Analysis of Population Structure in NHANES III

Christopher Sanders[1], Ajay Yesupriya[2] and Lester Curtain[1] [1]National Health and Nutrition Examination Surveys [2]National Office of Public Health Genomics

National Center for Health Statistics & National Office of Public Health Genomics • Centers for Disease Control and Prevention • U.S. Department of Health and Human Services

Objectives

Assess the population structure of 6,597 participants in the a nationally representative sample of National Health and Nutrition Examination Survey III (NHANES)

- Structure, a program that utilizes a Bayesian clustering approach, was used to infer the number of populations in our sample using multilocus genetic data
- Probability estimates were assessed for contributions of individuals' genetic background from one of three ancestral populations
- Highest probability estimate of ancestry for each participant was compared to their self-reported race/ethnicity
- Self-reported populations in which there was evidence of multiple distinct race/ethnic groups were analyzed for evidence of distinct demographic differences

Introduction

- Genetic association studies commonly stratify statistical analyses by race/ethnicity to compensate for inherent differences in genetic variations and disease susceptibility that is found between race/ethnic populations.
- 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.
- Probability estimates based on genetic variation data could provide a useful resource to assess the ancestry of self-reported race/ethnic groups

Methods

Survey Details – NHANES III

- Representative probability sample of the civilian non-institutionalized U.S. population
- Mission : Assess the health and nutrition of the U.S. population
- Collects thousands of variables from laboratory tests and questionnaires
- NHANES III DNA samples collected from Phase II (1991-1994)

Sample Details

Self-Reported Race/Ethnicity	# of Participants
Non-Hispanic White	2584
Non-Hispanic Black	1999
Mexican-American	2014
Total	6597

Table 1. Sample sizes for the NHANES III participants used in this study by Race/ethnicity.

Genetic Data

- 50 genetic variations used this study consisted of 15 short tandem repeats and 35 single nucleotide polymorphisms
- SNPs were selected based on two criteria
  - Not within one centimorgan of one another to avoid linkage disequilibrium
  - If multiple variations were within a one centimorgan region, the one with the highest relative unweighted F(st) value was chosen
- Genetic variations used in this study were originally genotyped not for the purposes of determining genetic ancestry. Therefore F(st) values may not be as high of those found in other studies

Assessment of the Number of Ancestral Populations

- We performed Bayesian clustering analyses using the program Structure 2.2 on 50 genetic variations to infer population structure for those who self-reported as non-Hispanic White, non-Hispanic Black or Mexican-American
- We assessed the number of predicted populations found in our sample, also known as *k* from log likelihood values of models that assess *k* between 2 and 6 using 30,000 burn-ins and 50,000 repetitions.
- Analyses suggested that there were three distinct genetic populations(*k*=3) within the NHANES III participants. Based on the self-reported race/ethnicity of our sample, these would correspond to European, African, and Amerindian ancestral populations.

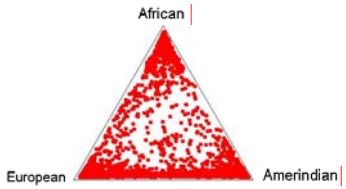

Figure 1. Population structure of NHANES III genetic component

Determination of Ancestry Population Probability Estimates

- We then used the Structure 2.2 to estimate the probability that participant's genome had contributions from one of these three ancestral populations; European, African and Amerindian
- Results indicated individuals that self-reported non-Hispanic White and non-Hispanic Black were primarily of European and African ancestries, respectively in greater than 93% of cases. However, we found that only 81% of self-reported Mexican-Americans were predicted to be of primarily of Amerindian ancestry

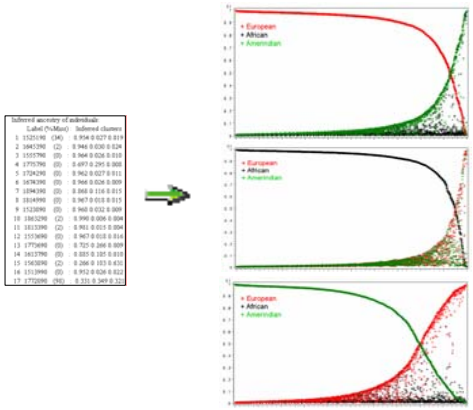

Figure 2. Probability estimates by self-reported race/ethnicity; Non-Hispanic White, Non-Hispanic Black and Mexican-American

Mean Ancestry Probability Estimates

	Non-Hispanic White	Non-Hispanic Black	Mexican-American
European	0.87	0.06	0.21
African	0.03	0.90	0.04
Amerindian	0.10	0.05	0.76

Table 2. Mean ancestry probability estimates for each NHANES self-reported race/ethnicity

Comparison of Probability Estimates to Self-Reported Race/Ethnicity

Self-Reported Race/Ethnicity	Matches to Highest Probability Estimate of Genetic Ancestry (%)
Non-Hispanic White	2406 / 2584 (93.1)
Non-Hispanic Black	1910 / 1999 (95.9)
Mexican-American	1625 / 2014 (80.7)

Table 3. Correlation of most likely categorical population structure predictions; European, African or Amerindian to self-reported Race/Ethnicity

	Categorical Ancestral Population Predicted		
Self-Reported Race/Ethnicity	European	African	Amerindian
Non-Hispanic White	2406	21	157
Non-Hispanic Black	61	1910	28
Mexican-American	373	16	1625

Table 4. Self-Reported Race/Ethnicity vs. Categorical Ancestral Population Prediction

Assessment of Demographic Differences

- Self-reported Mexican-Americans of primarily European ancestry and primarily Amerindian ancestry were compared for differences in key demographic variables
- These results suggest that between these two groups of self-reported Mexican-American groups there are significant differences that could create a scenario where population stratification could be a concern.

Demographic Variables	Primarily Amerindian Descent	Primarily European Descent	p-value
Mean Age	33.08 (0.49)	33.47 (1.21)	0.713
Male	51.67 (0.93)	50.14 (2.11)	0.538
Female	48.33 (0.93)	49.86 (2.11)	
Education <12 years	64.95 (2.50)	55.38 (3.62)	0.039
Education = 12 years	21.96 (1.63)	25.71 (2.71)	
Education > 12 years	13.10 (1.44)	19.44 (2.87)	
Census Region – Northeast	1.61 (0.84)	0.54 (0.55)	0.005
Census Region – Midwest	14.33 (6.65)	7.87 (3.30)	
Census Region – South	27.79 (10.72)	20.68 (9.75)	
Census Region – West	59.17 (12.39)	70.91 (11.23)	
Country of Origin – U.S.	41.21 (2.97)	53.16 (5.30)	0.032
Country of Origin – Other	58.79 (5.30)	46.84 (5.30)	
Income - Over poverty	36.14 (3.31)	27.62 (2.08)	0.061
Income - Under poverty	63.86 (3.31)	72.38 (2.08)	

Table 5. Assessment of key demographic variables in self-reported Mexican-Americans

Conclusions

- Overall high level of matching between self-reported race/ethnicity and predicted genetic ancestry of individuals
- Evidence that the self-reported Mexican-Americans are a highly admixed population with significant contributions from the Amerindian and European ancestries.
- Significant demographic differences exist between self-reported Mexican-Americans whose primary ancestry contributions were Amerindian or European.
- Ancestral informative markers may yield more accurate results and could be used to verify these analyses