` estimates is 0.066. Furthermore, `structure` estimated the time since admixing as 8.3 generations. In additional simulations, we amplified the magnitude of genetic drift by reducing the population sizes or by lengthening the time since admixing. Results from these simulations indicate that both the `structure` and EM algorithm produce unbiased estimates when a moderate number of markers are included.

Linkage Disequilibrium among Markers. In this simulation, markers from the same chromsomal region are linked, with an average intermarker distance of 0.99cM; as a result, the ancestry of alleles at neighboring markers may be correlated. On the other hand, the likelihood function in (1) ignores the dependency between linked markers and the bootstrap approach treats all markers as if they were independent. Therefore, we expect the BCa intervals to under-estimate the variability of the IA estimates; that is, the confidence interval tends to be too narrow. For the data presented in Simulation 4, the coverage probability of the 90% BCa interval is 0.86 suggesting weak LD in the simulated data. While the simulation parameters were chosen to approximate the population history of African-Americans (Falush *et al.* (2003)), Simulation 4 is potentially unrepresentative in two important aspects. First, necessitated by the programming and computational complexities in generating a large recombination graph that represents a genealogy, the 120 chromosomal regions in Simulation 4 were generated independently. As a result, the majority of marker pairs in the simulated data are unlinked. In contrast, the SNPs in the map of Smith *et al.* (2004) cover the genome in a continuous fashion such that many pairs are linked. Therefore, admixture linkage disequilibrium (ALD) decays faster in the simulated data. Second, the magnitude of ALD depends on the delta values of the markers. In the current simulation, many markers have relatively small delta values; in contrast, the map of Smith *et al.* (2004) focused on more informative markers. Therefore, we also consider here a far more extreme (and unrealistic) scenario in which admixing occurred 3 generations ago and markers are screened to have a delta value of at least 0.3. In this simulation, the coverage probability of 90% BCa intervals falls to 0.71, demonstrating that the likelihood function in (1) is invalid when most markers are in significant LD. In contrast, by incorporating the correct linkage information, the 90% credibility interval constructed by `structure` achieves a coverage probability of 88%. Linkage and LD among markers is modeled in the algorithms proposed in Falush *et al.* (2003), Hoggart *et al.* (2003), Zhu *et al.* (2004) and Patterson *et al.* (2004). However, while the assumption of marker independence in the ML approach can lead to underestimation of the variability under extreme and unrealistic conditions, the above results also indicate that the effect may not be very severe in situations likely to be encountered in practice (e.g. studying African Americans or Hispanics).

Full Likelihood versus Partial Likelihood Estimates. Here we compared the full versus partial likelihood estimates of IA using the simulated data (admixture 10 generations ago) which allowed for random deviations of the pseudo-ancestors from the true ancestors as described above. The partial likelihood method gave a mean IA of 0.25 while the full ML method still produced an unbiased estimate

(0.20). This illustrates that allowing the allele frequencies to be estimated jointly based on both the pseudo-ancestors and the admixed group can provide a greater level of robustness, particularly when the pseudo-ancestors do not derive precisely from the true ancestral populations.

DISCUSSION

The main contribution of this paper is to introduce a new, full likelihood based estimator of IA and to delineate its advantages, and more generally the advantages of a frequentist approach, for estimating individual admixture fractions using SNP or microsatellite genotype data. When computational efficiency and unbiased estimates are important considerations, our full likelihood approach is an attractive alternative to an existing ML estimator (Hanis *et al.* 1986) as well as Bayesian MCMC approaches, such as one implemented in program `structure`. In many situations, the MLE and Bayesian MCMC approaches produce similar estimates; no one method is universally preferable. While the full range of advantages and disadvantages of the frequentist vis-à-vis Bayesian approaches remain to be worked out, some practical rules of thumb may be gleaned based on the simulation presented here.

First, in an information-rich scenario in which the markers are informative and the ancestral groups are large and accurate (Simulations 1 and 2), all methods we examined perform well: The point estimates are nearly unbiased and the estimated confidence intervals have coverage probabilities close to the nominal level. Bayesian MCMC methods, `structure` and `ADMIXMAP`, may achieve a marginally smaller root mean square error (RMSE), but require substantially longer computer time. This computational difference can be quite important for computer-intensive applications, for example in simulations.

Second, in an information-poor scenario in which the markers are only moderately informative or the ancestral groups are small (Simulation 3), our full ML approach remains unbiased while `structure` estimates can be substantially biased. Additional simulations using small number of markers and pseudo-ancestors indicate that `structure` usually produces a slightly smaller sampling variance compared to the MLE, but at the expense of being substantially biased. The comparison of the RMSE for the two estimates therefore depends on the bias-variance trade-off (Hastie *et al.* 2001).

Third, a potential limitation in both the partial ML and the full ML approaches is the assumption of independence among markers. However, under the 10 generation simulation setting we considered in Simulation 4, the MLE remained unbiased and the coverage probability of the BCa intervals was robust because there was only weak LD among markers. On the other hand, when LD is stronger (as might

happen if admixing has occurred extremely recently), the full MLE does underestimate the variability; as a result, the BCa interval tends to be too narrow. In contrast, `structure` can incorporate known map information and model the LD; in this setting, the credibility intervals produced by `structure` should have more accurate coverage probability than the BCa intervals.

Fourth, a comparison between the full ML approach we present here to the partial ML estimator (Hanis *et al.* 1986) reveals that allowing the ancestral allele frequencies and IA to be jointly estimated provides a greater level of robustness against imprecision in selection of pseudo-ancestral groups.

Fifth, our results confirm that microsatellites are generally more informative than SNPs for IA estimation, but that pre-selecting SNPs to have δ-values greater than 0.3 achieves comparable power.

Finally, our simulation results argue for the importance of pseudo-ancestors. For example, in estimating continental-level admixture in African-Americans, Africans from West Africa and European Americans may be appropriate proxies. In the absence of these pseudo-ancestors, the ancestral populations may not be identifiable; in such situations, both Bayesian and frequentist approaches produce biased estimates. In theory, the identifiability problem does not arise if a number of admixed individuals in the study are almost "pure." However, in practice, we cannot assume, *a priori*, that such individuals will be present. For example, suppose we wish to estimate European admixture in African-Americans using a sample consisting entirely of individuals who identify themselves as African-American. It is unreasonable to assume any of them has no European admixture, nor is it reasonable to assume any of them has exclusively European ancestry.

Given the importance of including pseudo-ancestors in IA estimation, we believe it is infeasible to estimate the number of ancestral populations from a sample of admixed individuals. Furthermore, for the purpose of estimating continental-level gene-flow, sufficient knowledge is usually available. For example, African-Americans are largely an African group receiving European admixture; Puerto Ricans are typically a mix of European, Native American and African/African-American (Hanis *et al.* 1991), while Mexican/Mexican Americans are primarily of European and Native American ancestry with some African ancestry as well.

With the recent development of high throughput, array-based technology, it is now affordable to genotype 10,000 - 100,000 SNPs simultaneously (Kennedy *et al.* 2003). Systematically genotyping various indigenous populations using this or other similar platforms has at least two advantages: first, population-specific allele frequencies obtained from these studies enable all future association studies to genotype only the admixed individuals, which thereby facilitates the control for IA variation. Second,

informative SNPs with large δ-values between populations can be selected to improve the IA estimates in a future study in which large-scale genotyping is not practical. For example, Smith *et al.* (2004) recently developed a high-density map of informative markers for admixture mapping in African-Americans.

While it is tempting to simply use pre-existing databases for selecting informative markers and for estimating pseudoancestral allele frequencies, caution needs to be applied. Selecting SNPs from major inventories to have large δ-values leads to exaggeration of the allele frequency differences for those SNPs and distortion of allele frequencies in the ancestral individuals typed. Therefore, before such SNPs are used in practice, independent samples of ancestral populations need to be typed for those SNPs to obtain more accurate, unbiased allele frequency estimates.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health Grant P01 CA53996, and by a Stanford Graduate Fellowship (to Jie Peng and Pei Wang). We are grateful for Dr. Dan Schaid and the two anonymous reviewers who provided many insightful and valuable comments.

ELECTRONIC DATABASE INFORMATION

The URLs for data presented herein are as follows:

The SNP Consortium Diversity Panel genotype: http://snp.cshl.org/linkage_maps/

The World Human Diversity Panel genotype: http://www.marshfieldclinic.org/research/genetics/Freq/FreqInfo.htm

REFERENCES

Bonilla C, Shriver MD, Parra EJ, Jones A, Fernandez JR (2004) Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York city. Hum Genet 115:57-68.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL (2002) A human genome diversity cell line panel. Science. 296:261-2.

Chikhi L, Bruford MW, Beaumont MA (2001) Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. Genetics. 158:1347-62.

Clark AG, Nielsen R, Signorovitch J, Matise TC, Glanowski S, Heil J, Winn-Deen ES, Holden AL, Lai E (2003) Linkage Disequilibrium and Inference of Ancestral Recombination in 538 Single-Nucleotide Polymorphism Clusters across the Human Genome. Am J Hum Genet. 73:285-300.

Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B. 39:1-38.

Efron B, and Tibshirani RJ (1993) An Introduction to the Bootstrap. New York: Chapman Hall.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure: Extensions to linked loci and correlated allele freqeuncies. Genetics. 164:1567-87.

Hastie T, Tibshirani R, Friedman, J (2001) The Elements of Statistical Learning. Springer, New York.

Hanis CL, Chakraborty R, Ferrell RE, Schull WJ (1986) Individual admixture estimates: disease associations and individual risk of diabetes and gallbladder disease among Mexican-Americans in Starr County, Texas. Am J Phys Anthropol. 70:433-41.

Hanis CL, Hewett-Emmett D, Bertin TK, Schull WJ (1991) Origins of U.S. Hispanics. Implications for diabetes. Diabetes Care. 1991 14:618-27.

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.

Hudson, RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338.

Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips MS, Boyce-Jacino MT, Fodor SP, Jones KW (2003) Large-scale genotyping of complex DNA. Nat Biotechnol. 21:1233-7.

Kingman JFC (1982) The coalescent. Stoch. Proc. Appl. 13:235-248.

Long JC (1991) The genetic structure of admixed populations. Genetics. 127:417-28.

Matise TC, Sachidanandam R, Clark AG, Kruglyak L, Wijsman E, Kakol J, Buyske S, et al. (2003) A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. Am J Hum Genet 73:271–284.

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

Pfaff CL, Barnholtz-Sloan J, Wagner JK, Long JC (2004) Information on ancestry from genetic markers. Genet Epidemiol. 26:305-15.

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly MJ, Reich D (2004) Methods for high-density admixture mapping of disease genes. Am J Hum Genet. 74:979-1000.

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

Rambaut A and Grassly NC (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13: 235-238.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298: 2981-2985.

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet. 73:1402-22.

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 60:957-64.

Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, Malasky MJ, Scafe C, Le E, De Jager PL, Mignault AA, Yi Z, De The G, Essex M, Sankale JL, Moore JH, Poku K, Phair JP, Goedert JJ, Vlahov D, Williams SM, Tishkoff SA, Winkler CA, De La Vega FM, Woodage T, Sninsky JJ, Hafler DA, Altshuler D, Gilbert DA, O'Brien SJ, Reich D (2004). A high-density admixture map for disease gene discovery in african americans. Am J Hum Genet. 74:1001-13.

Wang J (2003) Maximum-Likelihood Estimation of Admixture Proportions From Genetic Data. Genetics 164:747–765.

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

Zhu X, Cooper RS, Elston RC (2004). Linkage analysis of a complex disease through use of admixed populations. Am J Hum Genet. 74:1136-53.

Ziv E, Burchard EG (2003) Human population structure and genetic association studies. Pharmacogenomics 4:431-41.

Table I: Results of Simulation 1 (SNP data, include pseudo-ancestors) using three approaches. A description of the data can be found in the caption of Figure 2. Bias is defined as mean($\hat{q}_i - q_i$); Length of CI denotes the median length of the 90% confidence (or credibility) intervals. Coverage probability refers to that of the 90% intervals. Computation was performed on a Dell workstation with a 2GHz processor.

	`structure`	`ADMIXMAP`	EM
Bias	0.011	0.004	0.001
RMSE	0.052	0.050	0.053
Length of CI	0.17	0.16	0.17
Coverage probability (%)	89	88.8	89.4
Computing time (hours)	5	NA	3.5

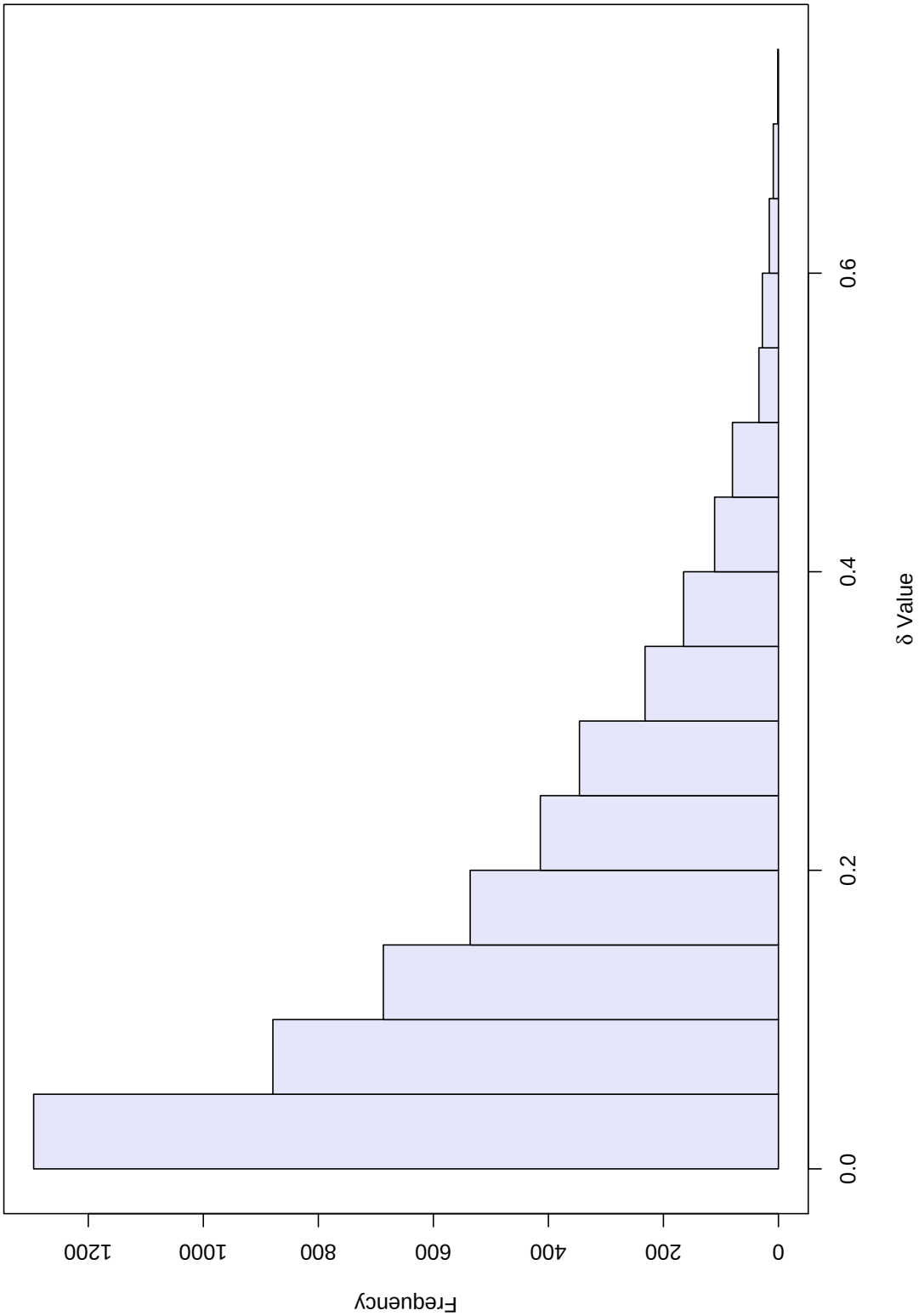
Figure Legends

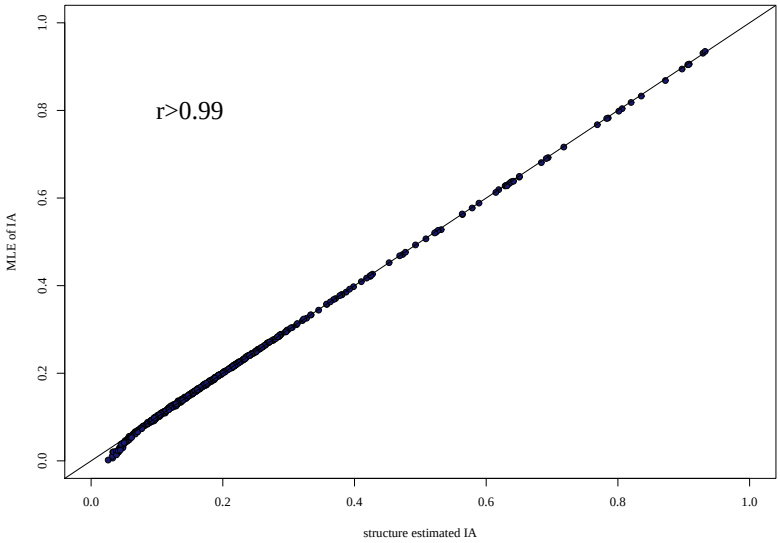
Figure 1. Distribution of δ-value for 4833 SNPs between African-Americans and Europeans American. SNPs were genotyped by the SNP Consortium. The panel includes 30 African-Americans and 30 European Americans. The mean δ-value is 0.16, and median 0.12.

Figure 2. Results of simulation 1: SNP data. (a) Comparison of IA estimates using program `structure` (x-axis) to the MLE (y-axis). (b) MLE of IA (points) with BCa intervals (vertical lines). Simulated data includes 500 admixed individuals and 250 pseudo-ancestors from each of the two ancestral populations. Subjects are genotyped at 200 SNPs with a minimum δ-value 0.3 (mean δ-value is 0.4).

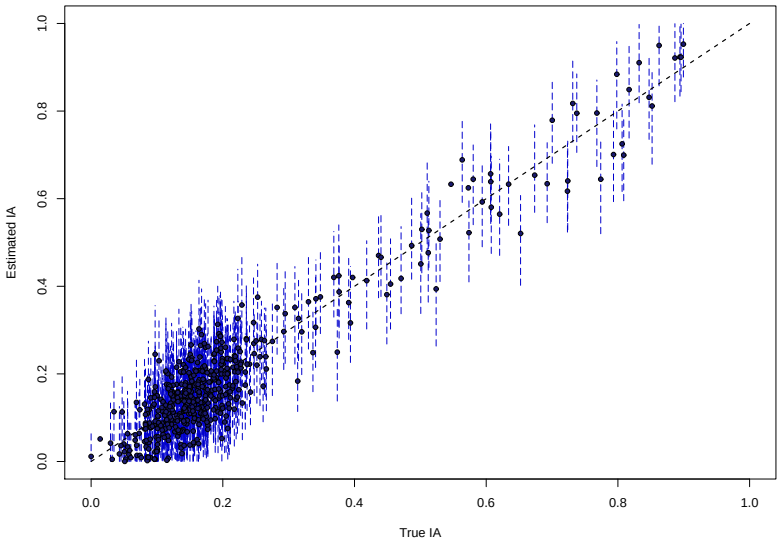
Figure 3. IA estimates using less informative markers. Data set includes 20 and 60 individuals from the two ancestral populations, respectively, and 500 admixed individuals. The true IA has a mean of 0.237. SNP markers are chosen to have a mean δ-value of 0.33. Each individual is genotyped at 60 SNPs.

Figure 4. Comparison of `structure`(x-axis) and MLE (y-axis) estimates of IA for the data of Simulation 4. IA for all individuals is near 0.2. Simulated data include 500 admixed individuals and 250 pseudo-ancestors from each ancestral populations; all individuals are genotyped at 421 SNP markers. The average δ-value between the two ancestral populations is 0.15.

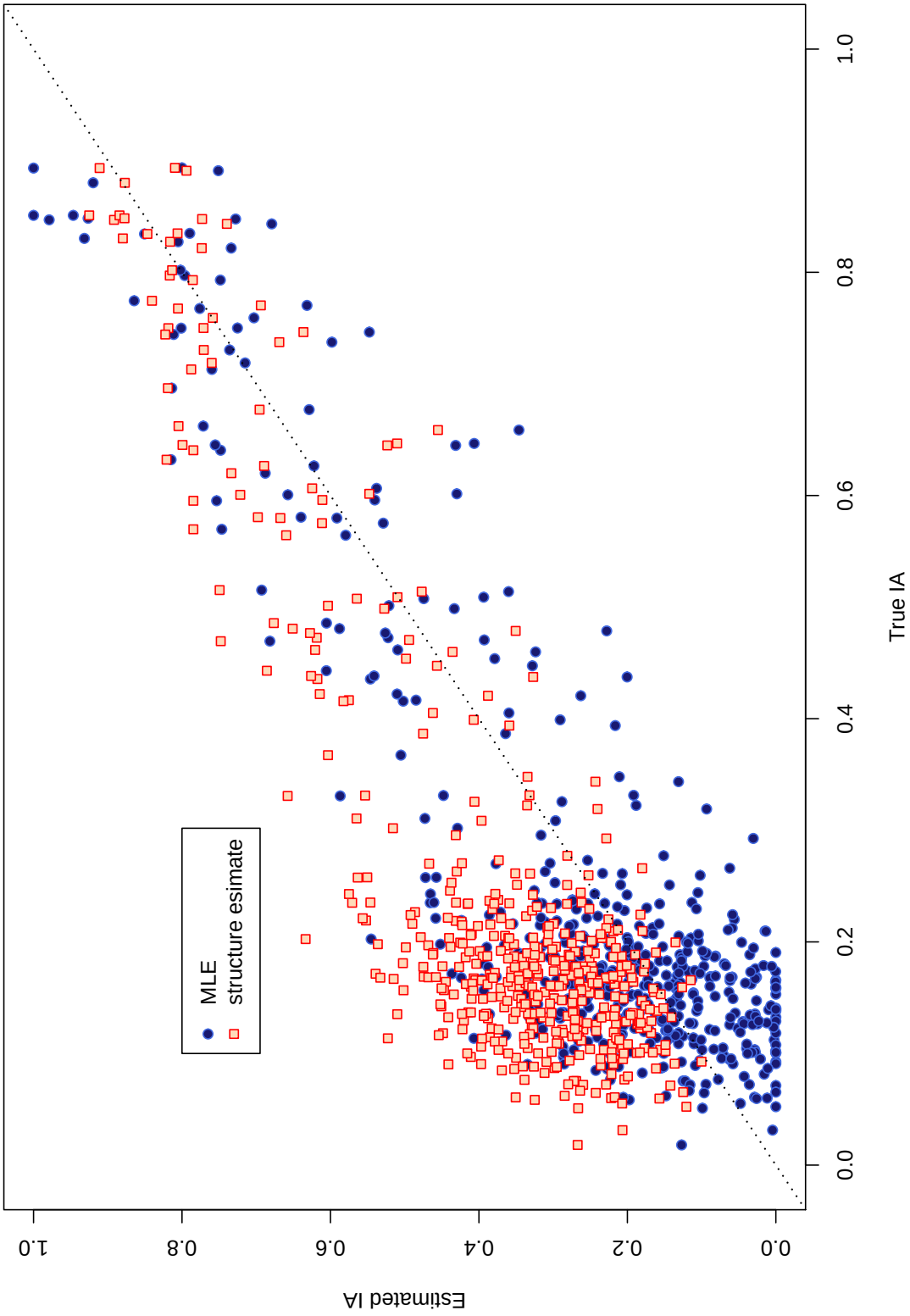

0
200
400
600
800
1000
1200
Frequency
0.0
0.2
0.4
0.6
δ Value

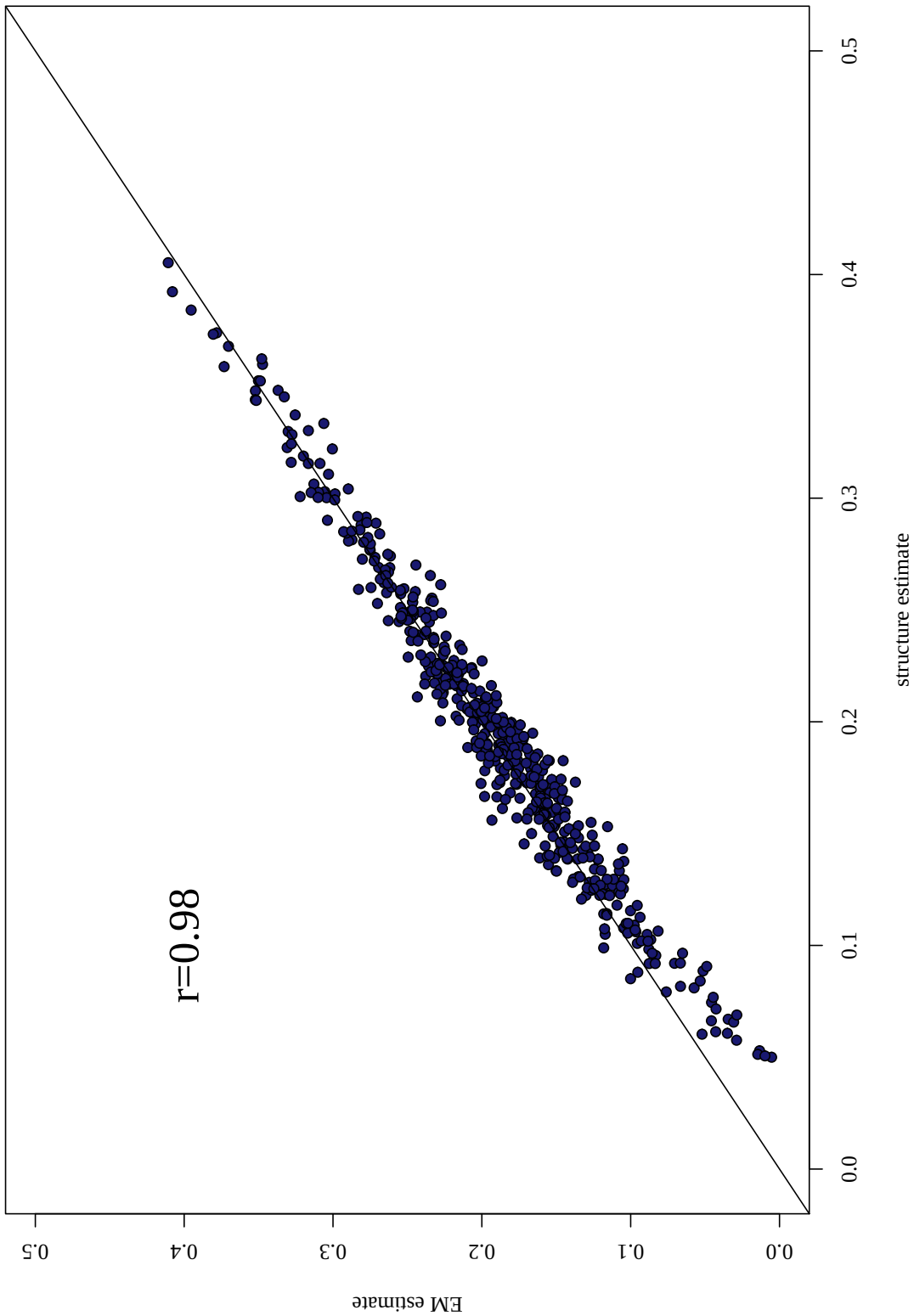

r>0.99
MLE of IA
structure estimated IA
0.0
0.2
0.4
0.6
0.8
1.0

(a)


Estimated IA
True IA
0.0
0.2
0.4
0.6
0.8
1.0

(b)


MLE
structure esimate
Estimated IA
True IA
0.0
0.2
0.4
0.6
0.8
1.0


r=0.98
EM estimate
structure estimate
0.0
0.1
0.2
0.3
0.4
0.5