Published in final edited form as:
Hum Genet. 2009 March ; 125(2): 199–209. doi:10.1007/s00439-008-0612-7.

Population Admixture Associated With Disease Prevalence in the Boston Puerto Rican Health Study

Chao-Qiang Lai[1], **Katherine L. Tucker**[2], **Shweta Choudhry**[3], **Laurence D. Parnell**[1], **Josiemer Mattei**[1], **Bibiana García-Bailo**[1], **Kenny Beckman**[4], **Esteban González Burchard**[3,5], and **José M. Ordovás**[1]

[1]Nutrition and Genomics Laboratory, JM-USDA Human Nutrition Research Center on Aging at Tufts University, Boston MA

[2]Dietary Assessment and Epidemiology Research Program, JM-USDA Human Nutrition Research Center on Aging at Tufts University, Boston MA

[3]Department of Medicine, University of California, San Francisco, San Francisco, CA

[4]Functional Genomics Core, Children's Hospital Oakland Research Institute, Oakland, CA. USA

[5]Department of Biopharmaceutical Sciences, University of California, San Francisco, San Francisco, CA

Abstract

Older Puerto Ricans living in the continental U.S. suffer from higher rates of diabetes, obesity, cardiovascular disease and depression compared to non-Hispanic White populations. Complex diseases, such as these, are likely due to multiple, potentially interacting, genetic, environmental and social risk factors. Presumably, many of these environmental and genetic risk factors are contextual. We reasoned that racial background may modify some of these risk factors and be associated with health disparities among Puerto Ricans. The contemporary Puerto Rican population is genetically heterogeneous and originated from three ancestral populations: European settlers, native Taíno Indians, and West Africans. This rich mixed ancestry of Puerto Ricans provides the intrinsic variability needed to untangle complex gene-environment interactions in disease susceptibility and severity. Herein, we determined whether a specific ancestral background was associated with either of four major disease outcomes (diabetes, obesity, cardiovascular disease and depression). We estimated the genetic ancestry of 1129 subjects from the Boston Puerto Rican Health Study, based on genotypes of 100 ancestry informative markers (AIMs). We examined the effects of ancestry on tests of association between single AIMs and disease traits. The ancestral composition of this population was 57.2% European, 27.4% African, and 15.4% Native American. African ancestry was negatively associated with type 2 diabetes and cardiovascular disease, and positively correlated with hypertension. It is likely that the high prevalence rate of diabetes in Africans, Hispanics, and Native Americans is not due to genetic variation alone, but to the combined effects of genetic variation interacting with environmental and social factors.

Keywords

population admixture; Puerto Ricans; ancestry informative markers

Correspondence to C.Q. Lai, Nutrition and Genomics Laboratory, JM-USDA Human Nutrition Research Center on Aging at Tufts University, 711 Washington St, Boston, MA 02111. Phone: 617-556-3206. E-mail chaoqiang.lai@ars.usda.gov.

Introduction

The history of the contemporary Puerto Rican population is rich and dates to the time of Christopher Columbus. Approximately 60,000-600,000 Taíno Indians lived on the island of Puerto Rico when Christopher Columbus and his crew arrived in 1493 (Fernandez-Mendez et al. 1970). Most of these Spanish explorers were men, who later settled and fathered the first generation of the newly admixed population between Europeans and Taíno in Puerto Rico. In 1508, the first African slaves were brought to the island (Alvarez-Nazario, 1974) and began to intermix. Due to conflict between Taíno Indians and new settlers, along with diseases and starvation, the Taíno, as a separate population, completely disappeared from Puerto Rico by 1524 (Rouse 1992). Thus, the contemporary population of Puerto Ricans is based on different phases of intermixing of the three constituent populations (Taíno Indians, Europeans, and Africans). After 1917, when Puerto Ricans became US citizens, they began to relocate to the US mainland, settling in major cities, including New York, Philadelphia, Chicago, Orlando, Miami, and Boston. By 2006, approximately 4 million Puerto Ricans, close to half of the entire Puerto Rican population, lived on the US mainland (US Census Bureau 2006).

Older Puerto Ricans living in the continental U.S. suffer from higher rates of diabetes, obesity, cardiovascular disease and depression compared to non-Hispanic White populations (Tucker 2005; Lai et al. 2008). Multiple factors contribute to the disproportionate health burden of elderly Puerto Ricans living in Massachusetts. First, large proportions of this group live below the poverty line, mainly in crowded, urban environments (Falcon and Tucker 2000). Their economic circumstances may limit their access to both health protective goods and health care. Second, as with many groups in such circumstances, their health-related behaviors may be inadequate, including very low levels of physical activity and poor dietary habits, which are likely contributing factors to their high prevalence of obesity and diabetes (Bermudez et al. 2000; Tucker et al. 2000a; Tucker et al. 2000b; Lai et al. 2008). However, these factors alone do not fully explain the excess prevalence of chronic disease and physical disability in this population, and it is likely that genetic components may place them at excess risk. Therefore, identifying both genetic and environmental factors that contribute to the health disparities in Puerto Ricans is needed for the development of effective strategies to prevent age-related diseases in this vulnerable population.

Genetic association studies of health outcomes in racially mixed populations (admixture) can be complicated. Admixture can result in genetic subgroups within a population. The existence of genetic subgroups or substructure in a population may lead to spurious associations if the subgroups are not equally represented in cases and controls (Li 1969). For example, if one subgroup has a higher prevalence of disease, then this subgroup will likely be over-represented among cases compared to controls. Therefore, any genetic variant (allele) that has a higher frequency in that subgroup may appear to be falsely associated with the disease. Theoretically, if cases and controls are matched by their genetic ancestry, then the confounding due to population stratification should be eliminated (Cardon and Palmer 2003). In practice, however, it may not be possible to precisely match cases and controls based on self-reported ancestry, especially in admixed populations in which individuals may not be completely aware of their precise ancestry (Ziv and Burchard 2003). To overcome the problem associated with population substructure, genomic control (Devlin and Roeder 1999), ancestry index (Pritchard et al. 2000,Falush et al. 2003) and principal component analysis (PCA, Price et al. 2006;Patterson et al. 2006;Paschou et al. 2007) have been developed and applied to association studies of admixed populations (Bonilla et al. 2004;Salari et al. 2005;Choudhry et al. 2006). To estimate individual ancestry, several panels of ancestry informative markers (AIMs) have been developed for Hispanic populations including those from Mexico and Puerto Rico (Salari et al. 2005;Halder et al. 2008). For the Puerto Rican population, we initially used a panel of 44 AIMs to estimate ancestry (Salari et al. 2005;Choudhry et al. 2006). To improve upon this

panel of AIMs, we determined the optimal number and the precise type of SNPs required to estimate ancestral proportions in Latino populations by using a combination of simulated and applied data: a panel of 100 AIMs were necessary to accurately estimate ancestral proportions in Latino populations (Tsai et al 2005). To identify Latino specific SNPs, we genotyped Puerto Rican founding populations using the Affymetrix 100K GeneChip (Choudhry et al. 2008). In this study, we have applied those results to an elderly Puerto Rican cohort recruited from the Boston metropolitan area. Specifically, we genotyped 100 Latino specific AIMs in the Boston Puerto Rican Health Study (BPRHS) population, estimated ancestry, and determined whether ancestry correlates with disease status and affects detection of association between genetic markers and disease phenotypes.

Methods and Materials

Recruitment of the study population

The study population comprised 337 men and 792 women who were self-identified Puerto Ricans living in the greater Boston metropolitan area with complete data records for demographic and biochemical characteristics, and for whom DNA samples were available. These subjects were recruited by investigators from the Boston Puerto Rican Center for Population Health and Health Disparities to participate in a longitudinal cohort study on stress, nutrition, health and aging - the Boston Puerto Rican Health Study (Tucker 2005), http://hnrcwww.hnrc.tufts.edu/departments/labs/prchd/. Participants were recruited primarily through door-to-door enumeration (approximately 84%), with additional participants identified randomly during major citywide activities (8%) or through referral from community organizations or contact through the media or flyers (8%). Enumeration was conducted using year 2000 Census blocks identified as containing 10 or more Hispanic individuals. After block enumeration, households with at least one Puerto Rican adult aged 45 to 75 years at the time of the first interview were identified and selected. All blocks were visited three to six times, including on weekend days. One qualified individual per household was invited to participate. Those who were unable to answer questions due to serious health conditions and/or advanced dementia were excluded. Of those invited, more than 85% agreed to participate.

Data collection and variable definition

Information on socio-demographics, health status, history and behavior, was collected by home interview, administered by bilingual interviewers. Cardiovascular disease (CVD) was defined as a positive response to the question “Have you ever been told by a physician that you have heart disease” or to similar questions on heart attack or stroke. Anthropometric and blood pressure measurements were collected using standard methods. Tobacco and alcohol use were determined by questionnaire and defined for this analysis as current, past or never smokers or drinkers. Using American Diabetes Association (ADA) criteria, subjects were classified as having type 2 diabetes mellitus (T2DM) if the fasting plasma glucose concentration was ≥ 126 mg/dl or use of insulin or diabetes medication was reported (American Diabetes Association 2007). Plasma glucose concentration was measured from blood collected from subjects after an overnight fast. Subjects with BMI ≥ 25 kg/m^2 were defined as being overweight, whereas a BMI ≥ 30 kg/m^2 was defined as being obese. Blood pressure was obtained at three points during the home interview and the average of the latter two measures was used. Hypertension was defined as blood pressure > 140 mm Hg for systolic and/or > 90 mm Hg diastolic or medication use for hypertension. Depressive symptomatology was assessed using the Center for Epidemiologic Studies Depression Scale (CES-D) (Moscicki et al. 1989; Radloff 1977). Physical activity was estimated as physical activity score based on the Paffenbarger questionnaire of the Harvard Alumni Activity Survey (Lee and Paffenbarger 1998). The Physical Activity Score is constructed by weighting time spent in various activities by factors that parallel increasing oxygen consumption rates associated with physical activity intensity,

and is categorized as follows: 0-29—sedentary, 30-39—light activity, 40-49—moderate activity and greater than 49—heavy activity. Socioeconomic status was determined based on total household income. Upper socioeconomic status was defined as total household income equal to or above $25,000, whereas lower socioeconomic status was defined as total household income below $25,000.

Selection of AIMs

One hundred AIMs were selected for this analysis and the process by which those AIMs were selected were described in Choudhry et al 2008. Briefly, AIMs were selected from genotype data generated using the Affymetrix Human Mapping 100K and Puerto Rican founder populations: West African, European, Native American (Choudhry et al 2008). The West African samples consisted of 37 people from West Africa (DNA samples kindly provided by Paul McKeigue). The European population consisted of 42 European American samples from Coriell's North American Caucasian panel. The Native American population consisted of 15 individuals who were Mayan and 15 who were Nahua of central Mexico (DNA samples kindly provided by Mark Shriver). Single nucleotide polymorphisms (SNPs) considered for this analysis all map to the somatic or sex chromosomes. No markers mapping to the mitochondrial genome were tested. Selection of AIMs was based on an iterative process and "informativeness" for ancestry in the Puerto Rican population. For each of the three possible pairs of ancestral populations, SNPs were selected if the difference in allele frequency (delta value) was at least 0.5 (scale from 0 to 1) between any two ancestral populations. Initially 112 AIMs were selected. Failure in genotyping and unavailability of ancestral population genotypes limited our final selection to 100 AIMs (see Sample QC and Genotyping). The selected 100 AIMs (see Supplemental Table 1) were adequately distributed across the genome, with sufficient physical distance between markers such that they were in linkage equilibrium in the three ancestral populations. The average distance between markers was about 2.4×10^7 bp.

DNA isolation and genotyping

Genomic DNA was isolated from buffy coats of peripheral blood using QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) according to the vender's recommended protocol.

Sample QC and Genotyping

Quality control was performed on all DNA using a two-part procedure. Quantitative QC (part 1) involved non-allelic quantitative real-time PCR using a single TaqMan probe in order to ensure ability to amplify the DNA samples. Qualitative QC (part 2) involved genotyping a balanced polymorphism present in most human populations (rs3818), in order to ensure that cross-contamination of samples had not occurred. Genotyping was performed using iPLEX reagents and protocols for multiplex PCR, single base primer extension (SBE) and generation of mass spectra, as per the manufacturer's instructions (for complete details, please see iPLEX Application Note, Sequenom, San Diego, USA). Four multiplexed assays containing 29, 29, 28, and 26 SNPs, respectively, for a total of 112 candidate ancestry informative markers. Of these 112 markers, 106 robustly generated call rates at 90% or higher, with typical call rates in excess of 99% of samples. Multiplexed PCR was performed in 5-μl reactions on 384-well plates containing 5 ng of genomic DNA. Reactions contained 0.5 U HotStarTaq polymerase (QIAGEN), 100 nM primers, 1.25× HotStar Taq buffer, 1.625 mM MgCl2, and 500 μM dNTPs. Following enzyme activation at 94 °C for 15 min, DNA was amplified with 45 cycles of 94 °C × 20 sec, 56 °C × 30 sec, 72 °C × 1 min, followed by a 3-min extension at 72 °C. Unincorporated dNTPs were removed using shrimp alkaline phosphatase (0.3 U, Sequenom). Single-base extension was carried out by addition of SBE primers at concentrations from 0.625 μM (low MW primers) to 1.25 μM (high MW primers) using iPLEX enzyme and buffers (Sequenom, San Diego) in 9-μl reactions. Reactions were desalted and SBE products measured

using the MassARRAY Compact system, and mass spectra analyzed using TYPER software (Sequenom, San Diego), in order to generate genotype calls and allele frequencies. Of 106 AIMs, genotype data for 100 markers were available for the three ancestral populations (37 Africans, 42 Europeans, 30 Native Americans) and the BPRHS population. Thus, these 100 AIMs were used for data analysis.

Population admixture

Individual ancestry was calculated based on the genotypes of 100 informative ancestral markers (Choudhry et al. 2008) in the BPRHS population using two programs: STRUCTURE 2.2 (Falush et al 2003, Prichard et al. 2000) and IAE3CI (Tsai et al. 2005, Parra et al. 2001), with reference to the three ancestral populations: West African, European, and Native American (Choudhry et al. 2006).

Principal component analysis (PCA) to control for population admixture

The EIGENSTRAT (Price et al. 2006) program implemented in HelixTree (Golden Helix, Bozeman, MT, USA) was used to calculate the principal components based on the genotypes of 100 AIMs in the BPRHS population. The estimated principal components from this analysis should reflect population ancestry (Price et al. 2006; Patterson et al. 2006). According to the principal component selection rule, a large gap between the major components and the rest indicates the major principal components can be retained whereas the others are discarded (Jollife 2002; Zhu and Ghodsi 2005).

Statistical analysis

Statistical analyses were performed using SAS 9.1. (Cary, NC, USA) and HelixTree. To compare ancestry differences between the case and control groups for a given disease, we conducted univariate analysis. We assessed the relationship between disease status, ancestry, and AIMs by regression analysis. For T2DM, hypertension, CVD, and depression, we employed logistic regression with disease status as the dependent variable and two of three ancestries as independent variables, while adjusting for potential confounders (age, sex, smoking, alcohol intake, BMI, medications for other diseases, physical activity). For example, for T2DM, we adjusted for age, sex, smoking, alcohol use, BMI, medications for hypertension and depression, and physical activity. For obesity, we used the same model and, except for BMI, adjusted for the identical set of potential confounders (age, sex, smoking, alcohol intake, medications for hypertension, depression, and physical activity). To examine associations between 100 AIMs and disease status, we conducted logistic regression analyses with disease status as dependent variables and AIMs as independent variables. Because BMI, the basic indicator of obesity, is potentially a risk factor of other diseases, we adjusted for BMI in all analyses except that for obesity. Furthermore, we adjusted for socioeconomic status based on total household income as this may also contribute to disease risk (Martinez-Marignac et al. 2007). In addition, all analyses were adjusted for population substructure using ancestral proportion and the first principal components estimated from EIGENSTRAT (Price et al. 2006). Correlation statistics were calculated as Pearson correlation coefficients. P values ≤ 0.05 were considered statistically significant.

Hardy-Weinberg equilibrium test and linkage disequilibrium

A Hardy-Weinberg equilibrium test was performed using the HelixTree program. Pair-wise linkage disequilibria among all AIMs were estimated as correlation coefficients using the same program.

Results

Characteristics of participants

Demographic characteristics of study participants are presented in Table 1. The percentage of individuals who reported smoking or drinking alcohol was significantly higher in men than in women ($P < 0.001$ for both). In contrast, the percentage of participants who were obese (BMI ≥ 30), or who reported depression symptomatology (i.e., depression) was significantly higher in women than men ($P < 0.001$ for both). Other demographic characteristics did not differ significantly by sex. The BPRHS population had high prevalence of T2DM (39.7, 39.2% for men and women, respectively), hypertension (69.7, 69.0%), obesity (39.9, 58.2%), and depression (42.1, 58.7%).

AIMs and population substructure

The minor allele frequencies and *P*-values of Hardy-Weinberg Equilibrium (HWE) tests for the 100 AIMs are provided in Supplemental Table 1. The mean frequency of minor alleles of these 100 AIMs is 0.32 with a range of 0.08 - 0.50. Twenty AIMs (20%) displayed a significant deviation from HWE with *P*-values varying from 0.0001 to 0.045. This is greater than would be expected under the null distribution and indicates that population substructure exists within the BPRHS population.

We estimated individual ancestry with reference to three ancestral populations: West African, European, and Native American using STRUCTURE2.2 (Falush et al 2003) and IBGA3IC (Parra et al. 2001; Tsai et al. 2005). Both methods gave similar results with correlation between estimates of 0.99 for all three ancestries. The estimated ancestral proportions of individuals are plotted in Figure 1. The ancestral composition is on average 57.2±15.2 (%) European with a range in individuals of 7.7 to 90.3%, 27.4±15.2 (%) African with a range of 3.7 to 84.6%, and 15.4±6.5 (%) Native American with a range of 3.8 to 57.9%. As depicted in Fig. 1, the ancestries of the Puerto Ricans studied here were mostly related to Europeans.

Based on PCA using EIGENSTRAT (Price et al. 2006), we also estimated population admixture as principal components. The first 20 eigenvalues (i.e., dimensions) are plotted in Figure 2. The first major principal component represents an eigenvalue of 70.4, 3.8 times the values of the subsequent components. We further found this component was highly correlated with European ancestry ($r = 0.88$, $P<0.0001$, $n=1129$) and American Indian ancestry ($r = 0.30$, $P<0.001$, $n=1129$), and negatively correlated with African ancestry ($r = -0.99$, $P<0.0001$, $n=1129$), all which were estimated using STRUCTURE2.2. Thus, this major principal component was used as a covariate in a logistic regression model to control for population admixture.

Correlation between ancestry and common diseases

We examined the ancestry difference according to disease status for the common diseases in the population. As shown in Table 2, individuals with T2DM or CVD showed significant differences in African and European ancestry. Subjects with T2DM had significantly lower African ancestry (0.26 vs 0.28, P=0.050) and higher European ancestry (0.58 vs 0.57, P=0.049). In addition, individuals with CVD had significantly lower African ancestry (0.26 vs 0.28, P=0.038) and higher Native American ancestry (0.16 vs 0.15, P=0.014). Furthermore, subjects with hypertension tend to have a higher African ancestry (0.28 vs 0.26, P=0.038). However, individuals with other diseases did not display any significant differences in ancestry. To further illustrate these observations, we conducted logistic regression analysis with disease status as dependent variables and adjusted for age, gender, smoking, alcohol use, medications, and physical activity. Multicollinearity of ancestral proportions dictates that only two ancestries are used in the linear regression models, whereas the European ancestry was treated

as the baseline. Results (see Table 3) demonstrated that African ancestry was inversely correlated with T2DM with an odds ratio (OR) of 0.33 with 95% confidence interval (CI) of 0.13 - 0.84 (P=0.021) when compared to European ancestry. Thus, higher African ancestry is correlated with a lower risk of T2DM. In addition, African ancestry was also negatively associated with CVD (OR=0.32, CI 0.10 – 1.00, P=0.049), whereas Native American ancestry was positively associated with CVD (OR=16.63, CI 1.34 – 211.20, P=0.029). In addition, African ancestry was significantly associated with hypertension (OR=2.94, CI 1.14 – 7.61, P=0.026). However, no significant association was observed between ancestry and obesity or depression symptomatology. As socioeconomic status may contribute to common disease risk (Martinez-Marignac et al. 2007), we further adjusted for socioeconomic status and the results remain basically the same (Table 3).

Association of AIMs with disease status after adjusting for population stratification

Next we determined whether any of the individual 100 AIMs were associated with common diseases in this population. Logistic regression analysis with and without adjustment for admixture estimates were conducted using individual ancestries or principal components (PCA). Only those AIMs with a P-value less than or equal to 0.05 are shown in Table 4. In total, 34 AIMs (34%) were significantly associated with one or more of the commonly occurring diseases of this population. For the diseases that were associated with ancestry, i.e., T2DM, CVD, hypertension (Table 3), the P-values of association changed substantially after adjustment for population stratification, estimated as ancestries or PCA. However, for those diseases not significantly associated with ancestry (obesity), adjustment for population admixture had little effect on the P-values of the association (Table 4). For example, for CVD, four AIMs (rs879780, rs4013967, rs10492585, and rs10484578), which were initially associated with CVD, showed no significant association after adjustment for population stratification using either ancestry or PCA. On the other hand, three other AIMs (rs10491097, rs1036543, rs12953952), which were not initially associated with CVD, became significantly associated with CVD (P=0.014, 0.041, and 0.026, respectively) after adjustment for ancestry estimated by PCA. Similar patterns were observed for T2DM and hypertension. Importantly, these results also suggest that there is both positive and negative confounding due to population stratification in this cohort.

We further examined the correlation between association P-values of non-adjustment and adjustment for population admixture for all 100 AIMs. As listed in Table 5, the correlation between association P-values of non-adjustment and adjustment (for ancestry or PCA) were weaker for those diseases that were significantly associated with ancestries (see Table 3, hypertension, CVD, and T2DM), than for those diseases that were not (Table 3, obesity). This observation underscores that adjustment for population stratification is particularly important when ancestry is associated with disease occurrence. In addition, the associated P-values between adjustments for ancestry and PCA were highly correlated with a mean coefficient of 0.98, suggesting that population admixture, estimated either by STRUCTURE or PCA, has similar effects on association tests.

Discussion

We estimated population admixture in 1129 Puerto Ricans living in Massachusetts, based on genotypes of 100 AIMs. Our estimates of mean ancestry of Puerto Ricans are consistent with those previously reported in other Puerto Rican populations (Bonilla et al. 2004; Salari et al. 2005; Choudhry et al. 2006). However, these estimates are strikingly different from the ancestral estimates based on analysis of mitochondrial DNA (mtDNA). Martinez-Cruzado et al (2001; 2005) used mtDNA to estimate the ancestry of Puerto Ricans at 61.3% from Native American, 27.2% from African, and 11.5% from European populations. As mtDNA is

maternally inherited, mitochondrial-based estimates of ancestry reflects the maternal lineage of Puerto Ricans. This observation is consistent with historical accounts of the founding of the island's modern population, where the majority of mtDNA originated from the Taíno Indian mothers (Fernandez-Mendez 1970). Furthermore, although both Puerto Ricans and Mexicans are considered Hispanic, the ancestry of the former is different from that of the latter because of their founding populations. The Mexican population (Burchard et al 2005) originated mainly from European (45.4%) and Native American (51.0%) ancestry with a small proportion of African ancestry (3.7%), whereas Puerto Rican recent ancestry mainly was derived from European (57.2%), and smaller but similar percentages from African (27.4%) and Native American (15.4%) lineages.

It is well established that Africans (including African-Americans), Hispanics and Native Americans have a high prevalence of T2DM compared to non-Hispanic Whites (Brancati et al. 2000; CDC and Prevention 2005) although the biological cause of such disparity is not well understood (Singh and Hiatt 2006). Thus, our observation that African ancestry is negatively associated with T2DM and CVD was not expected. Nonetheless, the onset and development of T2DM represents a complex process, which may be regulated by genomic variation, environmental factors and lifestyle factors such as diet and exercise. In addition, mtDNA dysfunction has been considered an important cause in the development of T2DM (Lowell and Shulman 2005; Manoli et al. 2007; Lai et al 2008). Ancestry analysis of Puerto Ricans based on mtDNA analysis (Martinez-Cruzado et al. 2005) indicates that the origin of Puerto Rican's mtDNA is strikingly different from that of the genomic DNA. Furthermore, mtDNA ancestry of Dominicans, who, like Puerto Ricans, have admixed ancestries of Native American, African, and European, was correlated with obesity and T2DM (Tajima et al 2004). Dominicans who were obese and had diabetes had a significantly higher mtDNA ancestry of African origin. Thus, the Puerto Rican population requires an examination of mtDNA ancestry and its correlation with the prevalence of T2DM and obesity in order to assess the relative contributions of nuclear and mtDNA variation on disease risk, particularly as influenced by environmental factors. On the other hand, ancestries in this popualtion did not show significant association with obesity and depression symptomatology. This suggests that there is no significant difference between ancestral proportions in terms of contribution to risk of these specific diseases in this population. However, for obesity, which is similar to and correlated with T2DM, the American Obesity Association (2005) reported that three ethnic groups of African Americans, Hispanics, and Native Americans exhibit a higher rate of obesity compared to Whites (Wang and Beydoun 2007). Given that we observed no strong correlation between ancestry and obesity in the Puerto Rican population, the higher prevalence of obesity in these three ethnic groups may be due to the combined effects of genomic and mtDNA variation and complex interactions with environmental factors (e.g., Western lifestyles of high-fat diet and low rate of exercise). One alternative explanation states that genetic variation contributing to these diseases is small in the face of environmental factors and genotype by environment interactions (e.g., acculturation). Finally, the AIMs selected in this study may not fully represent the Native American ancestry of our Puerto Rican group, as the tested markers were from Mayan and Nahua individuals; while the Native American ancestry of Puerto Rican is Taíno Indian. This could also mask any potential associations between ancestry and health outcomes.

Population stratification can cause false positive and negative associations in population studies (Knowler et al. 1988; Deng 2001; Marchini et al 2004; Tsai et al. 2005; Barholtz-Sloan et al. 2008). To overcome this problem, three approaches have been developed: genomic control, ancestry estimated by STRUCTURE, and principal component analysis. In this study, based on the genotypes of 100 AIMs, we estimated population admixture using STRUCTURE and PCA methods. We have examined individually the association of 100 AIMs with five common diseases. While none of the associations pass a Bonferroni correction (Bonferroni corrected

P=0.05/500= 0.00001), we did observe 30 associations with P-value <0.05. By chance we would expect to find 25 associations with P-value ≤ 0.05. Thus, some of these 30 associations should be statistically significant. Nevertheless, we observed that population stratification has both positive and negative confounding effects on association. When ancestry is associated with a disease, adjustment for ancestry has substantial effect on tests of association between genetic variants and disease traits. If ancestry is not associated with disease, adjustment for population stratification has little impact on the association test. In addition, our results also showed that adjustment for population stratification estimated either by STRUCTURE or by PCA have similar effects on the association test.

In conclusion, this study estimated individual ancestry of the BPRHS population, which can be used to adjust for population admixture when conducting association studies with other genetic variants. Using the individual ancestry data, we have been able to demonstrate that population admixture is associated with disease prevalence in the BPRHS population. In addition, our results showed that adjustment for individual admixture is particularly important when the ancestry is associated the disease; population admixture estimated either by STRUCTURE or PCA has a similar effect on tests of association.

Acknowledgments

This study was supported by the National Institutes of Health, National Institute on Aging, Grant Number 5P01AG023394-02, NIH/NHLBI grant number HL54776 and HL078885 and contracts 53-K06–5-10 and 58–1950-9–001 from the U.S. Department of Agriculture Research Service.

References

Álvarez-Nazario, M. El elemento afronegroide en el español de Puerto Rico. Instituto de Cultura Puertorriqueña; San Juan, Puerto Rico: 1974.

American Diabetes Association. Standards of medical care in diabetes. Diabetes Care 2007;30:S4–S41. [PubMed: 17192377]

American Obesity Association. Washington (DC): 2005. Obesity in minority populations. http://www.obesity.org/subs/fastfacts/Obesity_Minority_Pop.shtml

Barnholtz-Sloan JS, McEvoy B, Shriver MD, Rebbeck TR. Ancestry estimation and correction for population stratification in molecular epidemiologic association studies. Cancer Epidemiol Biomarkers Prev 2008;17:471–7. [PubMed: 18349264]

Bonilla C, Shriver MD, Parra EJ, Jones A, Fernandez JR. Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York city. Hum Genet 2004;115:57–68. [PubMed: 15118905]

Bermúdez OI, Falcón LM, Tucker KL. Intake and food sources of macronutrients among older Hispanic adults: association with ethnicity, acculturation, and length of residence in the United States. J Am Diet Assoc 2000;100:665–73. [PubMed: 10863569]

Brancati FL, Kao WH, Folsom AR, Watson RL, Szklo M. Incident type 2 diabetes mellitus in African American and white adults: the atherosclerosis risk in communities study. JAMA 2000;283:2253–2259. [PubMed: 10807384]

Burchard EG, Borrell LN, Choudhry S, Naqvi M, Tsai HJ, Rodriguez- Santana JR, Chapela R, Rogers SD, Mei R, Rodriguez- Cintron W, Arena JF, Kittles R, Perez-Stable EJ, Ziv E, Risch N. Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. Am J Public Health 2005;95:2161–2168. [PubMed: 16257940]

Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet 2003;361:598–604. [PubMed: 12598158]

Centers for Disease Control and Prevention. National diabetes fact sheet: United States, 2005. Atlanta (GA): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; 2005. Available from: URL: http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2005.pdf

Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, Tsai HJ, Naqvi M, Phong A, Ung N, Matallana H, Avila PC, Casal J, Torres A, Nazario S, Castro R, Battle NC, Perez-Stable EJ, Kwok PY, Sheppard D, Shriver MD, Rodriguez-Cintron W, Risch N, Ziv E, Burchard EG. Genetics of Asthma in Latino Americans GALA Study. Population stratification confounds genetic association studies among Latinos. Hum Genet 2006;118:652–64. [PubMed: 16283388]

Choudhry S, Taub M, Mei R, Rodriguez-Santana J, Rodriguez-Cintron W, Shriver MD, Ziv E, Risch NJ, Burchard EG. Genome-wide screen for asthma in Puerto Ricans: evidence for association with 5q23 region. Hum Genet 2008;123:455–68. [PubMed: 18401594]

Deng HW. Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. Genetics 2001;159:1319–1323. [PubMed: 11729172]

Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999;55:997–1004. [PubMed: 11315092]

Falcón LM, Tucker KL. Prevalence and correlates of depressive symptoms among Hispanic elders in Massachusetts. J Gerontol B Psychol Sci Soc Sci 2000;55:S108–16. [PubMed: 10794195]

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 2003;164:1567–1587. [PubMed: 12930761]

Fernandez-Mendez, E. Historia cultural de Puerto Rico. San Juan. PR: Ediciones “El Cemi.”; 1970.

Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat 2008;29:648–58. [PubMed: 18286470]

Jolliffe, IT. Principal Component Analysis. Vol. second. Springer; Berlin: 2002.

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 1988;43:520–6. [PubMed: 3177389]

Lai CQ, Tucker KL, Parnell LD, Adiconis X, García-Bailo B, Griffith J, Meydani M, Ordovás JM. *PGC-1α* (*PPARGC1A*) variations associated with DNA damage, diabetes and cardiovascular diseases: the Boston Puerto Rican Health Study. Diabetes 2008;57:809–816. [PubMed: 18162502]

Li CC. Population subdivision with respect to multiple alleles. Ann Hum Genet 1969;33:23–29. [PubMed: 5821316]

Lee IM, Paffenbarger RS Jr. Physical Activity and Stroke Incidence-The Harvard Alumni Health Study. Stroke 1998;29:2049–2054. [PubMed: 9756580]

Lowell BB, Shulman GI. Mitochondrial dysfunction and type 2 diabetes. Science 2005;307:384–387. [PubMed: 15662004]

Manoli I, Alesci S, Blackman MR, Su YA, Rennert OM, Chrousos GP. Mitochondria as key components of the stress response. TRENDS in Endocrinology and Metabolism 2007;18:190–198. [PubMed: 17500006]

Martínez-Cruzado JC, Toro-Labrador G, Ho-Fung V, Estévez-Montero MA, Lobaina-Manzanet A, Padovani-Claudio DA, Sánchez-Cruz H, Ortiz-Bermúdez P, Sánchez-Crespo A. Mitochondrial DNA analysis reveals substantial Native American ancestry in Puerto Rico. Hum Biol 2001;73:491–511. [PubMed: 11512677]

Martínez-Cruzado JC, Toro-Labrador G, Viera-Vera J, Rivera-Vega MY, Startek J, Latorre-Esteves M, Román-Colón A, Rivera-Torres R, Navarro-Millán IY, Gómez-Sánchez E, Caro-González HY, Valencia-Rivera P. Reconstructing the population history of Puerto Rico by means of mtDNA phylogeographic analysis. Am J Phys Anthropol 2005;128:131–55. [PubMed: 15693025]

Martinez-Marignac VL, Valladares A, Cameron E, Chan A, Perera A, Globus-Goldberg R, Wacher N, Kumate J, McKeigue P, O'Donnell D, Shriver MD, Cruz M, Parra EJ. Admixture in Mexico City: implications for admixture mapping of type 2 diabetes genetic risk factors. Hum Genet 2007;120:807–19. [PubMed: 17066296]

Mościcki EK, Locke BZ, Rae DS, Boyd JH. Depressive symptoms among Mexican Americans: the Hispanic Health and Nutrition Examination Survey. Am J Epidemiol 1989;130:348–60. [PubMed: 2750730]

Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, Drineas P. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet 2007;3 (9):e160.10.1371/journal.pgen.0030160

Parra EJ, Kittles RA, Argyropoulos G, Pfaff CL, Hiester K, Bonilla C, Sylvester N, Parrish-Gause D, Garvey WT, Jin L, McKeigue PM, Kamboh MI, Ferrell RE, Pollitzer WS, Shriver MD. Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. Am J Phys Anthropol 2001;114:18–29. [PubMed: 11150049]

Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet 2006;2:2074–2093.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 2006;38:904–909. [PubMed: 16862161]

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155:945–959. [PubMed: 10835412]

Radloff L. The CES-D scale: A self report depression scale for research in the general population. Applied Psychological Measurement 1977;1:385–401.

Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic structure of human populations. Science 2002;298:2381–5. [PubMed: 12493913]2002

Rouse, I. The Tainos: rise and decline of the people who greeted Columbus. Yale University Press; New Haven: 1992.

Salari K, Choudhry S, Tang H, Naqvi M, Lind D, Avila PC, Coyle NE, Ung N, Nazario S, Casal J, Torres-Palacios A, Clark S, Phong A, Gomez I, Matallana H, Pe´rez-Stable EJ, Shriver MD, Kwok PY, Sheppard D, Rodriguez-Cintron W, Risch NJ, Burchard EG, Ziv E, Genetics of Asthma in Latino Americans (GALA) Study. Genetic Admixture and Asthma-Related Phenotypes in Mexican American and Puerto Rican Asthmatics. Genetic Epidemiology 2005;29:76–86. [PubMed: 15918156]

Singh GK, Hiatt RA. Trends and disparities in socioeconomic and behavioural characteristics, life expectancy, and cause-specific mortality of native-born and foreign-born populations in the United States, 1979-2003. Int J Epidemiol 2006;35:903–919. [PubMed: 16709619]

Tajima A, Hamaguchi K, Terao H, Oribe A, Perrotta VM, Baez CA, Arias JR, Yoshimatsu H, Sakata T, Horai S. Genetic background of people in the Dominican Republic with or without obese type 2 diabetes revealed by mitochondrial DNA polymorphism. J Hum Genet 2004;49:495–499. [PubMed: 15368103]

Tang H, Coram M, Wang P, Zhu X, Risch N. Reconstructing genetic ancestry blocks in admixed individuals. Am J Hum Genet 2006;79:1–12. [PubMed: 16773560]

Tsai HJ, Choudhry S, Naqvi M, Rodriguez-Cintron W, Burchard EG, Ziv E. Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations. Hum Genet 2005;118:424–33. [PubMed: 16208514]

Tucker KL, Falcon LM, Bianchi LA, Cacho E, Bermudez OI. Self-reported prevalence and health correlates of functional limitation among Massachusetts elderly Puerto Ricans, Dominicans, and non-Hispanic white neighborhood comparison group. J Gerontol A Biol Sci Med Sci 2000a;55:M90–7. [PubMed: 10737691]

Tucker KL, Bermudez OI, Castaneda C. Type 2 diabetes is prevalent and poorly controlled among Hispanic elders of Caribbean origin. Am J Public Health 2000b;90:1288–93. [PubMed: 10937011]

Tucker KL. Stress and nutrition in relation to excess development of chronic disease in Puerto Rican adults living in the Northeastern USA. J Med Invest 2005;52(Suppl):252–258. [PubMed: 16366511]

U.S. Census Bureau. The American Community Survey 2006 - Hispanics in the United States. 2006. http://www.census.gov/population/www/socdemo/hispanic/hispanic.html

Wang YF, Beydoun MA. The obesity epidemic in the United States—gender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis. Epidemiol Rev 2007;29:6–28. [PubMed: 17510091]

Zhu M, Ghodsi A. Automatic dimensionality selection from the scree plot via the use of profile likelihood. Computational Statistics & Data analysis 2006;51:918–930.

Ziv E, Burchard EG. Human population structure and genetic association studies. Pharmacogenomics 2003;4:431–41. [PubMed: 12831322]

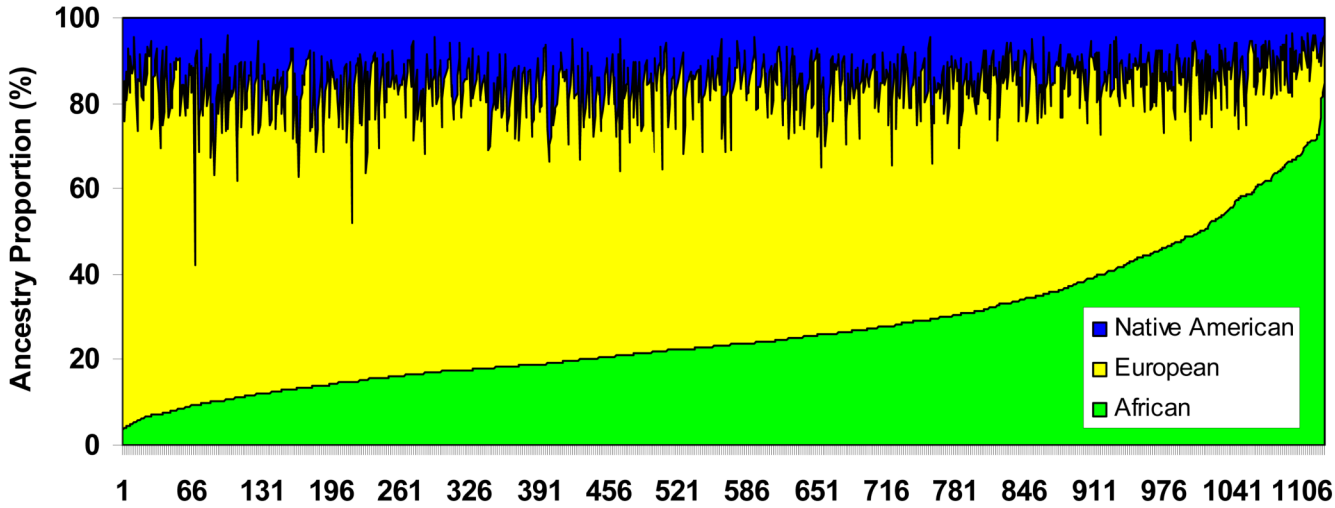

Fig.1.
Ancestry proportion of 1129 Puerto Ricans based on 100 AIMs using SRUCTURE 2.2 with reference to three ancestral populations: European, West African, and Native American.

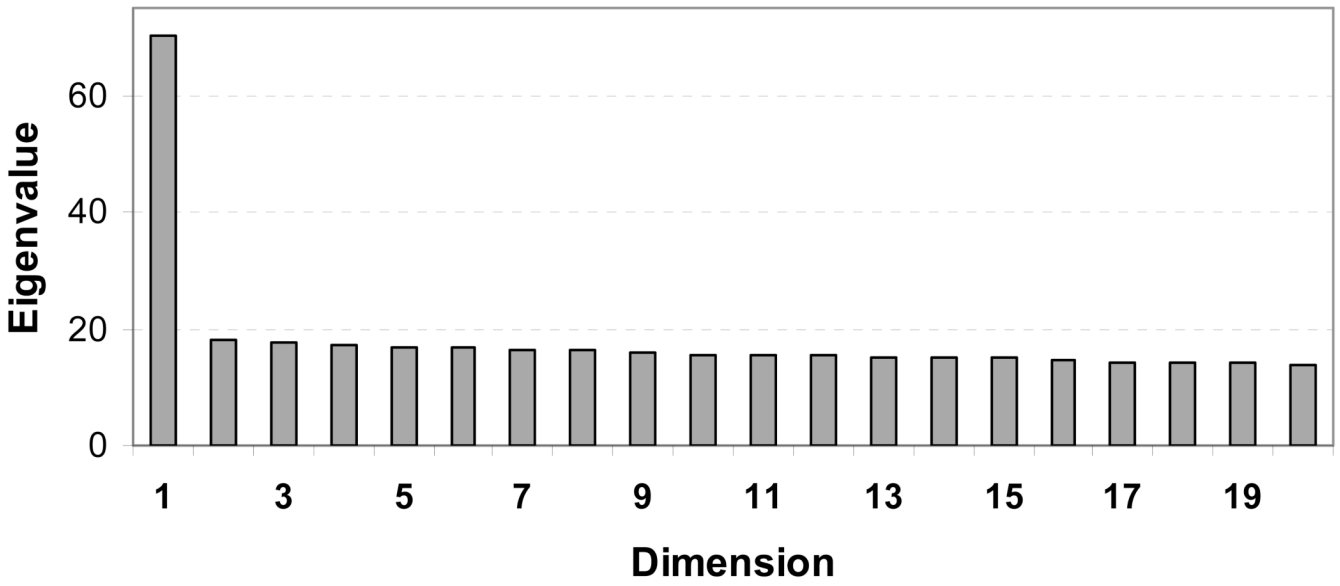

Fig.2.
The first 20 Eigenvalues of principal component analysis using the EIGENSTRAT (Price et al. 2006) based on the genotypes of 100 AIMs for 1129 Puerto Ricans.

Table 1

Demographic characteristics of participants according to genders

	Men (n=337)		Women (n=792)	
	Mean ± SD	**Range**	**Mean ± SD**	**Range**
Age (years)	57.3 ± 7.8	45 - 75	58.0 ± 7.3	45 - 76
* BMI (kg/m2)	29.5 ± 5.1	17.2 - 48.6	32.9 ± 7.0	17.0 - 63.8
Physical activity score	32.5 ± 5.8	24.3 - 62.6	32.5 ± 5.8	24.3 - 62.6
* Drinkers, n (%)	174 (50.0%)		299 (35.2%)	
* Smokers, n (%)	119 (34.3%)		174 (20.7%)	
Diabetes, n (%)	139 (39.7%)		331 (39.2%)	
Cardiovascular diseases, n (%)	84 (23.9%)		171 (20.2%)	
* Overweight n (%)	270 (76.5%)		705 (83.0%)	
*Obesity, n (%)	141 (39.9%)		494 (58.2%)	
Hypertension, n (%)	241 (69.7%)		581 (69.0%)	
* Depression symptomatology, n (%)	148 (42.1%)		498 (58.7%)	

* statistical significance between men and women at *P* -value <0.05

Table 2

Ancestry differences according to disease status[a]

	African			European			Native American		
Disease	**non-affected**	**affected**	***P*-value**[b]	**non-affected**	**affected**	***P*-value**[b]	**non-affected**	**affected**	***P*-value**[b]
Diabetes	0.281 ±0.006	0.263±0.007	**0.050**	0.565±0.006	0.583±0.007	**0.049**	0.154±0.002	0.154±0.003	0.925
Cardiovascular diseases	0.279±0.005	0.256±0.010	**0.038**	0.569±0.005	0.581±0.010	0.280	0.151±0.002	0.162±0.004	**0.014**
Obesity	0.273±0.007	0.276±0.006	0.763	0.571±0.007	0.572±0.006	0.957	0.156±0.003	0.152±0.003	0.397
Hypertension	0.259±0.008	0.280±0.006	**0.038**	0.582±0.008	0.569±0.005	0.167	0.159±0.003	0.151±0.002	0.080
Depression symptomatology	0.283±0.007	0.267±0.006	0.070	0.564±0.007	0.578±0.006	0.118	0.152±0.003	0.155±0.003	0.486

[a] Ancestries according to disease status were listed in frequency as means ± SE

[b] *P* -values were calculated using univariate analysis.

Table 3

Association between ancestry and disease status

	African					Native American				
Disease	**Estimate**	**Odds Ratio**	**95% Interval**	***P*-value**	***P*-value[b]**	**Estimate**	**Odds Ratio**	**95% Interval**	***P* -value**	***P*-value[b]**
T2DM	-1.11	0.33	0.13 - 0.84	**0.021**	**0.021**	-0.048	0.96	0.10 - 9.20	0.967	0.973
CVD	-1.143	0.32	0.10 - 1.00	**0.049**	**0.046**	2.823	16.83	1.34 - 211.20	**0.029**	**0.026**
Obesity[a]	0.048	1.05	0.46 - 2.40	0.909	0.759	-0.68	0.51	0.07 - 3.69	0.502	0.512
Hypertension	1.079	2.94	1.14 - 7.61	**0.026**	**0.021**	-1.622	0.20	0.02- 1.76	0.146	0.122
Depression symptomatology	-0.517	0.60	0.25 - 1.41	0.240	0.203	0.754	2.13	0.26 - 17.52	0.484	0.440

[a] Except Obesity, all analyses have been adjusted for gender, age, smoking, alcohol use, BMI, medications, and physical activity.

[b] *P*-values were calculated by adjusting for additional potential cofounder - socioeconomic status.

Table 4

Association *P*-values of AIMs with common diseases of the BPRHS population

Disease	AIM	Before adjustment	Adjust forancestry[c]	Adjust forPCA[d]
Type 2 diabetes[a]	rs1934393	0.0033	0.0129	0.0105
	rs2592888	0.0069	0.0015	0.0015
	rs1517634	0.0102	0.0117	0.0181
	rs10498919	0.0117	0.0126	0.0137
	rs10486576	0.0345	0.0316	0.0273
	rs2785279	0.0408	0.1040	0.1008
	rs10510791	0.0450	0.0708	0.0757
	rs10214949	0.1120	0.0287	0.0255
	rs1990745	0.0753	0.0561	0.0463
	rs4762106	0.2161	0.0471	0.0684
Hypertension[a]	rs1990745	0.0004	0.0017	0.0010
	rs708915	0.0007	0.0011	0.0021
	rs10214949	0.0153	0.0711	0.0680
	rs1451928	0.0254	0.0428	0.0240
	rs2829454	0.0263	0.0908	0.0702
Cardiovascular diseases[a]	rs2840290	0.0001	0.0003	0.0006
	rs1397618	0.0056	0.0007	0.0008
	rs10507688	0.0056	0.0112	0.0072
	rs6804094	0.0075	0.0188	0.0122
	rs2785279	0.0088	0.0301	0.0276
	rs1990745	0.0120	0.0319	0.0226
	rs879780	0.0211	0.0656	0.0704
	rs4013967	0.0312	0.0945	0.0808
	rs10492585	0.0333	0.1874	0.1877
	rs10484578	0.0437	0.1775	0.1621
	rs10491097	0.0728	0.0192	0.0136
	rs1036543	0.1323	0.0620	0.0414
	rs12953952	0.2147	0.0259	0.0256
Depression symptomatology[a]	rs2592888	0.0038	0.0087	0.0077
	rs4034627	0.0061	0.0148	0.0155
	rs1990745	0.0150	0.0255	0.0222
	rs4852696	0.0260	0.0599	0.0665
	rs2829454	0.0343	0.0671	0.0602
	rs1353251	0.0384	0.0152	0.0169
	rs10488172	0.0385	0.0688	0.0640
Obesity[b]	rs4852696	0.0122	0.0041	0.0027
	rs948360	0.0131	0.0049	0.0054

Disease	AIM	Before adjustment	Adjust forancestry[c]	Adjust forPCA[d]
	rs1397618	0.0145	0.0087	0.0084
	rs2569029	0.0101	0.0118	0.0112
	rs9292118	0.0133	0.0148	0.0141
	rs1353251	0.0156	0.0205	0.0176
	rs7535375	0.0275	0.0307	0.0309
	rs257748	0.0451	0.0384	0.0501
	rs10519979	0.0389	0.0527	0.0436

[a] *P* -values calculated using logistic regression models adjusted for age, sex, BMI, smoking, alcohol use, physical activity.

[b] *P* -values calculated using logistic regression models adjusted for age, sex, smoking, alcohol use, physical activity.

[c] Ancestry was estimated by STRUCTURE.

[d] Anscestry was estimated by PCA.

Table 5

Correlation of association *P* -values between non-adjustment and adjustment for population admixture for 100 AIMs in the BPRHS population

Disease	Non-adjustment and adjutment for ancestry[a]	Non-adjustment and adjustment for PCA[a]	Adjustment for ancestry and PCA[a]
Hypertension	0.537	0.533	0.948
Cardiovascular disease	0.631	0.659	0.974
Type 2 diabetes	0.688	0.714	0.988
Depression sysptomatology	0.811	0.818	0.993
Obesity	0.966	0.975	0.980

[a] All pair-wise correllations were calcucated as Pearson correlation coefficients and were highly signficant at $P < 0.0001$, n=100.