

Nan Yang · Hongzhe Li · Lindsey A. Criswell  
Peter K. Gregersen · Marta E. Alarcon-Riquelme  
Rick Kittles · Russell Shigeta · Gabriel Silva  
Pragna I. Patel · John W. Belmont · Michael F. Seldin

## Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine

Received: 5 March 2005 / Accepted: 1 June 2005 / Published online: 29 September 2005  
© Springer-Verlag 2005

**Abstract** We and others have identified several hundred ancestry informative markers (AIMs) with large allele frequency differences between different major ancestral groups. For this study, a panel of 199 widely distributed AIMs was used to examine a diverse set of 796 DNA samples including self-identified European Americans, West Africans, East Asians, Amerindians, African Americans, Mexicans, Mexican Americans, Puerto Ricans and South Asians. Analysis using a Bayesian clus-

tering algorithm (STRUCTURE) showed grouping of individuals with similar ethnic identity without any identifier other than the AIMs genotyping and showed admixture proportions that clearly distinguished different individuals of mixed ancestry. Additional analyses showed that, for the majority of samples, the predicted ethnic identity corresponded with the self-identified ethnicity at high probability ( $P > 0.99$ ). Overall, the study demonstrates that AIMs can provide a useful adjunct to forensic medicine, pharmacogenomics and disease studies in which major ancestry or ethnic affiliation might be linked to specific outcomes.

**Electronic Supplementary Material** Supplementary material is available for this article at <http://dx.doi.org/10.1007/s00439-005-0012-1>

N. Yang · H. Li · R. Shigeta · M. F. Seldin (✉)  
Rowe Program in Human Genetics, Departments of Biological  
chemistry and Medicine, University of California,  
One Shields Avenue, Davis, CA 95616-8669, USA  
E-mail: mfseldin@ucdavis.edu  
Tel.: + 530-754-6017  
Fax: + 530-754-6015

L. A. Criswell  
Rosalind Russell Medical Research Center for Arthritis,  
University of California, San Francisco, CA USA

P. K. Gregersen  
University of Uppsala, Uppsala, Sweden

M. E. Alarcon-Riquelme  
The Robert S. Boas Center for Genomics and Human Genetics,  
North Shore Lij for Medical Research, Manhasset, NY USA

R. Kittles  
Comprehensive Cancer Center, Ohio State University,  
Columbus, OH USA

G. Silva  
Obras Sociales Del Hermano Pedro, Antigua, Guatemala

P. I. Patel  
Institute for Genetic Medicine, University of Southern California,  
Los Angeles, CA USA

J. W. Belmont  
Department of Molecular and Human Genetics,  
Baylor College of Medicine, Houston, TX USA

### Introduction

The potential relationship of ancestry with biologic differences has engendered considerable controversy (Foster and Sharp 2002; Kaufman and Cooper 2001; Jones 2001; Cooper et al. 2003; Burchard et al. 2003; Royal and Dunston, 2004). On the other hand, there is little controversy regarding the observation that disease outcomes and even prevalence of disease may differ among various ethnic groups. Although the role of socioeconomic status and environmental factors must also be considered when disparate outcomes are observed in different ethnic groups (Stewart et al. 1999; Cooper et al. 2003), the linkage of ancestry to specific Mendelian disorders, e.g. sickle cell anemia, cystic fibrosis, and familial Mediterranean fever, or to specific drug toxicity, e.g. G6PD deficiency, or even to the pathogenesis of acquired immunodeficiency syndrome (AIDS) is well known (Su et al. 2000). The objective of the current study is to examine the potential practical utility of using DNA markers to determine ancestry and ethnic affiliation.

Genetic differences between major ancestral groups have been recognized for many years (reviewed in

Cavalli-Sforza et al. 2004). More recently, markers dispersed throughout the genome have been identified that can distinguish between the ancestral founders of African-Americans and Hispanic populations (Shriver et al. 1997; Smith et al. 2001; Collins-Schramm et al. 2002a; Weber et al. 2002; Rosenberg et al. 2002; Smith et al. 2004). These markers that have been termed Ancestry Informative Markers (AIMs) or Ethnic Difference Markers, includes microsatellites, short diallelic insertion–deletion polymorphisms and single nucleotide polymorphisms (SNPs). A recent study using several hundred microsatellite markers has suggested that the world's population can be grouped into six major ancestral groups (Rosenberg et al. 2002). The current study confirms and expands these observations. The results provide strong confidence in the ability to infer ancestry and ethnic affiliation in diverse populations including several recently admixed populations and a sub-continental South-Asian-derived population. We believe that these results have important implications for examining ancestry linkage to susceptibility or progression of different common diseases with complex inheritance, therapeutic response or toxicity and forensic studies.

---

## Materials and methods

### Populations studied

European American (EUA)(88 subjects), East Asian (EAS)(80 subjects), Mexican American (MAM) (89 subjects), Mexican (MXN)(94 subjects), South Asian (SAS)(88 subjects), Amerindian (AMI) (72 subjects), West African (AFR)(95 subjects), African American (94 subjects) and Puerto Rican (PRN)(96 subjects) populations were included in this study. These populations were based on self-identified ethnic affiliation. The EUA, MAM, and AFA were recruited from California as previously described (Collin-Schramm et al. 2004). The EAS subjects included those self-identified as Korean (4 subjects), Japanese (14 subjects) Filipino (26 subjects) and Chinese (35 subjects), and were recruited from California. The SAS group was composed of individuals who emigrated from India and included diverse state and language subgroups. These included at least 10 subjects with the following languages: Tamil, Gujarati, and Telugu. The majority of these SAS subjects were recruited from Houston, Texas and 8 subjects were recruited from California. The PRN subjects were recruited from New York city. The AMI subjects were self-identified as Mayan (Kachiquel language group) and were recruited from Chimaltenango, Guatemala. The AFR subjects were collected in Nigeria and were from the Edo (Bini) ethnic group. The Mexican subjects were recruited from Mexico City. Blood- or buccal-cell samples were obtained from all individuals, according to protocols and informed-consent procedures approved by institutional review boards, and were labeled with an anonymous code number.

### Ancestry informative markers (AIMs)

One hundred and ninety ninediallelic markers were used in this study. These included 165 SNPs assayed by TaqMan assays and 34 insertion–deletion markers assayed by size fragment length on an automated sequencer. For AIMs, SNPs an initial set of markers was chosen, based on allele frequency differences in the ABI database (USA Caucasian, African American, Chinese and Japanese). The insertion deletion markers were chosen from a primary screen with DNA pools as previously described (Collins-Schramm et al. 2004). Over 650 markers were then screened with an additional 24–48 chromosomes (comparing EUA, AFR, AMI, EAS and SAS). Markers were selected after initial screens for allele frequency differences in representatives of parental populations and in general, a wide chromosomal distribution (markers are distributed on each chromosome and the mean and median intra-chromosomal distances were 13.8 cM/13.1 Mb and 10.0 cM/8.6 Mb, respectively). Markers with evidence of departure from Hardy Weinberg equilibrium (HWE) expectations ( $P < 0.05$ ) during the initial screening, were removed from further studies and the final data set (should specify goodness-of-fit or exact test chi-square or P value criterion). All of the 199 markers included in the current data set were in H-W equilibrium in the putative continental populations and only three markers were not in H-W in one of the putative admixed populations (cv1945712 in the African American population, MID1348 in the MAM population and cv2966801 in the PRN population). Exclusion of these markers had no effect on the analyses, nor did the exclusion of three tightly linked markers. All markers including allele frequencies and chromosomal positions are provided in Web Tables A and B (see Electronic Supplementary Material).

### Statistical analyses

Fst was determined using FSTAT (see electronic database information) that applies the Weir and Cockerham (1984) algorithm and  $\delta$  was calculated by dividing the absolute value of the allele frequency difference between two populations in half. Population structure was examined using STRUCTURE v 2.1 (Pritchard et al. 2000, Falush et al. 2003). Each STRUCTURE analysis was performed without any prior population assignment and was performed at least five times with similar results (see Results), using  $> 5,000$  replicates and burn-in cycles under the linkage option and correlated allele frequency and uncorrelated allele frequency models and  $> 10,000$  replicates and burn-in cycles under the admixture model using the infer  $\alpha$  option (where  $\alpha$  is the Dirichlet parameter for degree of admixture). An additional run using 100,000 burn-in and 200,000 replicates under the admixture model showed similar results. Most runs were performed under the  $\lambda = 1$  option where  $\lambda$  parameterizes the allele frequency prior and is based on the Dirichlet

distribution of allele frequencies. When  $\lambda = 1$  a uniform prior distribution of allele frequencies over all loci is used. Runs using the infer  $\lambda$  option or setting  $\lambda = 1$ , showed similar results for a limited number of selected analyses.

Fishers linear discriminant analysis was performed using the R statistical package (Maindonald and Braun 2001) treating the logarithm of the estimated ancestry probabilities from STRUCTURE as variables.

## Results

### Selection of ancestry informative marker panel

In order to examine the ability to distinguish diverse ethnic groups, a set of diallelic AIMs was selected and screened using panels of 24 DNAs from self identified ethnic groups. These ethnic groups represented four different continental groups, namely European (represented by self identified European Americans from California), East Asian (self identified subjects from California), West African (from Nigeria) and Amerindian (Maya from Guatemala) (see Materials and methods for details). In addition, we also screened for AIMs that would distinguish South Asians (self identified subjects from Texas and California who have themselves immigrated from the Indian Subcontinent) from the continental groups. A final panel of AIMs was chosen based on large allele frequency differences between two or more continental groups, agreement with H-W equilibrium, chromosomal distribution, and inclusion of any marker with large allele frequency differences between European American and South Asian subjects ( $F_{st} > 0.15$ ). A summary of the inter-population differences shows that a total of 567 of the inter-population marker comparisons have very large frequency differences ( $\delta > 0.5$ ) between these self identified ethnic groups. At least 30 AIMs reached this allele frequency difference for each of the two-way comparisons with the exception of those distinguishing between European American and South Asian subjects, for which only three markers with  $\delta > 0.5$  were included in the panel (Table 1). Detailed information on all individual markers is provided in Web Tables A and B.

**Table 1** Summary of Interpopulation Information

	Mean $F_{st}$ / $\delta$ values between Populations <sup>a</sup>					Number of AIMs, $F_{st}$ values $> 0.5$ / $\delta > 0.5$ <sup>b</sup>				
	EAU	AFR	AMI	EAS	SAS	EAU	AFR	AMI	EAS	SAS
EAU	–	0.33	0.31	0.39	0.08	–	75	72	113	3
AFR	0.39	–	0.30	0.38	0.27	75	–	68	99	58
AMI	0.38	0.36	–	0.19	0.19	69	61	–	35	33
EAS	0.44	0.44	0.25	–	0.23	109	93	30	–	45
SAS	0.15	0.34	0.27	0.31	–	3	51	34	42	–

<sup>a</sup>The values above and below the diagonal are the mean  $F_{st}$ s and mean allele frequency differences ( $\delta$ ), respectively

<sup>b</sup>The numbers above and below the diagonal are the number of markers with  $F_{st} > 0.5$  and  $\delta > 0.5$ , respectively

**Table 2** Allele-frequency divergence among populations (Kullback-Leibler distance)

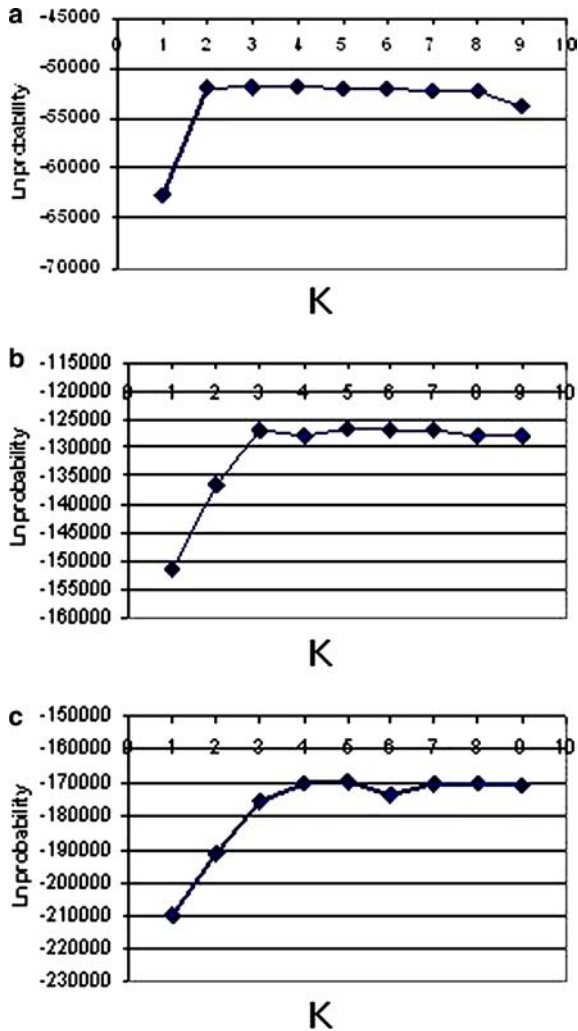
	1	2	3	4	5
1	–	0.77	0.74	1.09	0.1
2	0.8	–	0.72	0.97	0.55
3	0.7	0.65	–	0.43	0.37
4	0.98	0.89	0.35	–	0.51
5	0.13	0.62	0.46	0.68	–

Allele-frequency divergence is shown when each cluster group is compared with another cluster group. The major cluster group in each population is the same as that shown in Fig. 2

### Investigation of population structure

To examine the population structure of diverse populations, we genotyped a total of 796 subjects including the representatives of the four continental populations, South Asians, and four admixed populations (African Americans, Puerto Ricans, Mexican Americans, Mexicans). The program STRUCTURE (Pritchard et al. 2000; Falush et al. 2003) that applies a Bayesian clustering algorithm was used to analyze the genotyping data. This analysis was done without any pre-assignment of individuals to groups. The data was explored to examine the number of clusters that might best describe the population structure for various combinations of populations (Fig. 1). When only European, West African and African American subjects were examined, the Ln likelihood plateau is observed at two clusters ( $k = 2$ ). With the addition of the subjects of Amerindian, Puerto Rican, Mexican American and Mexican ethnicity the plateau shifts to  $k = 3$ . Addition of East Asian subjects (data not shown) or all subjects (plus South Asians) results in a further shift of the plateau to  $k = 4$ . These results suggest the number of major ancestral groups represented in these populations.

The population structure of each self identified group and each individual was examined under different models (see Materials and methods) and for a range of cluster numbers (Fig. 2). For  $k = 4$  and  $k = 5$ , with few exceptions, each individual (bottom section of fig. 2a, b) that purportedly represents a continental group which showed overwhelming contribution from a single distinct cluster (EUA, AFA, EAS and AMI in Fig. 2a, b). The exceptions were: one of 72 Amerindians, and two of



**Fig. 1** Probability estimations for the number ( $k$ ) of cluster groups (“ancestral” or founder populations) present using AIMs. The ordinate shows the Ln probability corresponding to the number of clusters ( $k$ ) (*abscissa*) when: **a** only European American (EUA), West African (AFA) and African American (AFR) are examined; **b** only EUA, AFR, AFA, Amerindian (AMI), Puerto Rican (PRN), Mexican (MXN), and MAM are examined; and **c** all populations are examined

80 East Asians that have a large contribution from the cluster that is predominant in the European American population (cluster 1 in Fig. 2), and one of 88 European Americans and one of 88 South Asians had a minor contribution from the cluster predominant in West Africans. Interestingly, the two East Asian subjects were both of Filipino origin (two of 28 Filipinos included in the self identified East Asian subject group).

The results of the STRUCTURE analyses showed little variation under either the linkage or admixture models for  $k < 7$ . For example at  $k = 5$  the standard deviation for each cluster assignment was  $< 0.02$  (comparing 10 different STRUCTURE runs under the admixture model). When analyses were performed under the admixture model using  $k = 5$  the individual

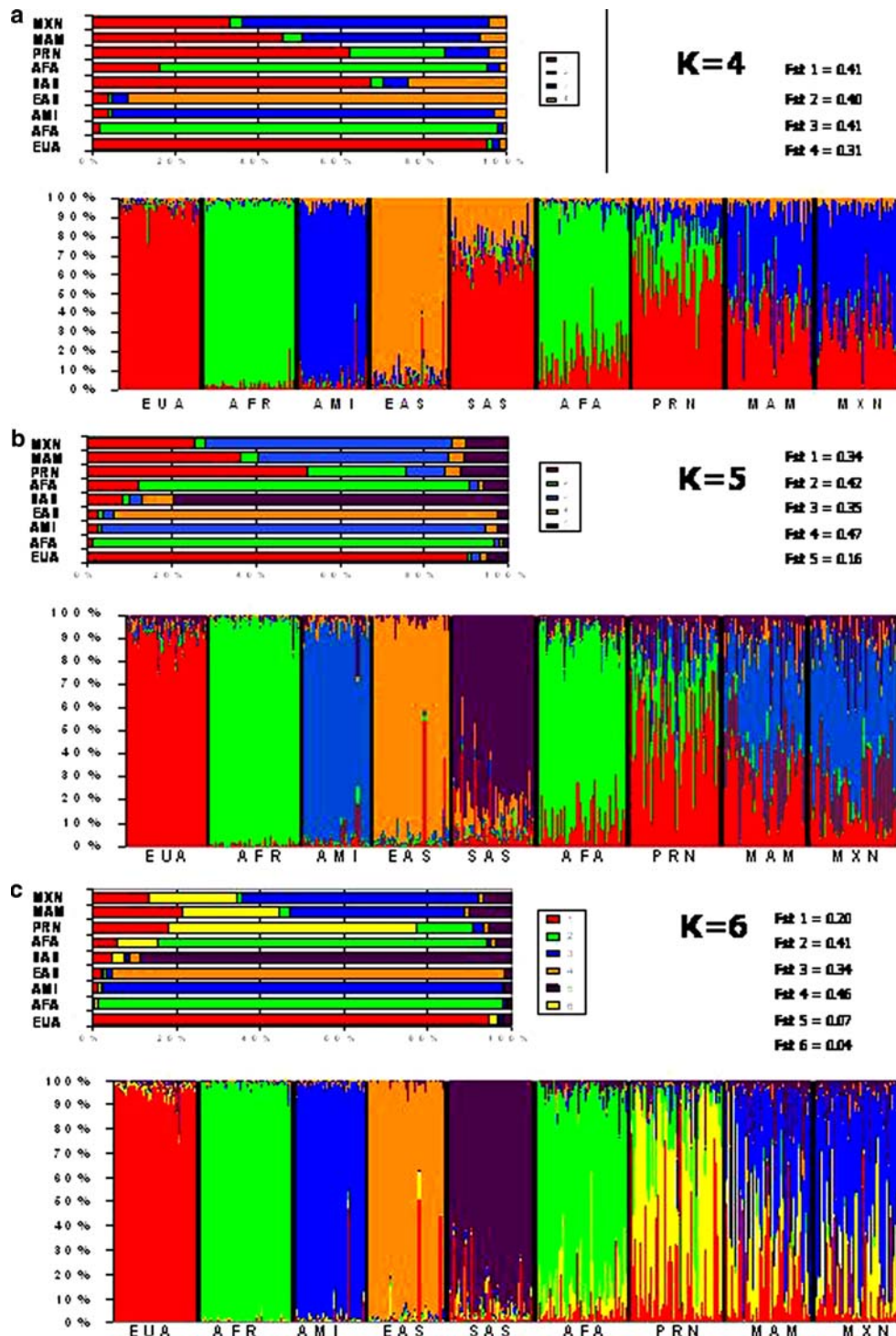
subject average 90% confidence limits for each cluster designation were: European American, 0.07; West African, 0.05; Amerindian, 0.07; East Asian, 0.07; South Asian, 0.15; Mexican American, 0.22; Mexican, 0.19; Puerto Rican, 0.22 and African American, 0.13. Under the linkage model the variation was slightly higher. When  $k > 6$ , we observed more variability in the results especially for the linkage model (see Discussion).

To further assess the robustness of these analyses, we examined several different conditions in which the sample size of different self-identified groups were decreased or removed from the analysis. These studies indicated that random reduction of any population group by 25% had no qualitative effect on the definition of the population structure i.e. the same population proportions were seen in each group (data not shown).

Cluster analysis shows separation of South Asian subjects from continental groups

As indicated above we were not able to identify large numbers of markers with large frequency differences between European American and South Asian subjects (Table 1). At  $k = 4$ , the South Asian subjects appear to be an admixed population with the major contribution from cluster 1 (predominant in European Americans) and a minor contribution from cluster 4 (predominant in East Asians). However, at  $k = 5$  or greater, the STRUCTURE analysis shows the presence of a new cluster (here designated cluster 5) in the South Asian subjects (Fig. 2b, c). This cluster was the predominant group in the South Asian population for each of the South Asian subjects and was present in the other populations only in small percentages. Similarly, the triangle plot descriptions of the different populations (Fig. 3), shows not only the clear separation of each of the continental populations but also the South Asian population from each of the other populations (Fig. 3c). This result was consistently observed in all analyses performed at  $k = 5$  using either the linkage or admixture models applied by the STRUCTURE program and was observed for all individuals that included members with diverse language dialects and states of Indian origin (see Materials and methods). This was observed despite the relatively small allele frequency divergence between cluster 5 (dominant in the South Asian population) and cluster 1 (dominant in the European American population) in contrast to the large frequency divergence present between the other clusters (Table 2). The data showing a distinctive clustering of subjects in the South Asian population, additional analyses at higher cluster numbers (presented below), and previous studies (Rosenberg et al. 2002), strongly suggests that the ability to examine population structure using AIMs and cluster algorithms extends beyond continental groups.

**Fig. 2** Examination of population structure in EUA, AFR, AMI, EAS, SAS, AFA, PRN, MAM and (MXN) populations. Analysis was performed without any prior population assignment. Results for **a**  $k = 4$ , **b**  $k = 5$ , and **c**  $k = 6$  are shown. The top section of each panel shows the average contribution of each color coded cluster as indicated by the proportion of the *horizontal bars*. The bottom section shows the proportion of each cluster (*ordinate*) for each consecutive individual (*abscissa*)

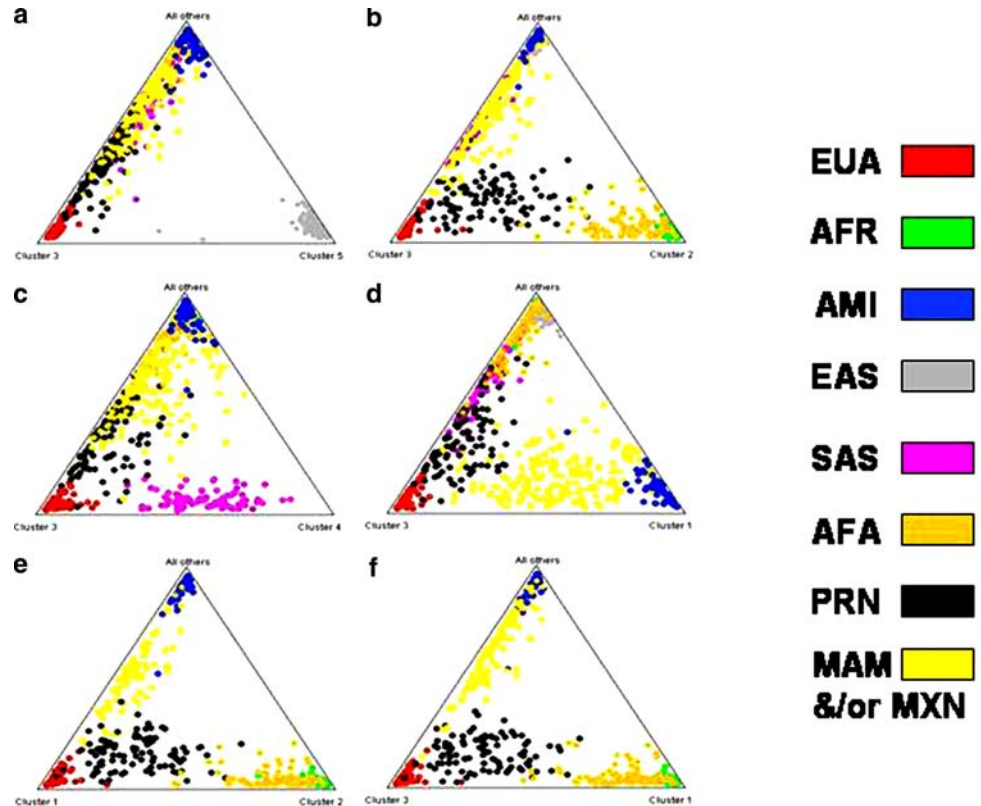


Difference in population structure in different Hispanic populations

Analysis of the presumptive admixed populations provides the potential ability of examining the parental population contributions. In the current study, self-identified African Americans, Puerto Ricans, Mexican Americans and Mexican subjects were examined. The relative contributions of the predominant clusters in the

continental populations (most clearly defined for  $k = 4$ ) show that the African Americans as expected have the largest contribution ( $>75\%$ ) from cluster group 2 (corresponding to West African population), with a contribution (15–20%) from cluster 1 (corresponding to the dominant cluster in European Americans) (Fig. 2a). In contrast, the Mexican American and Mexican populations shows large contributions from both cluster 1 (dominant cluster in European Americans) and cluster 3

**Fig. 3** Triangle plots of cluster groups corresponding to self identified population affiliations. **a, b, c** and **d** show all populations (EUA, AFR, AMI, EAS, SAS, AFA, PRN, MAM and MXN) analyzed with STRUCTURE under the condition of five clusters ( $k=5$ ). The Mexican American (MAM) and Mexican (MXN) populations were combined. **e** and **f** show an analysis of only EUA, AFR, AFA, PRN and MAM (**e**) or MXN (**f**) under the condition of three major populations ( $k=3$ )



(dominant cluster in Amerindians). These data also show that overall the contribution of cluster 3 is greater in the Mexican than in the Mexican American population (Fig. 2a). The Puerto Rican subjects representing a third Hispanic population shows a very different pattern, in that most of these subjects have minimal contributions from cluster 3 but a much more substantial contribution from cluster 2 (dominant cluster in West Africa) in addition to the large contribution from cluster 1 (Fig. 2a). The results in the Puerto Rican subjects is similar to that recently reported by Bonilla et al (2004).

Since these admixed populations are represented almost exclusively by clusters 1, 2, and 3, we have also separately analyzed just these populations (African American, Mexican, Mexican American and Puerto Rican) together with putative representatives of the parental populations, European American, West African and Amerindian. The triangle plots for  $k=3$  (Fig. 3e,f) show the similarity between the Mexican and Mexican American subjects and the minimal overlap between the individual subjects from the Puerto Rican population and those from these two other Hispanic populations.

#### Suggestion of substructure in European-derived populations

When the STRUCTURE analysis is performed under the condition of six clusters ( $k=6$ ), the appearance of

the new cluster (compared with  $k=5$ ) appears to derive almost entirely from a division of the major cluster (designated cluster 1) present in the European American population. This division was most evident in the Hispanic populations (Puerto Rican, Mexican American, and Mexican) in which the proportion of cluster 1 decreased to  $\sim 1/3$  the proportion seen when the analysis was performed at  $k=4$  or  $k=5$  with the difference corresponding to the new cluster (cluster 6). The proportion of the new cluster in the European American population is small with cluster 1 still predominating. The new cluster is also present in the African American population. These results were observed in 10 out of 10 analyses using the admixture model and 9 out of 10 analyses using the linkage model.

#### Analysis of ethnicity grouping

In order to assess the potential use of AIMs in estimating ethnic group membership, Fisher's linear discriminant analysis was used to examine the STRUCTURE output. The probability for membership in each self described group was performed using the leave-one-out cross validation analysis in which one individual was left out and the rest of the individuals with known ethnicity groupings were used to build the linear Fisher discriminant function. The membership probabilities were then calculated for the left-out individual and the individual was assigned to the group with the

**Table 3** Summary of Ethnic Affiliation Estimation from AIMs Analyses

Most Likely Ethnic Group Affiliation Based on Highest Probability <sup>a</sup>								
Self <sup>b</sup>	EUA	AFR	AMI	EAS	SAS	AFA	PRN	MXA
EUA	0.99–1.0	0	0	0	0	0	0–0.01	0
AFR	0	0.98–0.99	0	0	0	0.01–0.02	0	0
AMI	0	0	0.97–0.99	0	0	0	0	0.01–0.03
EAS	0	0	0	0.98–0.99	0	0	0	0.01–0.03
SAS	0	0	0	0	0.99–1.00	0	0–0.011	0
AFA	0	0.21–0.26	0	0	0	0.70–0.77	0.02–0.05	0
PRN	0.07–0.18	0	0	0	0–0.010	0.02–0.04	0.77–0.89	0–0.08
MXA	0.01–0.04	0	0.07–0.15	0	0.01–0.03	0.005	0.02–0.13	0.69–0.86

<sup>a</sup>The frequency range of most likely ethnic group affiliation is based on Fishers linear discriminant analysis of output from 20 independent STRUCTURE runs. At least three STRUCTURE runs were performed at each  $k > 3$ ;  $k < 10$  using the admixture model ( $> 10,000$  replicates, infer alpha option). Similar results were ob-

tained using output from STRUCTURE runs using the linkage model

<sup>b</sup>The results for each self identified ethnic group are shown on each row. All analyses were performed without any information other than that provided by the AIMs genotyping

highest probability. Since the original STRUCTURE analysis of the genotypes was performed without any pre-assignment of ethnicity grouping, this analysis is similar to using a reference panel (in both the STRUCTURE and subsequent linear discriminant analysis) to examine a large number of unknown samples and is based entirely on AIMs genotype results.

The analyses were performed using 20 separate STRUCTURE analytic runs (three each for  $k = 4$ ,  $k = 5$ ,  $k = 6$ , and  $k = 7$ , and four each for  $k = 8$  and  $k = 9$ ). Almost every individual representative of a continental population as well as the South Asian sub-continental group, the best ethnic affiliation (based on the highest probability of assignment to each of the self classified groups) was for the self described ethnic grouping (Table 3, Fig. 4a). In general, the results were consistent for each of the separate analyses of the different STRUCTURE outputs (the range of the most likely assignments by this analysis is shown in Table 3). In the presumptive admixed populations, there was more variability and assignment of some individuals to non-self identified ethnic groups. However, the variability in best ethnic group assignment was mostly where the probabilities for an individual were similar between two related ethnic groups and most of the “miss-assignments” may reflect the “true” ethnic affiliation, e.g., for African Americans, a large number could not be distinguished from West Africans and for Mexicans, multiple individuals could not be distinguished from Amerindians. The observation that fewer Mexican American individuals would be assigned to the Amerindian ethnic group by this analysis of AIMs typing data is also consistent with the likelihood that this type of forensic analysis is truly informative.

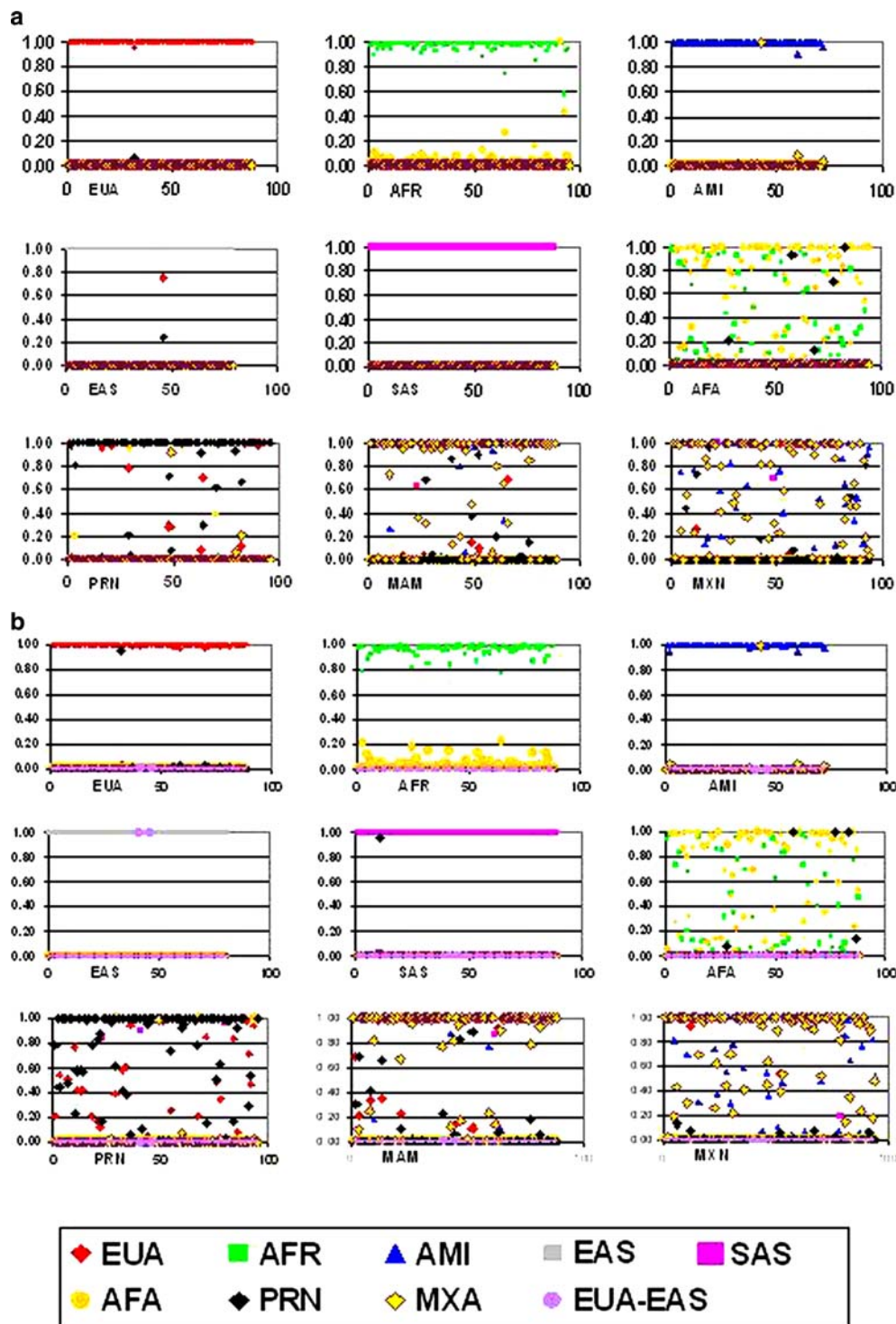
Since two of the self-identified East Asian subjects appeared to have substantial European admixture, an additional group (putative Eurasians) was simulated (see Materials and methods) to examine whether these methods might allow distinguishing additional population groups. The output from STRUCTURE analysis with the entire data set plus 75 simulated Eurasians

(performed at  $k = 5$ ) was examined using Fishers linear discriminant analysis. The results were consistent with the previous analyses with the exception of the high probability assignment of the two East Asian subjects in the new Eurasian group (Fig. 4b). Thus, this type of simulation may be useful in exploring ethnic groupings to provide additional information for the evidence of subgroups of admixed individuals that are not self-identified.

## Discussion

The current study extends the previous findings of Rosenberg et al. (2002) indicating that analysis of population structure using a non-hierarchical clustering algorithm can separate population groups based on DNA polymorphisms. We show here that both continental and sub-continental populations can be readily distinguished and that admixed populations can be examined in the context of the contributions of putative parental populations. These results were robust when  $k < 7$  were examined and were reproducible under many different models. In addition, the findings were not sensitive to exclusion of random groups of individuals, nor inclusion of large numbers of individuals from admixed groups. These findings and implications differ from those suggested by the recent studies of Serre and Paabo in which the microsatellite data set utilized by Rosenberg was reexamined (2004). These investigators suggested that observations of continental grouping in population structure analyses is due to the sampling methods and that there are gradients of human genetic diversity rather than discontinuities between the continents. In contrast, the current study using diallelic AIMs supports the conclusion that the continental population groups are relatively discrete and that such results are not due to limited sampling and exclusion of admixed populations. As discussed subsequently, the robust results observed in the current study may depend on the use of diallelic AIMs.

**Fig. 4** Individual ethnic affiliation probabilities determined from AIMs Genotyping. For each individual (consecutive subjects along the *abscissa*) the probability for ethnic affiliation for each of the eight possible ethnic groups is shown on the ordinate according to the color code key shown in the figure. Both panels show the results were grouped by self identification of the subjects. Probabilities were determined using Fisher's discriminate analysis of STRUCTURE output ( $k=5$ ) without any pre-assignment of the individual being examined. With the exception of the EAAS group [(b) only], each group was based on the self identified ethnicity of the subjects excluding the individual being tested. The EAAS group of 75 Eurasian individuals was simulated using estimated haplotypes from EUA and EAS genotypes (random EUA haplotype matched with random EAS haplotype for each individual)



The ability to clearly distinguish the South Asian subjects was observed despite the relative paucity of individual markers showing large allele frequency differences between this sub-continental population group and those derived from European ancestry. This feature of population groups may not only allow forensic identification but could also be used to test whether there is any association of a biologic property associated

with this grouping, for example, testing whether there is a specific drug toxicity or response in this population group. Although speculative, such patterns of allelic distribution could also underlie epistasis that is thought to be a common confounding variable in the study of complex genetic disease. Thus these results may allow the testing of such hypothesis albeit with additional considerations of non-genetic components.



Another intriguing result from the current analysis was the apparent presence of substructure within the presumed European contribution to Hispanic, and possibly African American populations, which was observed when the genotypes were analyzed under the condition of six major clusters ( $k=6$ ). We also observed this phenomenon when only the Hispanic populations were examined under  $k=4$  (data not shown). Although it is not clear whether this finding reflects real substructure, it is interesting to speculate whether this finding is due to a difference between southern European populations and northern European populations or possibly unique to Spanish populations. The latter might be explained by the Moorish or Arabian influence in this population. Ongoing studies of these populations may provide further insight into this finding. Additional substructure in several of the populations was suggested when analyses at higher cluster numbers were performed ( $k > 6$ ). However, the individual distributions in the clusters were not consistent over multiple analyses. In some analyses, there was further division within the presumed “European” clusters, whereas in others, division appeared in the “African”, “Amerindian” or “South Asian” clusters. With respect to the South Asian subjects there was no correlation with either the language or the state of Indian origin with any of the cluster divisions i.e. the present data did not identify any substructure within the South Asian population.

As indicated in the results the analyses at  $k < 6$  were highly consistent. It is less clear how accurately these STRUCTURE analyses reflect true ancestry. As we have previously shown, the majority of AIMs shows limited variation within West African or Amerindian groups (Collins-Schramm et al. 2002b, 2004). In fact, for many AIMs there is limited diversity within one or more continental groups, similar to Duffy (Livingstone 1984). Although, the AIMs are a limited set, i.e., they do not reflect true population diversity, their use in these analyses may be quite advantageous at least with respect to ascertainment of major ancestral components. Informal inspection of our results using AIMs in European American and Amerindian populations compared with those performed with random SNPs (Hinds et al. 2004) shows that the ancestry divisions are much more discrete using the AIMs. Although multi-allelic markers such as microsatellites may also be used for examination of population structure (Rosenberg et al. 2002), it is likely that these markers will show considerably more noise in the analysis due to the presence or absence of low frequency alleles and stutter polymorphisms. In addition, the higher mutation rate suggested for microsatellites as compared with SNPs may result in less correlation of such markers with ancestry. These differences may, in part, account for the robust results in the current study compared with those reported by Serre and Paabo (2004) using microsatellites. In addition, the number and geographic origin of the samples may account for differences between such studies.

The precise limitations of the current study and the optimization of AIMs analyses of diverse populations are uncertain. It is not clear whether certain clusters, present in small percentages in different population groups are accurate measurements or artifacts of the analytic method. However, reduction of the number of AIMs from 199 to 99 increased the 90% confidence limits by 36% (based on 10 random removals of 100 AIMs for  $k=5$  under the admixture model). It is also not apparent how to pre-optimize a set of AIMs given our finding that despite limited number of markers with high allele frequency differences distinguishing between the European American and South Asian populations, the analyses of population structure at  $k=5$  clearly identified differences. At present, we suggest initial selection of AIMs based on analysis of  $F_{st}$  or other measurements of informativeness, e.g.,  $I_N$  (Rosenberg et al. 2003), followed by empiric testing using clustering algorithms such as STRUCTURE.

Although our major interest in pursuing the current studies is directed towards the application of this approach to clinical epidemiology, the implications for forensic analyses are self evident. The current study demonstrates that a relatively modest set of selected diallelic markers can provide close correlation with self reported ethnicity for individuals from disparate groups. The current study utilized ethnic group sample sizes ranging from 72–96. When individual ethnic group sample sizes are reduced to less than 25 subjects in each group there was considerably more variation in the population structure analysis and confidence in assignment of ethnic group affiliation. Although the limitations will vary depending on the characteristics of each group, we would generally suggest using sample sizes of at least 50 subjects in each major ancestry group as a reference subset when examining population structure. The development of a larger panel of “reference subjects”, the removal of extreme outliers, as well as perhaps a larger and better set of AIMs may improve the current results.

This initial forensic type of analysis could also be utilized to remove particular outliers from specific case or control groups in the analysis of a therapeutic agent or to potentially control for genetic heterogeneity in genetic association tests. Finally, in regards to forensic identification of ethnic grouping, exploration of different analytic strategies including the direct use of genotyping data without an intermediate analysis of population structure (Baudouin et al. 2004) may also be useful.

While most of the genetic variation in humans may be independent of both the variation and patterns of variation that distinguish population groups (Lewontin 1972; Nei and Roychoudhry 1974; Latter 1980; and Barbujani et al. 1997), and as shown here, ethnic groups, the current results may have potentially profound implications for clinical epidemiology studies and forensic medicine. While examining the specific human sequence variation that is linked to clinical phenotypes will eventually provide the best information for the practice of molecular medicine (Royal and Dunston

2004); this is not yet possible. The ability to examine particular subject groupings based on ancestry information or even quantitative assessment of an individual's "ancestry" is, we believe, a practical reality worth testing. As presented in the introduction, and by others (Mountain and Risch 2004), the association of particular medical conditions with particular ethnic groups is clear. Although the importance of ethnic grouping with respect to most complex genetic diseases or therapeutics is far from clear, the differences in the effects of major histocompatibility region determinants and PTPN22 polymorphisms in rheumatoid arthritis (Begovich et al. 2004), CARD15 in Crohns disease (discussed in Burchard et al. 2003), or CCR5 polymorphisms in AIDs (Su et al. 2000) suggest that the phenomenon of ancestral association may be quite important. Thus, regardless of whether these ethnic groupings are largely due to social constraints, the potential biologic relevance of these groupings may be worth investigation in the context of many clinical studies. A recent clinical trial limited to African Americans underscores the relevance, potential value and controversy for using self identified ethnic information (Taylor et al. 2004; Bloche 2004). The extension of such studies to incorporate a closer correlate of ancestry, i.e. using AIMs as suggested by the current study, may answer the question of whether there are pathophysiologic differences linked to ancestry that can help direct to the appropriate therapeutic strategy. However, the same ethnic groupings defined solely by DNA markers may also, in many cases, correspond to particular groups of people who have differences in access to medical care, social conditions and even dietary differences, and it will be critical to include such covariates in these regression analyses.

There are multiple considerations for the potential practical application of DNA-defined ancestry and ethnic groups to clinical epidemiology. While few will advance arguments against controlling for population structure in genetic association tests, it may be uncomfortable to apply AIMs directly to clinical epidemiology studies for fear of racial stigmatization. However, analysis of Mendelian diseases suggests that this potential stigmatization is likely to be distributed among different ancestries and ethnic groups. Importantly, the ancestral linkage of particular gene variants linked to disease can also provide a potential method for mapping and positional cloning of the genes for some complex diseases (Hoggart et al. 2004; Patterson et al. 2004 and Seldin et al. 2004). Thus, clinical epidemiologic studies that suggest ancestral linkage may lead to specific efforts to identify individual-ancestry-linked gene variants. In summary, the present study suggests that the methods and tools are available for many of the additional studies needed to address more globally the role of ancestry-associated variations in human disease.

**Acknowledgments** Support for this research was provided by National Institute of Health grants U01-DK57249, AR44804, AR50267 and AR20684.

## References

- Barbujani G, Magagni A, Minch E, Cavalli-Sforza L (1997) An apportionment of human DNA diversity. *Proc Natl Acad Sci USA* 94:4516–4519
- Baudouin L, Piry S, Cornuet JM (2004) Analytical Bayesian approach for assigning individuals to populations. *J Hered* 95:217–224
- Begovich AB, Carlton VEH, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, Ardlie KG, Huang Q, Smith AM, Spoeerke JM, Conn MT, Chang M, Chang YP, Saiki RK, Catanese JJ, Leong DU, Garcia VE, McAllister LB, Jeffery DA, Lee AT, Batliwalla F, Remmers E, Criswell LA, Seldin MF, Kastner DL, Amos CI, Sninsky JJ, Gregersen PK (2004) A missense SNP in the protein tyrosine phosphatase PTPN22 is associated with rheumatoid arthritis. *Am J Hum Genet* 75:330–337
- Bonilla C, Shriver MD, Parra EJ, Jones A, Fernández 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
- Burchard E, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, Mountain JL, Perez-Stable EJ, Sheppard D, Risch N (2003) The importance of race and ethnic background in biomedical research and clinical practice. *N Engl J Med* 348:1170–1175
- Cavalli-Sforza L, Menozzi P, Piazza A. (1994) The history and geography of human genes. Princeton University Press
- Collins-Schramm HE, Phillips CM, Operario DJ, Lee JS, Weber JL, Hanson RL, Knowler WC, Cooper R, Li H, Seldin MF (2002a) Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. *Am J Hum Genet* 70:737–750
- Collins-Schramm HE, Kittles RA, Operario DJ, Weber JL, Criswell LA, Cooper RS, Seldin MF (2002b) Markers that discriminate between European and African ancestry show limited variation within Africa. *Hum Genet* 111:566–569
- Collins-Schramm HE, Chima B, Morii T, Wah K, Figueroa Y, Criswell LA, Hanson RL, Knowler WC, Silva G, Belmont JW, Seldin MF (2004) Mexican American Ancestry-Informative Markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. *Human Genetics* 114:263–271
- Cooper R, Kaufman J, Ward R (2003) Race and genomics. *N Engl J Med* 348:1166–1170
- Falush D, Stephens M, Pritchard J (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587
- Foster M, Sharp R (2002) Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. *Genome Res* 12:844–850
- Fstat (2002) A program to estimate and test gene diversities and fixation indices. Version 2.9.3, Lausanne, Switzerland
- Hinds DA, Stokowski RP, Patil N, Konvicka K, Kershenovich D, Cox DR, Ballinger DG (2004) Matching strategies for genetic association studies in structured populations. *Am J Hum Genet* 74:317–325
- 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
- Jones C (2001) Invited commentary: "Race," racism, and the practice of epidemiology. *Am J Epidemiol* 154:299–304
- Kaufman J, Cooper B (2001) Commentary: considerations for use of racial/ethnic classification in etiologic research. *Am J Epidemiol* 154:291–298
- Latter BDH (1980) Genetic differences within and between populations of the major human subgroups. *Am Nat* 116:220–237
- Lewontin R (1972) The apportionment of human diversity. *Evol Biol* 6:381–398
- Livingstone FB (1984) The Duffy blood groups, vivax malaria, and malaria selection in human populations: a review. *Hum Biol* 56(3):413–425

- Maindonald J, Braun J (2003) *Data analysis and graphics using R*. Cambridge University Press, Cambridge
- Mountain JL, Risch N (2004) Assessing genetic contributions to phenotypic differences among 'racial' and 'ethnic' groups. *Nat Genet* 36:S48–S53
- Nei M, Roychoudhury A (1974) Genetic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. *Am J Hum Genet* 26:421–443
- 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 J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- 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
- Royal CDM, Dunston GM (2004) Changing the paradigm from 'race' to human genome variation. *Nat Genet* 36:S5–S7
- Seldin MF, Morii T, Collins-Schramm HE, Chima B, Kittles R, Criswell LA, Li H (2004) Putative ancestral origins of chromosomal segments in individual African Americans: implications for admixture mapping. *Genome Res* 14:1076–1084
- Serre D, Paabo S (2004) Evidence for gradients of human genetic diversity within and among continents. *Genome Res* 14:1679–1685
- Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. *Am J Hum Genet* 60:957–964
- Smith MW, Lautenberger JA, Doo Shin H, Chretien J, 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
- Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kesing 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, Sninky 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* 74:1001–1013
- Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68:978–989
- Stewart A, Napoles-Springer A, Perez-Stable E, et al (1999) Interpersonal processes of care in diverse populations. *Milbank Qu* 77:305–339
- Su B, Sun G, Lu D, Xiao J, Hu F, Chakraborty R, Deka R, Jin L (2000) Distribution of three HIV-1 resistance-conferring polymorphisms (SDF1-3'A, CCR2-641, and CCR5-delta32) in global populations. *Eur J Hum Genet* 8:975–979
- Weber J, David D, Heil J, Fan Y, Zhao C, Marth G. (2002) Human diallelic insertion/deletion polymorphisms. *Am J Hum Genet* 71:854–862
- Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370