Annals of human genetics

doi: 10.1111/j.1469-1809.2011.00657.x

# Ancestral Heterogeneity in a Biethnic Stroke Population

Lynda D. Lisabeth[1,2]*, Lewis B. Morgenstern[1,2], David T. Burke[3], Yan V. Sun[1] and Jeffrey C. Long[4]

[1]*Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI, USA*
[2]*Stroke Program, University of Michigan, Ann Arbor, MI, USA*
[3]*Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA*
[4]*Department of Anthropology, University of New Mexico, Albuquerque, NM, USA*

## Summary

To test for and characterize heterogeneity in ancestral contributions to individuals among a population of Mexican American (MA) and non-Hispanic white (NHW) stroke/transient ischemic attack (TIA) cases, data from a community-based stroke surveillance study in south Texas were used. Strokes/TIA cases were identified (2004–2006) with a random sample asked to provide blood. Race-ethnicity was self-reported. Thirty-three ancestry informative markers were genotyped and individual genetic admixture estimated using maximum likelihood methods. Three hypotheses were tested for each MA using likelihood ratio tests: (1) $H_0$: $\mu i = 0$ (100% Native American), (2) $H_0$: $\mu i = 1.00$ (100% European), (3) $H_0$: $\mu i = 0.59$ (average European). Among 154 self-identified MAs, estimated European ancestry varied from 0.26 to 0.98, with an average of 0.59 (SE = 0.014). We rejected hypothesis 1 for every MA and rejected hypothesis 2 for all but two MAs. We rejected hypothesis 3 for 40 MAs (20 < 59%, 20 > 59%). Among 84 self-identified NHWs, the estimated fraction of European ancestry ranged from 0.83 to 1.0, with an average of 0.97 (SE = 0.014). Self-identified MAs, and to a lesser extent NHWs, from an established bi-ethnic community were heterogeneous with respect to genetic admixture. Researchers should not use simple race-ethnic categories as proxies for homogeneous genetic populations when conducting gene mapping and disease association studies in multi-ethnic populations.

Keywords: Stroke, ethnicity, ancestry

## Introduction

Mexican Americans (MA) are the largest subgroup of the largest minority group in the United States. Several health disparities have been identified for the MA population, including increased risk of complex neurologic diseases such as ischemic stroke, compared with non-Hispanic whites (NHW) (Morgenstern et al., 2004). Reasons for these health disparities are largely unknown but are likely multifactorial with environmental, social, and genetic underpinnings.

Characterization of health disparities among MAs has historically relied on self-reported race and ethnicity. Complicating an understanding of the observed health disparities in this population is an incomplete knowledge of what self-reported MA race-ethnicity represents from a genetic perspective. Recent advances in technology allow researchers to quantify race-ethnicity at the molecular level using ancestry informative genomic DNA markers (AIMs). Ancestry informative marker alleles provide quantitative estimates of the proportional contributions of African, European, and Native American ancestors to MA individuals and to the current MA population as a whole. Recent studies utilizing AIMs have reported that Native American ancestors contributed on average 35–52% of the genome to MA individuals (Bonilla et al., 2004a; Salari et al., 2005; Tang et al., 2006; Basu et al., 2008; Kosoy et al., 2009; Shtir et al., 2009).

Although it is possible to quantify ancestry independently of an individual's self-reported information using genetic markers, large-scale epidemiology studies are likely to continue to use self-reported race-ethnicity for several reasons. First, DNA is expensive to collect and genotype relative to acquiring self-report information. Second, self-reported race-ethnicity might indicate disease risk better than genetic ancestry alone because it is a proxy for lifestyle and other social factors as well as genetic inheritance. Third, we are uncertain how well ancestry from different populations serves as a proxy for disease risk, although recent studies have demonstrated associations of ancestry with subclinical cardiovascular disease

*Corresponding author: Lynda D. Lisabeth, 1415, Washington Heights, Ann Arbor, Michigan 48109. Tel: 734 936 9649; Fax: 734 764 3192; E-mail: llisabet@umich.edu

(Wassel et al., 2009) and complex neurologic diseases such as multiple sclerosis (Reich et al., 2005).

An understanding of ancestry at the molecular level in MAs would aid researchers trying to identify reasons for health disparities in this population through epidemiologic research by informing the degree to which self-reported race-ethnicity approximates genetic admixture. The objective of this study was to use previously identified AIMs to characterize and test for the heterogeneity in ancestral contributions to individuals among a population of self-identified MA and NHW stroke or transient ischemic attack (TIA) cases in southeast Texas.

## Methods

Participants in this study consist of $n = 154$ MAs and $n = 84$ NHWs from the Brain Attack Surveillance in Corpus Christi Project, a population-based stroke surveillance study in Nueces County, Texas. Detailed methods for this project have been published (Morgenstern et al., 2004; Smith et al., 2004). Nueces County is located in south Texas on the Gulf Coast, and has a population size of roughly 300,000. MAs comprise the majority of residents, at 56% of the population based on the 2000 U.S. Census. NHWs comprise 38% of the population, and other race-ethnicities comprise the remaining 6%. MAs in this county are primarily second- and third-generation U.S. citizens. We previously reported that 87% of MAs and 93% of NHWs were born in the United States. Mexico was the reported origin of all MA subjects not born in the United States. On average, these individuals had been living in the United States for 60 years (range 19–86 years) (Smith et al., 2003).

Stroke/TIA cases were identified among individuals $\geq$45 years seen at one of seven area hospitals located within Nueces County between June 2004 and June 2006. Cases were also identified through neurologists practicing in Nueces County. Cerebrovascular events were validated by board-certified neurologists based on published criteria and blinded to subjects' ethnicity and age (Asplund et al., 1988). A random sample was asked to participate in an in-person interview and to provide a blood sample. The response rate for the blood draw was 70% with no ethnic difference (MA: 73%, NHW: 65%; $P = 0.07$). All study participants signed an informed consent document and the study was approved by the Institutional Review Boards at the University of Michigan and all local hospitals.

Peripheral venous blood samples were collected by venipuncture from each participant by a trained phlebotomist. Clinical blood samples were sent to the National Institute of Neurological Disorders and Stroke (NINDS) Human Genetics Resource Center DNA and Cell Line Repository (http://ccr.coriell.org/ninds). According to established protocols, genomic DNA was extracted from the whole blood or lymphocyte cell pellets using the Qiagen Autopure method (Qiagen Inc., Valencia, CA, USA). Briefly, cells are lysed by addition of anionic detergent containing RNase and EDTA. After mixing, a salt solution is added and the insoluble cell debris is removed by centrifugation. An equal volume of isopropanol is added to the supernatant and the resulting DNA precipitate is collected by centrifugation. Following a brief rinse with 70% ethanol to remove residual salt, the DNA pellet is solubilized overnight in TE buffer (0.01 M Tris, pH 8.0/0.001 M EDTA). After extraction, the DNA proceeds through several processing steps and must meet specific criteria: 260/280 nm absorbance ratio is between 1.65 and 1.95, concentration is at least 0.1 mg/ml, sample contains less than 0.1 $\mu$g protein per microgram of DNA, and restriction enzyme digestion yields a broad-size distribution of DNA fragments. Amplification by PCR with microsatellite and amelogenin gene-specific primers must also produce amplicon sizes that bin into expected allele sizes, and give fragment peak heights that are at least three-fold above background. The amplified product allele peak heights are within 70% of each other, and there are not more than two allele peaks observed for each microsatellite locus.

Race-ethnicity was self-reported and collected as in the U.S. Census. MA ethnicity was defined as self-reported ethnicity "of Hispanic origin," either with race of "white" or with race "refused." Refused is included as it is common among this population to consider "Hispanic" or "Mexican American" as a race. NHW was defined by a self-reported race of "white" and ethnicity of "not of Hispanic origin." Individuals who reported a race-ethnicity other than MA or NHW were excluded due to small numbers ($n = 30$).

### Ancestry Informative Markers

We analyzed genotypes from 33 genomic single nucleotide polymorphisms (SNPs) dispersed across 17 chromosomes. The nearest physical distance between markers on the same chromosome was >1 million base pairs. This set of markers has been previously identified as being AIMs for estimating European and Native American contributions to admixed populations in the Americas (Seldin et al., 2007; Tian et al., 2007). The absolute value of the difference in allele frequency between two ancestral populations, $\delta$, is a simple measure of the effectiveness of a marker for estimating ancestry. Previous reports have used $\delta >$ 0.3 as the threshold for declaring an SNP as being "ancestry informative" (Bonilla et al., 2004a; Mao et al., 2007; Shtir et al., 2009). All markers used in this study (Table 1) had $\delta$ between Europeans and Native Americans $\geq 0.5$ (median = 0.8). For European and Native American parental population allele frequencies, we used published values (Seldin et al., 2007; Tian et al., 2007). The AIMs employed in this study are useful for analysis of Native American and European ancestral contributions because they show high allele frequency differences between indigenous populations from the Americas and Europe, and low allele frequency differences among local populations on the same continent.

### Genotyping Methods

The 33 AIMs were genotyped using oligonucleotide ligation (Barany, 1991) followed by electrophoresis using four main steps.

**Table 1** Ancestry informative markers (AIMs) and absolute value of the difference in allele frequency($\delta$) between ancestral populations (European and Native American).

| Locus | Allele | European | Native American | $\delta$ | Location |
|---|---|---|---|---|---|
| rs1951936 | A | 0.85 | 0.06 | 0.79 | 10p12 |
| rs11256014 | A | 0.05 | 0.55 | 0.5 | 10p14 |
| rs1638567 | C | 0.06 | 0.64 | 0.58 | 11q13 |
| rs11169154 | A | 0.95 | 0.15 | 0.8 | 12q13 |
| rs7995033 | C | 0.85 | 0.19 | 0.66 | 13q12 |
| rs9319336 | C | 0.04 | 0.89 | 0.85 | 13q12 |
| rs2324596 | C | 0.05 | 0.92 | 0.87 | 13q13 |
| rs1540979 | A | 0.89 | 0.21 | 0.68 | 13q31 |
| rs12102256 | A | 0.91 | 0.05 | 0.86 | 15q14 |
| rs1426654 | A | 1 | 0.05 | 0.95 | 15q21 |
| rs1950030 | A | 0.93 | 0.1 | 0.83 | 15q21 |
| rs6587216 | C | 0.8 | 0.2 | 0.6 | 17p11.2 |
| rs17638989 | C | 0.56 | 0.01 | 0.55 | 19p13.2 |
| rs1931059 | A | 0.19 | 0.77 | 0.58 | 1p35 |
| rs7504 | A | 0.22 | 0.95 | 0.73 | 1p36.1 |
| rs1407434 | C | 0.92 | 0.08 | 0.84 | 1q25 |
| rs6086473 | C | 0.22 | 0.84 | 0.62 | 20p12 |
| rs293553 | A | 0.67 | 0.02 | 0.65 | 20q11.2 |
| rs3755095 | A | 0.92 | 0.05 | 0.87 | 2p12 |
| rs3907854 | C | 0.99 | 0.21 | 0.78 | 2p13 |
| rs3827760 | C | 0.02 | 0.96 | 0.94 | 2q12 |
| rs7432238 | A | 0.93 | 0.1 | 0.83 | 3p24 |
| rs2700394 | C | 0.99 | 0.13 | 0.86 | 3q21 |
| rs2165139 | A | 0.88 | 0.04 | 0.84 | 3q22 |
| rs11725412 | A | 0.06 | 0.99 | 0.93 | 4p14 |
| rs12501010 | C | 0.06 | 0.93 | 0.87 | 4q26 |
| rs262838 | A | 0.92 | 0.21 | 0.71 | 5q36 |
| rs12662498 | A | 0.94 | 0.04 | 0.9 | 6p12 |
| rs9369677 | C | 0.07 | 0.88 | 0.81 | 6p12 |
| rs2439522 | A | 0.88 | 0.26 | 0.62 | 8q22 |
| rs4478653 | C | 0.36 | 1 | 0.64 | 9p21 |
| rs10809782 | A | 0.08 | 0.88 | 0.8 | 9p23 |
| rs7863917 | A | 0.01 | 0.8 | 0.79 | 9q31 |

First, we performed multiplex PCR amplification in batches of approximately 10 loci. Each locus was amplified using locus-specific primers. Second, to enrich the amplicon concentrations for all loci, we reamplified the products of step 1 using primers that are complementary to a universal tag sequence incorporated into the initial locus-specific PCR primer pairs. Third, we ligated fluorescently labeled oligonucleotides specific to SNP alleles to the PCR amplification products. The nucleotide lengths of the ligation oligonucleotide products yield size classes that allow unambiguous separation by gel electrophoresis. Finally, electrophoretic separation and detection of the ligated products occurred using a capillary DNA sequencer.

### Statistical Analysis

We tested deviations from Hardy–Weinberg equilibrium using likelihood ratio statistics, and measured the degree of departure from equilibrium using the within-locus intraclass allelic correlation, $F_1$, as defined by Risch et al. (2009). We tested for deviations from "linkage equilibrium" between all pairs of loci using $\chi^2$ tests based on the $r^2$ statistic (Weir, 1996). We estimated individual genetic admixture for each participant using the method of maximum likelihood (Chakraborty, 1986) based on two parental populations, European Americans and Native Americans. For each person, we evaluated the likelihood function $L(\mu_i)$, where $\mu_i$ represents the fraction of ancestors of that person who were of European origin. By this method, the estimate of individual ancestry is the value $\hat{\mu}_i$ that maximizes the likelihood function. For each estimate, $\hat{\mu}_i$, we estimated the standard error of the estimate $s_{\hat{\mu}_i}$ from Fisher's information criterion $I_{\hat{\mu}_i} = -(d^2/d\mu^2)\ln[L(\mu_i)]$ using the formula $s_{\hat{\mu}_i} = \sqrt{1/I_{\hat{\mu}_i}}$ (Edwards, 1992). We estimated average admixture for participants in each race-ethnic group using two methods: (1) the average of individual estimates described above and (2) the method of weighted least squares as implemented in ADMIX (Long, 1991).

Individual ancestry estimates from genetic markers have high standard errors that lead to wide confidence intervals. This presents two challenges: (1) showing that an individual deviates statistically from a predetermined reference point, such as 100% ancestry from either, or both, of the putative parental populations and (2) showing that individuals in a sample are heterogeneous, with respect to their true proportions of ancestry from the putative parental populations. We used the following likelihood ratio statistic to address these problems:

$$G = -2\left[\left(\ln L\left(\mu_1\right) - \ln L\left(\hat{\mu}_i\right)\right)\right],$$

where $\mu_1$ is a specified fraction of European ancestry and $\hat{\mu}_i$ is the ancestry fraction that maximizes the likelihood function for the $i$th individual. The null hypothesis is $H_0$: $\mu_i = \mu_1$. The $G$ is distributed asymptotically as a $\chi^2$ random variable with degrees of freedom one less than the number of parental populations (Edwards, 1992).

Finally, we compared the proportion of European ancestry with age at stroke onset and having a high school education using correlation coefficients and $t$-tests separately among MAs and NHWs.

## Results

Among the 238 stroke/TIA cases, mean age was 69 years ($\sigma$ = 13) and 49% were female. MAs were younger ($P < 0.0001$) and less likely to have a high school education ($P < 0.0001$) than NHWs (Table 2). Among the 154 participants of self-reported MA race-ethnicity, the range of estimated fraction of European ancestry was 0.259–0.975 (Table 3). The average of individual European ancestry estimates was 0.591 $\pm$ 0.014. Using weighted least squares method, we estimated the

**Table 2** Sociodemographic characteristics by self-reported race-ethnicity, Mexican American and non-Hispanic white ($n = 238$).

| Variable | Mexican American ($n = 154$) | Non-Hispanic white ($n = 84$) |
|---|---|---|
| Mean age (SD) | 66.3 (12.8) | 73.4 (12.6) |
| % Female ($n$) | 50.0 (77) | 47.6 (40) |
| % High school education ($n$) | 45.5 (70) | 79.8 (67) |

fraction of European ancestry for the group to be $0.589 \pm 0.011$, which agrees well with the average of individual estimates. Among the 84 participants of self-reported NHW race-ethnicity, the estimated fraction of European ancestry ranged from 0.827 to 1.00. The average of individual European ancestry was $0.968 \pm 0.014$. Using the weighted least squares method, we estimated the fraction of European ancestry for the group to be $0.963 \pm 0.014$, which also agrees well with the average of individual estimates for NHWs.

The next step was to document heterogeneity in ancestral contributions to the individual MA and NHW participants. We tested the following three hypotheses for each MA case, $H_0: \mu_i = 0$, $H_0: \mu_i = 1.00$, and $H_0: \mu_i = 0.591$. The first hypothesis establishes whether an MA participant differs significantly from a person who has 100% Native American ancestry. The second hypothesis establishes whether an MA participant differs significantly from a person who has 100% European Ancestry. The third hypothesis establishes whether an MA participant differs significantly from the average European ancestry for the group as a whole (i.e., 59% European). We rejected the "100% Native American ancestry" hypothesis for every MA case, and we rejected the "100% European ancestry" hypothesis for all but two MA cases. We rejected the hypothesis that a self-reported MA had ancestry consistent with the MA population (59% European ancestry) for 40 of the 154 MA cases. Twenty individuals were significantly higher and 20 significantly lower than the mean European ancestry (Fig. 1).

Given that there is significant heterogeneity in ancestral contributions to MA individuals, we expect to see an excess of homozygosity within loci, and linkage disequilibrium among pairs of loci (even unlinked loci). Our results confirm these expectations. At $\alpha = 0.05$ or less, we found a significant excess of homozygotes at six (18%) loci. The mean intraclass correlation, $F_{1,}$ was 0.035. At $\alpha = 0.05$ or less, we found significant linkage disequilibrium between 84 (16%) locus pairs.

To establish the extent of ancestral heterogeneity in the NHW sample, we tested the hypothesis that each person has 100% European ancestry, that is, $H_0: \mu_i = 1.00$. We rejected this hypothesis for 15 cases (Fig. 1). To follow-up on this result, we also tested the following two hypotheses: $H_0: \mu_i = 0.94$ and $H_0: \mu_i = 0.591$. The first hypothesis establishes whether an NHW participant differs significantly from a person who has the equivalent of one Native American great-great-grandparent. The second hypothesis establishes whether an NHW participant differs significantly from the average European ancestry for the MA sample. We rejected the $\mu_i = 0.94$ hypothesis for three NHWs, each of whom was estimated to have 100% European ancestry. We rejected the $\mu_i = 0.591$ hypothesis for all NHWs.

Given the small degree of heterogeneity in ancestral contributions to NHW individuals, we tested for excess of homozygosity within loci and for linkage disequilibrium among pairs of loci. At $\alpha = 0.05$ or less, we found a significant excess of homozygotes at three loci (9%). The mean intraclass correlation was 0.012. At $\alpha = 0.05$ or less, we found significant linkage disequilibrium between 42 locus pairs (8%) in NHWs.

Following further analysis, European ancestry was not found to be associated with age at stroke among MAs ($P = 0.93$) or NHWs ($P = 0.16$). European ancestry was also not associated with having a high school education, a proxy for socioeconomic status, among MAs ($P = 0.48$) or NHWs ($P = 0.93$).

## Discussion

Today many distinct populations live in the Americas with ancestry mixed between people who lived in Africa, Europe, or the Americas before the colonial era. Although many people refer collectively to mixed populations in the Americas as Hispanic, Latino, or Mestizo, geneticists recognize that these

**Table 3** Average contributions of European and Native American ancestry by self-reported race-ethnicity, Mexican American and non-Hispanic white ($n = 238$).

| | Mexican American ($n = 154$) | | | Non-Hispanic white ($n = 84$) | | |
|---|---|---|---|---|---|---|
| | European | Native American | SE | European | Native American | SE |
| WLS | 0.589 | 0.411 | 0.011 | 0.968 | 0.032 | 0.014 |
| Average $\mu_i$ | 0.591 | 0.409 | 0.014 | 0.963 | 0.037 | 0.005 |

WLS = weighted least squares, SE = standard error.

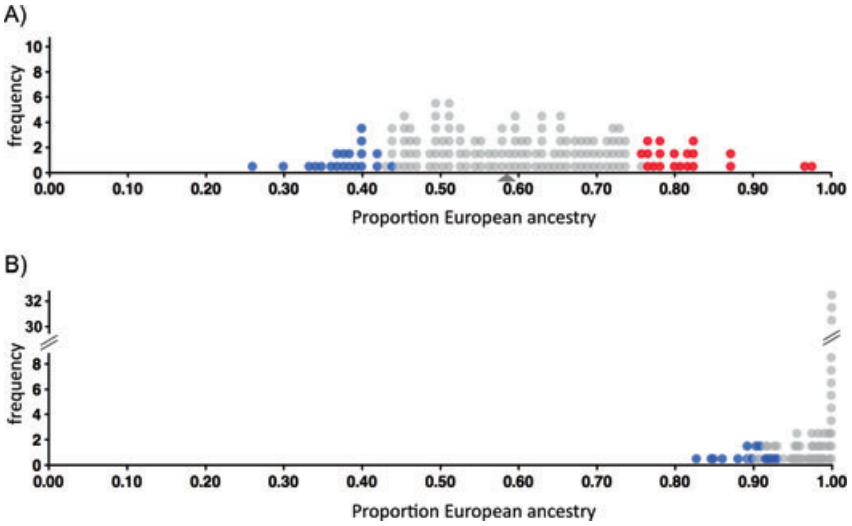


**Figure 1** (A) Proportion of European ancestry in the Mexican American sample ($n = 154$). Black triangle indicates mean European Ancestry for the Mexican American sample (59%). For Mexican Americans, blue dots represent individuals that had statistically lower European ancestry than the mean value ($n = 20$). Red dots represent individuals that had statistically greater European ancestry than the mean value ($n = 20$). (B) Proportion of European ancestry in the non-Hispanic white (NHW) sample ($n = 84$). For NHWs, blue dots represent individuals that had statistically lower European ancestry than 100% ($n = 15$). In total, 32 NHWs had genotype results with 100% European ancestry (32/84 = 38%).

groups have distinct gene pools that trace different proportions of ancestors to each of the three continental regions. Self-identified MAs living in different cities typically have 35–50% Native American ancestry and a trace component, 4–6%, of African ancestry, whereas the self-identified Puerto Rican population as a whole has 15–18% Native American ancestry but a more substantial component of African ancestry (~20%) (Tseng et al., 1998; Bonilla et al., 2004a; Collins-Schramm et al., 2004; Salari et al., 2005; Basu et al., 2008; Risch et al., 2009). In this study, where we compared the degree to which a sample of self-identified MAs approximated a random mating population in genetic equilibrium, we found that MAs were a heterogeneous group regarding genetic ancestry, with individual estimates of Native American ancestry ranging from 2% to 74%. While our estimated average of 41% Native American ancestry is consistent with recently reported estimates (Bonilla et al., 2004a; Salari et al., 2005; Tang et al., 2006; Basu et al., 2008; Kosoy et al., 2009; Risch et al., 2009; Shtir et al., 2009), we found that more than a quarter of the MA cases were significantly different from the average European ancestry in the MA population as a whole.

We also found some heterogeneity in the ancestry of NHWs, with individual estimates of Native American ancestry ranging from 0% to 17%. While we found that 18% of NHWs had significantly less than 100% European ancestry, the average Native American ancestry in the NHW sample was roughly equivalent to one Native American great-great-grandparent. Thus, we should recognize that people who consider themselves ethnically NHW may have ancestors who were Native American. Direct unions between Native Americans and NHWs may have introduced this ancestry, but it is also likely that unions between MA and NHWs introduced this Native American ancestry indirectly. Genetic marker analysis cannot resolve this issue, but questionnaires could provide some information about the patterns of gene flow. No matter the origin of Native American ancestry in the NHW sample, it is likely the cause of departures from Hardy–Weinberg equilibrium and linkage equilibrium in the sample. Our results confirm that individuals within both Hispanic and non-Hispanic white U.S. Census categories are heterogeneous with respect to European and Native American ancestry. While neither Census group in southeastern Texas constituted a genetic population, the NHW group was far more ancestrally homogeneous than the MA group.

Heterogeneity of population ancestry in other Hispanic communities has been reported including populations in New York (Bonilla et al., 2004b), southern Colorado (Bonilla et al., 2004a), the state of Guerrero, Mexico (Bonilla et al., 2005), Mexico City and San Francisco (Risch et al., 2009),

Population genetic principles show that random mating causes variation in ancestry to decrease from one generation to the next due to segregation and recombination. On this basis we expect that individuals in well-established admixed populations will be homogenous with respect to the composition of ancestral populations. Differences in ancestral contributions to individuals demonstrate some departure from random mating. Risch and colleagues recently found evidence that Mexicans in Mexico City and MAs in San Francisco prefer mates with similar ancestries (Risch et al., 2009). This assortative mating is one mechanism that can maintain interindividual heterogeneity in contributions from ancestral populations. Although Risch et al. did not formally test for heterogeneity, they observed a wide spread of Native American ancestry in their population with a mean of 0.44 ($\sigma = 0.14$), similar to our results. Also, paralleling our finding of 15% of unlinked locus pairs in linkage disequilibrium in our MA population, they reported that 10–16% of unlinked locus pairs in their samples were in linkage disequilibrium. In San Francisco, the mean correlation of alleles within loci was 0.015, whereas we observed 0.035. Together the results of these studies suggest that assortative mating may partially explain the observed interindividual heterogeneity in ancestry estimates demonstrated in MAs.

One-way gene flow from NHWs to MAs is another mechanism that could maintain Hardy–Weinberg and linkage disequilibrium in the MA population. In addition, recent migrants from Mexico may consist of individuals with lower European ancestry than the second- and third-generation U.S. Citizens that make up 87% of our Nueces County MA sample. Finally, differences in socioeconomic status may partially explain the observed heterogeniety if individuals of lower socioeconomic status have higher Native American ancestry. However, our analysis considering the association between ancestry and having a high school education, a proxy for socioeconomic status, did not support this hypothesis. We note that the mechanisms that can maintain heterogeneity in ancestry are not mutually exclusive. We currently lack the necessary data that could further distinguish among these possibilities.

Our findings are important to epidemiological studies because they show that researchers cannot use simple race-ethnic categories designed for the U.S. Census and other government purposes as proxies for homogeneous genetic populations when conducting gene mapping and disease association studies. Heterogeneity in ancestral contributions to individuals creates correlations among alleles within loci and among loci. Both sources of correlation can strengthen the association between linked genetic markers and complex diseases such as ischemic stroke, but correlations that result from nonrandom mating in populations are not always beneficial because they can create spurious associations between genetic markers and disease. Special care is needed to sort out the true nature of correlations in complex populations such as we have demonstrated for MAs in Texas.

The population studied consisted of stroke/TIA cases. Previous research in the study community has shown that stroke disproportionately affects MAs, especially at younger ages (Morgenstern et al., 2004). We have also demonstrated that having a first-degree relative with stroke increases one's risk of stroke particularly in MAs (Lisabeth et al., 2008). Siblings of MA ischemic stroke/TIA cases have roughly double the stroke risk compared to what would be expected based on national estimates of stroke prevalence in MAs. These findings together with the current finding of ancestral heterogeneity in the MA population suggest that ischemic stroke may be a suitable phenotype for admixture disequilibrium mapping to identify stroke susceptibility genes in this population.

Limitations of this work warrant discussion. Individual ancestry estimates were derived from 33 AIMs, which is a smaller set than other recent reports, which have characterized ancestry in MAs. This may have led to somewhat larger standard errors around our estimates. However, the AIMs used for this study were chosen such that the $\delta$ between Europeans and Native Americans was $\geq 0.5$. This criterion is stricter than most previous reports. More importantly, these 33 AIMs provided enough information for us to reject our null hypotheses of ancestry homogeneity in both the MA and NHW samples, and to show evidence for admixture-related departures from Hardy–Weinberg equilibrium and linkage equilibrium. Thus, they provide sufficient information to achieve the study's goals.

Our model for genetic admixture constructs MA ancestry in Nueces County, Texas using two parental populations, Europeans and Native Americans. However, several reports indicate that most MA populations also harbor a small proportion of African ancestry (~5%). We decided against a three-population model because the fraction of African ancestry is likely to be low and our AIMs are powerful only for distinguishing Native American ancestry from European ancestry. This is consistent with the typical marker selection strategy for admixture analyses in MAs based on the two populations contributing the most ancestry to MAs (Tian et al., 2007). In addition, African ancestry is unlikely to change our main findings that neither MAs nor NHWs in Nueces County, Texas are homogeneous populations with respect to their ancestral compositions.

The study population was limited to individuals with stroke/TIA. It is possible that this disease outcome influenced the estimates of genetic admixture, but it is unlikely. If a major gene contributes to stroke/TIA in the MA and NHW population, then it can influence admixture estimates through linkage disequilibrium with our AIMs, but our AIMs are unlinked and this would prevent linkage disequilibrium with a gene for stroke from having a large influence on

ancestry estimates. Moreover, since all of our participants in this study are stroke/TIA patients, the disease outcome cannot account for our major finding, that is, both the MA and NHW samples are heterogeneous with respect to Native American and European ancestry.

## Conclusion

We observed that self-identified MAs from a biethnic U.S. community were heterogeneous with respect to genetic admixture, with estimates of Native American ancestry ranging considerably among individuals. Our findings suggest that researchers should not use simple self-reported race-ethnic categories as proxies for homogeneous genetic populations when conducting gene mapping and disease association studies in this growing segment of the population. However, self-reported race-ethnicity is a proxy for lifestyle and social factors and thus retains importance in the study of complex diseases such as stroke.

## Acknowledgements

This study was funded by NIH K23 NS050161, R01 NS38916, and 5P30 AG024824-05. This study used samples from the NINDS Human Genetics Resource Center DNA and Cell Line Repository (http://ccr.coriell.org/ninds). NINDS Repository sample numbers corresponding to the samples used are:

## References

Asplund, K., Tuomilehto, J., Stegmayr, B., Wester, P. O. & Tunstall-Pedoe, H. (1988) Diagnostic criteria and quality control of the registration of stroke events in the MONICA project. *Acta Med Scand* **728**, 26–39.

Barany, F. (1991) The ligase chain reaction in a PCR world. *PCR Methods Appl* **1**, 5–16.

Basu, A., Tang, H., Zhu, X., Gu, C. C., Hanis, C., Boerwinkle, E. & Risch, N. (2008) Genome-wide distribution of ancestry in Mexican Americans. *Human Genet* **124**, 207–214.

Bonilla, C., Gutierrez, G., Parra, E. J., Kline, C. & Shriver, M. D. (2005) Admixture analysis of a rural population of the state of Guerrero, Mexico. *Am J Phys Anthropol* **128**, 861–869.

Bonilla, C., Parra, E. J., Pfaff, C. L., Dios, S., Marshall, J. A., Hamman, R. F., Ferrell, R. E., Hoggart, C. L., Mckeigue, P. M. & Shriver, M. D. (2004a) Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. *Ann Hum Genet* **68**, 139–153.

Bonilla, C., Shriver, M. D., Parra, E. J., Jones, A. & Fernandez, J. R. (2004b) Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York city. *Hum Genet* **115**, 57–68.

Chakraborty, R. (1986) Gene admixture in human populations – Models and predictions. *Yearb Phys Anthropol* **29**, 1–43.

Collins-Schramm, H. E., Chima, B., Morii, T., Wah, K., Figueroa, Y., Criswell, L. A., Hanson, R. L., Knowler, W. C., Silva, G., Belmont, J. W. & Seldin, M. F. (2004) Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. *Hum Genet* **114**, 263–271.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ND11106 | ND11487 | ND11653 | ND11806 | ND12197 | ND12207 | ND12208 | ND12415 | ND12536 | ND12538 |
| ND12539 | ND12624 | ND12715 | ND12784 | ND13111 | ND13128 | ND13388 | ND13389 | ND13430 | ND13431 |
| ND13465 | ND13518 | ND13600 | ND13885 | ND13886 | ND13887 | ND13994 | ND13995 | ND13996 | ND14146 |
| ND14229 | ND14279 | ND14376 | ND14377 | ND14378 | ND14632 | ND14809 | ND14810 | ND14812 | ND14849 |
| ND14894 | ND14895 | ND14896 | ND14897 | ND14898 | ND14930 | ND15032 | ND15133 | ND15190 | ND15223 |
| ND15335 | ND15522 | ND15524 | ND15525 | ND15601 | ND15602 | ND15603 | ND15628 | ND15629 | ND15630 |
| ND15631 | ND15649 | ND15728 | ND15757 | ND15758 | ND15786 | ND15787 | ND15792 | ND15793 | ND15810 |
| ND15849 | ND15850 | ND15851 | ND15988 | ND16030 | ND16076 | ND16077 | ND16079 | ND16081 | ND16092 |
| ND16237 | ND16238 | ND16239 | ND16240 | ND16242 | ND16243 | ND16245 | ND16310 | ND16311 | ND16350 |
| ND16351 | ND16353 | ND16354 | ND16355 | ND16404 | ND16535 | ND16536 | ND16537 | ND16538 | ND16564 |
| ND16565 | ND16566 | ND16599 | ND16600 | ND16602 | ND16635 | ND16636 | ND16637 | ND16669 | ND19236 |
| ND19237 | ND19238 | ND19240 | ND19241 | ND19242 | ND19288 | ND19330 | ND19487 | ND19488 | ND19489 |
| ND19491 | ND19532 | ND19533 | ND19603 | ND19632 | ND19740 | ND19792 | ND19793 | ND19797 | ND19854 |
| ND19856 | ND19858 | ND19861 | ND19922 | ND19981 | ND19985 | ND19986 | ND20030 | ND20099 | ND20148 |
| ND20195 | ND20196 | ND20197 | ND20394 | ND20396 | ND20397 | ND20398 | ND20399 | ND20400 | ND20401 |
| ND20402 | ND20469 | ND20514 | ND20515 | ND11828 | ND12045 | ND12414 | ND12535 | ND12537 | ND12716 |
| ND12717 | ND12847 | ND12978 | ND12979 | ND13108 | ND13209 | ND13599 | ND13766 | ND13858 | ND13859 |
| ND14033 | ND14145 | ND14147 | ND14424 | ND14633 | ND14687 | ND14688 | ND14689 | ND14746 | ND14807 |
| ND15132 | ND15224 | ND15521 | ND15523 | ND15604 | ND15648 | ND15675 | ND15676 | ND15677 | ND15719 |
| ND15756 | ND15784 | ND15791 | ND16029 | ND16032 | ND16033 | ND16080 | ND16093 | ND16235 | ND16236 |
| ND16244 | ND16309 | ND16312 | ND16329 | ND16331 | ND16352 | ND16403 | ND16598 | ND16667 | ND16668 |
| ND19179 | ND19235 | ND19239 | ND19283 | ND19284 | ND19285 | ND19286 | ND19328 | ND19406 | ND19407 |
| ND19423 | ND19486 | ND19490 | ND19602 | ND19791 | ND19853 | ND19855 | ND19857 | ND19919 | ND19920 |
| ND19921 | ND19987 | ND20033 | ND20145 | ND20147 | ND20245 | ND20440 | ND20516 | | |

Edwards, A. (1992) *Likelihood*. Baltimore, MD: Johns Hopkins University Press.

Kosoy, R., Nassir, R., Tian, C., White, P. A., Butler, L. M., Silva, G., Kittles, R., Alarcon-Riquelme, M. E., Gregersen, P. K., Belmont, J. W., De La Vega, F. M. & Seldin, M. F. (2009) Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat* **30**, 69–78.

Lisabeth, L. D., Peyser, P. A., Long, J. C., Majerisk, J. J., Smith, M. A. & Morgenstern, L. B. (2008) Stroke among siblings in a biethnic community. *Neuroepidemiology* **31**, 33–38.

Long, J. C. (1991) The genetic structure of admixed populations. *Genetics* **127**, 417–428.

Mao, X., Bigham, A. W., Mei, R., Gutierrez, G., Weiss, K. M., Brutsaert, T. D., Leon-Velarde, F., Moore, L. G., Vargas, E., Mckeigue, P. M., Shriver, M. D. & Parra, E. J. (2007) A genomewide admixture mapping panel for Hispanic/Latino populations. *Am J Hum Genet* **80**, 1171–1178.

Morgenstern, L. B., Smith, M. A., Lisabeth, L. D., Risser, J. M., Uchino, K., Garcia, N., Longwell, P. J., Mcfarling, D. A., Akuwumi, O., Al-Wabil, A., Al-Senani, F., Brown, D. L. & Moye, L. A. (2004) Excess stroke in Mexican Americans compared with non-Hispanic Whites: The Brain Attack Surveillance in Corpus Christi project. *Am J Epidemiol* **160**, 376–383.

Reich, D., Patterson, N., De Jager, P. L., Mcdonald, G. J., Waliszewska, A., Tandon, A., Lincoln, R. R., Deloa, C., Fruhan, S. A., Cabre, P., Bera, O., Semana, G., Kelly, M. A., Francis, D. A., Ardlie, K., Khan, O., Cree, B. A., Hauser, S. L., Oksenberg, J. R. & Hafler, D. A. (2005) A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat Genet* **37**, 1113–1118.

Risch, N., Choudhry, S., Via, M., Basu, A., Sebro, R., Eng, C., Beckman, K., Thyne, S., Chapela, R., Rodriguez-Santana, J. R., Rodriguez-Cintron, W., Avila, P. C., Ziv, E. & Gonzalez Burchard, E. (2009) Ancestry-related assortative mating in Latino populations. *Genome Biol* **10**, R132.

Salari, K., Choudhry, S., Tang, H., Naqvi, M., Lind, D., Avila, P. C., Coyle, N. E., Ung, N., Nazario, S., Casal, J., Torres-Palacios, A., Clark, S., Phong, A., Gomez, I., Matallana, H., Perez-Stable, E. J., Shriver, M. D., Kwok, P. Y., Sheppard, D., Rodriguez-Cintron, W., Risch, N. J., Burchard, E. G. & Ziv, E. (2005) Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. *Genetic Epidemiol* **29**, 76–86.

Seldin, M. F., Tian, C., Shigeta, R., Scherbarth, H. R., Silva, G., Belmont, J. W., Kittles, R., Gamron, S., Allevi, A., Palatnik, S. A., Alvarellos, A., Paira, S., Caprarulo, C., Guilleron, C., Catoggio, L. J., Prigione, C., Berbotto, G. A., Garcia, M. A., Perandones, C. E., Pons-Estel, B. A. & Alarcon-Riquelme, M. E. (2007) Argentine population genetic structure: large variance in Amerindian contribution. *Am J Phys Anthropol* **132**, 455–462.

Shtir, C. J., Marjoram, P., Azen, S., Conti, D. V., Le Marchand, L., Haiman, C. A. & Varma, R. (2009) Variation in genetic admixture and population structure among Latinos: The Los Angeles Latino eye study (LALES). *BMC Genet* **10**, 71.

Smith, M. A., Risser, J. M., Lisabeth, L. D., Moye, L. A. & Morgenstern, L. B. (2003) Access to care, acculturation, and risk factors for stroke in Mexican Americans: The Brain Attack Surveillance in Corpus Christi (BASIC) project. *Stroke* **34**, 2671–2675.

Smith, M. A., Risser, J. M., Moye, L. A., Garcia, N., Akiwumi, O., Uchino, K. & Morgenstern, L. B. (2004) Designing multi-ethnic stroke studies: The Brain Attack Surveillance in Corpus Christi (BASIC) project. *Ethn Dis* **14**, 520–526.

Tang, H., Jorgenson, E., Gadde, M., Kardia, S. L., Rao, D. C., Zhu, X., Schork, N. J., Hanis, C. L. & Risch, N. (2006) Racial admixture and its impact on BMI and blood pressure in African and Mexican Americans. *Hum Genet* **119**, 624–633.

Tian, C., Hinds, D. A., Shigeta, R., Adler, S. G., Lee, A., Pahl, M. V., Silva, G., Belmont, J. W., Hanson, R. L., Knowler, W. C., Gregersen, P. K., Ballinger, D. G. & Seldin, M. F. (2007) A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. *Am J Hum Genet* **80**, 1014–1023.

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 Anthropol* **106**, 361–371.

Wassel, C. L., Pankow, J. S., Peralta, C. A., Choudhry, S., Seldin, M. F. & Arnett, D. K. (2009) Genetic ancestry is associated with subclinical cardiovascular disease in African-Americans and Hispanics from the multi-ethnic study of atherosclerosis. *Circulation* **2**, 629–636.

Weir, B. (1996) *Genetic data analysis II: methods for discrete population genetic data*. Sunderland, MA: Sinauer Associates.

*Received*: 8 November 2010
*Accepted*: 17 March 2011