Published in final edited form as:
Hum Mutat. 2009 January ; 30(1): 69–78. doi:10.1002/humu.20822.

Ancestry Informative Marker Sets for Determining Continental Origin and Admixture Proportions in Common Populations in America

Roman Kosoy[1], **Rami Nassir**[1], **Chao Tian**[1], **Phoebe A White**[2], **Lesley M. Butler**[3], **Gabriel Silva**[4], **Rick Kittles**[5], **Marta E. Alarcon-Riquelme**[6], **Peter K. Gregersen**[7], **John W. Belmont**[8], **Francisco M. De La Vega**[2], and **Michael F. Seldin**[1,*]

[1]Rowe Program in Human Genetics, Departments of Biochemistry and Medicine, University of California Davis, Davis, CA 95616, USA

[2]Applied Biosystems, Foster City, CA 94404, USA

[3]Department of Public Health Sciences, University of California Davis, Davis, CA 95616, USA

[4]Obras Sociales del Hermano Pedro, Antigua, Guatemala

[5]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, Illinois 60637, USA

[6]Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden

[7]The Robert S. Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, North Shore LIJ Health System, Manhasset, NY 11030, USA

[8]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

Abstract

To provide a resource for assessing continental ancestry in a wide variety of genetic studies we identified, validated and characterized a set of 128 ancestry informative markers (AIMs). The markers were chosen for informativeness, genome-wide distribution, and genotype reproducibility on two platforms (TaqMan® assays and Illumina arrays). We analyzed genotyping data from 825 subjects with diverse ancestry, including European, East Asian, Amerindian, African, South Asian, Mexican, and Puerto Rican. A comprehensive set of 128 AIMs and subsets as small as 24 AIMs are shown to be useful tools for ascertaining the origin of subjects from particular continents, and to correct for population stratification in admixed population sample sets. Our findings provide general guidelines for the application of specific AIM subsets as a resource for wide application. We conclude that investigators can use TaqMan assays for the selected AIMs as a simple and cost efficient tool to control for differences in continental ancestry when conducting association studies in ethnically diverse populations.

[*]Corresponding author, Michael F. Seldin, mfseldin@ucdavis.edu, Rowe Program in Genetics, UC Davis, Tupper Hall Room 4453, Davis, California 95616, United States, Phone: 530-754-6016, Fax: 530-754-6015.

P.W. and F.D.L.V. declare competing financial interests.

Keywords

population structure; continental ancestry; population stratification; ancestry informative markers

Introduction

Analyses of population genetic structure have shown that continental population groups can be identified by examining differences in allele frequencies (Rosenberg, et al., 2005; Rosenberg, et al., 2002). Over the last several years studies have demonstrated that thousands of individual single nucleotide polymorphisms (SNPs) distributed through out the genome have very large differences in allele frequencies between two or more continental populations (Mao, et al., 2007; Price, et al., 2007; Smith, et al., 2004; Tian, et al., 2007; Tian, et al., 2006). These studies have set the framework for both admixture mapping and adjusting for population genetic structure in association testing. The latter is particularly important since differences in population genetic structure between cases and controls can confound SNP-disease associations leading to false positive or negative findings (Campbell, et al., 2005; Clayton, et al., 2005; Freedman, et al., 2004; Helgason, et al., 2005; Marchini, et al., 2004). Methods to measure, and therefore address differences in population structure in association testing have been developed (Epstein, et al., 2007; Hoggart, et al., 2003; Price, et al., 2006; Pritchard, et al., 2000b; Purcell, et al., 2007; Satten, et al., 2001). In the context of whole genome association (WGA) scans, these methods can be readily applied. However, for follow-up association studies to further define critical candidate regions in larger population sets, or for analyses of additional populations, a small set of ancestry informative markers (AIMs) is highly desirable.

While differences within continental populations, and population substructure, must also be considered (Bauchet, et al., 2007; Price, et al., 2008; Seldin, et al., 2006; Tian, et al., 2008), the larger difference in allele frequencies between continental populations potentially creates the greatest confounding problem in interpreting such association studies. At this point a large number of WGA studies have been conducted in populations of primarily or exclusively European ancestry. Thus, the issue of confounding by population stratification will become particularly evident as more genetic associations are conduced among multiethnic, and therefore substantially admixed populations, in order to evaluate ethnic disparities in disease risk. Addressing these differences in population structure is particularly relevant for extending genetic associations to underserved minority groups that include substantial admixture between continents.

The current study was undertaken to provide a resource for determining and quantifying differences in continental populations using the smallest numbers of SNPs possible as a cost and time efficient strategy. Previous studies by both our group and others, have shown that AIM sets of 200 markers or less have ability to discern continental structure (Parra, et al., 2004; Salari, et al., 2005; Yang, et al., 2005). However, the use of such markers has been sporadic, the validation of many of the markers incomplete, and in some cases have been limited to specific platforms that cannot be readily and inexpensively used by multiple laboratories. The current study utilizing the widely used TaqMan® platform provides a set of AIMs that distinguish continental groups that can be widely applied to genetic studies. In addition, the application of AIMs depends in part on availability of genotypes. Our study also provides genotypes of continental populations as a research community resource. Most importantly, the current study shows both the value and limitations of using smaller subsets of AIMs by providing guidance in practical application.

Methods

Population samples

DNA samples or genotypes used for population structure analyses were from 825 individuals that included: 128 European Americans (NYCPEA), 60 CEPH Europeans (CEU), 56 Yoruban African (YRI), 19 Bini West African, 23 Kanuri West African, 50 Mayan Amerindians, 26 Quechuan Amerindians, 29 Nahua Amerindians, 40 Mexican Americans (MAM), 26 Mexican (MXN), 28 Puerto Rican American (PRA), 43 Chinese (CHB), 43 Chinese American (CHAH), 43 Japanese (JPT), 8 Vietnamese American (VAH), 1 Korean American (KAH), 45 Filipino American (NYCPFA), 2 unspecified East Asian Americans (OEAS), 3 Japanese American (JAH), and 64 South Asian Indian Americans (SAS).

These populations were based on self-identified ethnic affiliation. The NYCPEA, NYCPFA and PRA were from New York City and were collected as part of the New York Cancer Project (Mitchell, et al., 2004). The Mayan samples were collected from two villages, Bola De Oro and Cienega Grande, from Chimaltenango Guatemala (provided by G.S. and J.B.), the Quechuan individuals were from Peru (provided by J.B.); the Nahua were from central Mexico (provided by M.EAR); the MXN were from Mexico City (provided by ME.AR.), the MAM and AFA were from California, and the CHAH, VAH, KAH, and SAS were from Houston (provided by J.B.). For the West African samples the Bini, are a Niger-Congo group of Bantu speakers from Edo State and the Kanuri, a group of Nilo Saharan speakers fro the Lake Chad region of northern Nigeria (provided by R.K.). The CEU and YRI were HapMap panel genotypes (Altshuler, et al., 2005) and the JPT and CHB were from the I-ControlDB (www.illumina.com/iControlDB, Illumina, San Diego, CA).

Additional genotypes used in modeling studies derived included 1) EURNIHLN genotypes (254 subjects) that were available from the NIH Laboratory of Neurogenetics at the Coriell Queue website, 2) East Asian genotypes from the iControlDB (198 subjects), 3) East Asian samples (85) genotyped at North Shore and 4) African American genotypes (1847 subjects) from the iControlDB. For the modeling studies we limited the genotypes to autosomal SNPs that were typed in >95% of each of the included subjects and that were in HWE ($p>0.001$) within a given self-identified group and in combined samples from a given continent.

The subjects studied were all healthy and not first-degree relatives of each other based on self-reporting. All DNA and blood samples were obtained according to protocols and informed-consent procedures approved by institutional review boards, and were labeled with an anonymous code number.

Statistical Methods

Fst was determined using Genetix software (Belkhir, et al., 2001) that applies the Weir and Cockerham algorithm (Weir and Cockerham, 1984) This algorithm defines Fst as (MSP-MSG)/[MSP +(n_c-1)MSG] where MSP denotes the observed mean square errors for loci between populations and MSG denotes the mean square errors for loci within populations. The pairwise Fst values thus provide a measurement of inter-population genetic variance in comparison to intra-population genetic variance. Hardy-Weinberg (HW) equilibrium was examined using an exact test implemented in the FINETTI software that can be accessed interactively at the internet address provided in the Web Resources section. Population admixture proportions were determined using the Bayesian clustering algorithms developed by Pritchard and implemented in the program m STRUCTURE v2.1 (Falush, et al., 2003; Pritchard, et al., 2000a). Informativeness between multiple population groups was determined using the I_n algorithm (Rosenberg, et al., 2003).

For STRUCTURE, unless otherwise noted in the results, each analysis was performed without any prior population assignment and was performed at least 3 times with similar results using > 10,000 replicates and 5000 burn-in cycles under the admixture model. For analyses using smaller marker sets (24 and 48 markers) longer runs were necessary to achieve similar results on multiple run comparisons. For 24 and 48 marker sets, 50,000 replicates and 10,000 burn-in cycles were used with the exception of 24 markers selected using I_n4 (four population informativeness). For these analyses, 100,000 replicates and 20,000 burn-in cycles were necessary. For all analyses reported we used the "infer α" option with a separate α estimated for each population (where α is the Dirichlet parameter for degree of admixture). Runs were performed under the λ = 1 option where λ parameterizes the allele frequency prior and is based on the Dirichlet distribution of allele frequencies.

Fst, I_n and allele Frequencies were determined using sets of 80 subjects representing European (EURA), West African (AFR), Amerindian (AMI) and East Asian (EAS) ancestry. These included the following distribution of subjects: 1) EURA, CEPH (17 subjects), NYCPEA (63 subjects); 2) AFR, YRI (45 subjects), Bini (17 subjects), and Kanuri (18 subjects); 3) AMI, Mayan (38 subjects), Nahua (23 subjects), and Quechuan (19 subjects); and 4) EAS, HCB (15 subjects), Filipino (16 subjects), 25 diverse ethnic Chinese American (25 subjects), JPT (15 subjects), Japanese American (1 subject), Korean American (1 subject), and Vietnamese Americans (7 subjects).

For modeling studies, association tests were performed using the EIGENSTRAT statistical package (Price, et al., 2006). False discovery rate statistics(Devlin and Roeder, 1999) were determined using HelixTree 5.0.2 software (Golden Helix, Bozeman, MT, USA).

Genotyping

TaqMan® SNP genotyping assays were developed for each of the SNPs used in the current study (Supplementary Table S1) and are commercially available (Applied Biosystems, Foster City, CA; cf. www.allsnps.com). Assays were performed with the TaqMan Genotyping Master Mix, using conditions recommended by the manufacturer, on an ABI 7900 Sequence Detection System (Applied Biosystems, Foster City, CA).

Genetic Map

For the current studies the deCODE (Kong, et al., 2002) genetic map was used. The position of each SNP was determined by interpolation using markers that were both on the genetic map and for which an unambiguous physical map position was available in NCBI build 35. Any markers that were not in the same relative order in both the genetic and physical maps were omitted as anchors for the interpolation of the genetic positions of the SNPs.

Ancestry Informative Markers

The SNPs chosen for inclusion were based on two large sets of previous genotyping results in our laboratory (Tian, et al., 2007; Tian, et al., 2006) were limited to those SNPs that overlapped with the 300K genome-wide Illumina SNP array. 250 SNPs were chosen selecting the best SNP in each 10 cM deCODE bin that met the criteria of a large allele frequency differences (>45%) between EURA and AMI groups and small allele frequency differences (<5%) between two disparate AMI groups (Pima and Mayan). Similarly, 250 SNPs with large frequency differences (>45%) between African and European groups were selected. From these 500 SNPs we reduced the number for testing to 184 based on the following criteria: 1) in silico design criteria for TaqMan assays; 2) genome-wide distribution pattern (minimum inter-marker distance = 8 cM on deCODE map); and 3) EAS differences based on HapMap results in JPT and CHB. TaqMan® SNP genotyping assays were designed for the 184 SNPs and tested using DNA panels. Of these, 128 SNPs passed

our quality filters demonstrating reproducible genotyping results in population samples of diverse origin, >90% complete typing results in each population and were in HW equilibrium ($p>0.01$) in the EURA group. A small number of SNPs were not in HW equilibrium in specific populations (2 SNPs in AFR, 3 SNPs AMI, and 3 SNPs EAS). These SNPs did not overlap between these groups and only 2 SNPs showed HW <0.005). Thus, these SNPs were not excluded, because recent admixture in these self-identified ethnic groups could result in departure from HW. Summary information for the final set of 128 SNPs is provided in Supplementary Table S1.

Identifying Subsets of Ancestry Informative Markers

Subsets of the 128 marker set were chosen using the In algorithm (Rosenberg, et al., 2003) with the goal of finding the most informative markers distinguishing one or more of the following: 1) four continental populations EURA, AFR, AMI, and EAS; 2) three continental populations (EURA, AFR, and AMI); or 3) two continental populations (EURA and AFR or EURA and AMI). Each subset was determined using 80 subjects from each ethnic group (described in Statistical Methods) and marker selection was based on the most informative set for each analysis (provided in Supplementary Table S2).

Modeling Studies

To test whether a limited number of AIMs can correct for false positive results observed in case-control studies due to population stratification we modeled three population specific loci as disease phenotypes. The modeling was done in the following step-wise manner independently for each surrogate phenotype: 1) surrogate cases and controls (with available SNP genotypes on Illumina 300 K platform) were chosen on the basis of genotypes for a population specific marker; 2) 200 K SNPs that passed quality control filters in the surrogate case-control sample sets were tested for association using the HelixTree software package; 3) significantly associated markers (by Armitage χ^2 test, $\chi^2 \geq 26.6$. $p \leq 0.05$ with Bonferroni correction for 200,000 tests) in or near the locus designating the surrogate phenotype are defined as true positive signal, while significantly associated SNPs outside the locus are defined as false positives; 4) six to ten SNPs with the strongest false positive associations and a similar number of true positive associations with χ^2 values comparable to the false positives were selected for further analysis; 5) the genotypes for the chosen true and false positively associated markers are combined with genotypes for the markers in the selected sets (all 200K SNP markers, 128 I_n4, 96 I_n4, 64 I_n4, 48 I_n4, and 24 I_n4), and were tested for association testing correcting for substructure by principal component analysis using EIGENSTRAT (Price, et al., 2006); 6) the positively associated markers were re-analyzed for association using correction for population stratification with an appropriate number of principal components (PC 1 or PC2 depending on the studies population, determined by the plateau of χ^2 values).

The surrogate phenotypes were assigned based on SNPs selected from haplotype analyses of three regions that contained genes with strong ancestry association. The models chosen were for the *SLC24A5*, lactase gene (*LCT*) and *ADH1B*. *SLC24A5*, coding for a K gated Na/Ca exchanger, is located on chromosome 15, and plays a role in human skin pigmentation (Lamason, et al., 2005). This study provided evidence that a non-synonymous genetic substitution (rs1426654, A/G 111) is under strong positive selection in Europeans, with allele A nearly fixed in various European populations (98.7 to 100%), whereas allele G is present at 97 to 100% frequency in African and East Asian HapMap populations (Lamason, et al., 2005). Since genotypes for rs1426654 was not available in our dataset, individuals homozygous for allele A of rs2675348, in complete linkage disequilibrium (LD) with allele A of rs1426654 (r^2 =1.00 in HapMap CEU samples), were designated as surrogate cases,

while individuals with A/G and G/G genotypes were designated as surrogate controls (Allele A is 1.0 in CEU, 0.5 in CHB, 0.589 in JPT, and 0.25 in YRI).

The second locus chosen for modeling a population specific phenotype is *LCT* located on chromosome 2. A variant within *LCT* gene, rs4988235 (C/T -13910), is associated with lactase persistence, leading to ability to digest milk in adults, and has been demonstrated to be under strong positive selection in Europeans (Bersaglieri, et al., 2004; Hamblin and Di Rienzo, 2000; Tishkoff, et al., 2007). Allele A is found at 0.75 frequency in HapMap European samples, but is absent in HapMap YRI, CHB and JPT samples. Since rs4988235 genotypes were not available for our sample set, an allele A for rs1446585, a nearby SNP in strong LD with allele T of rs4988235 (r^2=0.73 in HapMap CEU samples) was used for modeling. Individuals homozygous for allele A for rs1446585 were designated as surrogate cases, while individuals with A/G and G/G genotypes were designated as surrogate controls (Allele A is 0.792 in CEU, and 0.00 in CHB, JPT, and YRI).

The third locus is for the alcohol dehydrogenase *ADH1B* gene, where a nonsynonymous coding genetic variant rs1229984 (Arg47His) is reported to be under positive selection in East Asia (Han, et al., 2007). Allele A is found at 0.77 frequency in HapMap CHB and JPT, but is absent in CEU and YRI samples. Since genotypes for rs1229984 were not available for our sample set, allele A for rs10008281, a nearby SNP in strong LD with allele T for rs1229984 (r^2=0.53 in HapMap CEU samples) was used for modeling. [Note: since the trait is modeled on the proxy SNP, the performance of AIM sets should be unaffected by the r^2.] Individuals homozygous for allele A for rs10008281 were designated as surrogate cases, while individuals with A/G and G/G genotypes were designated as surrogate controls (allele A is 0.82 in CHB and 0.83 in JPT, and 0.28 in CEU and YRI).

Results

Small Ancestry Informative Marker Sets Distinguish Major Population Groups

A set of 128 SNPs selected on the basis of informativeness (I_n) between four continental groups (European, Amerindian, West African and East Asian) passed our initial quality filters (see **Methods**). Analysis of genotypes using this informative marker set (designated 128 I_n4) was first evaluated using Fst as a general measure of the ability to separate continental population groups. The markers showed large Fst differences between the continental populations and relatively small differences within large groups of disparate individuals within these continental groups (Table 1). The South-Asian group, not used in the marker selection, showed substantial differences with the European group consistent with previous observations that this sub-continental group is distinct (Yang, et al., 2005). In addition, there was a larger intercontinental difference among the Amerindian groups as previously observed (Price, et al., 2007;Tian, et al., 2007).

Population structure analyses using a Bayesian cluster analysis (STRUCTURE) showed a clear distinction between the continental population groups when the number of clusters was defined at 4 (K=4). The 128 I_n4 set consistently identified diverse individuals corresponding to European, West African, Amerindian, and East Asian population groups (Fig. 1a, Table 2). Adding an additional cluster (K=5), also allowed the identification of individuals from another genetically distinguishable population, that corresponding to a South Asian sub-continental group (Fig. 1b).

The ability of smaller sets of I_n4 markers (96, 64, 48 and 24) to discern population genetic structure was also examined. Here, the smaller sets were in each case the highest ranking I_n4 SNPs (Supplementary Table S1 and see Supplementary Table S2 **for additional summary information**). The individual estimation of continental ancestry was nearly identical when

128, 96 or 64 I_n4 markers were used (e.g. compare Fig. 1c with 1a). A summary of all the results shows that as few as 24 I_n4, could in fact identify the same general population clusters (Table 2). Specifically, for both West African and European ancestry the results are very consistent with similar proportion of population measurements seen even when comparing 128 I_n4 with 24 I_n4 results. For the Amerindian and East Asian continental population groups there is a modest fall-off in the concordance with self-identification as the numbers of markers decrease, for example, the cluster membership that corresponds best to self identified Amerindian ancestry (pop 4) decreased from 0.94 (128 I_n4) to 0.88 (24 I_n4) (Table 2). However, the difference is more pronounced for the estimated contribution from pop5 (corresponding to South Asian background) in the South Asian population (0.75/0.68/0.70/0.59/0.55). The increased uncertainty for South Asian contribution may be explained by the relatively low Fst values between South Asian and European/East Asian populations observed for the I_n4 markers (Table 1) that in turn reflects the selection criteria (see **Methods**).

The population structure analyses of different population groups are also influenced by which subjects are included. When the subject set is limited to only those individuals of particular self-identified backgrounds the results show more distinct cluster assignments. This is illustrated in Fig. 1d when East Asian and South Asian subjects are excluded from the analyses and the number of assumed population groups is defined as three (K=3). In addition, small numbers of markers chosen using other criteria may provide good distinction between two or three population groups but provide inaccurate information on other non-included population groups. The performance of subsets of markers selected using either European/West African informativeness or European/Amerindian informativeness is provided in Supplementary Table S3.

Ability to Exclude Subjects of Disparate Ancestry for Specific Studies

One practical aspect of utilizing continental AIMs is to identify sets of individuals corresponding to a particular continental group. The ability of I_n4 sets to exclude subjects from the different self identified groups is summarized in Table 3 using the predominant population group cluster membership as the standard for each continent. Two criteria, 10% non-membership and 15% non-membership are shown. In general, the 128 I_n4 AIMs and smaller sets showed nearly complete exclusion of individuals with other self-identified ancestries when considering any of the continental groups. However, for European, there was a large decrease in the performance of smaller maker sets (<64 markers) with respect to exclusion of South Asian subjects.

For both Amerindians and East Asians the exclusion criteria used in these analyses also would result in excluding a relatively large number of subjects for these specific ancestries. For example, 10% non Amerindian exclusion would result in excluding 17% of the Amerindian subjects using 128 I_n4. While this result is probably partially due to European admixture, there also is some difficulty in fully resolving AMI and EAS ancestry at this level. This issue is less severe when the criteria is set at 15% non-membership but is much more problematic when smaller I_n4 marker sets are used (Table 3). Nevertheless, investigators can use these criteria to improve analyses by excluding most subjects from disparate ancestry regardless of whether they are the result of miss-self-identification and/or due to mislabeling of samples.

Use of Ancestry Informative Markers for Admixture Studies

Another major use of continental AIMs is in admixture studies. The differences in admixture proportions estimated using the 128 I_n4 AIMs is illustrated in Fig. 1 and summarized in Table 2 for African Americans (AFA), Mexican Americans (MAM), Mexican (MXN) and

Puerto Rican (PRA) population groups. These results using STRUCTURE, similar to those with continental populations, are robust and yield consistent admixture proportions in multiple runs using appropriate analysis parameters (see Methods). The results also show that the overall admixture proportions of these groups, AFA, MAM, MXN and PRA can be ascertained with small numbers of I_n4 AIMs.

In order to further evaluate how consistently different subsets markers can estimate individual admixture we examined the correlation of ancestry assignments. Using the 128 I_n4 results as the standard we compared the estimated contribution of one of the ancestral parental populations contributing to each of three different admixed populations. These include West African contribution in AFA, European contribution in PRA, and Amerindian contribution in MAM and MXN. The latter two groups (MAM and MXN) were combined since the admixture proportions are similar. Marker sets chosen for their optimum ability to discriminate between four ancestral populations (I_n4 sets), and two ancestral populations (I_n2 sets) were examined (Fig. 2). The correlation values (r^2) for West African contribution in AFA are high, ranging between 0.988 for 96 I_n4 to 0.835 for 24 I_n4, suggesting that small number of markers are sufficient to identify West African contribution. Similar results in AFA were also observed using the marker sets selected specifically to distinguish European and West African (e.g. 0.976 for 48 I_n2 European/West African). As anticipated, the markers chosen for European/Amerindian differences did not accurately distinguish European/African admixture.

For Amerindian contribution in MAM and MXN the correlation values using I_n4 markers was also strong but did show a discernable decrease when 48 or 24 I_n4 markers were examined. For the I_n2 AIMs optimized for European/Amerindian differences, the results showed stronger correlations (e.g. 0.798 for 48 I_n2 European/Amerindian versus 0.733 for 48 I_n4). Similar results are also shown for the European contribution in PRA, however, the correlations were markedly lower. The correlations for European contribution in PRA population were 0.877, 0.587, 0.560, and 0.519 for 96 I_n4, 64 I_n4, 48 I_n4 and 24 I_n4.

The low correlation between estimates for European contribution in PRA may be explained by the fact that three ancestral populations, Europeans, Amerindians, and West Africans, have substantial contributions in the PRA population. This is unlike AFA and MAM/MXN, where there are two main contributing ancestral populations, West African and Europeans, and Amerindian and Europeans. Using $r^2 > 0.8$ as a threshold for high correlation, any of the I_n4 sets should be acceptable to estimate West African contribution in AFA, 128 I_n4, 96 I_n4 and 64 I_n4 are sufficient for Amerindian contribution in Mexican and Mexican American populations, and 128 I_n4 and 96 I_n4 sets should provide sufficiently accurate information for European contribution in PRA.

To further measure the precision of the ancestry estimation of individual subjects in admixed populations, we examined the 90% confidence intervals. For each individual the 90% Bayesian confidence interval was measured (STRUCTURE output). For each set of AIMs, the average size of this confidence interval was then calculated (Table 4). Comparison of these results shows the decrease in individual confidence intervals based on the number of markers and the dependency on the admixed population being analyzed. These confidence limits show that in studies of AFA, smaller sets can still provide good precision in individual admixture measurement. However, for MAM/MXN relatively larger numbers of AIMs are required. The confidence limits are smaller when I_n2 marker sets optimized for the particular admixed population are used. However, the 96 I_n 4 and 128 I_n4 set appear to perform very well in each of the admixed groups.

The ability to exclude subjects of other continental ancestry in admixed populations was also examined (Supplementary Table S4). For AFA, nearly all individuals of non-West African or European ancestry could be excluded at the 15% exclusion criteria while maintaining nearly all of the subjects of self-identified AFA ancestry using 64 or more I_n4 AIMs. However, for the MAM/MXN subjects much looser criteria (>30% non-Amerindian or European ancestry) were necessary to include >90% of self-identified MAM/MXN even with 96 I_n4 AIMs. This is probably due to the small West African contribution present in the MAM/MXN populations requiring a larger number of AIMs to enable good definition of this admixture component.

Performance of AIM sets in Association Studies

As another assessment of the performance of the AIM sets, we examined whether these AIMs could correct for false-positive association results in models for population specific disease susceptibility loci. Using 200K genotypes from the I-control database and additional genotypes available from other ongoing studies (see Methods) we specified specific genotypes as disease surrogates and identified true (located in a close genetic position to the modeled SNP) and false (unlinked) associated SNPs. These population sets included genotypes for each of the 128 I_n4 AIMs since each is included within the Illumina 300K array. Three disease gene models were specified using the surrogate phenotypes defined by SNPs in strong LD with 1) a nonsynonymous genetic substitution in *SLC24A5* on chromosome 15 under strong positive selection in Europeans, 2) lactase tolerance phenotype on chromosome 2 that is under strong positive selection in northern European populations and 3) a nonsynonymous coding variant in *ADH1B* under positive selection in East Asian populations (see **Method**s for additional details).

The surrogate phenotypes were specified in a sample set of 865 individuals primarily from three disparate continental populations, European (254 subjects), East Asia (283 subjects) and Africa (as represented by 328 African American subjects). In addition, the phenotype defined by *SLC24A5* was examined in 1847 African American subjects. For each of the phenotypes examined, both putative true positives (SNPs located close to the chromosomal position of the modeled genotype) and false positives, unlinked SNPs were found with strong association ($p < 0.01$ after Bonferroni correction) (Fig. 3 and Supplementary Table S5).

As expected, principal components analysis (PCA) using the entire 200K SNP sets were effective in correcting the false positive associations for each of the three surrogate phenotypes was examined in mixed population sets (Fig. 3a, b, c and Supplementary Table S5). The 128 I_n4 and 96 I_n4 AIM sets were nearly as effective in correcting the false positive associations. Smaller I_n4 sets also corrected most of the false positive results, however these sets failed on some of the analyses e.g. the false association for rs4871195 in the *LCT* model remained significant for 64 I_n4 and smaller sets. For the admixed AFA population group, similar results were observed (Fig. 3d and Supplementary Table S5). Here, the smallest set (24 I_n4) showed incomplete correction. Together, these analyses show that relatively small numbers of AIMs can correct for false positive results in these Mendelian models.

Discussion

The current study was undertaken to provide researchers with a set of validated AIMs for distinguishing continental populations. We believe that the results provide strong confidence that these 128 I_n4 AIMs and subsets of these SNPs can be used for characterizing sample sets from diverse population groups. These markers can be applied either to identify those individuals from a particular study that are members of one continental population group or alternatively used to adjust for population stratification due to differences in continental

population frequency in cases and controls. The former will reduce population heterogeneity that may also correspond to reducing genetic heterogeneity for specific traits. The latter can, as shown in our modeling studies, allow the reduction or elimination of false positive results.

Our analyses provide guidelines for application especially with regards using the program STRUCTURE (Falush, et al., 2003; Pritchard, et al., 2000a). Other computational programs including ADMIXMAP (Hoggart, et al., 2004) can also be applied with very similar results (data not shown). In general, as indicated in the methods section, the performance of smaller AIM SNP sets in STRUCTURE analyses is only consistently reproducible when very large numbers of iterations are used. This is not a major limitation since the computational time is not a major problem when small sets of markers are used even with large sample sets; several thousand samples will require <24 hours for 100,000 replicates using STRUCTURE and 48 markers. However, smaller marker sets (especially those <64) provide a poorer ability to exclude subjects of disparate continental ancestry and will provide less precision in the individual ancestry assignment. For larger studies (sample sizes of several thousand) the precision of individual assignments will be less consequential than for smaller studies in which the investigation will be more dependent on the accurate assessment of ancestry of each individual. Thus choice of the number of SNP AIMs depends on the populations being studied as well as practical aspects of genotyping. However, as shown in our study, the 96 I_n4 SNP AIMs perform well for each of the potential applications with only a very modest reduction of potential information compared with the 128 I_n4 set. Even smaller numbers perform adequately in particular situations but may require additional confidence in the prior information i.e. confidence in self identification of population membership.

A major application of SNP AIMs is to reduce false positives in association studies. For traits associated with continental ancestry our modeling studies found that relatively small numbers of SNP AIMs (64 or more) could adequately adjust for differences in ancestry stratification between cases and controls. It is notable that without the use of AIMs we observed many false positives even when the surrogate models used loci were not in complete linkage disequilibrium with the true ancestry associated trait (i.e. $r^2 = 0.73$ for model 2 and $r^2 = 0.53$ for model 3). This suggests that it is necessary to adjust for population structure for traits that are only partially association with continental ancestry and underscores the importance of the application of these or similar methods when subjects of mixed ancestry are studied. Our modeling studies also examined the use of AIMs in association tests for an admixed population (African Americans). Similar to the subject sets containing individuals from multiple continents, these studies showed that relatively small numbers of highly informative SNP AIMs (64 or more) can adequately adjust for population substructure and eliminate false positive results. Additional studies will be necessary to determine the efficacy of these AIMs in more complex sample sets and other population groups.

The identification of the ancestry groups using non-hierarchical clustering algorithms or for that matter PCA, is enhanced by the inclusion of representatives of the parental population groups. In the analyses performed in the current studies there were representatives of the different continental groups. The inclusion of these groups is particularly important when admixed populations are being examined. The inclusion of these groups, even without specifying population membership, allows more accurate cluster separation. In general, and specifically for the studies reported herein, we did not specify population membership, an available option in the STRUCTURE program. [Similar results are obtained using this option but with larger confidence intervals (data not shown)]. To facilitate the appropriate application of the AIMs described in this study, the genotypes of continental populations groups are provided as a resource to the scientific community (Supplementary Table S6).

Finally, for each of the SNP AIMs used in the current study a TaqMan® SNP genotyping assay is readily available (Supplementary Table S2). We also note that each of the SNPs is also part of the Illumina 300K array, that should enable inspection and utilization of genotypes that are provided in the I-control data base. A summary of the information for each SNP is provided in Supplemental Tables S2 and S6. In addition, since many researches may wish to use a smaller AIMs set we have optimized a panel of 96 SNPs for which robust TaqMan assays are available as a cost-effective format (see Supplementary Table S1 **footnote b**).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grants AR050267, DK071185, and P30 CA093373, and by Applied Biosystems. The researchers would also like to thank the Swedish Research Council support to MEAR and the Christine Landgraf Memorial Research Fund (LMB).

References

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. A haplotype map of the human genome. Nature. 2005; 437(7063):1299–1320. [PubMed: 16255080]

Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD. Measuring European population stratification with microarray genotype data. Am J Hum Genet. 2007; 80(5):948–956. [PubMed: 17436249]

Belkhir, K.; Borsa, P.; Chikhi, L.; Raufaste, N.; Bonhomme, F. Version 4.02. Montpellier, France: Laboratory Genome, Populations, Interactions CNRS UMR 5000, University of Montpellier II; 2001. GENETIX, software under Windows TM for the genetic of populations.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 2004; 74(6):1111–1120. [PubMed: 15114531]

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN. Demonstrating stratification in a European American population. Nat Genet. 2005; 37(8):868–872. [PubMed: 16041375]

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet. 2005; 37(11):1243–1246. [PubMed: 16228001]

Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55(4):997–1004. [PubMed: 11315092]

Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case-control studies. Am J Hum Genet. 2007; 80(5):921–930. [PubMed: 17436246]

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003; 164(4):1567–1587. [PubMed: 12930761]

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet. 2004; 36(4):388–393. [PubMed: 15052270]

Hamblin MT, Di Rienzo A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet. 2000; 66(5):1669–1679. [PubMed: 10762551]

Han Y, Gu S, Oota H, Osier MV, Pakstis AJ, Speed WC, Kidd JR, Kidd KK. Evidence of positive selection on a class I ADH locus. Am J Hum Genet. 2007; 80(3):441–456. [PubMed: 17273965]

Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K. An Icelandic example of the impact of population structure on association studies. Nat Genet. 2005; 37(1):90–95. [PubMed: 15608637]

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM. Control of confounding of genetic associations in stratified populations. Am J Hum Genet. 2003; 72(6):1492–1504. [PubMed: 12817591]

Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM. Design and analysis of admixture mapping studies. Am J Hum Genet. 2004; 74(5):965–978. [PubMed: 15088268]

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al. A high-resolution recombination map of the human genome. Nat Genet. 2002; 31(3):241–247. [PubMed: 12053178]

Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE, et al. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science. 2005; 310(5755):1782–1786. [PubMed: 16357253]

Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD, Leon-Velarde F, Moore LG, Vargas E, McKeigue PM, et al. A genomewide admixture mapping panel for Hispanic/Latino populations. Am J Hum Genet. 2007; 80(6):1171–1178. [PubMed: 17503334]

Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet. 2004; 36(5):512–517. [PubMed: 15052271]

Mitchell MK, Gregersen PK, Johnson S, Parsons R, Vlahov D. The New York Cancer project: rationale, organization, design, and baseline characteristics. J Urban Health. 2004; 81(2):301–310. [PubMed: 15136663]

Parra EJ, Kittles RA, Shriver MD. Implications of correlations between skin color and genetic ancestry for biomedical research. Nat Genet. 2004; 36(11 Suppl):S54–S60. [PubMed: 15508005]

Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scarnicci F, Ruiz-Linares A, Groop L, Saetta AA, Korkolopoulou P, et al. Discerning the ancestry of European Americans in genetic association studies. PLoS Genet. 2008; 4(1):e236. [PubMed: 18208327]

Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, Tandon A, Schirmer C, Neubauer J, Bedoya G, et al. A genomewide admixture map for Latino populations. Am J Hum Genet. 2007; 80(6):1024–1036. [PubMed: 17503322]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8):904–909. [PubMed: 16862161]

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000a; 155(2):945–959. [PubMed: 10835412]

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet. 2000b; 67(1):170–181. [PubMed: 10827107]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81(3):559–575. [PubMed: 17701901]

Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet. 2003; 73(6):1402–1422. [PubMed: 14631557]

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 2005; 1(6):e70. [PubMed: 16355252]

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic structure of human populations. Science. 2002; 298(5602):2381–2385. [PubMed: 12493913]

Salari K, Choudhry S, Tang H, Naqvi M, Lind D, Avila PC, Coyle NE, Ung N, Nazario S, Casal J, et al. Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. Genet Epidemiol. 2005; 29(1):76–86. [PubMed: 15918156]

Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet. 2001; 68(2):466–477. [PubMed: 11170894]

Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, Silva G, Belmont JW, Klareskog L, Gregersen PK. European Population Substructure: Clustering of Northern and Southern Populations. PLoS Genetics. 2006; 2(9):1339–1351.

Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, Malasky MJ, Scafe C, Le E, et al. A high-density admixture map for disease gene discovery in african americans. Am J Hum Genet. 2004; 74(5):1001–1013. [PubMed: 15088270]

Tian C, Hinds DA, Shigeta R, Adler SG, Lee A, Pahl MV, Silva G, Belmont JW, Hanson RL, Knowler WC, et al. A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. Am J Hum Genet. 2007; 80(6):1014–1023. [PubMed: 17557415]

Tian C, Hinds DA, Shigeta R, Kittles R, Ballinger DG, Seldin MF. A genomewide single-nucleotide-polymorphism panel with high ancestry information for african american admixture mapping. Am J Hum Genet. 2006; 79(4):640–649. [PubMed: 16960800]

Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, Klareskog L, Pulver AE, Qi L, Gregersen PK, et al. Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet. 2008; 4(1):e4. [PubMed: 18208329]

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007; 39(1):31–40. [PubMed: 17159977]

Weir B, Cockerham C. Estimating F-statistics for the analysis of population structure. Evolution. 1984; 38:1358–1370.

Yang N, Li H, Criswell LA, Gregersen PK, Alarcon-Riquelme ME, Kittles R, Shigeta R, Silva G, Patel PI, Belmont JW, et al. 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. Hum Genet. 2005; 118(3–4):382–392. [PubMed: 16193326]

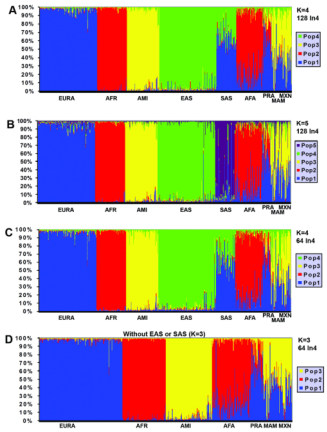

Figure 1. Analysis of population genetic structure using I_n4 AIMs

Each vertical line represents an individual subject. Along the abscissa each self identified population group is shown. The population groups include European American (EURA, 188 subjects), West African (AFR, 98 subjects), Amerindian (AMI, 88 subjects), East Asian (105 subjects), South Asian (SAS, 64 subjects) African American (88 subjects), Puerto Rican American (PRA, 28 subjects), Mexican American (MAM, 40 subjects) and Mexican (MXN, 26 subjects). Analyses were performed without any prior population assignment. Analyses for the128 I_n4 marker set are shown for 4 population groups (K=4) in (A), and K=5 in (B). Analyses for 64 I_n4 for K=5 in (C) and K=3 (without East or South Asian samples) in (D).

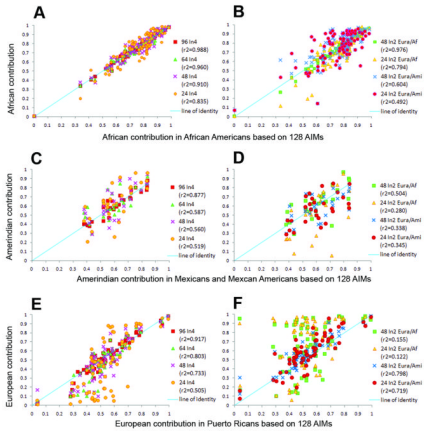

Figure 2. Correlation between the estimations of genetic contribution using different AIM sets and 128 I_n4 AIMs

The abscissa shows the 128 I_n4 result and the ordinate the result using the color coded AIM set. The individual for African contribution in African Americans [(A) and (B)], European contribution in Puerto Ricans [(C) and (D)], and Amerindian in Mexicans and Mexican Americans [(E) and (F)] are shown based on STRUCTURE analyses.

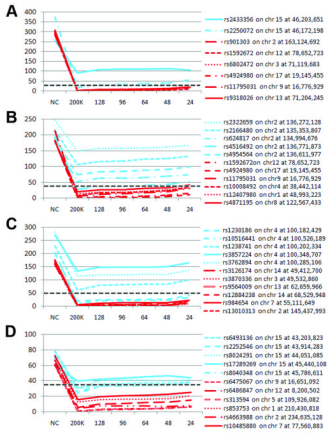

Figure 3. Correction of population stratification in association tests using different AIM sets

Three population specific alleles were used to model phenotypes prevalent in a particular population. The ordinate shows the χ^2 value with the first value showing the Armitage test result. The correction for false positive association tests (EIGENSTRAT analyses) using either 200K SNP markers, or the selected AIM sets are shown along the abscissa. The surrogate cases are defined by homozygosity for: (A) and (D) allele A for rs 2675348 in SLC24A5 locus; (B) allele A for rs1446585 in LCT locus; (C) allele A for rs100008281 in ADH1B locus. The surrogate cases are chosen in 865 samples from EURA, AFR, and EAS populations in (A), (B) and (C); and from 1847 African American samples in D). The dashed bold line represent nominal significance level (p=0.05) corrected for 200K independent tests: χ^2 = 26.6 (p=2.5e-7). The marker shade/color indicates the location of relative to the locus chosen to define the surrogate phenotype. The dark markers are located on chromosomes that do not contain the locus defining the surrogate phenotype while the lighter markers are located near the locus.

Table 1

Summary of Fst values between and within Ancestry Groups

	EURA[a]	AFR	AMI	EAS
EURA	0.00004[b]			
AFR	0.35558	0.00382		
AMI	0.40969	0.44161	0.03262	
EAS	0.21029	0.29647	0.17809	0.01454
SAS	0.0737	0.38736	0.35150	0.11988

[a] Populations are European American (EURA), West African (AFR), Amerindian (AMI) and East Asian (EAS).

[b] Fst values were determined by the Weir and Cockerham algorithm using the results genotypes for the 128 AIMs described in the text. The intrapopulation Fst was determined using two or three populations for the different continental populations. The population groups were: EURA (CEU and NYCP); AFR (YRI, Kanuri, and Bini); AMI (Mayan, Nahua, and Quechan); and East Asia (CHB, JPT and NYCPF). See Methods for further definition of population groups.

Table 2

Summary of Population Structure Results Using Markers Selected by Informativeness Between Four Continental Populations.

AIMs[a]		K = 5[b]					K = 4			
128		**Pop1**	**Pop2**	**Pop3**	**Pop4**	**Pop5**	**Pop1**	**Pop2**	**Pop3**	**Pop4**
	EURA (188)[c]	**0.96**	0.01	0.01	0.01	0.02	**0.97**	0.01	0.01	0.01
	AFR (98)	0.01	**0.97**	0.00	0.01	0.00	0.01	**0.97**	0.00	0.01
	AMI (105)	0.02	0.01	**0.94**	0.02	0.01	0.02	0.01	**0.94**	0.02
	EAS (188)	0.03	0.01	0.01	**0.93**	0.02	0.03	0.01	0.01	**0.94**
	SAS (64)	0.18	0.01	0.01	0.05	**0.75**	0.64	0.02	0.01	0.33
	AFA (88)	0.19	0.76	0.01	0.02	0.03	0.21	0.76	0.01	0.02
	PRA (28)	0.58	0.24	0.08	0.06	0.04	0.61	0.23	0.10	0.06
	MAM/MXN (66)	0.33	0.02	0.53	0.06	0.06	0.37	0.02	0.54	0.07
96	EURA	**0.95**	0.01	0.01	0.02	0.01	**0.97**	0.01	0.01	0.02
	AFR	0.01	**0.97**	0.00	0.01	0.00	0.02	**0.97**	0.00	0.01
	AMI	0.02	0.01	**0.93**	0.03	0.01	0.03	0.01	**0.94**	0.03
	EAS	0.03	0.01	0.02	**0.92**	0.02	0.04	0.01	0.02	**0.94**
	SAS	0.21	0.01	0.01	0.08	**0.68**	0.63	0.02	0.02	0.32
	AFA	0.19	0.75	0.01	0.03	0.02	0.21	0.75	0.01	0.03
	PRA	0.57	0.23	0.07	0.08	0.04	0.61	0.23	0.08	0.08
	MAM/MXN	0.32	0.03	0.51	0.10	0.05	0.36	0.02	0.52	0.09
64	EURA	**0.94**	0.01	0.01	0.01	0.02	**0.97**	0.01	0.01	0.02
	AFR	0.01	**0.97**	0.00	0.01	0.00	0.02	**0.97**	0.00	0.01
	AMI	0.02	0.01	**0.93**	0.03	0.01	0.03	0.01	**0.93**	0.03
	EAS	0.03	0.01	0.02	**0.91**	0.02	0.04	0.01	0.02	**0.93**
	SAS	0.21	0.01	0.01	0.06	**0.70**	0.67	0.02	0.02	0.29
	AFA	0.17	0.75	0.01	0.03	0.04	0.20	0.75	0.01	0.04
	PRA	0.56	0.23	0.07	0.05	0.10	0.64	0.23	0.07	0.06
	MAM/MXN	0.31	0.02	0.50	0.10	0.07	0.36	0.02	0.50	0.11
48	EURA	**0.93**	0.01	0.01	0.02	0.03	**0.96**	0.01	0.01	0.02
	AFR	0.01	**0.97**	0.01	0.01	0.00	0.02	**0.97**	0.01	0.01
	AMI	0.03	0.01	**0.92**	0.03	0.01	0.03	0.01	**0.92**	0.04

AIMs[a]		K = 5[b]					K = 4			
128		**Pop1**	**Pop2**	**Pop3**	**Pop4**	**Pop5**	**Pop1**	**Pop2**	**Pop3**	**Pop4**
	EAS	0.04	0.02	0.03	**0.90**	0.02	0.04	0.01	0.03	**0.91**
	SAS	0.29	0.01	0.02	0.08	**0.59**	0.71	0.02	0.02	0.25
	AFA	0.16	0.76	0.02	0.04	0.03	0.19	0.76	0.02	0.04
	PRA	0.55	0.22	0.07	0.06	0.10	0.63	0.22	0.07	0.08
	MAM/MXN	0.31	0.03	0.50	0.10	0.07	0.36	0.02	0.50	0.12
24	EURA	**0.86**	0.02	0.01	0.03	0.08	**0.95**	0.01	0.01	0.02
	AFR	0.01	**0.96**	0.01	0.01	0.00	0.02	**0.97**	0.01	0.01
	AMI	0.03	0.01	**0.88**	0.07	0.01	0.03	0.01	**0.89**	0.06
	EAS	0.04	0.02	0.05	**0.86**	0.03	0.06	0.02	0.05	**0.88**
	SAS	0.31	0.02	0.03	0.09	**0.55**	0.80	0.03	0.03	0.15
	AFA	0.13	0.75	0.03	0.05	0.04	0.17	0.75	0.03	0.05
	PRA	0.52	0.22	0.07	0.07	0.12	0.64	0.22	0.07	0.07
	MAMMXN	0.29	0.04	0.44	0.16	0.07	0.35	0.04	0.45	0.17

[a] The fraction of membership in each population group determined by STRUCTURE analyses is shown for different numbers of AIMs that were selected using I_n4 (see **Methods**). The number of AIMs is shown in the first column for each section of the table.

[b] The number of population groups (K) specified in the analysis.

[c] The number of subjects in each self identified group is provided for European American (EURA), West African (AFR), Amerindian (AMI), East Asian (EAS), South Asian (SAS), African American (AFA), Puerto Rican (PRA) and a combined group of Mexican (MXN) and Mexican American (MAM) populations. See **Methods** for additional self-identified population information.

Table 3

Comparison of the Ability of AIMs to Distinguish Different Continental Populations[a]

	number of AIMs (4 Pop In)[b]					number of AIMs (4 Pop In)				
	128	**96**	**64**	**48**	**24**	**128**	**96**	**64**	**48**	**24**
Subject Group	>10% non-EURA Ancestry[c]					>15% non-EURA Ancestry				
EURA (188)	0.04	0.05	0.03	0.09	0.09	0.01	0.01	0.02	0.02	0.05
AFR (98)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AMI (105)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
EAS (188)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SAS (64)	0.98	0.95	0.94	0.83	0.64	0.97	0.97	0.83	0.73	0.47
AFA (88)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
PRA (28)	1.00	1.00	0.96	0.96	0.89	1.00	0.96	0.93	0.86	0.79
MAM/MXN (66)	1.00	1.00	1.00	1.00	0.97	1.00	1.00	0.98	0.98	0.97
	>10% non-AFR Ancestry					>15 % non-AFR Ancestry				
EURA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AFR	0.04	0.04	0.02	0.09	0.07	0.03	0.03	0.01	0.02	0.03
AMI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
EAS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SAS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AFA	0.88	0.88	0.88	0.84	0.80	0.77	0.78	0.77	0.75	0.73
PRA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MAM/MXN	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	>10% non-AMI Ancestry					>15% non-AMI Ancestry				
EURA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AFR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AMI	0.17	0.23	0.22	0.26	0.31	0.10	0.08	0.12	0.13	0.19
EAS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SAS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AFA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PRA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MAM/MXN	0.95	0.97	0.95	0.97	0.95	0.95	0.95	0.92	0.91	0.83

	number of AIMs (4 Pop In)[b]					number of AIMs (4 Pop In)				
	128	**96**	**64**	**48**	**24**	**128**	**96**	**64**	**48**	**24**
	>10% non-EAS Ancestry					>15% non-EAS Ancestry				
EURA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AFR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AMI	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99
EAS	0.12	0.14	0.17	0.26	0.61	0.06	0.06	0.08	0.13	0.42
SAS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AFA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PRA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MAM/MXN	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98

[a] The ability to exclude subjects based on different numbers of AIMs is shown using both 10% and 15% non-Ancestry group membership.

[b] The top of each column indicates the number of AIMs in each set. The AIMs were selected based on informativeness in four population groups (see **Methods**). Analyses were performed using K = 4.

[c] For each set the criteria for exclusion is shown. The fraction of subjects that would be excluded for each criterion is indicated for each subject group.

Table 4

Summary of Confidence Intervals Using Different Marker Sets

	90% Confidence Intervals[a]		
AIMs	**AFA**	**MAM/MXN**	**AFA, PRA, MAM/MXN**
128 I_n4	0.167	0.187	0.163
96 I_n4	0.174	0.219	0.180
64 I_n4	0.196	0.267	0.219
48 I_n4	0.220	0.302	0.247
24 I_n4	0.274	0.418	0.305
64 I_n2[b]	0.171	0.209	N.A.
48 I_n2	0.188	0.232	N.A.
24 I_n2	0.240	0.304	N.A.

[a] The average of the individual subject 90% Bayesian confidence intervals (CI) was determined using the different AIM sets. For the AFA and MAM/MXA subject groups the CI were determined using K=2. For the combined admixed group (AFA, PRA, MAM/MXA) the CI was determined using K=3.

[b] For the 2PopIn marker sets, the CI was determined using the EURA/AFR for AFA or EURA/AMI for the MAM/MXN subject groups.