European Population Genetic Substructure: Further Definition of Ancestry Informative Markers for Distinguishing among Diverse European Ethnic Groups

Chao Tian,[1] Roman Kosoy,[1] Rami Nassir,[1] Annette Lee,[2] Pablo Villoslada,[3] Lars Klareskog,[4] Lennart Hammarström,[5] Henri-Jean Garchon,[6] Ann E Pulver,[7] Michael Ransom,[1] Peter K Gregersen,[2] and Michael F Seldin[1]

[1]Rowe Program in Human Genetics, Departments of Biochemistry and Medicine, University of California Davis, Davis, California, United States of America; [2]The Robert S Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, North Shore LIJ Health System, Manhasset, New York, United States of America; [3]Center for Applied Medical Research, University of Navarra, Pamplona, Spain; [4]Karolinska University Hospital, Stockholm, Sweden; [5]Karolinska Institute at KUS Huddinge, Stockholm, Sweden; [6]Institut Cochin, Inserm U567, University Paris Descartes, Paris, France; and [7]Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, Maryland, United States of America

The definition of European population genetic substructure and its application to understanding complex phenotypes is becoming increasingly important. In the current study using over 4,000 subjects genotyped for 300,000 single-nucleotide polymorphisms (SNPs), we provide further insight into relationships among European population groups and identify sets of SNP ancestry informative markers (AIMs) for application in genetic studies. In general, the graphical description of these principal components analyses (PCA) of diverse European subjects showed a strong correspondence to the geographical relationships of specific countries or regions of origin. Clearer separation of different ethnic and regional populations was observed when northern and southern European groups were considered separately and the PCA results were influenced by the inclusion or exclusion of different self-identified population groups including Ashkenazi Jewish, Sardinian, and Orcadian ethnic groups. SNP AIM sets were identified that could distinguish the regional and ethnic population groups. Moreover, the studies demonstrated that most allele frequency differences between different European groups could be controlled effectively in analyses using these AIM sets. The European substructure AIMs should be widely applicable to ongoing studies to confirm and delineate specific disease susceptibility candidate regions without the necessity of performing additional genome-wide SNP studies in additional subject sets.

Online address: http://www.molmed.org
doi: 10.2119/molmed.2009.00094

INTRODUCTION

Over the last several years there has been substantial progress in using genotypes to ascertain population genetic substructure and in applying this information to association testing (1–9). These studies have been advanced by the availability of efficient platforms for genotyping several hundred thousand SNPs, increased efforts in sampling various population groups, and application of both highly supervised (clustering algorithms) and largely unsupervised methods for analyzing high dimensional data (that is, genotypes in many individuals) (1,3,6). The results of such studies in European and European American populations have enabled the description of relationships among various European ethnic groups and the ability to use these relationships to better inform association studies by accounting for population stratification differences between cases and control groups without the necessity to utilize family-based methods such as transmission disequilibrium testing (10–13). The current study was undertaken to further define the relationships among European population groups and to use this information in identifying and testing SNP sets enriched for population substructure information.

Our initial description of European population substructure primarily defined a single axis of variation (north/south) using 7,000 genome-wide SNPs (10). Subsequently, studies by multiple groups, including our own, have utilized principal components analyses (PCA) or multidimensional scaling to further define European population substructure using several hundred thousand SNPs that are present in genome-wide panels

Address correspondence and reprint requests to *Michael F Seldin, Room 4453, Tupper Hall, Department of Biochemistry: Med, One Shields Avenue, University of California, Davis, Davis, CA 95616. Phone: 530-754-6016; Fax: 530-754-6015; Email: mfseldin@ucdavis.edu.*
Submitted July 21, 2009; Accepted for publication August 18, 2009; Epub (www.molmed.org) ahead of print August 27, 2009.

(11–16). These mathematical methods reduce higher dimensional data (each individual genotype) into lower dimensions based on patterns of identity by descent relationships. In particular, a recent study using PCA of ~200,000 SNP genotypes has shown that, with the exclusion of certain outliers, there is a remarkable concordance between the first two principal components and the geographical position of the country of origin for the included subject samples (15). The current study uses different subjects including population groups that were not included in previous studies, and adds further support to this observation. In addition, our study shows how inclusion or exclusion of particular ethnic groups and regions alters the PCA graphical description and evaluation of relationships among European populations. Furthermore, the current study shows that analyses of geographically or genotyped defined subregions (for example, southern European populations) may allow a clearer evaluation of ancestry information and differences that may be important in evaluating association studies or defining homogeneous ethnic groups. Notably, our analyses included Middle Eastern population groups. These population groups are more closely related to European populations than South Asian population groups (14,17), and are more closely related to southern European groups as shown here. In addition, individuals of origin in the Middle East or individuals with mixed European/Middle Eastern heritage may self identify as European or non-Hispanic white for different sample set collections or demographic characterizations.

The practical aspects of applying population substructure to association testing have been reviewed recently (18,19). For several different study designs, the ability to define population substructure using specific smaller sets of SNPs rather than tens or hundreds of thousands of genome-wide SNPs may be advantageous in pre-selecting population members before genome-wide SNP studies, limited testing of candidate SNPs, and fine mapping critical chromosomal regions using additional samples not included in initial genome-wide scans. Previously, several groups have identified European ancestry informative markers (AIMs) that can address population stratification partially. In the current study, we have identified and examined SNP sets for ascertaining more subtle population stratification in both northern and southern European populations. The results provide strong support for the application of these SNP marker sets to genetic studies of European populations.

MATERIALS AND METHODS

Populations Studied

For European substructure studies presented in Figures 1–3, the populations include those from the Human Genome Diversity Panel (HGDP), HapMap, the I-control database, Italian, Spanish, Swedish, and European Americans. The HGDP, HapMap, and I-control database genotypes were available from online databases. These included HapMap subjects (48 CEU), HGDP subjects (14 Orcadian, 28 Sardinian, 13 Northern Italian, 8 Tuscan) and 1,488 subjects from Children's Hospital of Philadelphia from the I-ControlDB (www.illumina.com/iControlDB, Illumina, San Diego, CA). Genotypes from other HGDP subjects (20 Druze, 23 Bedouin, 22 Palestinian, 13 Russian, 12 Basque, 12 Adygei) and 255 European American neurodevelopmental controls were obtained from the NIH Laboratory of Neurogenetics (http://neurogenetics.nia.nih.gov/paperdata/public/) (NIH, Baltimore, MD, USA). The sample set also used 591 Swedish genotypes (collected by LK). Additional samples that were genotyped included 12 Spanish samples (collected by PV), and 1,873 European Americans that were recruited as part of the New York Cancer Project (NYCP), a prospective longitudinal study (20). For a subset of the NYCP participants, complete four grandparent information was available and used in graphic depictions (see Figure 1 legend). All of the subjects (4,446) were included in Figure 1. Figure 2 (185 subjects) and Figure 3 (213 subjects) included only those individuals from specified groups as indicated in the legend.

For the testing of European substructure ancestry informative markers (ESAIMs) the study also included samples from 139 French (H-JG) and 240 Norwegian (collected by LH) subjects. Finally, genotypes from South Asian HGDP samples (7 Burusho and 15 Balochi) were used in some analyses and were obtained from NIH Laboratory of Neurogenetics.

For all European and European American subjects, blood 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 linked only to demographic information.

It should be noted that the current set of subjects has relatively few individuals of known Eastern European ancestry (11 subjects) and no subjects of known self-identified ancestry from many specific European countries or ethnicity (for example, southern Slavic population groups). However, the large number of subjects from both NYCP (1,832 individuals) and Children's Hospital of Philadelphia (CHOP) (1,487 individuals) are likely to provide good representation of most European population groups as well as individuals with mixed European heritage. For the NYCP participants, many have partial grandparent origin information and these show representation from all regions of Europe (10,12, and unpublished information).

Genotyping

Genotyping was performed using a 300K Illumina array according to the Illumina Infinium 2 assay manual (Illumina), as described previously (21).

Data Filters

SNPs and individual samples with less than 90% complete genotyping information from any data set were excluded from analyses. SNPs that showed ex-

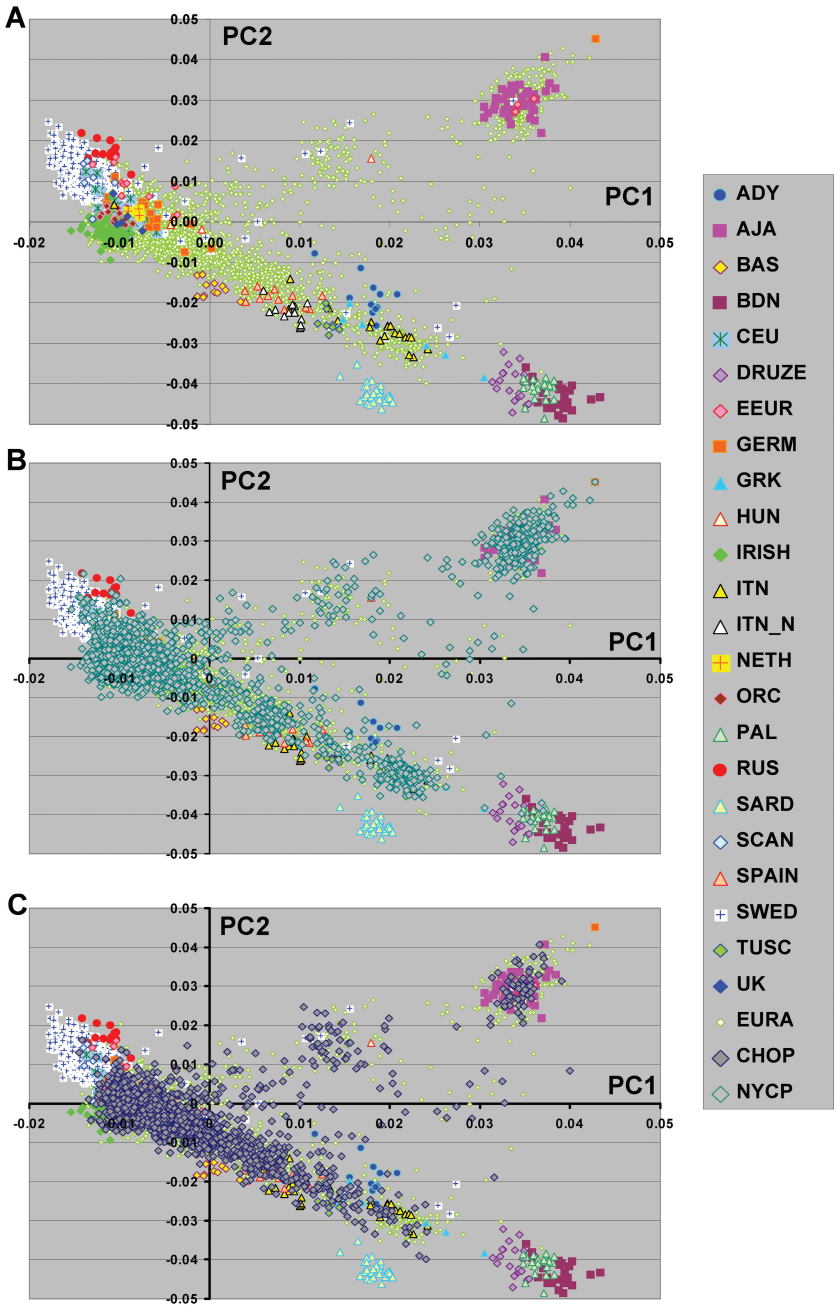

Figure 1. Principal component analyses of substructure in a diverse set of subjects of European descent. Graphic representation of the first two PCs based on analysis with >250,000 SNPs are shown. Color code shows subgroup of subjects for each population group. The subjects included Adygei (ADY, 12 subjects), Ashkenazi Jewish American (AJA, 40 subjects), Basque (BAS, 12 subjects), Bedouin (BDN, 23 subjects), CEPH European American (CEU, 48), Druze (20 subjects), Eastern European American (EEUR, 11 subjects), German American (GERM, 17 subjects), Greek American (GRK, 7), Hungarian American (HUN, 4), IRISH (84 subjects), Italian American (ITN, 20 subjects), northern Italian (ITN_N, 13 subjects), Dutch American (NETH, 3 subjects), Orcadian (ORC, 14 subjects), Palestinian (PAL, 22 subjects), Russian (RUS, 13 subjects), Sardinian (SARD, 28 subjects), Scandinavian American (SCAN, 6 subjects), Spanish (SPAIN, 12 subjects), Swedish (SWED, 591 subjects), Tuscan (TUSC, 8 subjects), and United Kingdom American (UK, 5 subjects). Each of the specific country or ethnic color-coded origins had consistent four grandparent origin information. The total number of individuals in this analysis was 4,446. (A) European Americans (EURA) without four grandparental information are shown (contains both NYCP and CHOP). (B) and (C) illustrate the distribution of the EURA from NYCP (1,873 subjects) and CHOP (1,488 subjects), respectively.

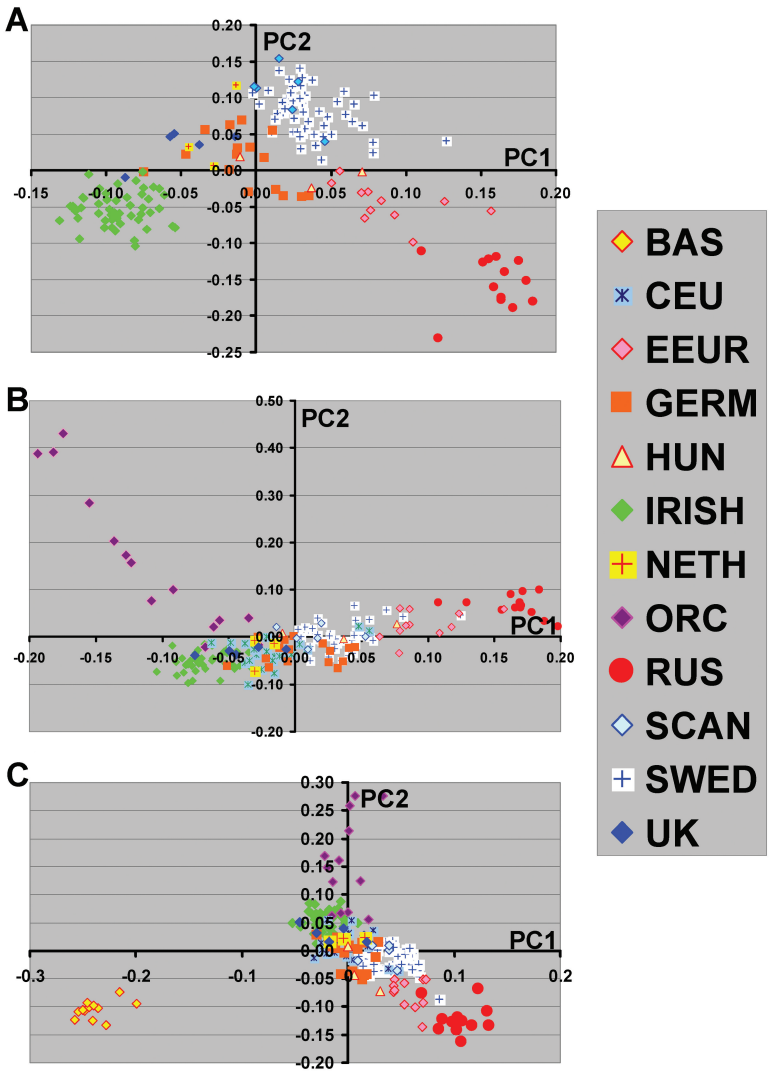

Figure 2. Principal component analyses of Northern European populations. The color-coded group membership is shown with the symbols corresponding to Figure 1 legend information. The subject sizes were as shown in Figure 1 with the exception of the Swedish group for which the sample size was reduced to 40 subjects. (A) Northern European population groups without inclusion of ORC, CEU, and BAS subjects. (B) and (C) show PCA results when either ORC and CEU, or BAS and CEU groups are added. Inclusion or exclusion of CEU did not affect the PCA pattern (data not shown).

treme deviation from Hardy–Weinberg (HW) equilibrium ($P < 0.00001$) in individual population groups were also excluded from analysis. The relatively strict HW criterion was chosen to minimize genotyping error rates and potential artifacts introduced by differences between array lots and genotyping sources (that is, data derived from multiple laboratories). These filters resulted in a total of 270,000 autosomal SNPs that were used for these studies. In addition, individuals with evidence of >10% contribution from other continents, with the exception of the South Asian subjects, were excluded from further study. This was performed using 128 continental AIMs (22). Samples also were filtered for possible cryptic relationships using the PLINK program (6).

Statistical Analyses

F_{st} and F_{is} was determined using Genetix software (23) that applies the Weir and Cockerham algorithm (24). A measure of informativeness for each SNP (I_n) was determined using an algorithm described previously (25). Hardy–Weinberg equilibrium was determined using HelixTree 5.0.2 software (Golden Helix, Bozeman, MT, USA). Linkage disequilibrium (LD) was determined using Haploview.

Population structure was examined using STRUCTURE v2.1 (1,9) using parameters and AIMs described previously (22). Briefly, each analysis was performed without any prior population assignment and was performed at least three times with similar results using >200,000 replicates and >100,000 burn-in cycles under the admixture model. 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 $\lambda = 1$ option where λ estimates the prior probability of the allele frequency and is based on the Dirichlet distribution of allele frequencies. This analysis was performed to exclude individuals with evidence of substantial continental admixture from Europe, Africa, or the American continent (see Data Filters). STRUCTURE analysis also was used for dividing the European population groups into Northern European and Southern European population groups for the correlation studies using 192 ns-ESAIMS (12). The STRUCTURE analysis of self-identified groups using these ESAIMs (Supplemental Figure 1) also corresponds to PCA results. For the correlation studies we used a cut-off of >0.8 membership in Pop 1 for Northern European population groups and <0.8 membership in Pop1 (or >0.2 membership in Pop 2) for the Southern European populations.

PCA was performed using the EIGENSTRAT statistical package (3). All analyses were performed after deleting the MHC region on chromosome 6, and known regions of common inversions on chromosomes 8, 11, and 17, since regions

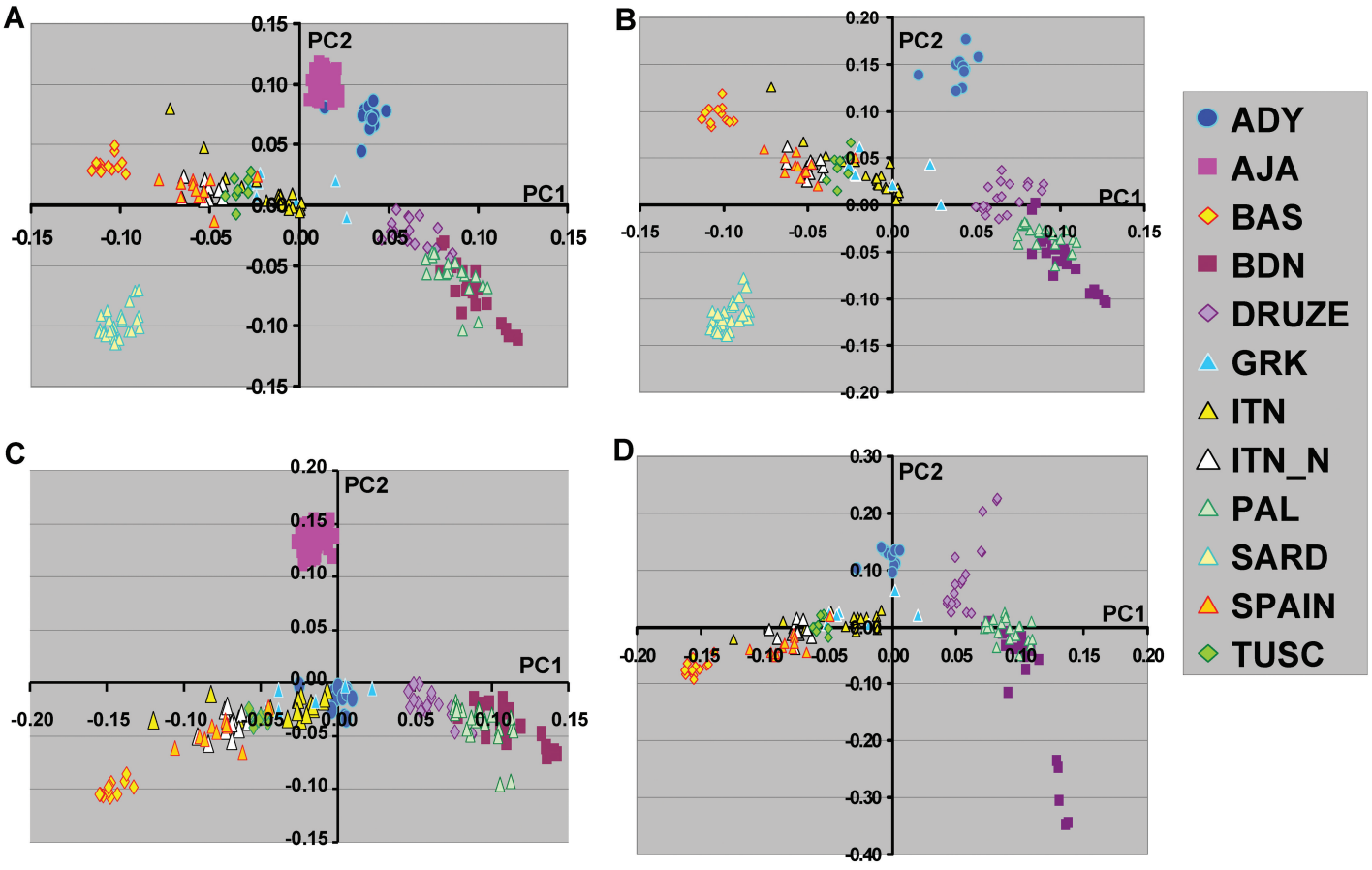

Figure 3. Principal component analyses of Southern European populations. (A) All subjects groups. (B) PCA analysis without Ashkenazi Jewish American (AJA). (C) PCA without the Sardinian (SARD) group. (D) PCA without AJA and SARD. The subject numbers were the same as those indicated in Figure 1.

of high linkage disequilibrium can overly influence PCA results (12).

Selection of European Substructure AIMs

Several strategies were investigated to identify SNPs for substructure information. These included using an algorithm for informativeness (I_n) (25), and SNP scores from PCA as well as using either the entire population studied or selecting particular disparate population groups. The most effective strategy in our hands was to combine sets of markers including those that were selected previously for distinguishing population groups along a northern/southern axis (12) with additional SNP sets chosen using I_n values between particular northern European groups or particular southern European groups. For the northern ESAIMs we combined a previous set of SNPs (12) with additional SNP sets chosen from Irish/Swedish or Irish/central Europe analyses. For these groups, the PCA results were used to select subjects that matched known four grandparent Irish, or four grandparent central European (German and French). Each of these selections included a minimum of 200 subjects from each group. For the southern European ESAIMs we utilized AJA/Arab subjects for the SNP selection. For this selection, the AJA included 200 subjects and the Arab subjects included 50 subjects (16 Druze, 16 Bedouin, and 18 Palestinian). Finally, each SNP set was screened to remove any SNPs with strong LD ($r^2 > 0.8$) and those were not compatible with I-select (Illumina) genotyping platform. The ESAIMs are provided in Supplemental Table 1. Note that the EUROset1 SNPs contain 85% (1,037/1,212) of the ESAIM SNPs used in our previous studies (12).

All supplementary materials are available online at www.molmed.org.

RESULTS

Population Differentiation Among European Populations

To examine similarities and differences in population differentiation among European and closely related populations,

Table 1. Paired Fst values for European populations.

Fst[a]	DRUZE	BDN	PAL	AJA	GRK	ITN	ADY	SPN	BAS	IRISH	GERM	EEURA	RUS	SWED	ORC	SARD
BDN[b]	0.0072															
PAL	0.0064	0.0056														
AJA	0.0088	0.0108	0.0093													
GRK	0.0052	0.0064	0.0057	0.0042												
ITN	0.0057	0.0079	0.0064	0.0040	–0.0001											
ADY	0.0092	0.0123	0.0108	0.0107	0.0054	0.0067										
SPN	0.0096	0.0103	0.0101	0.0056	0.0035	0.0010	0.0090									
BAS	0.0186	0.0204	0.0199	0.0144	0.0098	0.0084	0.0180	0.0060								
IRISH	0.0154	0.0187	0.0170	0.0109	0.0067	0.0048	0.0110	0.0037	0.0086							
GERM	0.0121	0.0147	0.0136	0.0072	0.0039	0.0029	0.0089	0.0015	0.0079	0.0010						
EEURA	0.0128	0.0149	0.0133	0.0068	0.0049	0.0040	0.0086	0.0033	0.0091	0.0034	0.0014					
RUS	0.0194	0.0211	0.0202	0.0137	0.0108	0.0088	0.0120	0.0079	0.0126	0.0038	0.0037	0.0029				
SWED	0.0167	0.0204	0.0191	0.0120	0.0084	0.0064	0.0117	0.0055	0.0100	0.0020	0.0007	0.0025	0.0036			
ORC	0.0194	0.0212	0.0201	0.0146	0.0103	0.0080	0.0136	0.0063	0.0124	0.0039	0.0048	0.0055	0.0092	0.0046		
SARD	0.0163	0.0183	0.0166	0.0131	0.0088	0.0072	0.0204	0.0071	0.0133	0.0140	0.0117	0.0132	0.0210	0.0155	0.0162	
TUSC	0.0086	0.0102	0.0096	0.0066	0.0005	0.0004	0.0094	0.0023	0.0084	0.0055	0.0032	0.0045	0.0108	0.0061	0.0098	0.0083

[a]The paired Fst value is the mean determined from three nonoverlapping sets of 3,500 SNPs using the Weir and Cockerham algorithm (24). Complete data including the standard deviation is provided in Supplemental Table 2.
[b]Population groups included Druze, Bedouin (BDN), Palestinian (PAL), Ashkenazi Jewish American (AJA), Greek (GRK), Italian (ITN), Adygei (ADY), Spanish (SPN), Basque (BAS), IRISH, German (GERM), Eastern European (EEUR), Russian (RUS), Swedish (SWED), Orcadian (ORC), Sardinian (SARD), and Tuscan (TUSC).

paired Fst values were determined between 18 population groups that were typed with genome-wide SNP arrays (see Materials and Methods). The studies included genotypes derived from HGDP (14), and samples collected in Europe and European American participants (see Materials and Methods). The Fst values (Table 1) were obtained using three random non-overlapping sets of 3,500 SNPs distributed over the autosomal genome (minimum of 50-kb distance between SNPs). The small differences in these independent samplings (mean SD = 0.0009; median SD = 0.0008) indicate that this approach resulted in good estimations of paired Fst values. In general, Fst values corresponded to geographical relationships with smaller values between population groups with origins in neighboring countries/regions (for example, Tuscan/Greek, Fst = 0.001) compared with those from very different regions in Europe (for example, Russian/Palestinian, Fst = 0.020) similar to previous studies (10). Ashkenazi Jewish participants showed smaller paired Fst values with southern European populations (for example, Ashkenazi/Italian, Fst = 0.004) than with northern populations (for example, Ashkenazi/Swedish, Fst = 0.0120). All of the intra-European paired Fst values were nearly an order of magnitude smaller than those observed in our previous studies when intercontinental groups (for example, Europe and East Asia) are determined (for example, Han Chinese/Swedish, Fst = 0.11) (17). However, there is substantial overlap in Fst values comparing European and Middle Eastern groups and those between European and South Asian population groups (17). For example, the Balochi and Burusho (two HGDP ethnic groups sampled in Pakistan) showed the following paired Fsts (Balochi/Palestinian, Fst = 0.016; Balochi/Swedish, Fst = 0.018; Burusho/Palestinian, Fst = 0.027; Burusho/Swedish, Fst = 0.029; Balochi/Burusho, Fst = 0.008). These paired Fsts overlapped with several inter-European Fsts (for example, Palestinian/Swedish Fst = 0.019, Basque/Bedouin Fst = 0.020). Thus, this measurement of population differentiation does not appear to provide a clear grouping of different ethnic groups (see Discussion).

Fst values also were determined for each population sample and did not indicate a strong inbreeding component for any of the tested sample groups with the exception of the Bedouin participants (Supplemental Table 3).

Principal Components Analyses Using >250,000 SNPs

To further explore the relationship among European population groups and examine population substructure, PCA was performed using the genotype results from a set of ~300,000 autosomal SNPs that was common to each of the populations examined. For most individuals with self-reported ethnic identities there was a general correspondence with the geographical location of origin (Figure 1A). For example, the relationship of Italian groups and the subjects from the island country of Sardinia shows a striking resemblance to maps of Europe. In addition, genotypes from the same or related population groups typed in different laboratories showed similar PCA results (for example, north Italian and Tuscan groups genotyped as part of HGDP overlapped with Italian American subjects).

For European American participants without self-identified ethnic affiliation

there was a wide distribution in PC1 and PC2 that overlapped with those individuals of known ethnic affiliation or country of origin. These included subjects from two large publicly available data sets (New York Cancer Project [NYCP] and Children's Hospital of Philadelphia [CHOP] genotypes in I-control database [www.illumina.com/iControlDB, Illumina, San Diego, CA]) (Figure 1B, C). Not surprisingly, these sample sets from collection sites in the New York City region and Philadelphia also showed differences in their relative distribution pattern observed in principal component (PC)1 and PC2. This is particularly evident with regards to both the Ashkenazi Jewish and Irish ethnic groups (Table 2).

Table 2. Distribution of NYCP and CHOP European subjects.

Presumed ethnicity[a]	PCA range[b] PC1	PCA range[b] PC2	Percentage of samples[c] NYCP	Percentage of samples[c] CHOP
Ashkenazi	0.030, 0.040	0.020, 0.040	13.0%	4.5%
North European	–0.016, –0.005	–0.010, 0.015	59.7%	45.4%
Italian/Greek	–0.002, 0.025	–0.035, –0.010	13.4%	16.7%
Irish	–0.016, –0.010	–0.010, 0.0001	14.0%	3.8%

[a]Presumed ethnicity corresponding to the PCA results. The Irish group is a subset of the Northern European larger group.
[b]The coordinate range for PC1 and PC2 for four different groups of subjects. The first value indicates the lower boundary, and the second value the upper boundary for each PC range.
[c]The percentage of subjects is shown for NYCP (1,832 subjects) and CHOP (1,487 subjects).

PCA of Northern and Southern European Population Groups

Our previous studies showed that further elucidation of the relationships among Northern European individuals was possible when analyses were performed excluding subjects of southern European origin. For the current study, we next examined Northern and Southern population groups separately and the effect of including or excluding different ethnic groups (Figures 2–3). We have included the Middle Eastern groups in the analysis of Southern European population groups due to the close relationship of these population. For subjects of Northern European origin, the inclusion of Orcadian subjects showed a substantial change in the PC2 relationships in which the other Northern European populations were no longer separated on this axis (compare Figure 2A with Figure 2B). The pattern observed with the Orcadian individuals also suggests that there may be ongoing admixture between this population and less isolated European populations. In addition, the inclusion of Basque subjects, a group that appeared intermediate between Northern and Southern European populations, changed both PC1 and PC2 results (Figure 2C). These differences in PCA results may be critical in assessing the ability of particular ancestry marker sets to define relevant population substructure for specific studies (see Discussion).

Similarly, studies were performed for the Southern European population groups (see Figure 3). Here the inclusion or exclusion of the Basque population group had no effect on the PC1/PC2 graphical representation (data not shown). However, the inclusion or exclusion of Sardinian and Ashkenazi Jewish population groups shifted the position of the Adygei subjects. It also is worth noting that the inclusion of the Arab population groups results in larger separation between northern Italian and southern Italian (and/or Greek) subjects and suggests that inclusion of the Arab population genotypes may be useful in analyses of southern European population groups (data not shown).

PCA of European and South Asian Population Groups

To further explore the relationship of population groups in Europe, we also performed PCA including the HGDP South Asian groups, Burusho, and Balochi. These South Asian population groups showed relatively small population differentiation with continental European subjects (14,17,26,27). Thus inclusion of these populations might further suggest clines, contributions, or relationships between European and these neighboring geographic ethnic groups. PCA analyses showed that these South Asian groups were different from any of the European groups (Figure 4) consistent with previous studies (14,26,27). This difference was less pronounced when only southern European in contrast to northern European groups were examined (Figure 4C). Interestingly, the Adygei, a population from the Caucasus, showed a closer relationship to these South Asian ethnic groups than other European groups, consistent with geographical relationships.

Identification and Testing of European Substructure Ancestry Informative Markers

AIMs that discern population substructure are likely to be useful in candidate gene, chromosomal position-based association studies, and defining homogeneous subject sets (28). Our previous studies identified two sets of European substructure AIMs (ESAIMs): ns-ESAIMs and north-ESAIMs. The ns-ESAIMs separated European populations on a single axis (north/south) that corresponds to the largest variation in European substructure in PCA (PC1). The north-ESAIMs provided distinction along an east/west gradient when only northern European population groups were studied. The availability of genotypes from many additional European population groups, as well as additional European American subjects, provided an opportunity to ascertain and test sets of SNP AIMs that might improve population substructure analyses. In addition, for the current study, we only included SNP

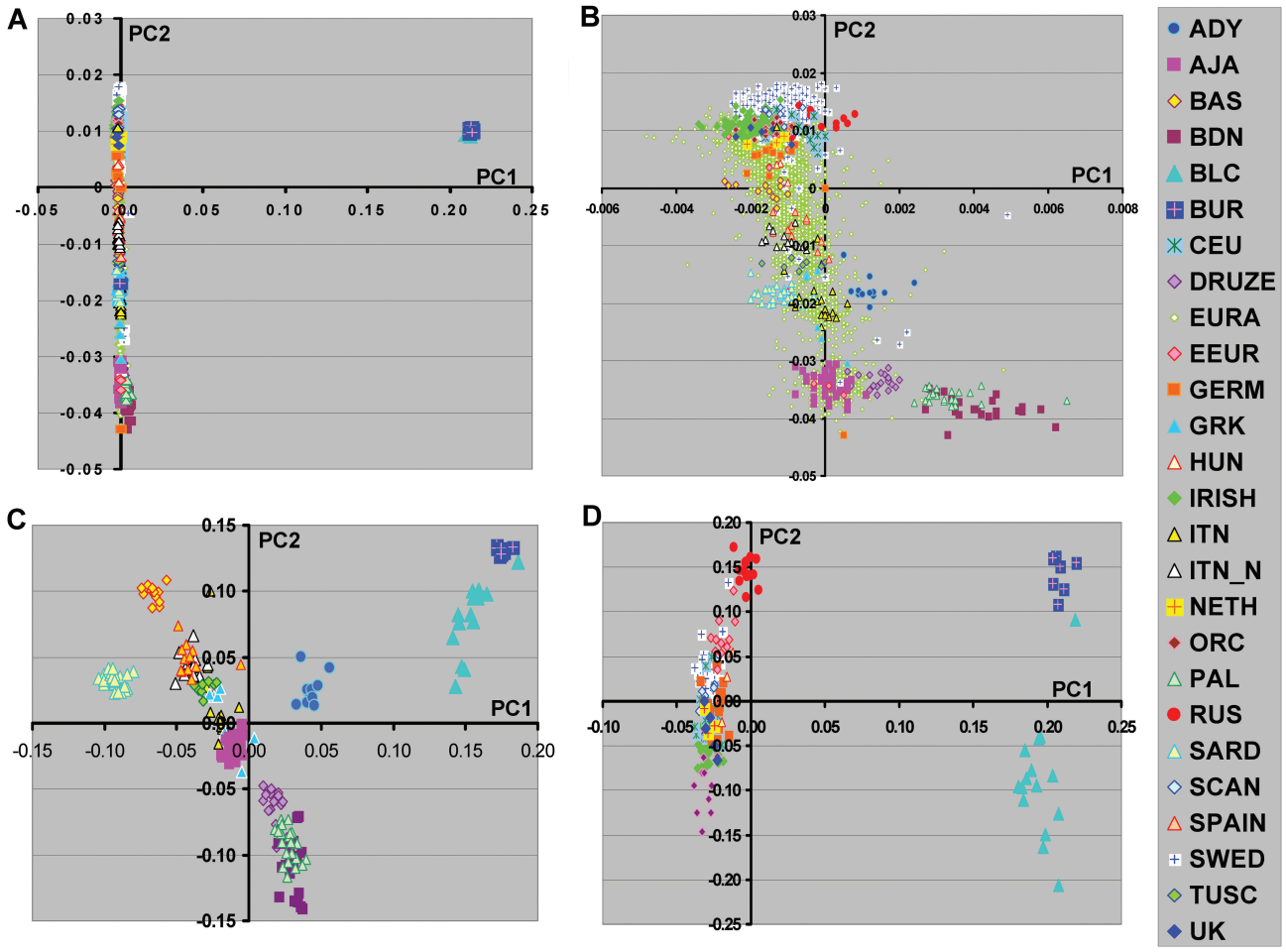

Figure 4. PCA analyses of European population groups together with two South Asian groups. (A) All European population groups together with Balochi (BLC, 15 subjects) and Burusho (BUR, 7 subjects). (B) Expanded view of European populations from PCA shown in A. (C) PCA results from southern European population groups analyzed together with BAL and BUR. (D) PCA results from northern European population groups analyzed together with BAL and BUR.

AIMs compatible for typing on the Illumina Infinium platform.

For north-ESAIMs we selected additional SNPs by using a measurement of informativeness (I_n) (25) using subjects not used in our previous studies. These additional SNPs were selected from those with the greatest I_n determined from analyses of Irish and Swedish genotypes and French and Swedish genotypes (see Materials and Methods for details). These SNPs were combined with our previously identified north-ESAIMs and ~300 north/south-ESAIMs. To assess the potential usefulness of these AIMs, we performed PCA using an independent set of samples. The results (Figure 5A) showed the clustering of self-identified groups in PC1 and PC2.

Similarly, for south-ESAIMs, we utilized a combination of Arab ethnic groups (Druze, Palestinians, and Bedouin) and Ashkenazi Jewish subjects to identify SNPs that could distinguish between these groups. These were combined with ~300 north/south AIMs that can separate Spanish, Italian, Greek, and Ashkenazi Jewish ethnic groups (12). The PCA results using independent sample (Figure 5B) showed the clustering of self-identified groups.

We also used the combined set of SNPs to examine the northern and southern European populations together. The PCA results (Figure 5C) show very similar patterns to that observed with 270,000 SNPs (Figure 5D).

As a measure of the ability of the ESAIMs to assess substructure accurately we examined the correlation (r^2) with 270,000 SNPs (Table 3). Northern and southern population groups were analyzed both separately and together, and the results also were compared with random marker sets. For each test population group (all, north and south subgroups), the complete ESAIM set

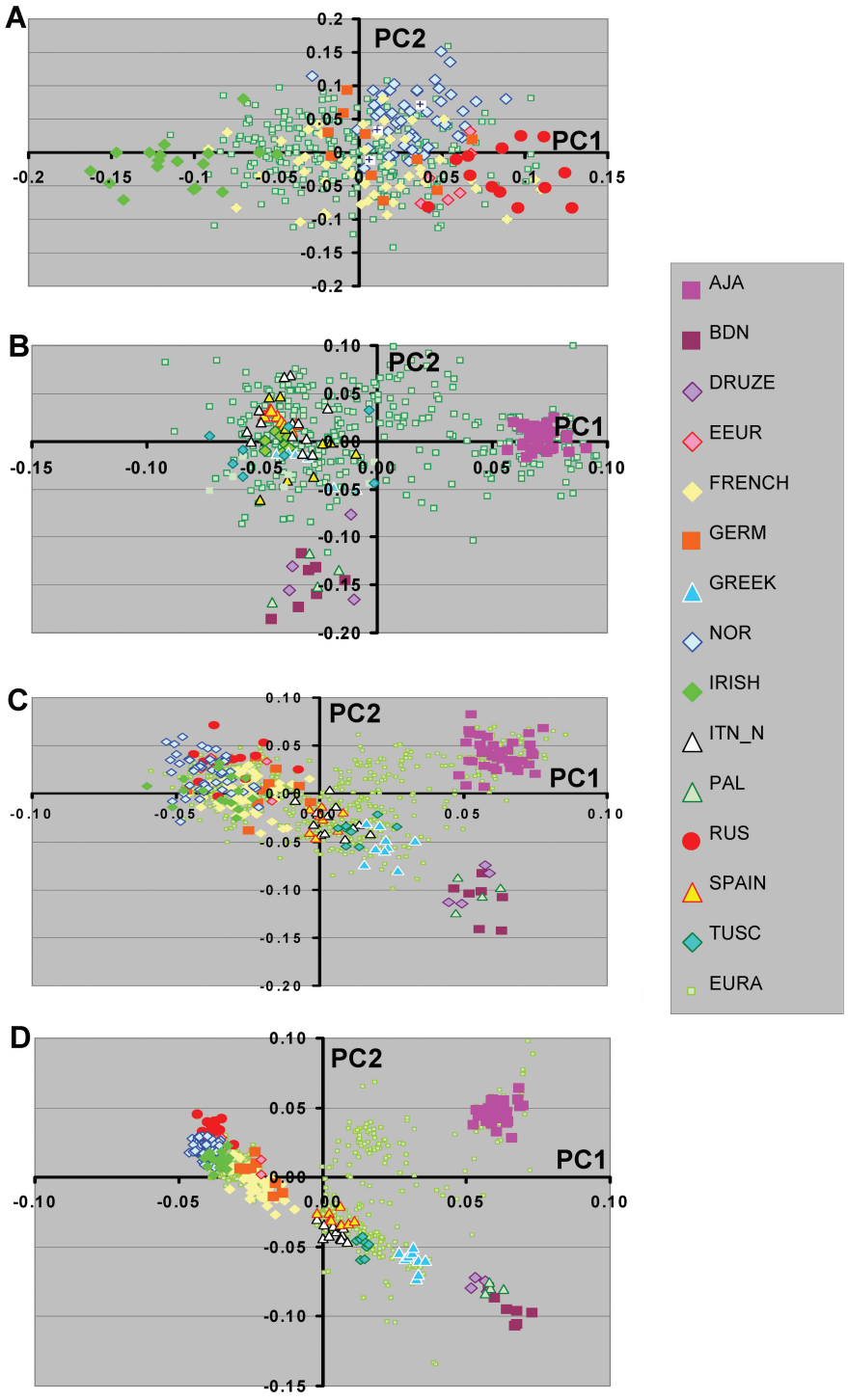

Figure 5. Ability of European Substructure AIMs to discern population substructure. (A) PCA analysis of northern Europeans with North AIMs set 2. (B) PCA analysis of southern Europeans with South AIMs. (C) PCA of all test population samples with EURO set2 ESAIMs. (D) PCA of all test populations with 270,000 SNPs. Note: The French subjects included only those identified as Northern European (~75% of subjects) and does not represent the diversity of all French subjects.

containing 3,519 SNPs (EUROset2 in Table 3) showed better correlation results than random sets of 6,000 SNPs. Smaller marker sets (for example, NORTHset2, and SOUTH) showed equivalent results when the test group was limited to either northern or southern European populations. In addition, although 3,000 random SNPs showed high correlation in the entire test group (both north and south together) for PC1 and PC2, the correlations for the northern and southern groups were much weaker when considered separately.

Performance of ESAIMs in Controlling for Population Substructure

To further assess the potential value and performance of ESAIMs we examined SNPs with large (10% to 25%) allele frequency differences between different European subgroups. The different subgroups were selected based on PCA results and included both individuals with self-identified ethnic affiliation and those without known self-identity information (see Supplemental Figure 2). This procedure allowed the separation of five distinct subgroups of individuals that correspond to different regions or ethnic groups. These included French, Irish/United Kingdom, Scandinavian, Ashkenazi, and other southern European including Italian and Greek (OSOUTH). We then identified those SNPs with the largest allele frequency difference between different pairs of subject groups. From these SNPs, we selected test SNPs for which the significant differences (in case/control association tests) between groups would most likely be due to differences in population substructure. This was determined by EIGENSTRAT (3) analyses using 270,000 SNP set (see Materials and Methods) and included markers for which EIGENSTRAT corrected *P* values were substantially different from baseline. We then examined the ability of different ESAIMs and random AIMs to adjust for population substructure differences that could account for the allele frequency differences of the test SNPs by examining Armitage χ^2

Table 3. Summary of correlations for European AIMs.

	ALL test group[b]		NORTH test group		SOUTH test group	
Marker set[a]	PC1	PC2	PC1	PC2	PC1	PC2
Random 3,000	0.92[c]	0.60	0.20	0.01	0.78	0.04
Random 6,000	0.96	0.77	0.55	0.05	0.93	0.36
Random 12,000	0.98	0.88	0.76	0.08	0.96	0.72
Random 24,000	0.99	0.95	0.88	0.37	0.98	0.87
Random 48,000	1.00	0.97	0.95	0.63	0.99	0.88
EUROset1 (2,795)	0.95	0.79	0.54	0.01	0.92	0.45
EUROset2 (3,519)	0.96	0.81	0.57	0.19	0.92	0.55
NORTHset1 (2,034)	NA	NA	0.53	0.02	NA	NA
NORTHset2 (2,762)	NA	NA	0.56	0.19	NA	NA
SOUTH (1,078)	NA	NA	NA	NA	0.87	0.40

[a]The random marker results are the mean from three independently selected marker groups for each category (3K, 6K, 12K, 24K, and 48K). The EUROset2 contained EUROset1 markers. Similarly, the NORTHset2 contained the NORTHset1 SNPs. There was a small overlap of SNPs in the NORTH and SOUTH SNP set (see Materials and Methods and Supplemental Table 1 for details).
[b]The ALL test group included 789 European subjects and did not include any subjects used in AIMs selection (see Materials and Methods). The NORTH (378 subjects) and SOUTH (411 subjects) were subsets of the ALL test group determined from STRUCTURE results (see Materials and Methods).
[c]All correlations are shown as r^2 values with PCA using 270,000 SNPs.

P values before and after EIGENSTRAT correction based on PCA (Table 4).

For certain population pairs (for example, ASH/OSOUTH and OSOUTH/FRENCH) most of the ESAIM sets and even 3,000 random SNPs performed well (Table 4). For example in the ASH/OSOUTH population pairing, the rs4859259 showed an initial uncorrected *P* value of 2.2E–10. After correction with any of the SNP sets, only very modest or insignificant *P* values were observed. However, for the other population pairs, the 3,000 random SNPs performed poorly in controlling for population structure. For one particular paired population group (FRENCH/IRISHUK) the most comprehensive ESAIMs (EUROset2) containing 3,519 selected SNPs showed substantially better ability to control for population substructure than 3,000 or 6,000 random SNP sets. It also is notable that the EUROset2 SNPs overall performed better than the EUROset1 SNPs that include both the majority of our previous set of north-ESAIMs (12) as well as additional SNPs (see Materials and Methods). For example, EUROset1 failed to show an appropriate correction for rs6723108 in the FRENCH/IRSHUK population pair (initial *P* value = 2.2E–10, EUROset1 corrected *P* value = 8.9E–08) whereas, when using the EUROset2 markers, the *P* value was adjusted (EUROset2 corrected *P* value = 4.8E–01).

Overall, the EUROset2 SNPs appeared to be comparable to 12,000 random SNP sets and only moderately less effective than the 270,000 SNPs. These data, together with the PCA correlation studies, support the use of the EUROset2 SNPs in future studies. The ESAIMs are provided in Supplemental Table 1.

DISCUSSION

The current study extends the definition of European population substructure and the relationships among diverse European ethnic groups. The results generally are consistent with the observation that the largest variations in PCA (those graphically depicted in PC1/PC2) correspond to geographical relationships. This result also is consistent with our recent studies of East Asia population substructure (29). While perhaps surprising, this result is consistent with migration (demic expansion) with geographic physical boundaries being the most critical or common factor in determining genotypic patterns in population groups within continents.

The current study extends the analysis of European population genetic structure to include additional southern European groups and Arab populations. Even within Italy, the relative position of northern Italians compared with subjects from Tuscany is consistent with the general geographic correspondence of PCA results. Interestingly, the majority of Italian Americans (NYCP four grandparents defined) appear to derive from southern Italy and overlap with subjects of Greek heritage. Both of these observations are consistent with previous historical information (30,31).

Possible exceptions to this observation of geographic correspondence include the Ashkenazi Jewish population. While the Ashkenazi are clearly of southern origin based on both PCA and STRUCTURE studies, in our analyses of diverse European populations (see Figure 1), this group appears to have a unique genotypic pattern that may not reflect geographic origins. Furthermore, the inclusion or exclusion of particular ethnic groups (that is, Ashkenazi Jewish, and Sardinian for southern European, and Orcadian for Northern European) shifted the relationships in PCA when southern or northern Europeans were examined separately. Similarly, the inclusion of South Asian populations (see Figure 4C) changes the relation-

Table 4. PCA correction for ancestry using ESAIMs.

	SNP	Chr	Mb	*P* value[a]	270,000[b]	EUROset1[c]	EUROset2	North	South	Ran 3,000	Ran 6,000	Ran 12,000
ASH/OSOUTH[d]	rs6587597	1	150	**8.2E–11**[e]	2.5E–01	6.5E–01	2.9E–01	1.5E–02	5.8E–01	2.2E–01	3.2E–01	1.7E–01
ASH/OSOUTH	rs1400452[f]	2	18	**3.0E–11**	1.3E–03	1.8E–03	9.1E–04	**2.5E–06**	1.1E–04	8.6E–03	4.3E–04	2.1E–02
ASH/OSOUTH	rs4859259	3	184	**2.2E–10**	5.3E–01	7.5E–01	7.5E–01	4.4E–01	3.7E–01	2.8E–02	1.1E–01	5.8E–01
ASH/OSOUTH	rs7456425	7	81	**4.7E–09**	7.5E–01	7.5E–01	6.5E–01	6.1E–02	8.3E–02	5.1E–02	5.3E–01	6.5E–01
ASH/OSOUTH	rs7139066	12	26	**1.3E–10**	6.9E–02	3.4E–01	2.5E–01	1.4E–02	4.0E–02	2.5E–01	1.6E–02	2.9E–01
OSOUTH/FRENCH	rs6723108	2	135	**5.2E–11**	6.5E–01	2.5E–01	4.6E–02	5.1E–02	7.8E–02	4.4E–03	1.1E–01	2.4E–01
OSOUTH/FRENCH	rs1584930	3	95	**1.3E–07**	3.4E–01	7.7E–03	2.4E–02	6.1E–02	4.5E–04	1.6E–02	1.3E–03	4.7E–03
OSOUTH/FRENCH	rs10515893	5	165	**2.9E–08**	2.4E–01	5.1E–02	1.7E–01	4.8E–02	8.2E–03	3.1E–04	1.4E–02	3.6E–02
OSOUTH/FRENCH	rs2072633	6	32	**2.9E–08**	1.3E–01	4.2E–03	6.2E–03	1.1E–03	1.1E–02	5.6E–04	7.8E–02	8.9E–02
OSOUTH/FRENCH	rs11663558	18	19	**5.9E–08**	1.0E + 00	1.6E–01	1.8E–01	6.5E–02	1.6E–02	8.9E–02	8.9E–02	1.7E–01
FRENCH/SCAN	rs4954564	2	137	**<E–12**	3.2E–01	3.7E–04	5.8E–01	4.8E–01	**4.9E–07**	1.6E–04	5.0E–04	1.0E–02
FRENCH/SCAN	rs2856718	6	33	**3.5E–12**	3.7E–01	2.2E–03	2.3E–02	1.4E–02	**4.6E–06**	**1.5E–05**	7.7E–03	1.5E–02
FRENCH/SCAN	rs11038910	11	47	**4.7E–09**	1.1E–01	4.1E–04	5.2E–03	6.2E–03	2.3E–04	**9.7E–05**	1.9E–02	1.0E + 00
FRENCH/SCAN	rs1146904	13	76	**2.2E–08**	6.5E–02	1.7E–02	4.9E–03	2.2E–03	4.1E–04	**7.4E–07**	2.3E–02	2.1E–02
FRENCH/SCAN	rs11631323[f]	15	96	**1.7E–09**	1.0E + 00	3.8E–03	2.1E–02	9.6E–03	8.7E–05	2.2E–03	1.1E–01	1.8E–01
SCAN/IRISHUK	rs10517522	4	39	**3.9E–08**	1.3E–02	3.0E–02	3.2E–02	9.1E–03	**5.7E–07**	**1.3E–05**	4.8E–04	6.5E–02
SCAN/IRISHUK	rs1265048	6	31	**6.9E–08**	1.8E–02	3.0E–03	2.5E–02	5.8E–03	**1.8E–06**	**8.3E–05**	1.8E–02	6.9E–03
SCAN/IRISHUK	rs511512	13	76	**8.3E–10**	2.7E–02	4.7E–03	6.5E–03	8.2E–03	**3.3E–08**	**2.7E–08**	**1.9E–08**	3.3E–04
SCAN/IRISHUK	rs255052	16	67	**1.5E–08**	1.7E–01	1.1E–01	4.4E–01	9.4E–02	**1.3E–07**	5.0E–04	1.8E–01	5.4E–02
SCAN/IRISHUK	rs9303363	17	50	**1.0E–08**	1.8E–01	3.4E–02	6.9E–03	3.4E–03	**1.5E–07**	6.9E–03	6.5E–03	4.8E–02
FRENCH/IRISHUK	rs6723108	2	135	**2.2E–10**	5.6E–04	**8.9E–08**	4.8E–01	6.5E–01	**4.2E–10**	**4.0E–10**	**5.1E–06**	8.2E–04
FRENCH/IRISHUK	rs1965299[f]	4	117	**1.2E–07**	2.8E–04	**9.2E–05**	1.1E–04	**1.8E–05**	**5.4E–07**	**3.4E–07**	**6.3E–05**	1.7E–03
FRENCH/IRISHUK	rs12498670	4	178	**4.9E–07**	1.5E–03	**2.9E–05**	**3.9E–05**	**3.7E–05**	**3.2E–07**	**8.7E–07**	**7.4E–05**	5.6E–04
FRENCH/IRISHUK	rs3135029	6	33	**7.4E–07**	7.4E–02	1.5E–03	1.4E–02	2.6E–03	1.6E–04	**4.1E–06**	**9.2E–05**	1.9E–03
FRENCH/IRISHUK	rs11645416	16	50	**6.5E–08**	3.5E–04	**2.9E–05**	1.5E–03	2.0E–04	**1.9E–07**	**8.9E–08**	**1.5E–05**	9.1E–04

[a]The *P* value was determined by the Armitage χ^2 test using the first population group as "case" and the second group as "control."
[b]The adjusted *P* value based on correction for PCA using EIGENSTRAT software is shown for 270,000 SNPs and each of the AIMs and random marker sets. The EUROset1 and EUROset2 are as defined in Table 2 and Supplemental Table 3.
[c]The EUROset1 contains 1,037 of 1,212 ESAIMs previously reported (12) and an additional 1,757 SNPs (see Materials and Methods).
[d]The five different population groups were defined by PCA (see Materials and Methods): Ashkenazi (ASH), 88 individuals; Other South (OSOUTH), 210 individuals; French, 444 individuals; and Irish United Kingdom (IRISHUK), 328 individuals.
[e]Bold *P* values highlight *P* values $<10^{-4}$.
[f]These SNPs were isolated significant values. For each of the other SNPs chosen, multiple closely linked SNPs (within 50 Kb) showed similar *P* values.

ships of the population groups with the Ashkenazi Jewish population appearing in the center of a presumed southern European cline. These findings are consistent with our previous observations (12), and show that PCA results are highly dependent on which population groups are included in the analysis. Thus, there should be some caution in interpreting these results and other results from similar analytic methods with respect to ascribing origins of particular ethnic groups. It is worth mentioning that two of the afore noted groups in our analyses, Sardinian and Orcadian, are either island populations, suggesting that relative isolation of particular groups may underlie some of the observed results. However, it also should be noted that the Bedouin population group with high inbreeding (Supplemental Table 3), is not represented differently in PCA than other Arab groups without evidence in high inbreeding.

The differences observed between measures of population differentiation and analyses based on either model-dependent clustering or PCA also are noteworthy. Although a common measurement of population differentiation (Fst) has been used in measurements of "distance/years" between population groups (32), the paired Fst values do not necessarily correspond to a measurement of the differences in genotypic patterns. This is illustrated in the current study with respect to the relationship of South Asian population groups and European groups. Although paired Fst values overlap between intra-European and inter-European/South Asian groups, PCA shows that the largest axis of variation

unambiguously separates European and South Asian groups. Presumably this is of practical importance since it is the patterns of genotypic variation that are critical in causing false positives (and perhaps false negatives) due to population stratification differences between case and control groups.

The current study focuses on the PC1 and PC2 results in analyses of European substructure. Additional PCs do show additional substructure, however, the amount of variation is relatively small compared with PC1 and PC2 when northern and southern European populations are considered separately (Supplemental Figure 3). In addition, it should be noted that the SNPs included in any of the population substructure studies excluded regions of high linkage disequilibrium and the ESAIMs were selected to exclude SNPs with $r^2 > 0.8$.

An important aspect of the current study was the identification and characterization of AIMs applicable to association studies in European and European American populations. The current availability of several thousand genotypes from a single platform (Illumina in the current study) enabled a more extensive selection of AIMs and the ability to test the performance of these selected markers in different sample sets. With respect to northern European populations, a direct comparison with our previous north-ESAIM set was not possible due to the absence of a portion of these SNPs in our current dataset. However, the currently recommended EUROset2 SNPs performed better than the EUROset1 SNPs containing 85% the north ESAIMs (Table 4).

Our strategy in selecting markers was based on using I_n (25) differences between paired population groups that showed large separations in PC1 and PC2 in either southern European, northern European, or combined analyses. Alternate approaches, including the use of PCA SNP scores, did not result in ESAIM sets of comparable ability to discern European substructure (data not shown).

In the current study, two analyses were presented that we believe demonstrate the ability and usefulness of these ESAIMs to ascertain and correct for population substructure. It should be acknowledged that one of these criteria (correlation with "whole genome" PCA results in PC1 and PC2) will depend in part on the diversity of the individuals and perhaps whether a particular population group is included within the sample set (that is, PC1 and PC2 are not necessarily fixed as discussed above). The inclusion of many samples from collections in NYC and Philadelphia that have a broad distribution of ethnic origin, albeit as defined by PCA, suggests that these markers will be of particular value in studies where ethnic origin matching by self reporting is not available or unreliable. However, the applicability of these or other ESAIM sets may vary depending on whether the current sample set is reasonably inclusive of subjects in a particular study. Our assessment of the ability of the marker sets to account for allele frequency differences between large European/European American subgroups supports the practical application of these ESAIMs (Table 4) and suggests that these ESAIMs will address population substructure in most, if not all, European/European American datasets. In conclusion, we believe that the current ESAIMs (EUROset2) will have wide applicability to complex genetic studies within European populations.

ACKNOWLEDGMENT

We thank Stephen Johnson and Robert Lundsten for informatics support on the New York Cancer Project samples. We also thank Anthony Liew and Houman Khalili for expert assistance with genotyping. We thank the volunteers from the different populations for donating blood samples. This work was supported by NIH grants DK071185, AR050267, and AR44422.

DISCLOSURE

The authors declare that they have no competing interests as defined by *Molecular Medicine*, or other interests that might be perceived to influence the results and discussion reported in this paper.

REFERENCES

1. Pritchard JK, Stephens M, Donnelly P. (2000) Inference of population structure using multilocus genotype data. *Genetics*. 155:945–59.
2. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. (2000) Association mapping in structured populations. *Am. J. Hum. Genet*. 67:170–81.
3. Price AL, *et al*. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet*. 38:904–9.
4. Satten GA, Flanders WD, Yang Q. (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet*. 68:466–77.
5. Hoggart CJ, *et al*. (2003) Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet*. 72:1492–504.
6. Purcell S, *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. 81:559–75.
7. Dawson KJ, Belkhir K. (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res*. 78:59–77.
8. Epstein MP, Allen AS, Satten GA. (2007) A simple and improved correction for population stratification in case-control studies. *Am. J. Hum. Genet*. 80:921–30.
9. Falush D, Stephens M, Pritchard JK. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*. 164:1567–87.
10. Seldin MF, *et al*. (2006) European population substructure: clustering of northern and southern populations. *PLoS. Genet*ics. 2:1339–51.
11. Bauchet M, *et al*. (2007) Measuring European population stratification with microarray genotype data. *Am. J. Hum. Genet*. 80:948–56.
12. Tian C, *et al*. (2008) Analysis and application of European genetic substructure using 300 K SNP information. *PLoS. Genet*. 4:e4.
13. Price AL, *et al*. (2008) Discerning the ancestry of European Americans in genetic association studies. *PLoS. Genet*. 4:e236.
14. Li JZ, *et al*. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. 319:1100–4.
15. Novembre J, *et al*. (2008) Genes mirror geography within Europe. *Nature*. 456:98–103.
16. Lao O, *et al*. (2008) Correlation between genetic and geographic structure in Europe. *Curr. Biol*. 18:1241–8.
17. Nassir R, *et al*. (2009) An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC Genet*. 10:39.
18. Tian C, Gregersen PK, Seldin MF. (2008) Accounting for ancestry: population substructure and genome-wide association studies. *Hum. Mol. Genet*. 17:R143–50.
19. McCarthy MI, *et al*. (2008) Genome-wide associa-

tion studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet*. 9:356–69.
20. Mitchell MK, Gregersen PK, Johnson S, Parsons R, Vlahov D. (2004) The New York Cancer project: rationale, organization, design, and baseline characteristics. *J. Urban Health*. 81:301–10.
21. Duerr RH, *et al*. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*. 314:1461–3.
22. Kosoy R, *et al*. (2009) Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum. Mutat*. 30:69–78.
23. Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F. (2001) GENETIX, software under Windows TM for the genetic of populations [computer program]. Version 4.02. Montpellier (France): Laboratory Genome, Populations, Interactions CNRS UMR 5000, University of Montpellier II.
24. Weir B, Cockerham C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*. 38:1358–70.
25. Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet*. 73:1402–22.
26. Rosenberg NA, *et al*. (2002) Genetic structure of human populations. *Science*. 298:2381–5.
27. Yang N, *et al*. (2005) 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*. 118:382–92.
28. Seldin MF, Price AL. (2008) Application of ancestry informative markers to association studies in European Americans. *PLoS. Genet*. 4:e5.
29. Tian C, *et al*. (2008) Analysis of East Asia genetic substructure using genome-wide SNP arrays. *PLoS ONE*. 3:e3862.
30. Mangione J, Morreale B. (1993) *LA Storia: Five Centuries of the Italian American Experience*. New York: HarperCollins. 580 pp.
31. Woodhead AG. (1962) *Greeks in the West*. New York: Praeger. 243 pp.
32. Cavalli-Sforza LL, Menozzi P, Piazza A. (1996) *The History and Geography of Human Genes*. Princeton (NJ): Princeton University Press. 413 pp.