Examining Population Stratification via Individual Ancestry Estimates versus Self-Reported Race

Jill S. Barnholtz-Sloan,[1,2] Ranajit Chakraborty,[3] Thomas A. Sellers,[1] and Ann G. Schwartz[4]

[1]Cancer Prevention and Control Program, H. Lee Moffitt Cancer Center and Research Institute; [2]Department of Interdisciplinary Oncology, University of South Florida College of Medicine, Tampa, Florida; [3]Center for Genome Research, Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio; and [4]Population Studies and Prevention Program, Karmanos Cancer Institute and Department of Internal Medicine, Wayne State University School of Medicine, Detroit, Michigan

Abstract

Population stratification has the potential to affect the results of genetic marker studies. Estimating individual ancestry provides a continuous measure to assess population structure in case-control studies of complex disease, instead of using self-reported racial groups. We estimate individual ancestry using the Federal Bureau of Investigation CODIS Core short tandem repeat set of 13 loci using two different analysis methods in a case-control study of early-onset lung cancer. Individual ancestry proportions were estimated for ''European'' and ''West African'' groups using published allele frequencies. The majority of Caucasian, non-Hispanics had >50% European ancestry, whereas the majority of African Americans had <20% European ancestry, regardless of ancestry estimation method, although significant overlap by self-reported race and ancestry also existed. When we further investigated the effect of ancestry and self-reported race on the frequency of a lung cancer risk genotype, we found that the frequency of the *GSTM1* null genotype varies by individual European ancestry and case-control status within self-reported race (particularly for African Americans). Genetic risk models showed that adjusting for individual European ancestry provided a better fit to the data compared with the model with no group adjustment or adjustment for self-reported race. This study suggests that significant population substructure differences exist that self-reported race alone does not capture and that individual ancestry may be confounded with disease status and/or a candidate gene risk genotype. (Cancer Epidemiol Biomarkers Prev 2005;14(6):1545–51)

Introduction

Most common complex phenotypes, such as cancer, show varying incidence rates, course of disease, and genetic susceptibility by race and ancestry (1, 2). For example, the world age-adjusted incidence rate of prostate cancer in Nigeria, with very little European admixture, is 23.3 per 100,000 men, whereas in Sweden, a relatively isolated population, the incidence rate is 90.9 per 100,000 men (3). For comparison, in the United States where individuals are mixtures of ancestral populations, the age-adjusted incidence rate of prostate cancer is 274.3 per 100,000 African American men and is 171.2 per 100,000 in Caucasian men (4). Although not adjusted for smoking status, sex-specific, age-adjusted lung cancer, incidence rates also differ significantly around the world. Specifically for males, the rate per 100,000 is 1.1 in Nigeria, 21.1 in Sweden, 120.7 for African Americans in the United States, and 82.3 for Caucasians in the United States (3, 4).

Studies of disease risk associated with candidate susceptibility genes are potentially vulnerable to bias due to population stratification. In order for important bias from population stratification to exist, the following must be true: (a) the frequency of the genotype of interest varies substantially by race and ancestry, (b) the disease rate varies substantially by race and ancestry, and (c) the disease rates and genotype frequencies vary together which occurs when the genotype is related to a true risk factor or is the true risk factor with a high attributable risk (5). Methods have therefore been developed to assess population stratification in case-control studies (6-12). Some of these methods involve using DNA markers to estimate genetic ancestry at the individual level, thereby allowing studies of its association with various complex traits (9, 13-16). Ancestry-informative markers generally show large allele frequency differences between ancestral populations (7).

Estimating individual ancestry requires genotyping an additional set of DNA markers for each individual in the study population. Whether this is necessary, or whether doing stratified analyses by self-reported race is adequate to control for population stratification, is unresolved (5). For example, it has been found that commonly used racial labels were insufficient and inaccurate representations of ancestral clusters and that genotype profiles, defined by the distribution of variants in drug-metabolizing genes (*CYP1A2, CYP2C19, CYP2D6, NAT2, GSTM1*, and *DIA4*), differed significantly among ancestral clusters (17). Individuals of mixed race are becoming increasingly more common in the United States (18). Most study subjects simply cannot report what percent of their genome originated from Europe, West Africa, or Native American ancestry via questionnaire. Populations in the United States are generally formed from more recent admixture, which causes interindividual differences in genetic ancestry to become more pronounced (19, 20). The West African contribution to ancestry for African Americans in the United States is on average 80%, but it can range from 20% to 100% and ~30% of United States Caucasian, non-Hispanics have >90% European ancestry (7, 21). African Americans also have significant admixture from European and Native American ancestral populations, whereas United States Caucasian, non-Hispanics have significant admixture from West African and Native American populations (22).

To assess the utility of using individual genetic ancestry estimates to better understand population stratification in a standard epidemiologic case-control study, we genotyped early-onset lung cancer cases (i.e., diagnosed before age 50)

Received 11/11/04; revised 2/24/05; accepted 3/28/05.

Grant support: National Cancer Institute grants K07 CA91849 (J.S. Barnholtz-Sloan), R01 CA60691 (A.G. Schwartz), and N01 PC35145 (A.G. Schwartz).

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 Barnholtz-Sloan, Cancer Prevention and Control Program, H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, Tampa, FL 33612. Phone: 813-745-6531; Fax: 813-632-1334. E-mail: barnhojs@moffitt.usf.edu

and population-based controls for a panel of ancestry informative markers. We then estimated individual ancestry from these markers using two different methods and used these estimates to assess population stratification within this case-control sample. We used the glutathione *S*-transferase μ (*GSTM1*) locus, a candidate gene for lung cancer risk (23-27), as an example of how using individual ancestry estimates versus self-reported race can affect estimates of disease risk associated with genotype in groups of individuals.

Materials and Methods

Study Population. Cases with early-onset lung cancer were identified between September 15, 1990 and November 30, 2003 through the metropolitan Detroit Cancer Surveillance System, a participant in the National Cancer Institute's Surveillance, Epidemiology and End Results program. This study was approved by the local institutional review board and all subjects provided written informed consent. Case eligibility criteria included an incident primary, malignant cancer of the lung or bronchus, <50 years of age at diagnosis, and a resident of the Detroit tri-county area (Wayne, Macomb, and Oakland) at the time of diagnosis. Population-based controls were ascertained concurrently with the cases via random digit dialing and were frequency matched to cases by race, sex, 5-year age group, and county of residence. Over 98% of the eligible, successfully contacted controls agreed to participate. Seven hundred forty-six (N_{total} = 746) cases and population-based controls with available extracted normal DNA via a blood or a cheek swab sample and who self-reported their race as Caucasian, non-Hispanic, or African American were used in this analysis (n_{cases} = 252 and $n_{controls}$ = 494).

Genotyping. Each individual was genotyped for the lung cancer candidate gene, *GSTM1* (null or present; refs. 28, 29). In addition, all individuals were genotyped for the U.S. Federal Bureau of Investigation CODIS Core short tandem repeat (STR) set of 13 loci for analysis of individual ancestry (30). A list of these 13 loci with the chromosomal location and the number of alleles are shown in Table 1. The 13 CODIS loci were tested for Hardy-Weinberg equilibrium and linkage disequilibrium and were found to not violate Hardy-Weinberg equilibrium within loci or show linkage disequilibrium between loci (data not shown; P = 0.08-0.75 for tests of Hardy-Weinberg equilibrium and linkage disequilibrium). The average of German and Polish parental frequencies were used to represent *European* (31, 32) and the average of Rwandan and Nigerian parental frequencies to represent *West African* (32, 33), for the maximum likelihood estimations (MLE). Detroit, MI was originally settled by the Polish and Germans, with African ancestral populations settling in over time (34), making these parental populations appropriate for this study population for estimation of individual ancestry.

Table 1. Composite deltas for 13 STR loci between ancestral groups

CODIS loci name	Chromosomal location (no. alleles)	Overall composite δ (δ_c)*, European versus West African
CSF1PO	5q33.3-34 (14)	0.18
D13S317	13q22-q31 (12)	0.29
D16S539	16q22-24 (10)	0.19
D18S51	18q21.3 (20)	0.31
D21S11	21q21.1 (34)	0.25
D3S1358	3p21 (12)	0.16
D5S818	5q21-q31 (12)	0.17
D7S820	7q (18)	0.14
D8S1179	8q24.1-24.2 (11)	0.27
FGA	4q28 (31)	0.30
THO1	11p15-15.5 (10)	0.31
TPOX	2p23-2pter (10)	0.26
vWA	12p12-pter (12)	0.15

*δ_c is the composite δ calculated as half the sum across all loci pairs of the allele frequencies in two different populations when there are multiple alleles at a locus; European = average Polish and German, West African = average Nigerian and Rwandan.

Individual Ancestry Estimation. We estimated individual ancestry using two methods: (a) MLE (16, 22) and (b) Bayesian clustering techniques as implemented in the STRUCTURE 2.1 program (9, 35). For the first method, using the contemporary published allele frequencies mentioned above as the parental frequencies, the individual maximum likelihood ancestral proportions for the two parental populations, European and West African, were calculated for all early-onset lung cancer cases and population-based controls, using each individual's CODIS Core STR loci genotypes.

Considering a population that was formed by admixture between two genetically distinct ancestral populations (this can be easily extended to any number of populations), the frequency of the kth allele at the gth locus in the admixed population, A, is

$$p_{gAk} = m_1 p_{g1k} + m_2 p_{g2k} = p_{g2k} + m_1 \delta_{g1k} \quad \text{(A)}$$

where the two ancestral contributions, m_j, i = 1 and 2 sum to 1.0, and the δ coefficients (i.e., allele frequencies differences between parental populations) are defined as $\delta_{g1k} = p_{g1k} - p_{g2k}$. The constraint $\sum_i m_i = 1.0$ ensures that the outcome of analysis is unaffected by the way parental populations are numbered, or which population is subtracted from the others. The log-likelihood function for an individual is,

$$\ln L = \sum_g \sum_k \ln(p_{gAk}) = \sum_g \sum_k \ln(p_{g2k} + m_1 \delta_{g1k}) \quad \text{(B)}$$

Equation B applies to all alleles at all loci. Estimates of individual admixture were obtained by treating each individual as a sample of size one because the same likelihood applies to samples of any size.

Maximum likelihood estimates for the ancestral contributions were obtained from the log-likelihood function by setting the partial derivatives, with respect to m_j,

$$\frac{\partial \ln L}{\partial m_j} = \sum_g \sum_k \frac{\delta_{gik}}{p_{gAk}} \quad \text{(C)}$$

equal to zero, and solving simultaneously for $\hat{m}_1$, using the Newton-Rhapson method (36). The MLE of $\hat{m}_2$ equals $1 - \hat{m}_1$ (37).

For the second method, individual ancestry for two ''clusters'' (i.e., ancestral European and West African populations), using each individual's CODIS Core STR loci genotypes was calculated. The STRUCTURE method assigns each individual to clusters by calculating a posterior probability that an individual belongs to a cluster, given the observed marker genotypes (i.e., the CODIS STR genotypes). The number of clusters can either be inferred by the program or can be given as an initial variable. In this case, we set the number of clusters to two to compare with the MLE estimates. In the presence of admixture and hence correlated allele frequencies, the STRUCTURE method also estimates the proportion of an individual's genome that derives from each of the two cluster subpopulations.

Statistical Analysis. Composite delta (δ_c) was calculated for each of the 13 CODIS loci for the European and West African ancestral combination. Composite δ was calculated as half the

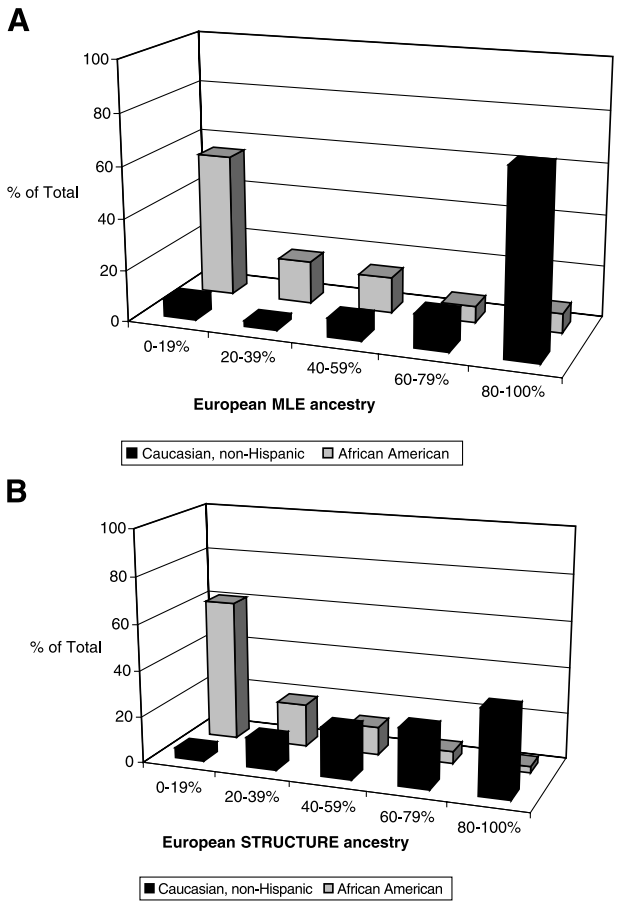

Figure 1. A. Histogram of European individual MLE ancestry by self-reported race (χ^2, $P < 0.0001$). **B.** Histogram of European individual STRUCTURE ancestry (cluster 2) by self-reported race (χ^2, $P < 0.0001$).

sum across all loci pairs of the allele frequencies in two different populations when there are multiple alleles at a locus. Spearman correlation coefficients were calculated for MLE individual ancestry compared with STRUCTURE individual ancestry. Only European ancestry estimates were used for further analyses because the West African estimates were equal to one minus the European estimates. Median European MLE and STRUCTURE ancestry were compared within self-reported racial group by case-control status using a *t* test; the frequency of the *GSTM1* null genotype was also compared within self-reported racial group by case-control status using a χ^2 test. Histograms of individual European ancestry for both MLE and STRUCTURE estimates by self-reported race were generated. To assess differences in ancestry between cases and controls related to the *GSTM1* null risk genotype, histograms were generated to compare the frequency of the risk genotype by case-control status within European MLE or STRUCTURE ancestral group, stratified by self-reported race. Unconditional logistic regression models were used to estimate odds ratios and 95% confidence intervals to measure the association between early-onset lung cancer and the *GSTM1* null genotype. Potential confounders, including gender, age at diagnosis for cases or age at interview for controls (continuous), family history of lung cancer, and pack-years of smoking (continuous) were included in all models. To test the effects of self-reported race and individual ancestry on genetic risk, models were additionally adjusted for self-reported race or individual European MLE or STRUCTURE ancestry and were compared with the general model using the likelihood ratio test. Additionally, models were compared using the Akaike Information Criterion that adjusts the −2 log-likelihood for the model by twice the number of estimated variables in the model (38). All statistical analyses were done using SAS version 9.1 (39).

Results

A total of 555 self-reported Caucasian, non-Hispanics and a total of 191 self-reported African Americans were available for analysis. The 13 CODIS STR loci allele frequencies varied between the ancestral groups used in this study and were also highly multiallelic markers (Table 1). The δ_c values for the majority of the CODIS loci were >0.2 making them appropriate for ancestry estimation analysis (40, 41). Because the ancestral allele frequencies were used in the MLE, it was clear which of the two MLE estimates correlated with which ancestral group; however, this correlation was less clear with the STRUCTURE results. Spearman correlation coefficients showed that the European individual MLE ancestry estimates were highly positively correlated (+0.80) with the cluster 2 estimates from STRUCTURE, whereas the West African individual MLE ancestry estimates were highly positively correlated (+0.80) with the cluster 1 estimates from STRUCTURE. Therefore, we denoted the cluster 2 STRUCTURE estimates as European and the cluster 1 STRUCTURE estimates as West African, to compare with the MLE results in subsequent analyses.

There were no significant differences in median individual European ancestry estimates within self-reported racial group by case-control group or *GSTM1* null genotype frequency (Table 2). However, the distribution of European ancestry values was significantly different by self-reported race, whether using MLE or STRUCTURE estimates (Fig. 1A and B). The *GSTM1* null genotype frequency in Caucasian, non-Hispanic controls was 48.3%, whereas in African American controls it was 28.8% (Table 2).

To further investigate the effects of self-reported race or individual European ancestry on case-control status and the *GSTM1* ''risk'' genotype (i.e., the *GSTM1* null genotype), histograms of the *GSTM1* null genotype by case-control status were generated by ancestry, stratified by self-reported

Table 2. Median European MLE and STRUCTURE ancestry estimates and percentage of *GSTM1* null genotypes by self-reported race and case-control status

	Caucasian, non-Hispanic			African American		
	Cases (n = 192)	Controls (n = 363)	P*	Cases (n = 60)	Controls (n = 131)	P*
Median European MLE	0.99	1.0	0.65	0.20	0.16	0.69
Median European STRUCTURE (cluster 2)	0.65	0.71	0.31	0.15	0.13	0.71
GSTM1 null (column %)	43.6	48.3	0.29	27.1	28.8	0.81

*P represents a test for differences between cases and controls within self-reported racial group (t test for median ancestry values; χ^2 test for *GSTM1* null genotype).

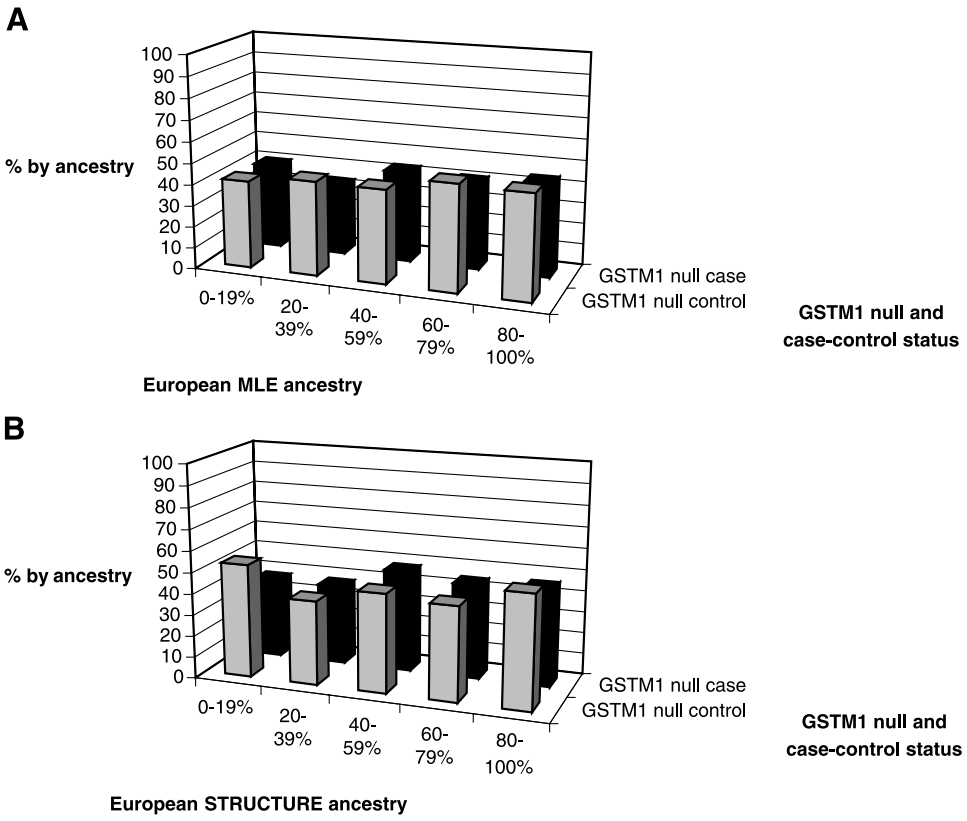

Figure 2. A. *GSTM1* null genotype by case-control status within European MLE ancestry group for Caucasian, non-Hispanics only. **B.** *GSTM1* null genotype by case-control status within European STRUCTURE ancestry group for Caucasian, non-Hispanics only.

race (Figs. 2 and 3). Although the ancestral distributions differed by estimation method, the distribution of the *GSTM1* null genotype showed similar patterns within self-reported racial group. Among individuals who self-reported as African American, the *GSTM1* null genotype frequency showed greater variability by ancestry and case-control status, regardless of ancestry estimation method, compared with Caucasian, non-Hispanics. Risk of early-onset lung cancer associated with the *GSTM1* genotype, although not significant, increased when adjusting for individual European ancestry compared with the model adjusting by self-reported race. Adjusting for self-reported race did not significantly affect the risk estimate, whereas adjusting for individual ancestry did (LRT P for race adjusted model = 0.74; LRT P for individual ancestry adjusted models = 0.001 or <0.0001; Table 3). Using the Akaike Information Criterion value to compare genetic risk models, the models adjusted for individual ancestry (MLE or STRUCTURE) clearly did better than the model adjusted for self-reported race. This information provides evidence that individual ancestry may confound disease/candidate gene associations and provides a better measure of ancestral background than self-reported race.

Discussion

In epidemiologic studies, racial differences are commonly investigated by doing analyses stratified by self-reported race. Self-reported race, however, is not always a reliable measure of one's ancestral make-up. To investigate the relationship between self-reported race and ancestry, we used genetic markers to estimate individual ancestry and population structure. Using study participants from a case-control study of early-onset lung cancer, we also examined the effects of this population structure on the distribution of the risk genotype for a lung cancer candidate gene. Individual European ancestry did not correlate completely with self-reported race in a Metropolitan Detroit population. Moreover, there was significant overlap by individual ancestry between the Caucasian, non-Hispanics and African Americans in this study sample. We observed that within self-reported racial group, the frequency of a lung cancer susceptibility genotype varied by European ancestry and case-control status. Models adjusting for individual European ancestry, regardless of the estimation method, better explained genetic risk associated with early-onset lung cancer risk, compared with a model adjusting for self-reported race. Given that incidence rates of early-onset lung cancer vary worldwide (4, 42), the distribution of the risk genotype in cases and controls varied by ancestry within self-reported racial group, and estimates of risk of disease associated with the risk genotype were significantly affected when adjusting by ancestry and not by self-reported race, we conclude that population stratification could be a significant issue in this Detroit population.

Genetic ancestry estimation is not commonly used in studies of complex diseases, because of the difficulty and expense of genotyping additional markers. However, many studies use race as an eligibility criterion. Genetic ancestry proportions seem to not only vary between groups of individuals who would self-identify to the same racial group but also among individuals within a group (22). Assortative mating, patterns of linkage and linkage disequilibrium among loci, and random genetic drift can all contribute to variability in ancestry among individuals (43). Allele frequencies have been shown to vary substantially across populations that have mixed ancestry from different continents (44) and within the same continent (45). Even if common variants are shared among racial groups, the frequencies can often differ substantially (46). Although the error caused by population stratification seems stronger for rare variants compared with common ones, the greatest bias and type I error caused by population stratification is when there is a hidden subpopulation of at least 50% in the study

population of interest and the allele frequencies differ by at least 25% (47). Therefore, it has been recommended that information on ethnic origin be collected in the greatest detail possible (48).

Few studies have analyzed empirical (i.e., nonsimulated data) data to compare self-reported race with individual ancestry estimates to assess population substructure. A study by Wacholder et al. (5) examined *NAT2* and incidence of bladder and breast cancers in relation to ancestry. They concluded that genetic markers of ancestry were unlikely to create a better proxy than self-reported race for cultural practices that could strongly affect cancer risk. In simulation studies, it has been shown that errors that occur from using self-reported race instead of genetic ancestry would be more problematic in large studies searching for susceptibility loci with small effects (19), as the adverse effects of this stratification seem to increase with increasing sample size (49, 50). Although not a case-control design, a recent study by Wilson et al., observed that frequency of risk genotypes in six drug metabolizing genes, including *GSTM1*, varied by ancestral group and that self-reported race was an insufficient and inaccurate representation of these ancestral clusters (17). The conclusions from the study by Wilson et al. and from our study suggest that using individual ancestry information could enhance the validity of epidemiologic studies and improve precision of estimated effects.

The present study suffers from limitations that are common to all studies involving ancestry estimation. The precise estimation of the ancestral proportions is highly dependent on four factors: (a) choice of parental populations, (b) choice of markers for ancestry estimation, (c) precise estimation of the parental allele frequencies, and (d) choice of method for ancestry estimation. Studies show that human populations worldwide can be subdivided based on parental/ancestral population combinations from five continents: Africa, Europe and the Middle East, Asia, Pacific Islands, and America (Native American; ref. 51). Groups of Caucasian and African American individuals in the United States today, like those used in this study, have been shown to have a combination of parental/ancestral genes from West Africa and Europe (7, 16, 21). We chose to estimate these two ancestral proportions for each individual. However, to choose proper exact parental populations, with available allele frequencies for the ancestry markers, a well-described history of the immigration and migration of the study population is needed and is not always readily available. The settlement history of Detroit is well described and is available through the Center for Michigan Studies (34). Detroit was first settled by the Polish followed by the Germans and still has a large population of individuals who identify themselves as having Polish or German ancestry (34, 52). It is estimated that 20% to 30% of African Americans in the United States originated from Nigeria and it is believed to be the most homogeneous group in Western Africa (53), making it a rational choice for the estimation of African ancestry. Rwandans were also used because this country is in the most populated area in sub-Saharan Africa having the same linguistic affiliation as many other groups in Africa, indicating recent mixture of this group with other groups in Africa (33).

The choice of markers for ancestry estimation depends on the marker's informativeness for ancestry, which has generally been thought to depend solely on allele frequency differences between parental/ancestral populations or δ (7, 40, 41, 54). However, recent studies show that informativeness for ancestry can also depend on other population genetic events (55), such as which population is the contributor of genes and which population is the acceptor of genes (56). Because there currently is no ''standard'' set of markers available for ancestry estimation, we chose to use the Federal Bureau of Investigation CODIS Core set of 13 STR loci. This set of markers is readily available in an easy to use, reasonably priced laboratory kit. These markers show considerable allele frequency variation among racial and ancestral groups from around the world (57-59). In particular, for the two ancestral

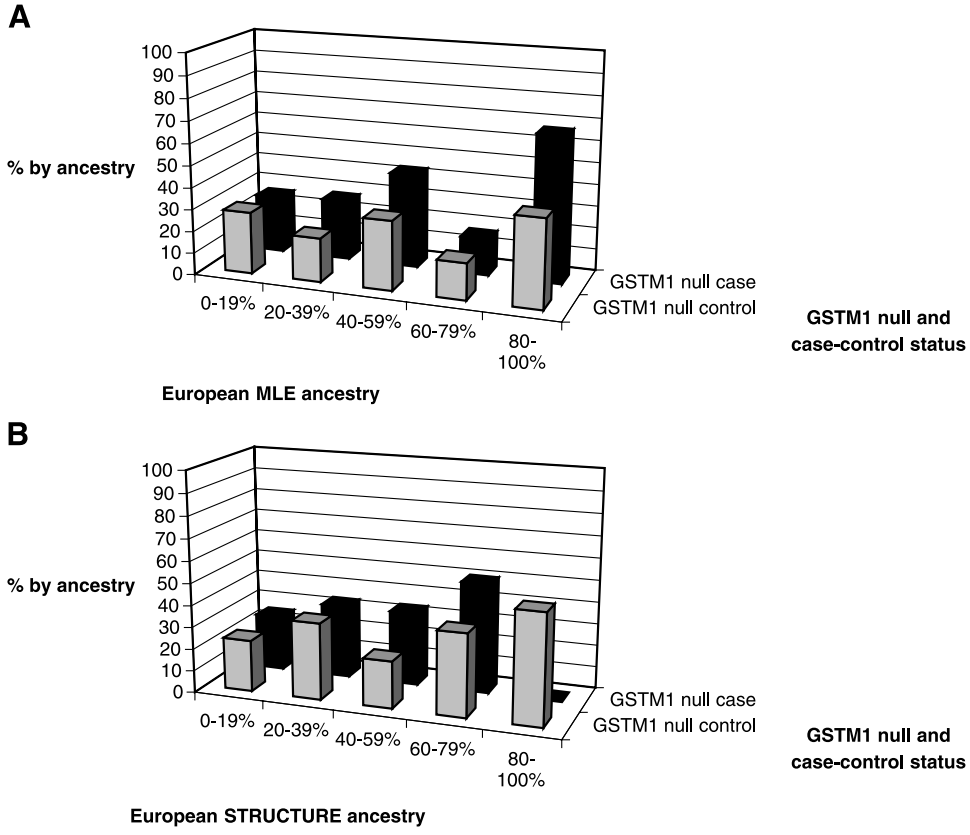

Figure 3. A. *GSTM1* null genotype by case-control status within European MLE ancestry group for African Americans only. **B.** *GSTM1* null genotype by case-control status within European STRUCTURE ancestry group for African Americans only.

Table 3. Logistic regression models of early-onset lung cancer risk for the *GSTM1* null genotype, comparing adjustments for self-reported race versus individual ancestry

Model	Odds ratio (95% confidence interval) for *GSTM1* null genotype	−2 log-likelihood	No. variables	LRT χ^2 (P)	AIC
Base*	1.11 (0.77-1.61)	732.58	5	—	742.58
Base* + self-reported race	1.12 (0.77-1.64)	732.47	6	0.11 (0.74)	744.47
Base + MLE†	1.13 (0.77-1.65)	722.45	6	10.13 (0.001)	734.45
Base + STRUCTURE‡	1.26 (0.86-1.84)	704.52	6	28.06 (<0.0001)	716.52

Abbreviations: LRT, likelihood ratio test (comparing all models to the Base model); AIC, Akaike's Information Criterion.
*Model is adjusted for age at diagnosis for cases or age at participation in study for controls, pack-years of smoking, gender, and family history of lung cancer.
†Model is additionally adjusted for MLE individual European ancestry.
‡Model is additionally adjusted for STRUCTURE individual European ancestry.

groups used in this analysis, the majority of the 13 markers had composite δ values of $\geq$0.2. In addition, the CODIS loci were unlinked to each other (59) and were unlinked to the candidate gene of interest in this study, *GSTM1*, which is a key assumption for ancestry analysis (8, 10).

If the size of the parental population is small, then the precision of the estimation of the allele frequencies is poor (22). Estimation of the allele frequencies from the parental populations used in this study, however, was based on large parental sample sizes.

Controversy about the best method to use to estimate individual ancestry still exists. Therefore, individual ancestry was estimated using both MLE and STRUCTURE to compare the ability of each estimation technique to assess population structure. From the MLE estimates, it was clear which estimates were the European and West African, because ancestral allele frequencies were specified to perform the MLE calculations. Although the STRUCTURE estimates eventually showed similar population structure results compared with MLE estimates, interpretation of the individual cluster proportion estimates in terms of which ancestral population each cluster was describing was difficult, because prior ancestral allele frequency information is not used in the estimation algorithm. However, the STRUCTURE estimates did give a better fit to the data for the modeling of genetic risk compared with the MLE estimates, possibly because it did not rely on prior allele frequency information for the estimation of ancestry.

In summary, this study is one of the first to evaluate the association of individual genetic ancestry with self-reported race in a case-control study of cancer. We found that individual European ancestry did not completely correlate with self-reported race and that there was significant overlap by individual ancestry between the Caucasian, non-Hispanics and African Americans. We also observed that the frequency of a risk genotype, *GSTM1* null, varied substantially within self-reported racial group by individual ancestry and case-control status, thereby affecting models of disease risk. We conclude that individual ancestry may confound associations between disease status and a candidate gene risk genotype and could have a direct effect on accuracy of risk estimation for early-onset cancer in this Detroit population.

Acknowledgments

We thank Thomas Dyer, Ph.D. for his computational support.

References

1. Wiencke JK. Opinion: impact of race/ethnicity on molecular pathways in human cancer. Nat Rev Cancer 2004;4:79–84.
2. Holden C. Race and medicine. Science 2003;302:594–6.
3. Ferlay J, Bray F, Pisani P, Parkin DM. GLOBOCAN 2002: cancer incidence, mortality and prevalence worldwide; IARC Cancer Base No. 5. version 2.0. Lyon: IARC Press; 2004.
4. Ries LAG, Eisner MP, Kosary CL, et al. SEER cancer statistics review, 1975-2000, http://seer.cancer.gov/csr/. Bethesda (MD): National Cancer Institute; 2003.
5. 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.
6. 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.
7. 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.
8. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65:220–8.
9. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155:945–59.
10. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 2001;60:155–66.
11. Stephens JC, Briscoe D, O'Brien SJ. Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. Am J Hum Genet 1994;55:809–24.
12. Hoggart CJ, Parra EJ, Shriver MD, et al. Control of confounding of genetic associations in stratified populations. Am J Hum Genet 2003;72:6.
13. Bonilla C, Shriver MD, Parra EJ, Jones A, Fernandez JR. Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York City. Hum Genet 2004;115:57–68. Epub 2004 Apr 30.
14. Shriver MD, Parra EJ, Dios S, et al. Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 2003;112:387–99.
15. 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.
16. Chakraborty R, Kamboh MI, Nwankwo M, Ferrell RE. Caucasian genes in American Blacks: new data. Am J Hum Genet 1992;50:145–55.
17. Wilson JF, Weale ME, Smith AC, et al. Population genetic structure of variable drug response. Nat Genet 2001;29:265–9.
18. USCB. U.S. Census Bureau, Population Division: annual estimates of the population by race alone or in combination and Hispanic or Latino origin for the United States and States. Washington (DC): U.S. Government Printing Office; 2003.
19. Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic structure of human populations. Science 2002;298:2381–5.
20. 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.
21. 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.
22. Chakraborty R. Gene admixture in human populations: models and predictions. Yearbook of Physical Anthropology 1986;29:1–43.
23. Benhamou S, Lee WJ, Alexandrie AK, et al. Meta- and pooled analyses of the effects of glutathione *S*-transferase M1 polymorphisms and smoking on lung cancer risk. Carcinogenesis 2002;23:1343–50.
24. Houlston RS. Glutathione *S*-transferase M1 status and lung cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev 1999;8:675–82.
25. Seidegard J, DePierre JW, Pero RW. Hereditary interindividual differences in the glutathione transferase activity towards *trans*-stilbene oxide in resting human mononuclear leukocytes are due to a particular isozyme(s). Carcinogenesis 1985;6:1211–6.
26. Seidegard J, Pero RW. The hereditary transmission of high glutathione transferase activity towards *trans*-stilbene oxide in human mononuclear leukocytes. Hum Genet 1985;69:66–8.
27. Taioli E, Gaspari L, Benhamou S, et al. Polymorphisms in CYP1A1, GSTM1, GSTT1 and lung cancer below the age of 45 years. Int J Epidemiol 2003;32:60–3.
28. Lo YM, Lau TK, Chan LY, Leung TN, Chang AM. Quantitative analysis of the bidirectional fetomaternal transfer of nucleated cells and plasma DNA. Clin Chem 2000;46:1301–9.

29. Chen CL, Liu Q, Relling MV. Simultaneous characterization of glutathione *S*-transferase M1 and T1 polymorphisms by polymerase chain reaction in American whites and blacks. Pharmacogenetics 1996;6:187–91.
30. Budowle B, Moretti TR, Niezgoda S, Brown BL. CODIS and PCR-based short tandem repeat loci: law enforcement tools. Second European Symposium on Human Identification 1998. Madison (WI): Promega Corporation; 1998. p. 73–88.
31. Pepinski W, Janica J, Skawronska M, Koc-Zorawska E, Niemcunowicz-Janica A, Soltyszewski I. Population genetics for the CODIS core STR loci in the population of Northeastern Poland. J Forensic Sci 2003;48:1197–8.
32. Sun G, McGarvey ST, Bayoumi R, et al. Global genetic variation at nine short tandem repeat loci and implications on forensic genetics. Eur J Hum Genet 2003;11:39–49.
33. Tofanelli S, Boschi I, Bertoneri S, et al. Variation at 16 STR loci in Rwandans (Hutu) and implications on profile frequency estimation in Bantu-speakers. Int J Legal Med 2003;117:121–6. Epub 2003 Feb 15.
34. MIEPIC. Michigan EPIC: about immigration in Michigan. Grand Rapids (MI): Center for Michigan History Studies; 2004.
35. 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.
36. Lange K. Mathematical and statistical methods for genetic analysis. New York: Springer-Verlag, Inc.; 1997.
37. Edwards AWF. Likelihood. Baltimore: Johns Hopkins University Press; 1992.
38. Akaike H. Information theory and an extension of the maximum likelihood principle. In: CF Petrov BN, editor. Second international symposium on information theory. Budapest: Akademiai Kiado; 1973. p. 267–81.
39. SAS. Statistical analysis software, version 9.1. Cary (NC); 2003.
40. 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.
41. 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.
42. Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB. Cancer incidence in five continents. Volume VIII. IARC Scientific Publication No. 155. Lyon: IARC Press; 2003.
43. Hartl DL, Clark AG. Principles of population genetics. 2nd ed. Sunderland (MA): Sinauer Associates, Inc.; 1989.
44. Cavalli-Sforza LL, Menozzi P, Piazza A. The history and geography of human genes. Princeton: Princeton University Press; 1994.
45. Sokal RR, Harding RM, Oden NL. Spatial patterns of human gene frequencies in Europe. Am J Phys Anthropol 1989;80:267–94.
46. Bamshad M, Wooding S, Salisbury BA, Stephens JC. Deconstructing the relationship between genetics and race. Nat Rev Genet 2004;5: 598–609.
47. Khlat M, Cazes MH, Genin E, Guiguet M. Robustness of case-control studies of genetic factors to population stratification: magnitude of bias and type I error. Cancer Epidemiol Biomarkers Prev 2004;13:1660–4.
48. 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.
49. 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.
50. Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 2001;20:4–16.
51. Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nat Genet 2003;33 Suppl: 266–75.
52. USCB. U.S. Census Bureau (USCB). Current population reports, Series. Population profile of the United States: 1997. Washington (DC): U.S. Government Printing Office, 1998. p. 23–194.
53. Reed TE. Caucasian genes in American Negroes. Science 1969;165:762–8.
54. 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.
55. Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 2003;73:6.
56. Pfaff CL, Barnholtz-Sloan J, Wagner JK, Long JC. Information on ancestry from genetic markers. Genet Epidemiol 2004;26:305–15.
57. Gill P. Population genetics of short tandem repeat (STR) loci. Genetica 1995; 96:69–87.
58. Budowle B, Shea B, Niezgoda S, Chakraborty R. CODIS STR loci data from 41 sample populations. J Forensic Sci 2001;46:453–89.
59. Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B. The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. Electrophoresis 1999;20: 1682–96.