Hypothesis/Commentary

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

Jill S. Barnholtz-Sloan,[1] Brian McEvoy,[2] Mark D. Shriver,[3] and Timothy R. Rebbeck[4]

[1]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, Ohio; [2]Trinity College Dublin, Dublin, Ireland; [3]Department of Anthropology, Pennsylvania State University, University Park, Pennsylvania; and [4]Center 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

Cancer Epidemiol Biomarkers Prev 2008;17(3):471–7

Received 5/30/07; revised 12/7/07; accepted 12/12/07.

Grant support: NIH grants P50-CA105641 and R01-CA08574.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Jill S. Barnholtz-Sloan, Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, 11100 Euclid Avenue, Cleveland, OH 44106-5065. Phone: 216-368-1506; Fax: 216-844-7832. E-mail: jsb42@case.edu

doi:10.1158/1055-9965.EPI-07-0491

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 frequencies 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, λ; λ 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 χ^2 association test statistic by this factor, where the new χ^2 / λ test statistic still has a χ^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 χ^2 association test statistic follows a noncentral χ^2 distribution (52). If the noncentral variable is truly small, then adjusting by the estimated inflation factor λ is a good approximation to this distribution; however, if the noncentrality variable is truly large, then adjusting for the estimated inflation factor λ 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). Genomic 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 hypervariables. 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 δ) 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 ($\sim$40 AIMs) are needed when the average allele frequency difference between ancestral populations (denoted δ) of the panel of markers is $\geq$0.6 (4, 15, 75). However, the minimal δ 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 mixed in equal contributions. 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 δ (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

Type of markers	Population studied	Total no. individuals genotyped	Reference	No. AIMs	Web site
SNPs and diallelic insertion/deletions	European American African American Hispanic African Jamaican	>1,000	Shriver et al. (39) and Parra et al. (11)	~75-100	dbSNP database (http://www.ncbi.nlm.nih.gov/SNP), keyword: PSUANTH
Short tandem repeats	African American Hispanic European American Asian	175	Smith et al. (38)	744	Laboratory of Genomic Diversity (http://lgd.nci.nih.gov)
Microsatellites and diallelic insertion/deletions	European American Mexican American African American	DNA pooling used	Collins-Schramm et al. (81, 82)	151 for Mexican American and 97 for African American	University of California-Davis, Rowe Program (http://roweprogram.ucdavis.edu)
	Amerindian African				University of California-Santa Cruz Human Genome Project Center (http://genome.ucsc.edu)
SNPs	European American African American African Chinese Amerindian	>300	Smith et al. (83)	3,011	Laboratory of Genomic Diversity (http://lgd.nci.nih.gov)
SNPs	European American Mexican American Japanese	>500	Collins-Schramm et al. (84)	123	University of California-Davis, Rowe Program (http://roweprogram.ucdavis.edu)
	Amerindian				The SNP Consortium Allele Frequency Project (http://snp.cshl.org)
SNPs	European American African American Asian American	71	Hinds et al. (85)	1,586,383	Perlegen Genome Browser (http://www.hapmap.org/cgi-perl/gbrowse/gbrowse) Haplotype data (http://research.calit2.net/hap/wgha)
SNPs	European American African American Asian	85	Miller et al. (86)	1,410	The SNP Consortium Allele Frequency Project (http://snp.cshl.org)
SNPs	African European American Chinese Japanese	269	Altshuler et al. and The International HapMap Consortium (87)	877,351 polymorphic in all three groups 75,997 monomorphic across all three groups	The HapMap Project (http://www.hapmap.org)
SNPs	12 worldwide population samples	203	Shriver et al. (21)	11,555	Shriver Laboratory (http://www.anthro.psu.edu/biolab/index.html)
SNPs	6 European populations European American Ashkenazi Jewish Asian American African American Amerindians	>1,000	Seldin et al. (19)	400-800	University of California-Davis, Rowe Program (http://roweprogram.ucdavis.edu)
SNPs	European American	>300	Tian et al. (88)	>4,000	University of California-Davis, Rowe Program (http://roweprogram.ucdavis.edu)
	Centre d'Etude du Polymorphisme Humain Europeans				

(*Continued on the following page*)

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

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

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.

References

1. Linneaus C. Systemae naturae (The system of nature). Stockholm (Sweden): Laurentii Salvii, Holmiae; 1758.
2. Jacobson MF, Whiteness of a different color: European immigrants and the alchemy of race. Cambridge (MA): Harvard University Press; 1998.
3. Snowden FM, Before color prejudice: the ancient view of blacks. Cambridge (MA): Harvard University Press; 1983.
4. Risch N, Burchard E, Ziv E, Tang H, Categorization of humans in biomedical research: genes, race and disease. Genome Biol 2002;3:1–12.
5. Tang H, Quertermous T, Rodriguez B, et al. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet 2005;76:268–75.
6. Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic structure of human populations. Science 2002;298:2381–5.
7. Foster MW, Sharp RR. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. Genome Res 2002;12:844–50.
8. Williams DR. Race and health: basic questions, emerging directions. Ann Epidemiol 1997;7:322–33.
9. Helgadottir A, Manolescu A, Helgason A, et al. A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. Nat Genet 2006;38:68–74.
10. Parra EJ, Kittles RA, Argyropoulos G, et al. Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. Am J Phys Anthropol 2001;114:18–29.
11. Parra EJ, Marcini A, Akey J, et al. Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 1998;63:1839–51.
12. Hanis CL, Hewett-Emmett D, Bertin TK, Schull WJ. Origins of U.S. Hispanics. Implications for diabetes. Diabetes Care 1991;14:618–27.
13. U.S. Census 2000: The Hispanic population census 2000 brief; 2001.
14. Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet 2004;36:388–93. Epub 2004 Mar 28.

15. Choudhry S, Coyle NE, Tang H, et al. Population stratification confounds genetic association studies among Latinos. Hum Genet 2006;118:652–64.
16. Salari K, Choudhry S, Tang H, et al. Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. Genet Epidemiol 2005;29:76–86.
17. Hanis CL, Chakraborty R, Ferrell RE, Schull WJ. Individual admixture estimates: disease associations and individual risk of diabetes and gallbladder disease among Mexican-Americans in Starr County, Texas. Am J Phys Anthropol 1986;70:433–41.
18. Bauchet M, McEvoy B, Pearson LN, et al. Measuring European population stratification with microarray genotype data. Am J Hum Genet 2007;80:948–56.
19. Seldin MF, Shigeta R, Villoslada P, et al. European population substructure: clustering of northern and southern populations. PLoS Genet 2006;2:e143.
20. Campbell CD, Ogburn EL, Lunetta KL, et al. Demonstrating stratification in a European American population. Nat Genet 2005; 37:868–72.
21. Shriver MD, Mei R, Parra EJ, et al. Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Hum Genomics 2005;2:81–9.
22. 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;37:90–5.
23. Chakraborty R, Weiss KM. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci U S A 1988;85:9119–23.
24. Deng HW. Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. Genetics 2001;159:1319–23.
25. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet 2004;36:512–7. Epub 2004 Mar 28.
26. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 2000;92:1151–8.
27. Wacholder S, Rothman N, Caporaso N. Counterpoint: Bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev 2002;11:513–20.
28. Wang Y, Localio R, Rebbeck TR. Evaluating bias due to population stratification in case-control association studies of admixed populations. Genet Epidemiol 2004;27:14–20.
29. Wang Y, Localio R, Rebbeck TR. Evaluating bias due to population stratification in epidemiologic studies of gene-gene or gene-environment interactions. Cancer Epidemiol Biomarkers Prev 2006;15:124–32.
30. Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 2001; 20:4–16.
31. Dean M. Approaches to identify genes for complex human diseases: lessons from Mendelian disorders. Hum Mutat 2003;22:261–74.
32. Burnett MS, Strain KJ, Lesnick TG, de Andrade M, Rocca WA, Maraganore DM. Reliability of self-reported ancestry among siblings: implications for genetic association studies. Am J Epidemiol 2006; 163:486–92.
33. Ziv E, Burchard EG. Human population structure and genetic association studies. Pharmacogenomics 2003;4:431–41.
34. Allison DB. Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet 1997;60:676–90.
35. Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 1996;59: 983–9.
36. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–16.
37. Thomas DC, Witte JS. Point: Population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 2002;11:505–12.
38. Smith MW, Lautenberger JA, Shin HD, et al. Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. Am J Hum Genet 2001;69:1080–94.
39. Shriver MD, Smith MW, Jin L, et al. Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 1997;60: 957–64.
40. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. Genome Res 2002;12:1805–14.
41. Bacanu SA, Devlin B, Roeder K. The power of genomic control. Am J Hum Genet 2000;66:1933–44.
42. Bacanu SA, Devlin B, Roeder K. Association studies for quantitative traits in structured populations. Genet Epidemiol 2002;22:78–93.
43. Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999;55:997–1004.
44. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 2001;60:155–66.
45. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 2003;164:1567–87.
46. Hoggart CJ, Parra EJ, Shriver MD, et al. Control of confounding of genetic associations in stratified populations. Am J Hum Genet 2003; 72:1492–504.
47. Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM. Design and analysis of admixture mapping studies. Am J Hum Genet 2004;74:965–78. Epub 2004 Apr 14.
48. McKeigue PM. Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. Am J Hum Genet 1997;60:188–96.
49. McKeigue PM. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am J Hum Genet 1998;63: 241–51.
50. McKeigue PM, Carpenter JR, Parra EJ, Shriver MD. Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. Ann Hum Genet 2000;64:171–86.
51. Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. Theor Popul Biol 2001;60:227–37.
52. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65:220–8.
53. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155:945–59.
54. 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: 466–77.
55. Zhang S, Zhao H. Quantitative similarity-based association tests using population samples. Am J Hum Genet 2001;69:601–14.
56. Zhang S, Zhu X, Zhao H. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genet Epidemiol 2003;24:44–56.
57. Zhu X, Zhang S, Zhao H, Cooper RS. Association mapping, using a mixture model for complex traits. Genet Epidemiol 2002; 23:181–96.
58. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet 2006;2:e190.
59. Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case-control studies. Am J Hum Genet 2007;80:921–30.
60. Nievergelt CM, Libiger O, Schork NJ. Generalized analysis of molecular variance. PLoS Genet 2007;3:e51.
61. 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;38:904–9.
62. Chen HS, Zhu X, Zhao H, Zhang S. Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Ann Hum Genet 2003;67:250–64.
63. Williams RC, Long JC, Hanson RL, Sievers ML, Knowler WC. Individual estimates of European genetic admixture associated with lower body-mass index, plasma glucose, and prevalence of type 2 diabetes in Pima Indians. Am J Hum Genet 2000;66:527–38.
64. Fernandez JR, Shriver MD, Beasley TM, et al. Association of African genetic admixture with resting metabolic rate and obesity among women. Obes Res 2003;11:904–11.
65. Gower BA, Fernandez JR, Beasley TM, Shriver MD, Goran MI. Using genetic admixture to explain racial differences in insulin-related phenotypes. Diabetes 2003;52:1047–51.
66. Barnholtz-Sloan JS, Chakraborty R, Sellers TA, Schwartz AG. Examining population stratification via individual ancestry estimates versus self-reported race. Cancer Epidemiol Biomarkers Prev 2005; 14:1545–51.
67. Ziv E, John EM, Choudhry S, et al. Genetic ancestry and risk factors for breast cancer among Latinas in the San Francisco Bay Area. Cancer Epidemiol Biomarkers Prev 2006;15:1878–85.
68. Reiner AP, Ziv E, Lind DL, et al. Population structure, admixture, and aging-related phenotypes in African American adults: the Cardiovascular Health Study. Am J Hum Genet 2005;76:463–77.
69. Wilson JF, Weale ME, Smith AC, et al. Population genetic structure of variable drug response. Nat Genet 2001;29:265–9.

70. 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:e70.
71. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 1992;131: 479–91.
72. Smouse PE, Long JC. Matrix correlation analysis in anthropology and genetics. Am J Phys Anthropol 1992;35:187–213.
73. Chakraborty R, Kamboh MI, Nwankwo M, Ferrell RE. Caucasian genes in American blacks: new data. Am J Hum Genet 1992;50: 145–55.
74. Chakraborty R. Gene admixture in human populations: models and predictions. Yearbook Phys Anthropol 1986;29:1–43.
75. Tsai HJ, Choudhry S, Naqvi M, Rodriguez-Cintron W, Burchard EG, Ziv E. Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations. Hum Genet 2005;118:424–33.
76. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol 2005;28:289–301.
77. Pfaff CL, Barnholtz-Sloan J, Wagner JK, Long JC. Information on ancestry from genetic markers. Genet Epidemiol 2004;26:305–15.
78. Barnholtz-Sloan JS, Pfaff CL, Chakraborty R, Long JC. Informativeness of the CODIS STR loci for admixture analysis. J Forensic Sci 2005;50:1322–6.
79. Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 2003;73:6.
80. Wright S. The genetic structure of populations. Ann Eugen 1951;15: 323–54.
81. Collins-Schramm HE, Kittles RA, Operario DJ, et al. Markers that discriminate between European and African ancestry show limited variation within Africa. Hum Genet 2002;111:566–9.
82. Collins-Schramm HE, Phillips CM, Operario DJ, et al. Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. Am J Hum Genet 2002;70:737–50.
83. Smith MW, Patterson N, Lautenberger JA, et al. A high-density admixture map for disease gene discovery in African Americans. Am J Hum Genet 2004;74:1001–13. Epub 2004 Apr 14.
84. Collins-Schramm HE, Chima B, Morii T, et al. Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. Hum Genet 2004;114:263–71. Epub 2003 Nov 20.
85. Hinds DA, Stuve LL, Nilsen GB, et al. Whole-genome patterns of common DNA variation in three human populations. Science 2005; 307:1072–9.
86. Miller RD, Phillips MS, Jo I, et al. High-density single-nucleotide polymorphism maps of the human genome. Genomics 2005;86:117–26.
87. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. A haplotype map of the human genome. Nature 2005; 437:1299–320.
88. 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;79:640–9.
89. Tain C, Hinds DA, Shigeta R, et al. A genomewide single-nucleotide polymorphism panel for Mexican American admixture mapping. Am J Hum Genet 2007;80:1014–23.
90. Price AL, Patterson N, Yu F, et al. A genomewide admixture map for Latino populations. Am J Hum Genet 2007;80:1024–36.
91. Mao X, Bingham AW, Meui R, et al. A genomewide admixture mapping panel for Hispanic/Latino populations. Am J Hum Genet 2007;80:1171–8.
92. Price AL, Butler J, Patterson N, et al. Discerning the ancestry of European Americans in genetic association studies. PLOS Genet 2008;4:9–17.

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

Jill S. Barnholtz-Sloan, Brian McEvoy, Mark D. Shriver, et al.

Cancer Epidemiol Biomarkers Prev 2008;17:471-477.