Hypothesis/Commentary

Ancestry Estimation and Correction for Population Stratification in Molecular Epidemiologic Association Studies

Jill S. Barnholtz-Sloan,1 Brian McEvoy,2 Mark D. Shriner,3 and Timothy R. Rebbeck4

1Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, Ohio; 2Trinity College Dublin, Dublin, Ireland; 3Department of Anthropology, Pennsylvania State University, University Park, Pennsylvania; and 4Center for Clinical Epidemiology and Biostatistics, Department of Biostatistics and Epidemiology, and Abramson Cancer Center, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania

Race, Ethnicity, Ancestry, and Genetics

Historically, racial groups have been defined by common geographic origins and shared physical characteristics, such as skin color, facial features, and hair texture. Linnaeus’s Systema Naturae (1) described four racial groups (Europeanus, Asiaticus, Americanus, and Africanus), which were subdivided in 1775 by Johann Blumenbach into Caucasian, Mongolian, Ethiopian, American, and Malay. Current commonly used racial/ethnic categories are those defined by the U.S. Census Bureau (http://www.census.gov). Although racial/ethnic classifications are not systematically or uniformly defined or applied (2, 3), genetic studies using polymorphic loci have shown that self-identified race or ethnicity correlate with ancestral population of origin (4-6). Aside from ancestry, cultural and behavioral factors influence an individual’s self-identified race/ethnicity (7, 8). Thus, race/ethnicity should be recognized as a complex composite variable. Here, we define race/ethnicity as a self-identified concept of ancestry, culture, and behavior, such as an individual may report to the U.S. census or a research study.

Classifying individuals into classes that represent heterogeneous racial/ethnic groups may simplify data collection and analysis, but it may also misclassify a person’s actual ancestral background (that is, the origins of their familial lineage; ref. 9) and limit assessment of variation within racial/ethnic groups that is relevant for understanding disease risk or outcome. For example, regional estimates of European ancestry among African Americans vary widely from 3.5% among the Gullah Sea Islanders of South Carolina (10) to 22.5% among African Americans in New Orleans (11). Using self-reported race/ethnicity as a proxy for ancestral background is even more problematic in Latinos, who show substantial variation based on country of birth or nationality; estimates of the proportion of African, European, and Native American ancestry are 37%, 45%, and 18% in Puerto Ricans and 8%, 61%, and 31% in Mexicans (12). Furthermore, research methods that allow choices of only one racial/ethnic group may be inadequate because many persons can trace their ancestry to multiple ancestral populations. More than 2.5% of United States residents reported that they belonged to more than one racial/ethnic group in the 2000 Census (13).

Here, we provide an overview of the issue of population stratification and how to test for and adjust for it using ancestry estimation techniques. Population stratification refers to differences in allele frequencies between cases and controls due to systematic differences in ancestry rather than association of genes with disease (14). We also discuss how to choose the appropriate genomic markers for ancestry estimation.

Ancestry and Bias in Molecular Epidemiologic Association Studies

Much variation in genetic ancestry can exist within or between racial/ethnic groups, thereby causing significant population stratification to be present not only in recently admixed populations like African Americans and Latinos (15-17) but also in European American populations (18-21) and historically isolated populations including Icelanders (22). A consequence of population stratification is the potential for increased allelic associations and deviations from Hardy-Weinberg equilibrium (23). Another consequence of population stratification is bias in the estimate of genetic associations, which can lead to incorrect inferences as well as inconsistency across reports (24). In order for bias due to population stratification to exist, both of the following must be true: (a) the frequency of the marker genotype of interest varies significantly by race/ethnicity and (b) the background disease prevalence varies significantly by race/ethnicity. If either of these is not fulfilled, bias due to population stratification cannot occur. Bias due to population stratification can induce both false-positive and false-negative associations (24, 25). This bias has been shown in some studies to be small in magnitude (26-28) and bounded by the magnitude of the difference.
in background disease rates across the populations being compared (29). Simulation studies have shown that the adverse effects of population stratification increase with increasing sample size (25, 30). An unresolved question is how large the difference in disease rates or genotypes must be for meaningful bias to arise.

When race/ethnicity can be accurately described in terms of actual ancestry and there is ancestral homogeneity in a study population, standard epidemiologic approaches of matching or statistical adjustment by race/ethnicity may be sufficient to remove or reduce bias due to population stratification. Controlling for self-reported race has generally been thought to suffice (31); however, self-reported race/ethnicity and/or ancestry can be quite unreliable. Burnett et al. (32) showed that only 49% to 68% of non-Hispanic European American siblings agreed on their ancestry. Recent data show that matching on ancestry is more robust. However, in many populations, whether recently admixed or not, individuals cannot accurately report their precise ancestry (32, 33).

Other approaches exist that account for ancestry and minimize the potential for bias due to population stratification. The transmission-disequilibrium test has been shown to be the most robust test with respect to controlling for population stratification (34-36). However, because it requires data from parent-child triads, it may be too expensive or impractical to implement for late-onset complex diseases. Therefore, other methods have been developed to test for and/or adjust for population stratification in case-control studies, although no true consensus has been reached as to which method is best (27, 37). These methods all use genotype information either from a set of random markers and/or from a set of selected ancestry informative markers (AIM). AIMs are defined as markers that show large allele frequency differences between ancestral populations (21, 38-40). These methods for testing for and/or adjusting for population stratification can be broadly classified into three classes: (a) genomic control (30, 41-44), (b) structured association (45-57), and (c) other (58-62).

Genomic control was one of the first methods developed to adjust for population stratification (41-44). The genomic control technique uses a set of noncandidate, random markers (sometimes called null markers) to estimate an inflation factor, \( \lambda \). \( \lambda \) is equal to 1 if there is no population stratification present. This inflation is assumed to be caused by population stratification and the genomic control method corrects the standard \( \chi^2 \) association test statistic by this factor, where the new \( \chi^2 / \lambda \) test statistic still has a \( \chi^2 \) distribution. Therefore, genomic control performs a uniform adjustment to all association tests assuming the same inflation factor. One of the main assumptions of this method is that if the study population comes from a larger population made up of a mixture of subpopulations with different disease prevalences and disease allele frequencies, then the \( \chi^2 \) association test statistic follows a noncentral \( \chi^2 \) distribution (52). If the noncentral variable is truly small, then adjusting by the estimated inflation factor \( \lambda \) is a good approximation to this distribution; however, if the noncentrality variable is truly large, then adjusting for the estimated inflation factor \( \lambda \) will not be sufficient to prevent false-positive associations and loss of statistical power (62). This method considers group-level population stratification only (as defined by racial/ethnic category) and can help to control against false-positive associations but not against false-negative associations. If AIMs are used instead of random markers, more false-positive associations will result simply because the AIMs show large population differences in allele frequencies and there will be a tendency towards overcorrection (62).

Genetic control, in general, is a relatively computationally easy method to implement and interpret.

Structured association methods use Bayesian techniques to assign individuals to clusters or subpopulation classes using information from a set of noncandidate, unlinked loci under a model of admixture (45-57). The structured association methods use a Bayesian, Monte Carlo Markov Chain approach to simultaneously estimate two pieces of information: (a) a multidimensional vector of all allele frequencies for all subpopulations at all loci and (b) a vector of populations of ancestral origin for every allele for every individual. Assumptions are made that these vectors are from separate Dirichlet distributions with different hyperparameters. These models originally assumed both linkage equilibrium and Hardy-Weinberg equilibrium but have now been modified for situations where linkage disequilibrium is present (45). Tests for association within each cluster or subpopulation class are then undertaken using these markers. This method considers both individual-level and group-level population stratification. In structured association approaches, genotype information from sets of random markers or AIMs may be used. The most commonly used implementations of structured association are the programs STRUCTURE (45, 51-53) and ADMIXMAP (46-50). These programs use similar structured association methods to estimate individual-level and/or group-level ancestry, but ADMIXMAP can also simultaneously model the association between a candidate genotype and the trait of interest allowing for the error associated with estimating ancestry to be included in the association test. However, unresolved issues with structured association techniques still exist that include deciding on the optimal clustering similarity metric, distinguishing the optimal number of ancestral clusters, and determining the biological meaning of the clusters.

The estimation of genomic ancestry at the individual or group level and the use of this information in genotype-disease association studies in place of race/ethnicity to measure stratification (63-68) can also be considered a structured association technique. The utility of using individual genetic ancestry estimates for understanding complex disease risk has recently been shown in genetic association studies of asthma (15, 16), cardiovascular disease-related phenotypes (68), insulin-related phenotypes (65), and early-onset lung cancer (66). Wilson et al. observed that frequency of risk genotypes in six drug-metabolizing genes varied by genetically defined ancestry and that self-reported race/ethnicity was an insufficient and inaccurate representation of these ancestral clusters (69).

Other techniques that can be used to correct for the effects of population stratification include principal component methods (58, 61, 62), a latent variable approach using a stratification score (59), and an approach based on molecular analysis of variance (60).
The principal components approaches use genotype data to estimate axes of variation that can be interpreted as describing continuous ancestral heterogeneity within a group of individuals (70). These axes of variation are defined as the top eigenvectors of a covariance matrix between individuals in the study population that was formed using genotype information from random markers or AIMs. Then, the association between genotypes and phenotypes can be adjusted for the association attributable to ancestry along each axis. This method is insensitive to the number of inferred axes and can be easily done on a genome-wide scale. In addition, the appropriate number of axes of variation can be formally tested. The latent variable approach (59) assigns each individual to an ancestral strata using a stratification score that is created from a latent variable using information from additional genotyped markers (random or AIMs). This latent variable is created using a generalized partial least-squares approach and it is assumed that using this latent variable to stratify the data will estimate the true association between disease and candidate genotype. Tests for association between the disease locus and candidate locus are done within each stratum. Generalized partial least squares approach is similar to principal components methods, except that it is able to model variability in both the marker data and the trait at once. This method requires fewer assumptions than genomic control, structured association, and principal component methods, can accommodate multilocus haplotypes, and is computationally simple. A final approach (60) constructs a genotype similarity matrix, from genotype information from random markers or AIMs, and then tests the relationship between any grouping factor or quantitative measure and the variability in the genotypic similarities of individuals. This approach is similar to AMOVA (71) and the Mantel-based test statistic (72) in that differences by various factors of interest between groups of individuals or population with adjustment for diversity in ancestral genetic background can be systematically tested. This method can be easily adapted to be used in multiple regression-like test settings and shows excellent power for low levels of subpopulation variation.

Because of the vast number of options now available for assessing and controlling for population stratification, care must be taken to ensure that all assumptions of the method are being met and that the method of choice is actually testing the intended hypothesis.

**AIMs and Ancestry Estimation**

Estimation of genetic ancestry can be achieved by genotyping AIMs. As defined above, AIMs are unlinked markers found throughout the genome that show large allele frequency differences (denoted \( \delta \)) between the relevant ancestral populations (21, 38-40). The two most commonly used methods for ancestry estimation from AIMs are maximum likelihood estimation (73, 74) and structured association clustering techniques as implemented in STRUCTURE (45, 51-53) and ADMIXMAP (46-50). These methods have been shown to be comparable in terms of accuracy (50, 52, 75), but their validity is dependent on the informativeness of the panel of AIMs being used as well as the availability of allele and genotype frequency data (76).

Simulation studies were first used to show that 50 to 100 AIMs are needed to accurately assign one’s individual ancestry; fewer markers (~40 AIMs) are needed when the average allele frequency difference between ancestral populations (denoted \( \delta \)) of the panel of markers is >0.6 (4, 15, 75). However, the minimal \( \delta \) needed can vary from study to study. Hence, multiple investigators have proposed information calculations on the informativeness for ancestry analyses of specific markers (77-79). Fisher’s information is the inverse of the maximum likelihood estimation of the ancestral proportion and therefore has a direct relationship to the precision of the ancestral proportion estimate (77). Rosenberg et al. (79) developed three information statistics, which produce similar results to each other and to the Fisher’s information statistic but may produce upwardly biased estimates in small samples. Other measures that have been used include Wright’s \( F_{ST} \) (80), expected heterozygosity, or the number of alleles present by subpopulation. Wright’s \( F_{ST} \) is only useful if there are two subpopulations that have familial distributions. This assumption may not be appropriate in situations of continuous gene flow, as may be acting in U.S. populations (49). The information statistics proposed by Rosenberg et al. (79) and the Fisher’s information statistic are relevant and useful for multiple reasons: (a) they both allow for multiple alleles at a locus [and thus can used for microsatellites or single nucleotide polymorphisms (SNP) so these types of markers can be compared directly for ancestry informativeness], (b) they both use information on allele frequencies within an ancestral population and the absolute differences in allele frequencies by pairs of ancestral populations, and (c) they both take into account multiple mixing ancestral populations in a single analysis (77). Therefore, using either Fisher’s information or one of new information measures proposed by Rosenberg et al. (79) is likely to provide the most useful approach to determine the choice of a panel of AIMs.

There are currently several existing AIMs panels that can be implemented in genetic association studies (Table 1). Most of these panels consist of SNPs, although some include microsatellites. The choice of markers depends on the marker’s ancestry informativeness, which depends on the value of \( \delta \) (38, 39, 81, 82). The choice can also depend on other population variables (79), such as the relative ancestral proportional contributions from each of the parental populations (77) and how many ancestral populations have mixed. A practical understanding of the history of the immigration and migration history of the study population is critical to accurately select an appropriate panel of AIMs. Knowledge of this history is also critical to establish the analytical models that require knowledge of how many and which of the ancestral parental populations should be considered for robust ancestry estimation.

Not all AIM panels are equivalent. For example, an AIMs panel assembled for Mexican Americans might be inappropriate for use in a Puerto Rican sample, because the level of African ancestry differs between these populations. Thus, estimation of ancestral proportions is highly dependent on (a) knowledge of parental populations, (b) choice of markers for ancestry estimation (that is, informativeness for ancestry analyses), (c) estimation of the parental allele frequencies, (d) method for ancestry estimation, and (e) level of...
Table 1. Published genome-wide panels of AIMs appropriate for ancestry analyses

<table>
<thead>
<tr>
<th>Type of markers</th>
<th>Population studied</th>
<th>Total no. individuals genotyped</th>
<th>Reference</th>
<th>No. AIMs</th>
<th>Web site</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs and diallelic insertion/deletions</td>
<td>European American</td>
<td>&gt;1,000</td>
<td>Shriver et al. (39) and Parra et al. (11)</td>
<td>~75-100</td>
<td>dbSNP database (<a href="http://www.ncbi.nlm.nih.gov/SNP">http://www.ncbi.nlm.nih.gov/SNP</a>), keyword: PSUANTH</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>African</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jamaican</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short tandem repeats</td>
<td>African American</td>
<td>175</td>
<td>Smith et al. (38)</td>
<td>744</td>
<td>Laboratory of Genomic Diversity (<a href="http://lgd.nci.nih.gov">http://lgd.nci.nih.gov</a>)</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>European American</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microsatellites and diallelic insertion/deletions</td>
<td>European American</td>
<td>DNA pooling used</td>
<td>Collins-Schramm et al. (81, 82)</td>
<td>151 for Mexican American and 97 for African American</td>
<td>University of California-Davis, Rowe Program (<a href="http://roweprogram.ucdavis.edu">http://roweprogram.ucdavis.edu</a>)</td>
</tr>
<tr>
<td></td>
<td>Mexican American</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amerindian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>African</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNPs</td>
<td>European American</td>
<td>&gt;300</td>
<td>Smith et al. (83)</td>
<td>3,011</td>
<td>Laboratory of Genomic Diversity (<a href="http://lgd.nci.nih.gov">http://lgd.nci.nih.gov</a>)</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Native American</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNPs</td>
<td>European American</td>
<td>&gt;500</td>
<td>Collins-Schramm et al. (84)</td>
<td>123</td>
<td>University of California-Davis, Rowe Program (<a href="http://roweprogram.ucdavis.edu">http://roweprogram.ucdavis.edu</a>)</td>
</tr>
<tr>
<td></td>
<td>Mexican American</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Japanese</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amerindian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNPs</td>
<td>European American</td>
<td>71</td>
<td>Hinds et al. (85)</td>
<td>1,586,383</td>
<td>Perlegen Genome Browser (<a href="http://www.hapmap.org/cgi-perl/gbrowse/gbrowse">http://www.hapmap.org/cgi-perl/gbrowse/gbrowse</a>)</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asian American</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNPs</td>
<td>European American</td>
<td>85</td>
<td>Miller et al. (86)</td>
<td>1,410</td>
<td>The SNP Consortium Allele Frequency Project (<a href="http://snp.cshl.org">http://snp.cshl.org</a>)</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNPs</td>
<td>African</td>
<td>269</td>
<td>Altshuler et al. and The International HapMap Consortium (87)</td>
<td>877,351 polymorphic in all three groups 75,997 monomorphic across all three groups</td>
<td>The HapMap Project (<a href="http://www.hapmap.org">http://www.hapmap.org</a>)</td>
</tr>
<tr>
<td></td>
<td>European American</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Japanese</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNPs</td>
<td>European American</td>
<td>12 worldwide population samples</td>
<td>Shriver et al. (21)</td>
<td>11,555</td>
<td>Shriver Laboratory (<a href="http://www.anthro.psu.edu/biolab/index.html">http://www.anthro.psu.edu/biolab/index.html</a>)</td>
</tr>
<tr>
<td>SNPs</td>
<td>European American</td>
<td>&gt;1,000</td>
<td>Seldin et al. (19)</td>
<td>400-800</td>
<td>University of California-Davis, Rowe Program (<a href="http://roweprogram.ucdavis.edu">http://roweprogram.ucdavis.edu</a>)</td>
</tr>
<tr>
<td>SNPs</td>
<td>European American</td>
<td>&gt;300</td>
<td>Tian et al. (88)</td>
<td>&gt;4,000</td>
<td>University of California-Davis, Rowe Program (<a href="http://roweprogram.ucdavis.edu">http://roweprogram.ucdavis.edu</a>)</td>
</tr>
</tbody>
</table>

(Continued on the following page)
population stratification in the admixed population. Applying generic AIM sets developed in one population to an ancestrally different population may be suboptimal. Therefore, we propose three principles for choosing AIMs for a specific study: (a) markers should have a \( \delta \geq 0.6 \); (b) a measure of informativeness (77, 79) for multiple possible combinations of ancestral proportions should be calculated and those markers that are informative across multiple different ancestral proportion combinations should be prioritized; and (c) knowledge of immigration/migration patterns in the region from which the study population was drawn should inform choice of ancestral parental populations and the number of ancestral parental populations.

**Summary**

Explanations for observed differences within and between populations in disease incidence and outcome are an important area of research. To maximize the potential for epidemiologic association studies to identify meaningful, reproducible genetic associations in large studies of common diseases, it is imperative that careful consideration be given to population stratification. In some situations, self-reported race/ethnicity may be sufficient to alleviate concerns about bias due to population stratification. However, in many situations, genotype-based estimates of group and/or individual ancestry using AIMs may be required to properly account for ancestry, admixture, and bias due to population stratification in association studies.

**Table 1. Published genome-wide panels of AIMs appropriate for ancestry analyses (Cont'd)**

<table>
<thead>
<tr>
<th>Type of markers</th>
<th>Population studied</th>
<th>Total no. individuals genotyped</th>
<th>Reference</th>
<th>No. AIMs</th>
<th>Web site</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs</td>
<td>West African (including Yorubans) African Americans 5 Different Amerindian populations European American Japanese Chinese</td>
<td>&gt;700 Tian et al. (89)</td>
<td>&gt;8,000</td>
<td>University of California-Davis, Rowe Program (<a href="http://roweprogram.ucdavis.edu">http://roweprogram.ucdavis.edu</a>)</td>
<td></td>
</tr>
<tr>
<td>SNPs</td>
<td>Latino</td>
<td>&gt;700 Price et al. (90)</td>
<td>&gt;4,100</td>
<td>Reich Laboratory (<a href="http://genpath.med.harvard.edu/~reich/">http://genpath.med.harvard.edu/~reich/</a>)</td>
<td></td>
</tr>
<tr>
<td>SNPs</td>
<td>European Native American (North and South America) European American 4 Amerindian populations West African Japanese Chinese European Americans 21 European and worldwide populations</td>
<td>&gt;300 Mao et al. (91)</td>
<td>&gt;2,000</td>
<td>Shriver Laboratory (<a href="http://www.anthro.psu.edu/biolab/euroaims.pc1.xls">http://www.anthro.psu.edu/biolab/euroaims.pc1.xls</a>)</td>
<td></td>
</tr>
<tr>
<td>SNPs</td>
<td>European Americans</td>
<td>297 Bauchet et al. (18)</td>
<td>1,200</td>
<td>Shriver Laboratory (<a href="http://www.anthro.psu.edu/biolab/euroaims.pc1.xls">http://www.anthro.psu.edu/biolab/euroaims.pc1.xls</a>)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;4,000 Price et al. (92)</td>
<td>300</td>
<td>Reich Laboratory (<a href="http://genpath.med.harvard.edu/~reich/">http://genpath.med.harvard.edu/~reich/</a>)</td>
<td></td>
</tr>
</tbody>
</table>

**References**

Population Stratification in Epidemiologic Studies
