Analysis of Population Structure and Stratification in NHANES III Self-Reported Race/Ethnicities

Christopher L. Sanders, Ajay J. Yesupriya, Lester R. Curtin

Background and objectives: It is common in genetic association studies to stratify statistical analyses by race/ethnicity to compensate for inherent differences in genetic variation frequencies and disease susceptibility that is found between population subgroups. However, when genetic models are stratified by self-reported race/ethnicity, this could lead to spurious gene-disease associations caused by the presence of multiple distinct subpopulations with each self-reported category because there are differences in both disease and genetic variation frequencies that are not related to one another. Race and ethnicity variables in the Third National Health and Nutrition Examination Survey (NHANES III) are primarily self-reported with a small number derived from interviewer observations. The NHANES III survey, conducted prior to OMB directive 15 on collecting multiple race information, requested that each participant report only one race/ethnicity. This could lead to potential misclassification of this important variable when used in genetic association studies.

Population structure analyses use variations in multiple genetic loci to infer the probabilities that an individual's genome originates from one or more ancestral populations. The results of these probability estimates are used here to assess the correlation of self-reported race/ethnicity in the three major populations found in NHANES III; non-Hispanic Whites, non-Hispanic Blacks and Mexican-Americans as well as to assess the heterogeneity of a population's genetic ancestry. Ancestry estimates determine whether there is any evidence of multiple populations with distinct genetic ancestry within a single self-reported group; this is termed population stratification. While the potential for population stratification exists in all populations, the Mexican-American population is particularly of concern given recent mixing of European and American-Indian populations, also termed admixture (Seldin et al, 2007). In previous studies, population stratification based on self-reported race/ethnic groups resulted in inaccurate gene-disease associations (Cardon, 2003). An often cited example is a study of diabetes in a self-reported Pima and Papago Indian population where spurious associations were attributed to a confounding effect by the presence of distinct subpopulations (Knowler et al, 1988).

In this paper, we estimate population structure as the vector of probabilities that participants are from one or more ancestral populations. We assess the correlation of self-reported race/ethnicity relative to these probabilities to determine whether there is evidence of multiple distinct genetic populations within a single self-reported race/ethnicity group which would suggest population stratification. When multiple populations within a single race/ethnicity were detected, we assessed whether these populations significantly differed by looking at certain demographic variables.

Methods: Using multilocus genetic data, we assessed the population structure of 6,597 participants from NHANES III, a nationally representative population based sample. We used a total of 50 genetic variations in this study that consisted of 15 short tandem repeats (STRs) and 35 single nucleotide polymorphisms (SNPs). SNPs were selected based on two criteria; they were not within a chromosomal distance of one centimorgan of one another to avoid linkage disequilibrium (LD) and secondly, if multiple variations were within a one centimorgan region, then the variation with the highest unweighted F(st) value was used as it was more likely to be the most informative marker.

These genetic variations were used in the software Structure 2.2, a program that utilizes a Bayesian clustering algorithm, to infer the number of populations in our sample as well as estimate the probabilities that an individual's genetic background was from one or more ancestral populations.

Results: Analysis began with the determination of the overall population structure of the 6,811 individuals in the sample. This first step assessed the number of predicted populations in the

NHANES sample which is also known as *k*. For this, we performed Bayesian clustering analyses on 50 genetic variations using the program Structure to infer the log likelihood value of *k* for 2 through 6 populations. These results estimated that the appropriate *k* for our study was three, meaning that the structure of the NHANES sample consisted of three distinct genetic populations. We continued with further analysis under the assumption that these ancestral populations represented European (Caucasian), African and Amerindian (American Indian) based on the three major self-reported populations found in NHANES III.

Analyses suggested that there were three distinct genetic populations within the NHANES III participants. Based on the self-reported race/ethnicity of our sample, these correspond to European, African, and Amerindian populations. Results indicated that individuals that self-reported non-Hispanic White and non-Hispanic Black were overwhelmingly of European and African ancestries, respectively. However, we found evidence of two distinct populations in self-reported Mexican-Americans that were primarily of either European or Amerindian ancestry. Evidence of significant demographic differences between these two subpopulations suggests population stratification that could confound statistical analyses involving genetic markers in self-reported Mexican-Americans.

Next, we used Structure 2.2 to estimate the probabilities that a participant's genome had contributions from the European, African and Amerindian ancestral populations that had been predicted. We used the highest probability estimate for each participant to create a categorical variable that represented the single most likely genetic population for that individual. This variable was then compared to self-reported race/ethnicity where we found that they were in agreement in 90.1% of the participants across all self-reported race/ethnicities. Within self-reported race/ethnicities we found a 93.1% match between self-reported non-Hispanic Whites who were estimated to be of primarily of European ancestry and 95.9% of self-reported non-Hispanic Blacks who were estimated to be of primarily of African ancestry. Yet, we only found that 80.7% of those that self-reported Mexican-Americans who were estimated to be of primarily of Amerindian ancestry

We assessed any evidence of population stratification that would not be corrected by statistical models that stratify by self-reported race/ethnicity. To identify population stratification within race/ethnicities, we looked for evidence of demographic differences between participants whose most likely ancestral population probability did not match their self-reported race/ethnicity. These analyses showed statistically significant differences for key demographic variables between these two distinct self-reported Mexican-American groups but no evidence of stratification in the other two self-reported race/ethnicities.

Discussion/Conclusion: This study provides evidence that potentially two distinct ancestral populations with significant demographic differences exist within the self-reported Mexican-Americans found in NHANES III. These results suggest that the population stratification in this group could lead to confounding in some statistical analyses. Therefore, it may be useful to develop methods to account for population stratification to avoid confounding that could occur in genotype-phenotype analyses of the Mexican-American population.