

# Practical Population Group Assignment with Selected Informative Markers: Characteristics and Properties of Bayesian Clustering via STRUCTURE

Bao-Zhu Yang,<sup>1,2</sup> Hongyu Zhao,<sup>3</sup> Henry R. Kranzler,<sup>4</sup> and Joel Gelernter<sup>1,2\*</sup>

<sup>1</sup>Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut

<sup>2</sup>VA Connecticut Healthcare Center, West Haven, Connecticut

<sup>3</sup>Departments of Epidemiology and Public Health and Genetics, Yale University School of Medicine, New Haven, Connecticut

<sup>4</sup>University of Connecticut School of Medicine, Department of Psychiatry, Farmington, Connecticut

Population stratification, which is caused by population genetic substructure (PGS), is a critical issue for the design and interpretation of genetic association studies. Methods to address this problem have been devised, but little is known at this point about practical genotyping requirements for resolving PGS based on different marker characteristics. In this report, we seek to (1) identify a small, practical marker set to differentiate African Americans (AAs) from European Americans (EAs), and (2) assess the impact of marker efficiency and sample size on clustering individuals into subgroups by the methods of STRUCTURE (Pritchard et al., [2000a] *Genetics* 155:945–959). A panel of 37 markers was genotyped for 865 individuals (640 EAs and 225 AAs) from the Northeastern United States. Among EAs, the assignment accuracy reached >99% using only the 4 most efficient markers. Among AAs, the assignment accuracy exceeded 95% when using the 6 most informative markers. Smaller sample size increased the variance in population differentiation, rather than degrading the results consistently. We conclude that the use of marker-efficiency measures for marker selection yielded a relatively small set of STR markers that were effective at differentiating EA and AA populations. The number of markers required is much lower than has been suggested in previous studies. *Genet. Epidemiol.* 28:302–312, 2005. © 2005 Wiley-Liss, Inc.

**Key words:** STRUCTURE; population stratification; case-control; European American; African American; short tandem repeats

Contract grant sponsor: NIH; Contract grant numbers: MH14276, MH01387, DA12690, DA12849, DA12468, AA11330, AA12870, AA13736, AA03510, GM59507, and RR06192; Contract grant sponsor: U.S. Department of Veterans Affairs; Contract grant sponsor: VA CT REAP.

\*Correspondence to: Joel Gelernter, MD, Yale University School of Medicine, VA CT Healthcare Center, Psychiatry 116A2, 950 Campbell Avenue, West Haven, CT 06516. E-mail: joel.gelernter@yale.edu

Received 11 October 2004; Revised 20 December 2004; Accepted 4 January 2005

Published online 22 March 2005 in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/gepi.20070

## INTRODUCTION

The case-control association approach has been advocated as a possible solution to the limited power of genetic linkage studies in mapping genes for complex traits [see, e.g., Risch, 2000]. However, the limitations of this approach have also been widely recognized. One of the most concerning of these limitations (because it has historically been difficult both to detect and to address analytically) is the problem of population stratification caused by population genetic substructure (PGS).

Human PGS is the consequence of events in population history, such as migration, isolation, genetic drift, non-random mating, and bottle-

necks. PGS cannot be observed directly; the common conventional surrogate of PGS, race or ethnicity, may not be adequate to identify PGS, especially for admixed populations. Moreover, self-identification does not take into account the inherently admixed nature of many important population groups, such as African Americans.

PGS, most notoriously, may lead to the generation of spurious results from genetic association studies [Cardon and Palmer, 2003; Freedman et al., 2004]. The mechanism of these spurious effects originates from differential sampling from structured populations of study participants in case and control groups, if underlying cryptic PGS exists and both trait prevalence and allele frequencies of candidate loci vary among constituent

subpopulations. Certain populations may have higher disease susceptibility; if such populations are over-represented in the case group, any marker locus with higher or lower allele frequencies tends to associate with the disease. Although the impact of this effect in generating spurious results has been debated [Thomas and Witte, 2002; Wacholder et al., 2002], it remains a clear problem for studying recently admixed populations (such as African Americans and Hispanic Americans) [Thomas and Witte, 2002; Hoggart et al., 2003, 2004]. Thanks to newly available statistical methods to evaluate population structure (discussed below), it is now possible to determine the extent to which this phenomenon influences the design and interpretation of genetic association studies.

In recent years, several methods have been proposed to allow valid analysis of genetic association in case-control samples by detecting and accounting for population structure [Devlin and Roeder, 1999; Pritchard et al., 2000a,b; Reich and Goldstein, 2001; Ripatti et al., 2001; Satten et al., 2001; Sillanpaa et al., 2001; Zhang and Zhao, 2001; Zhang et al., 2003; Pfaff et al., 2002; Chen et al., 2003; Falush et al., 2003; Hoggart et al., 2003]. Among these methods, the structured association (SA) method [Pritchard et al., 2000a,b] has received considerable attention owing to the elegance of its solution to the problem of stratification, its relative preservation of power, its accommodation of numerous genetic marker types (including both SNPs and STRs), and the availability of well-designed, easily available software programs to implement the method.

Several published studies have used Pritchard's method to detect PGS. A study of congestive heart failure [Small et al., 2002] used 9 markers ostensibly to rule out a false-positive due to population stratification. Romualdi et al. [2002] used 21 random biallelic Alu markers in 32 populations and concluded that there was little evidence of a clear subdivision of humans into biologically defined groups. Wilson et al. [2001] used 16 (+23 X-linked microsatellite) markers to infer human PGS in the evaluation of drug safety and efficacy, and to understand the relationship between PGS and drug response. It is unclear if the number, and even more importantly, the nature, of the markers used for these studies provided a sufficient basis to detect underlying PGS.

Some larger-scaled applications of these methods for PGS detection or inference have gained attention recently [Rosenberg et al., 2002, Bamshad et al., 2003, Turakulov and Easteal 2003]. Rosen-

berg et al. [2002] used 377 autosomal microsatellite loci that apparently were not selected because of their population genetic characteristics in 1,056 individuals from 52 populations, and employed the geographic (continental) origin of individuals to identify five genetic clusters corresponding to these geographic regions, out of six main genetic clusters inferred by STRUCTURE. Bamshad et al. [2003] used 100 Alu biallelic markers and 60 microsatellite markers, and implemented the models of admixture and uncorrelated allele frequencies in STRUCTURE. They found that a set of about 100 markers with characteristics similar to those of the markers they used were required to differentiate populations via STRUCTURE. Turakulov and Easteal [2003] concluded that more than 65 random SNP markers are required for identifying distinct geographically separated populations. The conclusions regarding number of markers required to evaluate PGS from these two studies [Bamshad et al., 2003; Turakulov and Easteal 2003] were derived using the bootstrap procedure; therefore, these conclusions need to be explained and evaluated carefully in that light.

Partitioning the ancestry of individuals accurately is an important step in the structured association method for adjusting bias caused by PGS. The program-generated partition accuracy can be evaluated by comparison with subjects' self-identification of population group. Assignment accuracy depends on many factors, such as marker efficiency [Rosenberg et al., 2003] and number of markers. In light of various studies discussed above, we sought to delineate practical genotyping requirements for inferring PGS by the methods of STRUCTURE. We used a commercially available marker set often used in forensic applications, and also developed a new marker panel designed to include markers with high  $\delta$  to distinguish between EAs and AAs.

## MATERIALS AND METHODS

### SUBJECTS

Eight hundred sixty-five individuals recruited in Connecticut, at Yale University School of Medicine or the University of Connecticut Health Center, were studied and classified as 640 EAs and 225 AAs. The classification was based on self-identification and verified by interviewers' observation. All subjects provided informed consent as approved by the appropriate institutional review boards.

## MARKERS AND GENOTYPING

Two different sets of STR markers were used. First, we used the AmpFLSTR Identifiler PCR Amplification Kit [Applied Biosystems (ABI), Foster City, CA], which provides data from a set of 16 loci useful for forensic purposes (D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S443, vWA, TPOX, D18S51, D5S818, FGA, and amelogenin). Amelogenin is used for sex identification rather than for polymorphism content, so information from that locus was not included in any analyses. The markers in this set are all co-amplified in a single PCR reaction. Second, we selected 21 markers known to have high  $\delta$ , a measure of marker efficiency (more detail below in "Classification of Marker Efficiency"), between EAs and AAs, and in some cases Hispanic and Asian populations, based on the report of Smith et al. [2001]. This marker panel includes markers D1S196, D1S2628, D2S162, D2S319, D5S407, D5S410, D6S1610, D7S640, D7S657, D8S272, D8S1827, D9S175, D10S197, D10S1786, D11S935, D12S352, D14S68, D15S1002, D16S3017, D17S799, and D22S274. We selected markers that could be genotyped in a single lane on the ABI PRISM 3100, redesigning primer sequences when necessary. Thus, the 36 STR markers studied can be genotyped in 2 lanes. All STR markers were analyzed on an ABI PRISM 3100 semiautomated capillary fluorescence sequencer. Data were scored using Genemapper (ABI). We also genotyped marker FY, added to the 36 STRs because of its known value in identifying individuals of primarily African ancestry [see, e.g., Lautenberger et al., 2000]. This SNP marker was genotyped via PCR-RFLP. The average percentage of missing genotypes for all 37 markers for the entire sample was 3.27%.

## CLASSIFICATION OF MARKER EFFICIENCY

Marker efficiency is defined by  $\delta$  [Smith et al., 2001], a measure of marker information that reflects cumulative allele frequency differences, and, therefore, efficiency in statistically separating populations. The estimate of  $\delta$  for each marker locus was calculated as the sum over all  $L$  alleles of the absolute value of allelic frequency difference between two populations divided by 2, i.e.,

$$\delta = \frac{1}{2} \sum_{i=1}^L |p_i^A - p_i^B| \quad (1)$$

where  $p_i^A$  and  $p_i^B$  are the allele frequencies for the  $i$ th allele in population  $A$  and  $B$ , respectively. It is

straightforward to find  $0 \leq \delta \leq 1$ .  $\delta$  is zero when the marker locus in two populations contains exactly the same distribution of allele frequencies. On the other hand,  $\delta$  is 1 when two populations do not share any common allele for the marker. The larger the value of  $\delta$ , the more informative the marker is for differentiating two populations.

## ANALYSIS WITH STRUCTURE

We used STRUCTURE 2.0 (March 2002) to evaluate the optimal number of clusters and to assign each individual to a subgroup without using the predefined racial information. STRUCTURE adopts a Bayesian cluster approach under the main assumptions of HWE within populations and complete linkage equilibrium between loci within populations for markers not in admixture linkage disequilibrium.

For our STRUCTURE analyses, a bootstrap procedure of random selection of increasing numbers of markers (up to all 37 available markers) was carried out to compare models of correlated (correlations between markers) and independent (independent markers) allele frequencies, both with the software-specific "ancestry" model of admixture. Each run for the set of randomly selected markers was repeated for 100 iterations, and used 30,000 burn-in and 20,000 Markov chain Monte Carlo (MCMC) iterations (or 100,000 burn-in and 100,000 MCMC iterations for 1 to 15 markers).

To assess the impact of marker efficiency and the number of markers on assignment accuracy, markers were added to the analysis one by one based on the magnitude order of  $\delta$  in an ascending or descending manner, with 100,000 burn-in and 100,000 MCMC iterations and appropriate STRUCTURE modeling.

If  $K$  denotes the number of clusters modeled, using the estimated proportions of ancestry for the  $K$  ancestral origins, an individual was considered to be assigned accurately to a group when the greatest proportion of ancestry identified the same ethnicity as the pre-defined population group of the individual (by self-identification). Assignment accuracy in each population group was defined as the proportion of correctly assigned ethnicities.

The effect of sample size on clustering was examined by an empirical approach. Equal-sized subsamples of 5, 10, 15, 20, 25, 50, and 225 individuals were drawn randomly without replacement from the predefined populations of 640

EAs and 225AAs, and then combined into 7 data sets of sample size 10, 20, 30, 40, 50, 100, and 450 subjects. These 7 random samples were submitted to STRUCTURE using all 37 markers with models of admixture and allele frequencies correlated, and 100,000 burn-in and 100,000 MCMC iterations. These procedures were repeated 100 times for each sample size.

## RESULTS

The primary runs of STRUCTURE for the dataset of 865 subjects without using racial self-identification showed that the optimal number of clusters,  $K$ , was 2 based on the estimated posterior probabilities (data not shown). For determination of the appropriate model to be implemented in STRUCTURE, bootstrap random selection of markers, i.e., sampling markers with replacement, was used to compare the “independent” and

“correlated” allele frequencies models with the ancestral model of admixture. The results (Fig. 1) showed no major difference between the two models in assignment accuracy for EAs or AAs. The 95% confidence interval shown in Figure 1 was based on the bootstrap percentile. The upper 97.5 percentile of assignment accuracy indicates that only five markers are needed to achieve >95% assignment accuracy for the model of “allele frequencies correlated.” On the other hand, more markers are required to achieve 95% accuracy for the lowest 2.5 percentile of assignment accuracy. Thus, as expected, assignment accuracy depends in part on the number of markers considered.

Figure 2 presents the assignment accuracy, which changed as the replicates proceeded, for varying sample size. For a sample size as small as 10 subjects (5 in each population group), 42 and 59% of the 100 replicates reached an assignment accuracy of 100% for EAs and AAs, respectively.

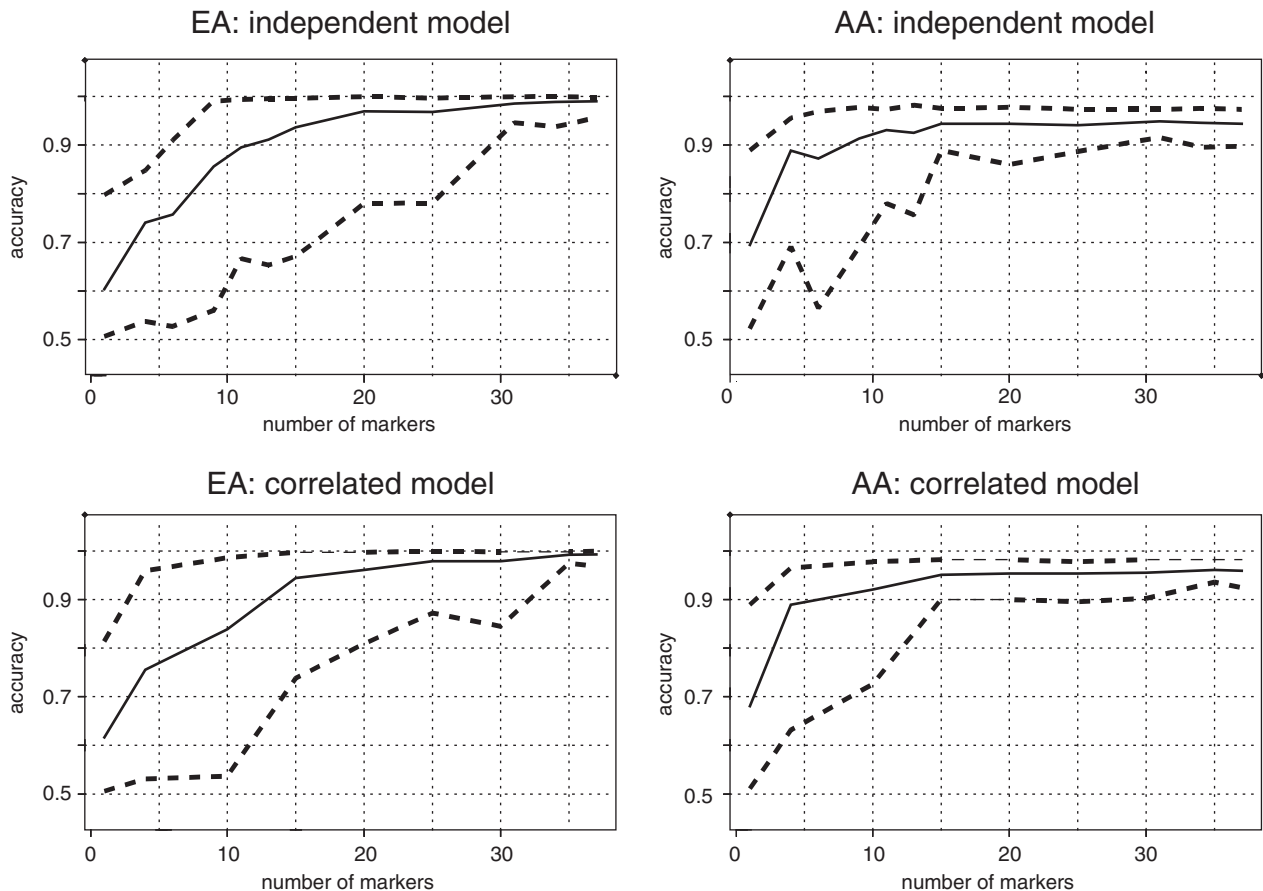
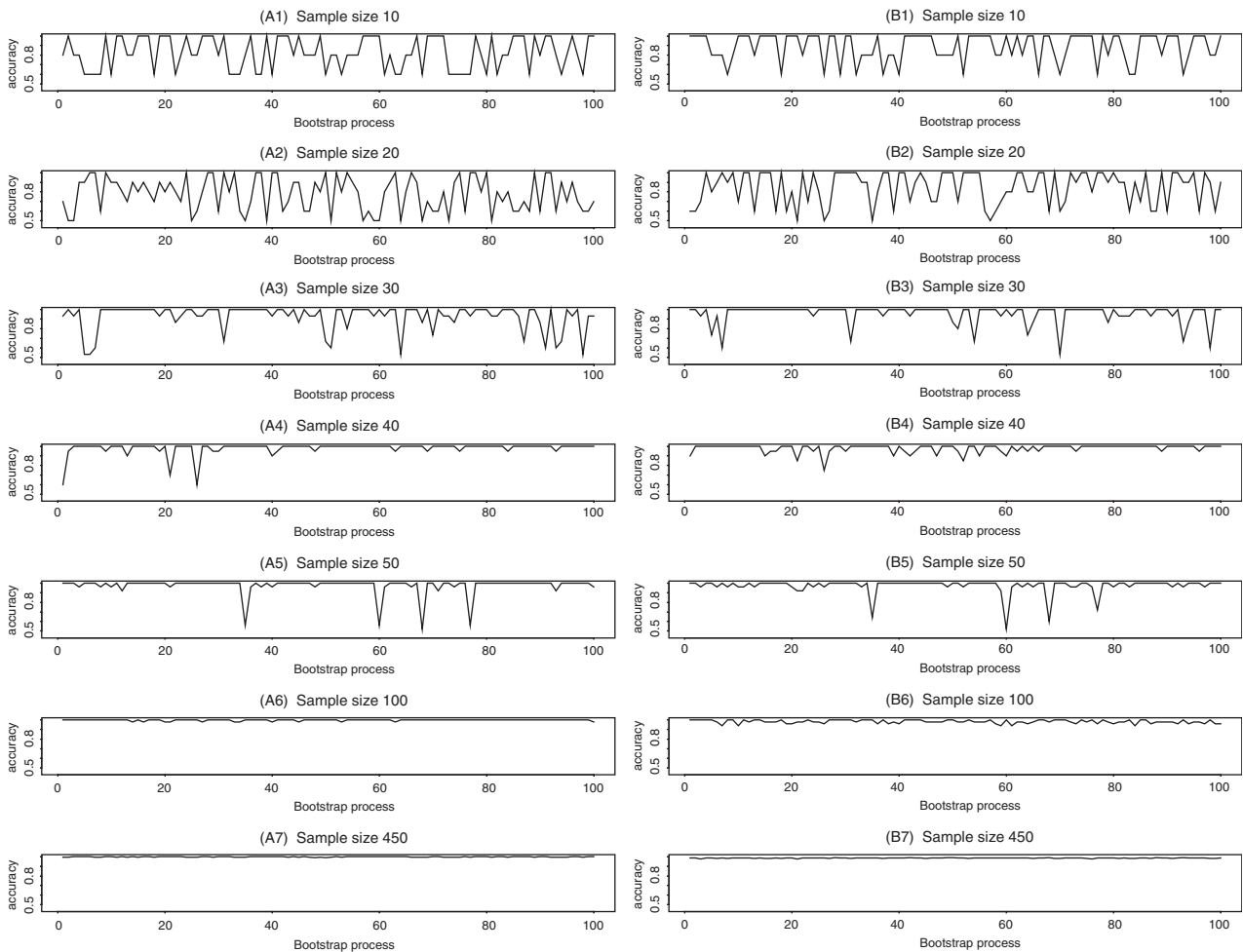


Fig. 1. Assignment accuracy in EA and AA samples. Comparison of allele frequencies correlated model and allele frequencies independent model in assignment accuracy by bootstrap procedure of 100 replicates implemented in STRUCTURE. The solid lines represent the bootstrap means, while the dashed lines represent the 2.5 and 97.5 percentiles.



**Fig. 2. Sample size impact on assignment accuracy.** (A1)–(A7) are for EA. (B1)–(B7) are for AA. Empirical random selection of equal size from EA and AA in each batch of total sample sizes from 10, 20, 30, 40, 50, 100, up to 450.

For a sample size of 20 (i.e., 10 in each group), assignment accuracy  $\geq 90\%$  was observed for 41 and 60% of analysis runs for EAs and AAs, respectively. Here the accuracy decreased when we doubled the sample size to 20. (Possibly, when the sample size is not large enough to yield a clear clustering pattern, the sparser data points tend to distort clustering performance.) As the sample size increased to 40 or 50 subjects (20 or 25 in each group), the assignment accuracy increased as well with few instances of lower assignment accuracy (Fig. 2, A5 and B5). When the sample size increased to a total of 100 subjects, the randomness of assignment accuracy diminished and the accuracy reached close to 100% for all 100 replicates. The same results were observed for a sample size of 450, which included the whole AA sample of 225 and a random sample of 225

EAs (Fig. 2, A1 compared to A7, B1 compared to B7).

In order to ascertain the smallest marker set effective at differentiating the populations under study from the markers used herein and to determine how many markers are needed to resolve PGS, we took the approach of selecting markers based on their efficiency. The estimated efficiencies of each marker,  $\delta$ , (calculated per equation (1)) are listed in Table I. It should be noted that, even for samples of subjects representing the same ethnicity, the rank of these markers might vary slightly among different study populations due to sampling variation and geographical location. The Spearman rank correlation coefficient is 0.82 between the deltas for our populations and those originally reported for the markers (ABI, Foster City, CA) [Smith et al., 2001].

TABLE I.  $\delta$  for the panel of 36 STR and FY markers studied

Order <sup>a</sup>	Locus	$\delta$ (EA/ AA)	Source <sup>b</sup>	Rank/ study <sup>c</sup>	Rank/ reference <sup>d</sup>
1	FY	0.824	L	1	1
2	D11S935	0.541	S	2	7
3	D15S1002	0.499	S	3	10
4	D7S657	0.496	S	4	2
5	D9S175	0.449	S	5	6
6	D10S1786	0.448	S	6	12
7	D1S2628	0.423	S	7	3
8	D5S410	0.406	S	8	5
9	D17S799	0.405	S	9	13
10	D8S1827	0.366	S	10	15
11	D6S1610	0.349	S	NA <sup>c</sup>	NA <sup>e</sup>
12	D16S3017	0.348	S	11	17
13	D8S272	0.342	S	12	11
14	D5S407	0.336	S	13	16
15	D7S640	0.323	S	14	4
16	TH01	0.322	A	15	21
17	D8S1179	0.290	A	16	26
18	D13S317	0.284	A	17	28
19	D2S162	0.280	S	18	9
20	D2S1338	0.278	A	19	27
21	D19S433	0.274	A	20	22
22	D2S319	0.271	S	21	24
23	D14S68	0.262	S	22	8
24	D18S51	0.260	A	23	23
25	D10S197	0.259	S	24	18
26	D12S352	0.233	S	25	20
27	TPOX	0.214	A	26	25
28	vWA	0.212	A	27	33
29	D22S274	0.203	S	28	14
30	D1S196	0.197	S	29	19
31	D21S11	0.192	A	30	34
32	D16S539	0.182	A	31	31
33	D7S820	0.177	A	32	36
34	D5S818	0.168	A	33	30
35	CSF1PO	0.155	A	34	29
36	D3S1358	0.152	A	35	35
37	FGA	0.151	A	36	32

<sup>a</sup>The order represents the descending magnitude of  $\delta$ .

<sup>b</sup>S, A, and L denote the source from Smith et al. [2001], ABI [2001], and Lautenberger et al. [2000], respectively.

<sup>c</sup>Rank by deltas of the study populations.

<sup>d</sup>Rank by the deltas of the reference populations. <sup>e</sup>This marker was excluded (see Results [Figure 3]).

The Spearman rank correlation coefficient between  $\delta$  and  $F_{st}$  in our study is 0.55 (data not shown), which indicates a moderate correlation. However, near equivalence (Spearman correlation coefficient=0.9998; data not shown) is found between  $\delta$  and the “ $I_n$ ” index, informativeness for assignment, based on the equation (4) in Rosenberg et al. [2003]. Note that Rosenberg et al. [2003] use a different definition of  $\delta$  than that used in the present article. Thus,  $I_n$  adds

minimal, if any, information to the simpler  $\delta$  measure.

The relative assignment accuracy for STRUCTURE in clustering EAs and AAs was evaluated by adding markers one by one up to 36 markers (one marker was deleted from this analysis due to the allele frequency being inconsistent with that in other reference populations), with the order of  $\delta$  either descending or ascending; the results are shown in Figure 3. (When we considered the effects of  $F_{st}$  as a means to order markers by information content, it proved to be less informative than  $\delta$ ; data not shown.) FY was the most informative marker, and due to its unique value in distinguishing the EA and AA populations under study, we performed analyses separately either including or excluding this marker. In EAs (Fig. 3, (1)), the assignment accuracy reached >99% using the four most efficient markers including FY, and using the 10 most efficient markers excluding FY. In contrast, it would take 30 markers to reach >99% assignment accuracy when the least efficient markers were selected or the six most efficient markers were omitted (note that each marker in our set was selected either because of known high  $\delta$  or because of forensic value in distinguishing among individuals, and, therefore, this set as a whole should represent a relatively informative set for distinguishing populations, compared to a random set of markers). When FY was excluded and the best of the 35 remaining markers was used instead (i.e., starting with the second most efficient marker, D11S935), the initial assignment accuracy dropped 34%. This drop reflects the difference of assignment accuracy between FY ( $\delta=0.824$ ), and D11S935 ( $\delta=0.541$ ). In AAs (Fig. 3, (2)), the assignment accuracy could reach >95% when using at least the 6 most efficient markers. When FY was excluded, the assignment accuracy dropped 11%. These results showed that FY is powerful in distinguishing EAs from AAs.

In order to address the generalizability of our results, we also evaluated assignment accuracy by adding the markers from the most efficient ones to the least using  $\delta$  values calculated from the previously reported allele frequencies (ABI) [Smith et al., 2001]. In EAs, the assignment accuracy reached >99% using the six most efficient markers including FY, while in AAs, it took seven most efficient markers to reach 95%. These results are very close to those obtained by adding markers according to the  $\delta$ s of our own data. Using the previously reported  $\delta$ s rather than

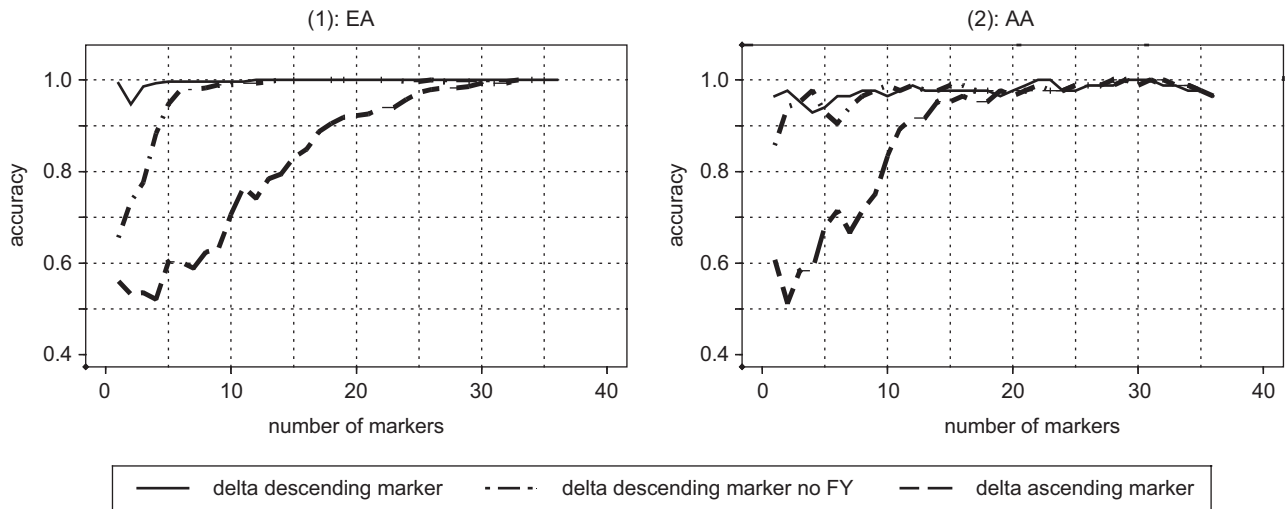


Fig. 3. Assignment accuracy of STRUCTURE. The markers are adding one by one either by  $\delta$  descending or ascending. Assignment accuracy without FY, the most efficient marker in the panel studied, was also evaluated.

the ones derived from our samples, the set of seven markers used to classify the AAs substitutes two other markers for those ranked 3th and 6th best using our own data. The markers substituted for these, actually rank 10th and 12th, if our own data are used to classify them (Table I). This represents a relatively small loss of information.

## DISCUSSION

### GENERAL COMMENTS

Through systematic investigation, we demonstrated that delineating PGS depends on the efficiency of the markers, the number of markers sampled, the extent of differentiation among population genetic subgroups (i.e., difference in magnitude of allele frequencies), and the extent of admixture ( $\delta$  proved to be a very useful indicator of marker efficiency, and it is also extremely easy to compute). However, a small sample size in each population increases the variance in population differentiation rather than degrading the results consistently. We demonstrated that through careful marker selection, the number of markers required for accurate population group assignment can be greatly decreased compared to prior estimates. Rosenberg et al. [2002] for example, noted that marker sets of 100–200 would often be necessary to distinguish populations (depending on their genetic distance), but also noted that reductions in marker number would be possible through selection of markers

according to their information content. We provide herein a first evaluation of how much increased information (with consequent reduction in genotyping requirement) can be provided through marker selection.

We demonstrated that using markers with high efficiency for distinguishing the populations (e.g., FY with  $\delta=0.82$ ) can dramatically enhance assignment accuracy; the increase in accuracy of separating EAs and AAs attained by adding this one marker was 34 and 11%, respectively. (We have exploited this unusual property of FY to approximate admixture potential in three previous studies [Lappalainen et al., 2002; Luo et al., 2004; Covault et al., 2004].) Consequently, the number of markers needed for reasonably accurate clustering is decreased by optimal marker selection; e.g., only four markers are needed to achieve >99% accuracy in EAs, which is considered to be a more homogeneous group genetically, and 6 markers are needed to achieve >96% accuracy for AAs. The number of markers needed according to our analyses is much less than what was reported by Bamshad et al. [2003]. Optimal resolution can be achieved by selection of markers that are most efficient in detecting PGS [Rosenberg et al., 2001; Bamshad et al., 2003; Watkins et al., 2003]. It is clear that fewer markers were required for accurate clustering in our study because we pre-selected many of our markers on the basis of their high  $\delta$  computed for the populations comparable to those we sought to resolve.

Results were similar when we used  $\delta$  values calculated from the previously reported allele frequencies rather than values computed from our own sample. Most markers did not change rank order by more than a few places, when our computed values are compared to previously published values (Table I). Using a less-optimal marker order results in a relatively small loss of information (keeping in mind that most of these markers were originally selected for published high  $\delta$ ), and suggests that our marker set should be generally useful for at least EA and AA populations. Further evidence supporting this inference is that the markers we used were selected according to published literature initially, and not from our own allele frequency data, yet they proved highly efficacious at classifying our particular sample. This marker set may, therefore, prove useful for investigators who wish to apply the structured association method to EA and AA populations.

#### EVALUATING THE IMPACT OF SAMPLE SIZE ON THE ACCURACY OF THE SUBPOPULATION ASSIGNMENT

When the effect of sample size on assignment accuracy was evaluated previously, it was found that the sample size required to reach >90% accuracy varied with the number of markers used [Rosenberg et al., 2001]. This evaluation was done by one-time random sampling of various sample sizes with the number of markers varied; in contrast, our evaluation was implemented through a 100-time random process of selecting individuals in various numbers with all 37 markers used. The results from our empirical assessment to select subsamples of subjects for comparing sample size effects suggest that the effect of sample size in each subpopulation depends on the differentiation between population genetic subgroups. If the random samples of individuals from each subpopulation yield two new subpopulations for which genetic patterns are sufficiently distinct, then as few as five persons in each group can yield close to 100% assignment accuracy in individual iterations. In small samples, it is possible that only a very small number (even zero, one, or two) of observations for some alleles in some of the markers in some of the population groups are available. In such cases, allele frequency estimates would be based on very little data. Even though some allele frequencies might not be estimated accurately for small

sample sizes, the Bayesian clustering is based on 37 dimensions (markers), which increase the chance to provide enough information on the boundary for differentiating groups. With small samples, STRUCTURE has very limited information available to classify subjects correctly, resulting in some inconsistency; either there is enough information in the data to locate the correct clusters and STRUCTURE can classify most subjects correctly, or it fails in classification globally. A larger sample size stabilizes the representation of the two sampled clusters, i.e., it reduces the variation of the two sample clusters. A larger sample size in each population could only be methodologically beneficial if increasing sample size broadens the differentiation between population genetic subgroups. If optimal efficiency and number of markers are achieved for stabilized or fixed differentiation between population genetic subgroups, smaller sample size should not decrease assignment accuracy. That is why, with a sample size as low as five persons in each population, over 40% of the 100 random samples yielded an assignment accuracy of ~100%. This simulation was based on equal sample sizes in the two populations. However, the ratio of sample sizes between the two clustered populations should not be a factor impacting the assignment accuracy because our study samples contain unbalanced sample sizes, specifically, 640 EAs and 225 AAs, and the assignment accuracy for AAs was almost as good as that for EAs.

STRUCTURE currently has the advantage of providing probabilistic population assignments for each individual in a sample suitable for use as input to the program STRAT (which is used for structured association analysis) [Pritchard et al., 2000b], making the process of structured association analysis straightforward. Furthermore, ancestral proportions from STRUCTURE can be used as covariates in adjusting for confounding caused by population stratification in a regression framework for case control association studies, if the panel of markers is appropriate and informative enough to cluster individuals accurately. Since assignment accuracy is defined by ancestral proportions, all factors affecting assignment accuracy also affect the estimate of ancestral proportions.

Some issues with STRUCTURE need to be handled carefully: convergence of MCMC, estimating the number of clusters, and missing data. The convergence issue was detected when we

assessed the impact of sample size. In Figure 2 (A5) and (B5), the apparent spikes for sample size 50 were further investigated. The random samples that generated those spikes were run 100 times under the same models and the same parameter set and increasing the length of MCMC in STRUCTURE. About 3% of the 100 replicates yielded assignment accuracy as low as 64%, with the rest of the replicates being 100%. This means that occasionally the algorithm is converging to a different clustering solution and longer runs do not remedy this problem because once the Markov chain reaches a particular mode, it is extremely difficult for it to exit it. For estimating the number of clusters, it is necessary to take the posterior estimation as a crude guide and incorporate other analytical or biological information. There are other recently developed Bayesian methods and software programs available to estimate the number of populations, allele frequencies, and individual assignments simultaneously [Dawson and Belkhir, 2001; Corander et al., 2003, 2004]. However, the performance of these methods and programs is unknown. STRUCTURE also assumes that data are missing at random, i.e., not informative about the actual allele, and the missing data are ignored in all calculations (personal communication, Dr. Jonathan Pritchard, August 12, 2003). In our analysis, when compared to the whole dataset with an overall rate of 3.27% of missing data, STRUCTURE performed better when we selected a subsample restricted only to subjects with complete data, i.e., no missing genotypes. It is, therefore, important to check whether data are missing at random and what impact that will have on performance before STRUCTURE is implemented.

We also compared our results with those in Bamshad et al. [2003]. They used the “allele frequencies independent” model in STRUCTURE. They concluded that, derived from the bootstrap procedure, at least 100 loci were needed to achieve 99–100% assignment accuracy for the continent of origin of Africa, Asia, or Europe and that more informative markers defined by  $F_{st}$  could improve the assignment accuracy. The bootstrap procedure is implemented by random sampling with replacement. As a result of the replacement, the same marker could be probabilistically selected more than once and submitted to analysis in differentiating PGS. If the most efficient markers were selected, the estimated assignment accuracy would be improved, compared to an unselected marker set. On the other hand, if the least efficient

markers were selected, the estimated assignment accuracy would be reduced. Therefore, the markers derived from the bootstrap procedure represent a wide range of marker diversity, and were not of as high a value of delta as our selected marker panel.

Our results used the subjects’ self-identification, confirmed by interviewer observation, for initial population group classification. Indeed, it would have been preferable to have reports of population background for all four grandparents of each subject, for example. However, this information is not available for many research studies. Rosenberg et al. [2002] noted that “...self-reported ancestry can facilitate assessments of epidemiological risks but does not obviate the need to use genetic information in genetic association studies” (p. 2381). They used pre-defined population origins as a “gold standard.” Bamshad et al. [2003] also used predefined population groups as a “gold standard” in assessing the accuracy of clustering by STRUCTURE. Although we acknowledge that this is a practical limitation, we note that this limitation is shared by nearly every published genetic association study that considers population group at all, and is a limitation that the present study aims to address. Moreover, if self-identification and observation were markedly inaccurate, we would have expected much more discrepancy with group clustering based on genetic marker data. Finally, we note that there is, unfortunately, no more valid method (than subject self-assessment combined with observation) available to us to assign a preliminary population group membership with which to compare STRUCTURE group assignments.

As noted above, the use of STRUCTURE and related approaches provides a significant advance in our ability to understand and discern PGS, apparent or occult. For optimal use of these tools, we need a thorough understanding of the effects of various parameters on the outcome, and, ideally, a specific set of markers to yield the best population structure information for the populations to be resolved. The use of  $\delta$  for marker selection yielded a set of STR markers that we demonstrated to be effective at differentiating EA and AA populations. Our experience with other EA and AA populations using the same strategy suggests that this is generally a reasonable strategy. By means of careful marker selection, excellent clustering may be achieved with a marker set markedly smaller than that proposed in several previous reports. The marker set

described here may be genotyped in a total of only 2 sequencer lanes, which results in low cost and high data collection efficiency. We demonstrated that sample size, within the range generally used for case-control association designs, should not be an impediment to achieving high-accuracy clustering. These results should aid in the practical application of population clustering methods, particularly for EA and AA samples.

## ELECTRONIC-DATABASE INFORMATION

Pritchard Lab, <http://pritch.bsd.uchicago.edu/> (for the software program STRUCTURE, used to detect PGS and infer the population assignment). Applied Biosystems (2001) AmpFLSTR® Identifier™ PCR Amplification Kit: User's Manual., User manual number 4323291B. <http://docs.appliedbiosystems.com/search.taf> (for allele frequencies of the reference group ABI).

## ACKNOWLEDGMENTS

Greg Dalton-Kay provided excellent technical assistance; Dr. Jonathan Pritchard provided helpful comments; Drs. Michael Bamshad and Lynn Jorde kindly provided their data set, and also both provided helpful comments on the manuscript. This work was supported in part by grants from NIH: MH14276 (Biological Sciences Training Program support to B.Z.Y.), MH01387, DA12690, DA12849, and DA12468, AA11330, AA12870, AA13736, AA03510, GM59507, and RR06192 (University of Connecticut General Clinical Research Center), as well as the U.S. Department of Veterans Affairs (the VA Medical Research Program [Merit Review to J.G.], and the VA CT REAP [Research Enhancement Award Program]).

## REFERENCES

- ABI, Applied Biosystems. 2001. AmpFLSTR® Identifier™ PCR Amplification Kit: User's Manual., User manual number 4323291B. Foster City, CA: Applied Biosystems.
- Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB. 2003. Human population genetic structure and inference of group membership. *Am J Hum Genet* 72:578–589.
- Cardon LR, Palmer LJ. 2003. Population stratification and spurious allelic association. *Lancet* 361:598–604.
- Chen HS, Zhu X, Zhao H, Zhang S. 2003. Qualitative semi-parametric test to detect genetic association in case-control design under structured population. *Ann Hum Genet* 67: 250–264.
- Corander J, Waldmann P, Sillanpaa MJ. 2003. Bayesian analysis of genetic differentiation between populations. *Genetics* 163:367–374.
- Corander J, Waldmann P, Marttinen P, Sillanpaa MJ. 2004. BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* 20:2363–2369.
- Covault J, Gelernter J, Hesselbrock V, Nellissery M, Kranzler HR. 2004. Allelic and haplotypic association of GABRA2 with alcohol dependence. *Am J Med Genet* 129B:104–109.
- Dawson KJ, Belkhir K. 2001. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet Res* 78:59–77.
- Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* 55:997–1004.
- Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587.
- Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D. 2004. Assessing the impact of population stratification on genetic association studies. *Nat Genet*. 36:388–393.
- 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.
- Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM. 2004. Design and analysis of admixture mapping studies. *Am J Hum Genet* 74:965–978.
- Lappalainen J, Kranzler HR, Malison R, Price LH, Van Dyck C, Rosenheck RA, Cramer J, Southwick S, Charney D, Krystal J, and Gelernter J. 2002. A functional neuropeptide Y Leu7Pro polymorphism is associated with alcohol dependence in a large population sample from the US. *Arch Gen Psychiatry* 59:825–831.
- 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.
- Luo X, Klempan TA, Lappalainen J, Rosenheck RA, Charney DS, Erdos J, van Kammen DP, Kranzler HR, Kennedy JL, Gelernter J. 2004. NOTCH4 gene haplotype is associated with schizophrenia in African-Americans. *Biol Psychiatry* 55:112–117.
- Pfaff C, Kittles R, Shriver MD. 2002. Adjusting for population structure in admixed populations. *Genet Epidemiol* 22:196–198.
- Pritchard JK, Stephens M, Donnelly P. 2000a. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000b. Association mapping in structured populations. *Am J Hum Genet* 67:170–181.
- Reich DE, Goldstein DB. 2001. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* 20:4–16.
- Ripatti S, Pitkaniemi J, Sillanpaa MJ. 2001. Joint modeling of genetic association and population stratification using latent class models. *Genet Epidemiol* 21(Suppl 1):S409–414.
- Risch NJ. 2000. Searching for genetic determinants in the new millennium. *Nature* 405:847–856.
- Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G. 2002. Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res* 12:602–612.

- Rosenberg NA, Burke T, Elo K, Feldman MW, Freidlin PJ, Groenen MA, Hillel J, Maki-Tanila A, Tixier-Boichard M, Vignal A, Wimmers K, Weigend S. 2001. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* 159:699–713.
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. *Science* 298:2381–2385.
- Rosenberg NA, Li LM, Ward R, Pritchard JK. 2003. Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* 73:1402–1422.
- Satten GA, Flanders WD, Yang Q. 2001. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 68:466–477.
- Sillanpaa MJ, Kilpikari R, Ripatti S, Onkamo P, Uimari P. 2001. Bayesian association mapping for quantitative traits in a mixture of two populations. *Genet Epidemiol.* 21(Suppl 1):S692–699.
- Small KM, Wagoner LE, Levin AM, Kardia SL, Liggett SB. 2002. Synergistic polymorphisms of beta1- and alpha2C-adrenergic receptors and the risk of congestive heart failure. *N Engl J Med* 347:1135–1142.
- 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.
- Thomas DC, Witte JS. 2002. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev* 11:505–512.
- Turakulov R, Eastal S. 2003. Number of SNPS loci needed to detect population structure. *Human Hered* 55:37–45.
- Wacholder S, Rothman N, Caporaso N. 2002. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomark Prev* 11:513–520.
- Watkins WS, Rogers AR, Ostler CT, Wooding S, Bamshad MJ, Brassington AE, Carroll ML, Nguyen SV, Walker JA, Prasad BV, Reddy PG, Das PK, Batzer MA, Jorde LB. 2003. Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. *Genome Res* 13:1607–1618.
- Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG, Bradman N, Goldstein DB. 2001. Population genetic structure of variable drug response. *Nature Genet* 29:265–269.
- Zhang S, Zhao H. 2001. Quantitative similarity-based association test using population samples. *Am J Hum Genet* 69:601–614.
- Zhang S, Zhu X, Zhao H. 2003. On a semi-parametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genet Epidemiol* 24:44–56.