doi: 10.1111/j.1529-8817.2005.00255.x

The role of Self-Defined Race/Ethnicity in Population Structure Control

X-Q. Liu[1], A. D. Paterson[1,2], E. M. John[3] and J. A. Knight[2,4]*

[1]*Program in Genetics and Genomic Biology, The Hospital for Sick Children, Toronto, Ontario, Canada;*
[2]*Department of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada;*
[3]*Northern California Cancer Center, Fremont, California, USA;*
[4]*Prosserman Centre for Health Research, Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, Ontario, Canada*

Summary

Population-based association studies are powerful tools for the genetic mapping of complex diseases. However, this method is sensitive to potential confounding by population structure. While statistical methods that use genetic markers to detect and control for population structure have been the focus of current literature, the utility of self-defined race/ethnicity in controlling for population structure has been controversial. In this study of 1334 individuals, who self-identified as either African American, European American or Hispanic, we demonstrated that when the true underlying genetic structure and the self-defined racial/ethnic groups were roughly in agreement with each other, the self-defined race/ethnicity information was useful in the control of population structure.

Keywords population stratification bias, association study, gene mapping, genetic structure, skin pigmentation

Introduction

Population-based association studies are powerful tools for the genetic mapping of complex disease loci (Risch, 2000). However, this method is sensitive to potential confounding by population structure (or population stratification). It can cause false results if cases and controls are not well matched for their population origins, or in the case of recent admixed populations if they are not well matched for their proportions of ancestry, and different frequencies of both marker alleles and disease are present in different population groups (Lander & Schork, 1994). The effects of population structure are more serious when the sample size increases (Pritchard & Rosenberg, 2001).

Two approaches have been developed to detect and control population structure in association studies. The Genomic Control approach uses a group of markers that are unlinked to the candidate locus to obtain a correct null distribution on which the observed test statistic is based (Pritchard & Rosenberg, 1999; Devlin & Roeder, 1999; Reich & Goldstein, 2001). In contrast, the Structured Association approach uses unlinked markers to estimate the number of subpopulations and individuals' ancestral proportions, and then performs the association test conditioning on the inferred population structure (Pritchard *et al.* 2000a; Satten *et al.* 2001).

While the detection of and control for population structure have focused on genetic markers, the utility of self-defined race/ethnicity in controlling for population structure has been controversial. In some studies, it was observed that genetically inferred clusters correspond poorly to self-defined race/ethnicity (Wilson *et al.* 2001); in other studies, the authors concluded that given sufficient numbers of markers and sample sizes, self-defined race/ethnicity could adequately represent the inferred genetic clusters (Risch *et al.* 2002; Bamshad *et al.* 2003). Furthermore, studies have shown that by matching cases and controls on

*Correspondence: J. A. Knight, Ph.D., Prosserman Centre for Health Research, Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, 60 Murray Street, Toronto, Ontario, Canada M5G 1X5. Phone: (416) 586-8701. Fax: (416) 586-8404. Email: knight@mshri.on.ca

self-defined race/ethnicity, it may be possible to control for large-scale population structure (Ardlie *et al.* 2002). For certain phenotypes, however, mild population structure is very difficult to control for using self-defined race/ethnicity alone (Freedman *et al.* 2004).

In this study, we evaluated the utility of self-defined race/ethnicity versus the individual ancestry estimated by *Structure* using 15 microsatellite markers in controlling for population structure. We demonstrated this by examining the association between the markers and constitutive skin pigmentation, a phenotype that differs between some racial/ethnic groups (Parra *et al.* 2004).

Materials and Methods

Markers and Genotyping

The DNA samples were from female breast cancer cases and controls who participated in a population-based case-control study of breast cancer carried out in the San Francisco Bay Area (John *et al.* 2003). A total of 1,500 subjects were genotyped with a multiplex short tandem repeat (STR) system of 15 autosomal markers and one sex chromosome marker, amelogenin (AMEL), from the AmpFLSTR Identifiler kit (PE Applied Biosystems). The DNA amplifications were undertaken according to the manufacturer's protocol and separated using an ABI-3700 automated sequencer. The allele calls were made using the Genotyper software. Samples containing allelic ladders allowed consistent allele scoring between plates. In addition to the positive and negative controls, 79 blinded duplicates were included among the samples. Genotyping was repeated for 187 samples with poor quality genotypes. Samples were excluded for the following reasons: more than one genotype discrepancy for the blinded duplicates or repeats (n = 51); discrepancies between the genetic sex-type from the AMEL marker and the self-defined gender (n = 2); or missing genotypes due to PCR failures (n = 113). Among the blinded duplicates, the average error rate was 2%. Genotyping was performed blind to self-defined race/ethnicity and case-control status. This study was approved by the Mount Sinai Hospital Research Ethics Board and the Northern California Cancer Center Institutional Review Board.

Populations and Phenotypes

After the exclusions described above, there were 1,334 individuals (615 cases, 719 controls), including 392 African Americans, 439 European Americans and 503 Hispanics, in the analysis. Race/ethnicity was based on self-report from a categorical list obtained in an initial telephone screening interview and a subsequent in-person interview that included questions on the country of birth of parents and grandparents.

Constitutive skin pigmentation was measured at the middle of the upper inner arm (a site that is generally not exposed to sunlight) using a skin reflectometer (Minolta chromameter CR300). Two measurements at adjacent sites were taken for each individual, and the mean was used in the analysis. Higher values indicate lighter skin. The subjects' height (in metres) and weight (in kilograms) were also measured, and body mass index (BMI) was calculated as weight/(height)^2. For individuals who declined the measurements, self-reported height and weight were used to calculate BMI.

Statistical Methods

Three measures of marker information content of ancestry were calculated: 1) F_{st}, which is the proportion of the total variance that is due to between-subpopulation variance (Wright, 1965); 2) the allele frequency differential (δ), which is the absolute value of the difference of allele frequencies between populations (Chakraborty & Weiss, 1988); and 3) informativeness for assignment (I_n), which is the difference between the log-likelihood of drawing an allele randomly from a set of populations and the log-likelihood of drawing an allele from an 'average' population whose allele frequencies are equal to the mean across the populations (Rosenberg *et al.* 2003). F_{st}, expected heterozygosity, and departure from Hardy-Weinberg equilibrium (HWE) were calculated using the computer program *Arlequin* Version 2 (Schneider *et al.* 2000). Chi-square (χ^2) tests were applied to test for differences in allele frequencies between racial/ethnic groups, and for the association between the marker alleles and breast cancer. Alleles were removed from these tests if they had a frequency of less than 1%. Other descriptive values, δ, and I_n were calculated using SAS (Version 8.2, Cary, NC).

Table 1 Locations, number of alleles, and heterozygosities for the 15 markers

		Number of alleles				Heterozygosity			
Marker	Chromosomal location	African Americans	European Americans	Hispanics	All	African Americans	European Americans	Hispanics	All
CSF1PO	5q33.1	10	8	8	10	0.79	0.72	0.73	0.75
D2S1338	2q35	14	11	11	14	0.90	0.87	0.86	0.88
D3S1358	3p21.31	8	9	8	9	0.75	0.79	0.74	0.77
D5S818	5q23.2	11	9	9	11	0.75	0.71	0.69	0.73
D7S820	7q21.11	7	8	9	9	0.78	0.82	0.79	0.80
D8S1179	8q24.13	11	11	10	11	0.79	0.81	0.80	0.81
D13S317	13q31.1	7	8	8	8	0.72	0.79	0.82	0.79
D16S539	16q24.1	8	8	7	8	0.80	0.78	0.79	0.79
D18S51	18q21.33	18	14	16	20	0.88	0.88	0.88	0.88
D19S433	19q12	16	15	13	18	0.84	0.79	0.83	0.83
D21S11	21q21.1	17	16	15	22	0.84	0.84	0.83	0.84
FGA	4q31.3	21	18	14	26	0.88	0.87	0.87	0.87
TH01	11p15.5	8	7	6	8	0.75	0.78	0.77	0.79
TPOX	2p25.3	8	6	9	9	0.76	0.63	0.65	0.68
vWA	12p12	10	8	10	10	0.82	0.80	0.78	0.81
Mean ± SD*		12 ± 5	10 ± 4	10 ± 3	13 ± 6	0.80 ± 0.05	0.79 ± 0.07	0.79 ± 0.07	0.80 ± 0.06

*SD, standard deviation.

Of the 15 markers, two marker pairs (CSF1PO and D5S818, D2S1338 and TPOX) are on the same chromosome (Table 1). The genetic distances between the markers are approximately 26 cM and 218 cM, respectively, which are too large to expect any linkage disequilibrium (LD) under normal circumstances. The LD tests were performed using *Mendel* Version 5.0 (Lange *et al.* 2001) to assess if there was any association between markers.

Structure (Version 2.0) was applied to determine population clustering (Pritchard *et al.* 2000b). This program uses multilocus genotype data to infer population structure and assign individuals to ancestral populations with a model-based Bayesian approach. It assumes HWE and no LD between all markers within each cluster. Because most African Americans and Hispanics are from admixed populations with ancestors of African, European and Native American origins (Tseng *et al.* 1998; Shriver *et al.* 2003), we performed the analysis assuming that the individuals originated from more than one ancestral population and that the allele frequencies were correlated between populations. Depending on the degree of admixture, there may be detectable substructures within each of these two populations. For this reason, we chose K, the number of populations, to be from 1 to 6 so it could cover a reasonable number of populations (or subpopulations).

To estimate K present in our sample, each estimation was repeated 10 times for K from 1 to 6. The posterior probabilities of K, Pr(X|K), where X was the observed genotypes, were compared to choose the best K. Then, each individual was assigned an estimated admixture proportion (Q) for each of the K clusters. We ran *Structure* with 100,000 burn-in length and 100,000 Monte Carlo Markov Chain (MCMC) repeats. The agreement between the *Structure* assignments based on the Q values and self-defined race/ethnicity was measured by *kappa* for each racial/ethnic group (Cohen 1960) using SAS PROC FREQ (Version 8.2, Cary, NC).

When *Structure* assignments correspond to already known population or geographic information, *Structure* can incorporate this information in its analysis in addition to the information derived from genetic markers, and identify individuals who are misclassified (Pritchard *et al.* 2000b). After the estimation of K and Q values in our sample, we applied this method to identify individuals whose genetic constitution was not consistent with their self-defined race/ethnicity (population outliers) by incorporating self-defined race/ethnicity information in the analysis, in addition to the 15 genetic markers.

An evaluation of population structure control was carried out by testing the association between each of the 15 markers and constitutive skin pigmentation. The tests were performed for each racial/ethnic group separately

and for the three racial/ethnic groups combined. Associations were tested with and without adjustment for the Q values estimated by *Structure* based on the 15 microsatellite markers. In addition, adjustment for self-defined race/ethnicity was carried out in the combined data. ANOVA was used to test for association between markers and skin pigmentation when no adjustment was applied. When adjusting for Q values and self-defined race/ethnicity, linear regression was first applied to generate residuals that were then used to test for associations between individual markers and skin colour using non-parametric Kruskal-Wallis tests. Since none of the 15 markers is known to be associated with skin colour, we assumed that any associations would be due to population structure. The statistical analyses were performed using SAS (Version 8.2, Cary, NC).

Results

Description of the Markers

Table 1 lists the locations, number of alleles, and heterozygosities for the 15 highly polymorphic markers. For the three racial/ethnic groups combined there were on average 13 alleles per marker, and the heterozygosities ranged from 0.68 to 0.88 (mean 0.80). When the markers were analyzed separately by race/ethnicity, the number of alleles and heterozygosity measures were not significantly different between groups, though African Americans had the largest total number of alleles ($n = 174$ vs. $n = 156$ in European Americans, and $n = 153$ in Hispanics), as well as the largest number of unique alleles ($n = 24$ vs. $n = 9$ in European Americans, and $n = 5$ in Hispanics). The unique alleles were rare, with mean combined frequencies of 0.33%, 0.27%, and 0.20% for African Americans, European Americans, and Hispanics, respectively. The allele frequencies were significantly different between African Americans and European Americans, and between African Americans and Hispanics, with p-values less than 0.0001 for all the markers. For European Americans and Hispanics the allele frequencies for D18S51 were not significantly different ($p = 0.56$), while the allele frequencies for the other markers were significantly different with p-values less than 0.01.

We tested whether any of these markers were associated with breast cancer risk. For the tests in each racial/ethnic group (45 independent tests with 15 tests for each group), three markers (D13S317, D19S433, and D7S820) in the African American group had p-values less than 0.05, a lowest p-value of 0.006. In the combined data, marker D19S433 had a p-value of 0.03. Given the number of independent tests this number of low p-values would be expected by chance alone. These findings indicate that these 15 markers were not likely to be associated with disease status, and therefore cases and controls were pooled in the subsequent analyses.

We also tested for departure from HWE and the presence of LD in each of the racial/ethnic groups. Markers D8S1179 ($p = 0.03$) in African Americans, D21S11 ($p = 0.03$) and FGA ($p = 0.02$) in European Americans, and D21S11 ($p = 0.01$) in Hispanics showed marginal departure from HWE. For the LD tests there were 9 marker pairs in African Americans, 3 pairs in European Americans, and 5 pairs in Hispanics that had p-values less than 0.05, with the lowest p-value being 0.002 (for the marker pair D13S317 and FGA in African Americans). Given the large number of tests performed (45 tests for HWE and 315 tests for LD), the number of p-values less than 0.05 would be expected by chance or might indicate moderate population substructure within some of the racial/ethnic groups.

Table 2 lists the I_n, F_{st}, and δ values for the 15 markers. For the three racial/ethnic groups combined I_n ranged from 0.022 to 0.061 with a mean of 0.042; F_{st} ranged from 0.005 to 0.034 with a mean of 0.015. When combining two racial/ethnic groups the I_n, F_{st}, and δ values for African Americans and European Americans, and for African Americans and Hispanics, were always more significantly larger than the corresponding values for European Americans and Hispanics (with all $p < 0.05$). These results indicate that the genetic distance between European Americans and Hispanics was smaller than the distances between African Americans and European Americans, and between African Americans and Hispanics.

Population Structure

For the three racial/ethnic groups combined, the posterior probabilities of K obtained from *Structure*, $\Pr(X \mid K)$,

Table 2 Measures of the marker information content for ancestry

	Informativeness for assignment (I_n)[a]				F_{st}				Allele frequency differential (δ)		
Marker	AA-EA[b]	AA-HA	EA-HA	All	AA-EA	AA-HA	EA-HA	All	δ_{AA-EA}	δ_{AA-HA}	δ_{EA-HA}
CSF1PO	0.047	0.031	0.010	0.041	0.0115	0.0118	0.0030	0.0085	0.35	0.32	0.15
D2S1338	0.047	0.035	0.019	0.046	0.0179	0.0124	0.0083	0.0125	0.45	0.44	0.32
D3S1358	0.023	0.012	0.018	0.024	0.0150	0.0101	0.0188	0.0148	0.31	0.27	0.32
D5S818	0.029	0.068	0.028	0.059	0.0117	0.0514	0.0202	0.0284	0.54	0.56	0.34
D7S820	0.014	0.015	0.020	0.022	0.0072	0.0105	0.0126	0.0103	0.26	0.30	0.33
D8S1179	0.046	0.026	0.007	0.035	0.0264	0.0127	0.0024	0.0129	0.50	0.34	0.18
D13S317	0.035	0.062	0.019	0.052	0.0190	0.0394	0.0132	0.0236	0.35	0.56	0.30
D16S539	0.024	0.009	0.018	0.023	0.0173	0.0054	0.0091	0.0103	0.36	0.20	0.26
D18S51	0.053	0.051	0.005	0.049	0.0193	0.0180	0.0004	0.0120	0.52	0.52	0.14
D19S433	0.067	0.050	0.027	0.066	0.0151	0.0109	0.0052	0.0099	0.42	0.42	0.28
D21S11	0.021	0.044	0.014	0.037	0.0025	0.0178	0.0055	0.0088	0.23	0.43	0.23
FGA	0.033	0.022	0.028	0.040	0.0047	0.0019	0.0067	0.0045	0.27	0.24	0.32
TH01	0.052	0.030	0.018	0.045	0.0581	0.0278	0.0196	0.0338	0.59	0.44	0.36
TPOX	0.049	0.066	0.015	0.061	0.0279	0.0373	0.0066	0.0234	0.38	0.56	0.23
vWA	0.023	0.022	0.014	0.027	0.0125	0.0137	0.0134	0.0132	0.35	0.31	0.28
Mean ± SD[c]	0.038 ± 0.015	0.036 ± 0.020	0.017 ± 0.007	0.042 ± 0.014	0.0177 ± 0.0132	0.0187 ± 0.0140	0.0097 ± 0.0064	0.0151 ± 0.0083	0.39 ± 0.11	0.39 ± 0.12	0.27 ± 0.07

[a]Informativeness for assignment (I_n) is the difference between the log-likelihood of drawing an allele randomly from a set of populations and the log-likelihood of drawing an allele from an 'average' population whose allele frequencies are equal to the mean across the populations (Rosenberg *et al.* 2003). F_{st} is the proportion of the total variance that is due to between-subpopulation variance (Wright, 1965). Allele frequency differential (δ) is the absolute value of the difference of allele frequencies between populations (Chakraborty & Weiss, 1988).

[b]AA, African American; EA, European American; HA, Hispanic.

[c]SD, standard deviation.

were close to 1 when K = 3 and close to 0 for all other tested K values. This means that the optimal number of clusters was three. Each individual was assigned three Q values corresponding to the ancestry proportion estimates of the three clusters which corresponded to the three racial/ethnic groups. No subpopulation was detected within any of the racial/ethnic groups. For all the subsequent analyses using the combined data, K = 3 was used.

When comparing the *Structure* assignment with self-defined race/ethnicity, the agreement measures (*kappa* values) were 0.72 with a 95% confidence interval (0.68-0.76), 0.52 (0.47-0.57), and 0.53 (0.48-0.58) for African Americans, European Americans, and Hispanics, respectively (Figure 1A). According to the *kappa* guidelines from Altman (1991), the agreements for European Americans and Hispanics were moderate (*kappa* between 0.41 and 0.60), while the agreement for African Americans was good (*kappa* between 0.61 and 0.80). We then performed the analysis incorporating the self-defined race/ethnicity information in addition to the genetic data to identify population outliers. Out of 1,334 individuals *Structure* identified 10 individuals whose highest Q value (ancestry proportion), based on both the marker information and self-defined race/ethnicity, did not correspond with their self-defined race/ethnicity (Figure 1B). We checked the parental and grandparental birthplaces of the 10 outliers, and found that four of them probably had mixed ethnicities. For the other six individuals no additional information could be derived from the birthplace information.

Association of Markers with Skin Pigmentation

Table 3 lists the traits (age, skin pigmentation, weight, height and BMI) for the 1,293 individuals with constitutive skin pigmentation measurements. There was no significant difference in skin pigmentation between the breast cancer cases and controls, whereas the differences were statistically significant between all pairs of racial/ethnic groups, as expected. Overall about 69% of the variability in skin pigmentation could be explained by self-defined race/ethnicity ($p < 0.0001$).

The variances of skin pigmentation for these three racial/ethnic groups were also significantly different from each other, with African Americans having the largest variance ($p < 0.0001$ when compared to the variances of European Americans and Hispanics), European Americans having the smallest, and Hispanics having the intermediate variance (Table 3). We also found linear associations between ln(age) and ln(skin pigmentation) for

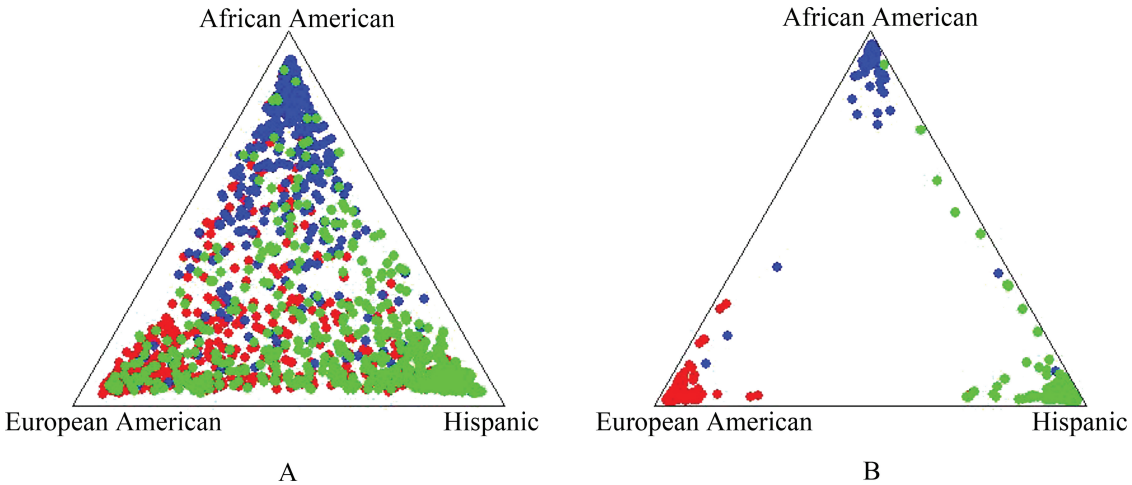

Figure 1 Triangle plots of the clustering results from *Structure*. The ancestry of individuals was estimated under a model of three subpopulations (K = 3). The locations of the points were decided by the distance (estimated admixture proportion) to one side of the triangle. An individual's self-defined race/ethnicity was represented by colors: blue (African American), red (European American) and green (Hispanic). A. without self-defined race/ethnicity information; B. with self-defined race/ethnicity information.

Table 3 Phenotypes in the three racial/ethnic groups (Mean ± SD)[a]

Race/ethnicity	Sample size	Age (years)	Skin pigmentation	Weight (kg)	Height (m)	BMI (kg/m^2)[b]
African Americans	370	55 ± 12	21.09 ± 6.58	85 ± 20	1.64 ± 0.07	32 ± 7.2
European Americans	426	59 ± 12	38.93 ± 3.62	73 ± 16	1.63 ± 0.07	28 ± 5.9
Hispanics	497	54 ± 11	34.46 ± 4.22	75 ± 18	1.57 ± 0.07	30 ± 6.8
All	1293	56 ± 12	32.11 ± 8.70	78 ± 19	1.61 ± 0.07	30 ± 6.8

[a]SD, standard deviation. [b] BMI: body mass index was calculated by BMI = weight/(height)^2.

African Americans ($p = 0.0005$, $R^2 = 0.03$, $\beta = 0.27$), and between BMI and skin pigmentation for European Americans ($p = 0.009$, $R^2 = 0.02$, $\beta = 0.08$). In subsequent analyses, skin pigmentation adjusted for age and BMI was used.

For the association tests between genetic markers and skin pigmentation in individual racial/ethnic groups, two markers in African Americans and two markers in Hispanics had p-values less than 0.05 (Table 4). The lowest p-value was 0.01. For 45 independent tests this could happen by chance, or be caused by mild population structure in African Americans and Hispanics. Adjustment for the Q values (ancestry proportions) estimated by *Structure* did not change the results.

For the association tests between the markers and skin pigmentation in the combined data, before the adjustment for individual ancestry, 14 of the 15 markers showed highly significant associations with skin pigmentation (Table 4). After the adjustment for the Q values (only the African and Hispanic ancestry estimates were used in the adjustment, since the European ancestry estimates were equal to one minus the other two estimates, given that the estimates are proportions) based on the marker information, the number of markers with p-values less than 0.05 was reduced to 3. No marker was significantly associated with skin pigmentation after the adjustment for self-defined race/ethnicity.

Discussion

In this study we found that the major population structure (African American, European American, and Hispanic) could be detected, and partially controlled, with 15 microsatellite markers whose genotypes could be obtained from a single PCR reaction from a small amount of DNA. However, self-defined race/ethnicity was sufficient in controlling the population structure in our sample from California, as adjustment for self-defined

Table 4 Association between individual markers and skin pigmentation[a] (p-values) in each racial/ethnic group and combined data

	Individual racial/ethnic group			All racial/ethnic groups combined		
Marker	European Americans	African Americans	Hispanics	Before adjustment	After adjustment for Q based on genetic markers	After adjustment for self-defined race/ethnicity
CSF1PO	0.36	0.24	0.27	0.08	0.46	0.84
D13S317	0.63	0.01*	0.15	<0.0001*	0.02*	0.14
D16S539	0.30	0.14	0.64	<0.0001*	0.60	0.07
D18S51	0.41	0.96	0.98	<0.0001*	0.57	0.84
D19S433	0.82	0.76	0.91	<0.0001*	0.87	0.82
D21S11	0.91	0.17	0.53	0.03*	0.17	0.55
D2S1338	0.36	0.22	0.07	<0.0001*	0.04*	0.18
D3S1358	0.93	0.09	0.04*	<0.0001*	0.32	0.16
D5S818	0.08	0.17	0.02*	0.009*	0.81	0.97
D7S820	0.53	0.18	0.85	0.009*	0.54	0.56
D8S1179	0.61	0.98	0.53	<0.0001*	0.63	0.63
FGA	0.35	0.07	0.92	0.007*	0.21	0.52
TH01	0.68	0.02*	0.16	<0.0001*	0.002*	0.13
TPOX	0.48	0.80	0.91	<0.0001*	0.28	0.92
vWA	0.35	0.22	0.77	<0.0001*	0.15	0.38

[a]Skin pigmentation was adjusted for age and BMI. *p-values < 0.05/

race/ethnicity alone eliminated all associations between each of the markers and skin pigmentation. In addition, using both self-defined race/ethnicity and genetic information can help detect the population outliers whose genetic constitution was not consistent with their self-defined race/ethnicity.

We did not detect any population substructure within any of the three racial/ethnic groups using the 15 markers. One reason may be that the 15 markers are not sufficient to detect the substructure within each of the racial/ethnic groups. The power to detect population structure mainly depends on markers' information content of ancestry, number of markers, and sample size (Pritchard *et al.* 2000a; Bamshad *et al.* 2003; Rosenberg *et al.* 2003; Rosenberg *et al.* 2001). Our 15 markers (total 193 alleles) were not specifically selected based on their different allele frequencies in the three racial/ethnic groups. Compared to other microsatellite markers, the informativeness of these markers was in the top 20 to 50 percent according to their average δ values (Risch *et al.* 2002). The number of markers genotyped in this study was relatively small, but the sample size was reasonably large.

Another reason that we did not detect any substructure within any of the racial/ethnic groups may be that there is a low degree of population substructure within the individual racial/ethnic groups in our sample. The association with skin pigmentation was significant for 4 markers (two in African Americans, two in Hispanics, and none in European Americans) out of 45 independent tests (Table 4). This may be a chance finding or due to mild population structure in African Americans and Hispanics. For admixed populations such as African Americans and Hispanics, the admixture process usually results in a population that has various individual admixture levels. However, as pointed out by Ardlie *et al.* (2002), subtle admixture needs to be distinguished from population structure. In association studies the results will be valid, and not be affected by stratification bias, as long as the admixture is distributed equally in both cases and controls.

The role of self-defined race/ethnicity in controlling for population structure has been controversial (Ardlie *et al.* 2002; Barnholtz-Sloan *et al.* 2005; Wacholder *et al.* 2000; Wilson *et al.* 2001). A recent study by Barnholtz-Sloan *et al.* (2005) investigated population structure via individual ancestry estimates versus self-defined race/ethnicity, using lung cancer as a phenotype. They concluded that 'significant population substructure differences exist that self-reported race alone does not capture.' Comparing our study with the study by Barnholtz-Sloan *et al.* (2005), both used forensic STR markers (15 marker panel vs. 13 marker panel) and both included European Americans (426 individuals vs. 555 individuals) and African Americans (370 individuals vs. 191 individuals). However, the population samples were from different geographic locations in the United States (San Francisco vs. Detroit), and the study by Barnholtz-Sloan *et al.* (2005) did not include any Hispanics. In the United States studies have shown that African Americans or Hispanics from different geographic locations have different degrees of admixture (Hanis *et al.* 1991; Parra *et al.* 1998). For the 13 markers that were common between these studies, we compared the δ values of African American versus European American with our δ values and found the correlation was not statistically significant ($p = 0.25$). Based on this finding, differences in population samples may contribute to the different results in the two studies.

Overall, we demonstrated in this population sample from California that self-defined race/ethnicity could play an important role in the control for population structure. Our conclusion was based on adjustment for the three major self-defined racial/ethnic groups. However, it may be extended to individual racial/ethnic groups when more detailed racial/ethnic information is available for each group, and when more markers are unable or unavailable to control for the subtle population structure. In the study by Ardlie *et al.* (2002) it was shown that the effect of population structure could be reduced by removing recent immigrants (non-U.S. born). We observed similar results in our data: when individuals who were not born in the U.S. were excluded from the analysis, none of the markers was significantly associated with skin pigmentation in each racial/ethnic group. Traditional epidemiological information, such as self-defined race/ethnicity, geographic location and birthplace, may be important in population structure control because they represent common cultural and environmental exposures. An analysis that adjusts for self-defined race/ethnicity may eliminate some of the confounding effects caused by these factors (Wacholder *et al.*

2002). Therefore, in addition to our growing knowledge of the role of genetic data in detection and control for population structure, the importance of this epidemiological information should not be ignored.

Acknowledgments

We would like to thank all the participants who provided blood samples and personal information to this study. This work was funded by the Canadian Cancer Etiology Research Network of the National Cancer Institute of Canada. The parent study was funded by the US National Cancer Institute (R01 CA77305 to E.M.J.). X-Q.L. and A.D.P. were supported by Genome Canada through OGI. A.D.P. holds a Canada Research Chair (Tier II) in the Genetics of Complex Diseases.

References

Altman, D. G. (1991) Practical statistics for medical research. London: Chapman & Hall.

Ardlie, K. G., Lunetta, K. L. & Seielstad, M. (2002) Testing for population subdivision and association in four case-control studies. *Am J Hum Genet* **71**, 304–311.

Bamshad, M. J., Wooding, S., Watkins, W. S., Ostler, C. T., Batzer, M. A. & Jorde, L. B. (2003) Human population genetic structure and inference of group membership. *Am J Hum Genet* **72**, 578–589.

Barnholtz-Sloan, J. S., Chakraborty, R., Sellers, T. A. & Schwartz, A. G. (2005) Examining population stratification via individual ancestry estimates versus self-reported race. *Cancer Epidemiol Biomarkers Prev* **14**, 1545–1551.

Chakraborty, R. & Weiss, K. M. (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci USA* **85**, 9119–9123.

Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas* **20**, 37–46.

Devlin, B. & Roeder, K. (1999) Genomic control for association studies. *Biometrics* **55**, 997–1004.

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., Pato, M. T., Petryshen, T. L., Kolonel, L.N., Lander, E.S., Sklar, P., Henderson, B., Hirschhorn, J.N. & Altshuler, D. (2004) Assessing the impact of population stratification on genetic association studies. *Nat Genet* **36**, 388–393.

Hanis, C. L., Newett-Emmett, D., Bertin, T. K. & Schull, W. J. (1991) Origins of U.S. Hispanics. Implications for diabetes. *Diabetes Care* **14**, 618–627.

John, E. M., Horn-Ross, P. L. & Koo J. (2003) Lifetime physical activity and breast cancer risk in a multiethnic population: The San Francisco Bay Area Breast Cancer Study. *Cancer Epidemiol Biomarkers Prev* **12**, 1143–1152.

Lander, E. S. & Schork, N. J. (1994) Genetic dissection of complex traits. *Science* **265**, 2037–2048.

Lange, K., Cantor, R., Horvath, S., Perola, M., Sabatti, C., Sinsheimer, J. & Sobel, E. (2001) Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Am J Hum Genet* **69**(supplement), A1886.

Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R., Forrester, T., Allison, D. B., Deka, R., Ferrell, R. E. & Shriver, M. D. (1998) Estimating African American admixture proportions by use of population specific alleles. *Am J Hum Genet* **63**, 1839–1851.

Parra, E. J., Kittles, R. A. & Shriver, M. D. (2004) Implications of correlations between skin color and genetic ancestry for biomedical research. *Nat Genet* **36**, S54-S60.

Pritchard, J. K. & Rosenberg, N. A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* **65**, 220–228.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. (2000a) Association mapping in structured populations. *Am J Hum Genet* **67**, 170–181.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000b) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Pritchard, J. K. & Rosenberg, N. A. (2001) Case-control studies of association in structured or admixed populations. *Theor Popul Biol* **60**, 227–237.

Reich, D. E. & Goldstein, D. B. (2001) Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* **20**, 4–16.

Risch, N. J. (2000) Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856.

Risch, N., Burchard, E., Ziv, E. & Tang, H. (2002) Categorization of humans in biomedical research: genes, race and disease. *Genome Biol* **3**, 1–12.

Rosenberg, N. A., Burke, T., Elo, K., Feldman, M. W., Freidlin, P. J., Groenen, M. A. M., Hillel, J., Mäki-Tanila, A., Tixier-Boichard, M., Vignal, A., Wimmers, K. & Weigend, S. (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* **159**, 699–713.

Rosenberg, N. A., Li, L. M., Ward, R. & Pritchard, J. K. (2003) Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* **73**, 1402–1422.

Satten, G. A., Flanders, W. D. & Yang, Q. (2001) Accounting for unmeasured population structure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* **68**, 466–477.

Schneider, S., Roessli, D. & Excoffier, L. (2000) Arlequin ver. 2.000: A software for population genetics data analysis.

Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Shriver, M. D., Parra, E. J., Dios, S., Bonilla, C., Norton, H., Jovel, C., Pfaff, C., Jones, C., Massac, A., Cameron, N., Baron, A., Jackson, T., Argyropoulos, G., Jin, L., Hoggart, C. J., McKeigue, P. M. & Kittles, R. A. (2003) Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum Genet* **112**, 387–399.

Tseng, M., Williams, R. C., Maurer, K. R., Schanfield, M. S., Knowler, W. C. & Everhart, J. E. (1998) Genetic admixture and gallbladder disease in Mexican Americans. *Am J Phys Anthrop* **106**, 361–371.

Wacholder, S., Rothman, N. & Caporaso, N. (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst* **92**, 1151–1158.

Wacholder, S., Rothman, N. & Caporaso, N. (2002) 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* **11**, 513–520.

Wilson, J. F., Weale, M. E., Smith, A. C., Gratrix, F., Fletcher, B., Thomas, M. G., Bradman, N. & Goldstein, D. B. (2001) Population genetic structure of variable drug response. *Nat Genet* **29**, 265–269.

Wright, S. (1965) The interpretation of population structure by F-statistics with special regards to systems of mating. *Evolution* **19**, 395–420.

Received: 11 May 2005
Accepted: 15 September 2005