doi: 10.1111/j.1469-1809.2006.00270.x

Application of an African Ancestry Index as a Genomic Control Approach in a Brazilian Population

V. M. Zembrzuski[1], S. M. Callegari-Jacques[1,2] and M. H. Hutz[1,*]

[1]*Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil*
[2]*Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil*

Summary

Ten ancestry informative markers were investigated in 101 coronary artery disease patients and 102 healthy controls from a Southern Brazilian population, in order to determine if stratification occurs in this population. The degree of African admixture detected in this population was estimated to be as high as 6%, but no differences between cases and controls were observed. Using an African Ancestry Index (AAI) that estimates admixture at the individual level it was possible to remove from the samples those individuals with evidence of African admixture. Therefore we have shown that it is possible to control for population stratification by choosing individuals, without the loss of statistical power that occurs with the use of other methods of genomic control.

Introduction

Association studies have been widely used to help understand the genetic basis of quantitative traits, such as the susceptibility to complex diseases. However there has been much debate about the impact of population stratification on case-control association studies, and the fraction of associations that are attributable to stratification is still unknown (Ardlie *et al.* 2002; Freedman *et al.* 2004).

Ethnic population stratification exists when the total population has been formed by admixture between subpopulations, and when admixture proportions vary between individuals (Hoggart *et al.* 2003). The most likely source of confounding in association studies is ethnicity, whereby allele frequencies vary according to ethnicity and cases and controls are not adequately matched in terms of ethnicity (Pritchard & Rosenberg, 1999; Risch, 2000). As long as the heterogeneity is equivalent in case and control subjects (i.e., the two groups have the same mix of ethnic/genetic subgroups) stratification bias will not occur (Ardlie *et al.* 2002). Statistical methods, including the use of genomic markers, have frequently been suggested to control for population stratification in genetic association studies (Reich & Goldstein, 2001; Chen *et al.* 2003).

In order to verify if the degree of physical appearance of an individual is related to his/her African genomic ancestry Parra *et al.* (2003), using the same set of ancestry informative markers (AIMs) described by Parra *et al.* (1998), created an individual index of African ancestry that allowed them to estimate African genomic ancestry at the individual level. Parra *et al.* (2003) also showed that overall, based on self-classification, Brazilians have an intermediate African ancestry index (AAI) between Europeans and Africans.

In Brazil there might be a strong cultural bias toward claiming European ancestry, so we used morphological classification based on skin colour and morphological traits instead of self-classification for ethnic ascertainment in association studies (Rios *et al.* 2003; Mattevi *et al.* 2004). Within a country of continental size, such as Brazil, population composition varies widely among regions. As a whole the population is highly mixed, as pointed out by Parra *et al.* (2003), but the southern Brazilian population differs from this general pattern as

[*] Corresponding author: Prof. Mara H. Hutz, Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Caixa postal 15053, 91501-970 Porto Alegre, RS, Brazil. Tel. 55 51 3316-6720, Fax. 55 51 3316-6727. E-mail: mara.hutz@ufrgs.br

shown by admixture quantification studies. This population is mainly of Portuguese descent, although Italians, Spaniards and Germans have also contributed to its gene pool. A review about the ethnic admixture in Brazilian and other Latin American populations was recently published by Salzano & Bortolini (2002). These authors also showed a good agreement between morphologic classification, based on skin colour and morphological traits, and genetic estimates of admixture.

The current study was undertaken to examine if population stratification in this European derived population would be a confounding factor in association studies, and to test if the African ancestry index derived by Parra *et al.* (2003) would be useful to control for population stratification in case-control association studies.

Material and Methods

Population Sample

The sample comprised 101 individuals with coronary artery disease (cases) and 102 healthy controls. All subjects were of European ancestry as ascertained by skin colour and morphological characteristics. Clinical and demographic characteristics of this sample have been described elsewhere (Rios *et al.* 2003). All subjects gave informed consent to participate in the study.

Genotyping

The markers studied in this study are those defined by Parra *et al.* (1998), with the exception of ICAM1 which was replaced by APOA1*83 as described by Shriver *et al.* (2003). The polymorphisms investigated are listed in Table 1; they were also used by Parra *et al.* (2003). These genetic markers were chosen because they possess differences of frequency (δ) between the parental populations (European and African) that are higher than 45% or are population-specific. All markers except APOA1*83 were genotyped as previously described by Parra *et al.* (1998), whereas the primers and protocols from Wang *et al.* (1996) were used for APOA1*83.

Statistical Analyses

The estimates of the allele frequencies, tests for Hardy-Weinberg equilibrium, and comparisons between cases and controls for these frequencies, were performed using the GENEPOP v.3.1d program (Raymond & Rousset, 1995). The proportions of admixture in the samples were estimated using Long's (1991) weighted least squares (WLS) method (ADMIX program), and the approach of Chakraborty (1985) which provides least-squares estimates using gene identity probabilities. The calculations for this method were performed by the Admix routine written by R. Chakraborty, modified and adapted for Windows by B. Bertoni (available at http://www.genetica.fmed.edu.uy/software.htm). The AAI was calculated for each individual using the allelic frequencies obtained from the literature, as described by Parra *et al.* (2003). The AAI comparison between samples was performed with the Mann-Whitney U test using SPSS® version 8.0. Additionally, the *Structure* program (Pritchard *et al.* 2000) was used to search for subpopulations of genetically similar individuals (Pritchard & Donnelly, 2001). This program was run with k = 2 as the predefined setting for the number of populations, with 30,000 iterations for the burn-in period and 100,000 additional iterations to obtain parameter estimates. The Mann-Whitney U test was also used to compare the individual probabilities of assignment to the inferred cluster 1 (Q values) between cases and controls.

To further test the feasibility of using AAI values for stratification control we simulated a situation in which

Table 1 Allele frequencies of the 10 Ancestry Informative Markers analyzed in cases and controls

Group	APO[1]	APOAI83[1]	AT3[1]	FY-NULL[1]	LPL[1]	OCA2[1]	RB2300[1]	Sb19.3[1]	GC-F	GC-S
Cases	0.942	0.947	0.306	0.990	0.505	0.699	0.257	0.854	0.296	0.534
Controls	0.951	0.956	0.235	0.936	0.500	0.681	0.368	0.819	0.275	0.549
Europeans[2]	0.927	0.925	0.279	1.000	0.486	0.769	0.333	0.910	0.156	0.607
Africans[2]	0.441	0.420	0.874	0.000	0.973	0.098	0.920	0.425	0.824	0.078

[1] Presence of *Alu* insertions and absence of the polymorphic restriction sites.

[2] European and African frequencies were described by Parra *et al.* (1998) and Shriver *et al.* (2003) (APOAI83).

a population of 16,000 control individuals, with an African admixture of about 14% and a hypothetical locus associated with coronary artery disease, was generated. A random sample of 102 controls was then selected from this artificial population, and the corresponding AAI values calculated. Subsequently, individuals with AAI values indicative of African ancestry (greater than −4.86) were excluded. Odds ratios (OR) and respective 95% confidence intervals (0.95 CI) for the hypothetical association were calculated before and after stratification controlling, using SPSS® version 8.0.

Results

Allele frequencies for the ten polymorphisms investigated are shown in Table 1. The genotype frequencies for all markers were in agreement with those expected under Hardy-Weinberg equilibrium (data not shown). No significant differences between cases and controls in the frequency of each marker were observed.

African admixture estimates with both methods showed low levels of admixture in this European derived population (Table 2). The introgression of African genes was lower in cases (2%) than in controls (6%), but considering the standard errors of these estimates they were of the same order of magnitude.

Table 2 African admixture estimated by two different methods based on the frequencies of 10 Ancestry Informative Markers in cases and controls

Group	WLS Method[a]	Chakraborty's Method[b]
Cases	0.022 ± 0.020	0.046 ± 0.001
Controls	0.060 ± 0.024	0.065 ± 0.001

[a] Long, 1991.
[b] Chakraborty, 1985.

Table 3 presents information about AAI in cases and controls. The AAI median for controls (−10.78) was somewhat higher than that estimated for cases (−10.89), but the difference was not statistically significant ($p = 0.392$) (Figure 1A). The median values are more similar to those described for the Portuguese sample (−11.73), as reported by Parra *et al.* (2003), than to the median observed for southern Brazilians (−9.11) in the same study (Table 3). The highest AAI (−4.86) observed in that Portuguese sample was used as a cut-off point to define European ancestry in our samples. In order to test the use of this index to select cases and controls in association studies the AAI was recalculated, removing individuals with an index higher than the highest Portuguese value (−4.86: 9 controls and 2 cases, values ranging from −4.77 to −0.78). The resulting medians for the controls (−10.92) and cases (−10.98) became practically identical ($p = 0.999$) (Table 3; Figure 1B). It should be noticed that the lowest value described by Parra *et al.* (2003) for Africans was +2.86, therefore no overlap with the African AAIs was observed in this population.

To corroborate the absence of structure in this European-derived population the *Structure* program was used to identify clusters of individuals in the whole sample of 203 subjects. Again, no evidence of stratification in the total sample was observed, and no significant differences between cases and controls were detected when Q values were compared (cases: mean ± SE: 0.507 ± 0.023; controls: 0.510 ± 0.024; Mann-Whitney *U* test, $p = 0.903$).

We also simulated a new control sample with 102 individuals randomly selected from a population with 14% African admixture, and investigated whether an apparent association between a hypothetical gene A and coronary artery disease was due to stratification. The

Table 3 Ancestry African Index (AAI) values in cases and controls

	With outliers[a]		Without outliers[a]	
Group	N	AAI Median (range)	N	AAI Median (range)
Cases	101	−10.89 (−17.66 to −4.38)	99	−10.98 (−17.66 to −5.65)
Controls	102	−10.78 (−17.66 to −0.78)	93	−10.92 (−17.66 to −5.25)
Southern Brazilians[b]	52	−9.11	-	-
Portuguese[b]	20	−11.73	-	-

[a] Mann-Whitney *U* test between cases and controls (with outliers: $z = 0.855$, $p = 0.392$; without outliers: $z = 0.001$, $p = 0.999$).
[b] Parra *et al.* (2003).

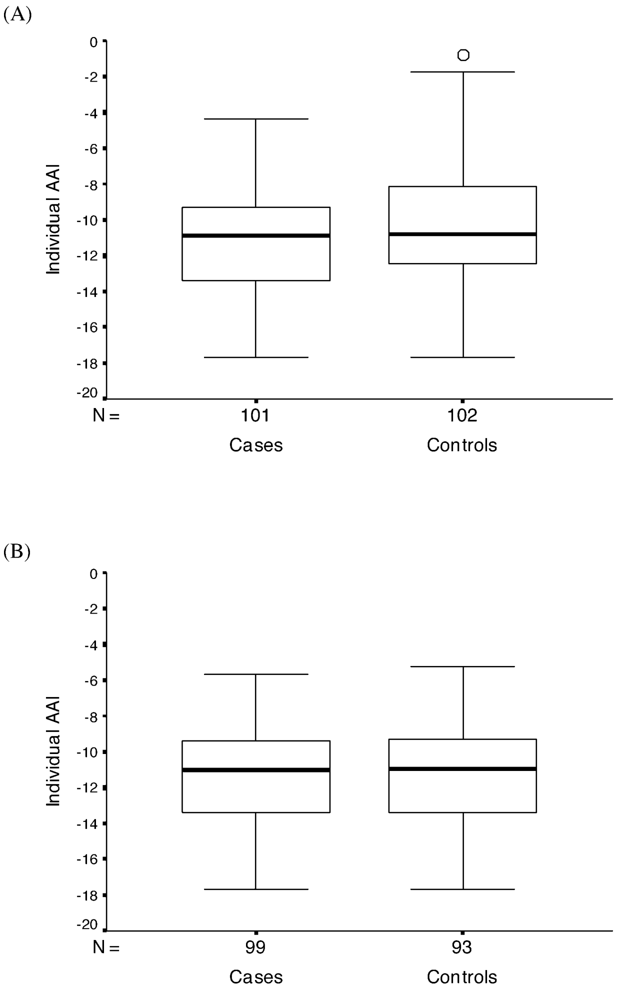

Figure 1 Distribution of individual Ancestry African Index (AAI) values in the samples of cases and controls. Each group is represented by a box whose top and bottom are at the lower and upper quartiles, with a line at the median; thus, the box contains the middle half (50%) of the scores of the distribution. Vertical lines outside the box extend to the largest and the smallest observations within 1.5 interquartile ranges from the box. Open circles represent extreme values. (A) Ascertained samples; (B) Only individuals with AAI lower than −4.86.

Table 4 Simulation of a control sample derived from a population with 14% admixture, and association of a hypothetical gene A with coronary artery disease using the African Ancestry Index to control for population stratification

	Total sample		Without outliers[a]	
Group	Cases N = 101	Controls N = 102	Cases N = 99	Controls N = 79
Allele frequencies				
A1	0.342	0.216	0.328	0.240
A2	0.658	0.784	0.672	0.760
OR (0.95 CI)	1.89 (1.212 − 2.937)		1.54 (0.964 − 2.470)	
AAI median	−10.89	−9.58	−10.98	−10.94
P value[b]	0.001		0.272	

[a] Sample without outliers: removing individual with AAI > −4.86, the highest Portuguese index
[b] Mann-Whitney *U* test.

sample of cases was the same used in our previous analysis. The AAI median for controls (−9.58) was now statistically lower than that estimated for cases (−10.89) ($p = 0.001$). We also simulated a statistically significant association between a hypothetical gene A and coronary artery disease (OR = 1.89; 0.95 CI: 1.21-2.94; $p = 0.007$). After the removal of individuals with an AAI higher than −4.86 (23 controls and two case individuals) medians for the controls (−10.89) and cases (−10.98) were no longer different ($p = 0.272$), and the association became non-significant (OR = 1.54; 0.95CI: 0.96-2.47; $p = 0.090$; Table 4, Figure 2).

Discussion

Admixed populations are an important resource that can be used to study the genetics of complex disorders. A prerequisite to this application is a better understanding of the admixture proportions and dynamics of the admixture process. Several investigations have been conducted using AIMs in order to describe the admixture process and population dynamics in American admixed populations (Parra *et al.* 1998; Shriver *et al.* 2003; Bonilla *et al.* 2004; Collins-Schramm *et al.* 2004). These studies provided evidence that AIMs have applicability in admixture mapping, as well as to control for structure in association tests. Although these data provide support for these practical applications, they were only validated in few populations.

The emerging picture is that populations do generally cluster by broad geographic regions that correspond with common racial classifications; thus knowledge of ethnicity is important for proper design of case-control association studies, and for identifying disease predisposing alleles that may differ across ethnic groups. Several methods have been described to correct for population stratification (Devlin & Roeder, 1999; Pritchard & Rosenberg, 1999; Pritchard *et al.* 2000; Reich & Goldstein, 2001). These methods are quite general, and they mathematically correct for the degree of stratifi-

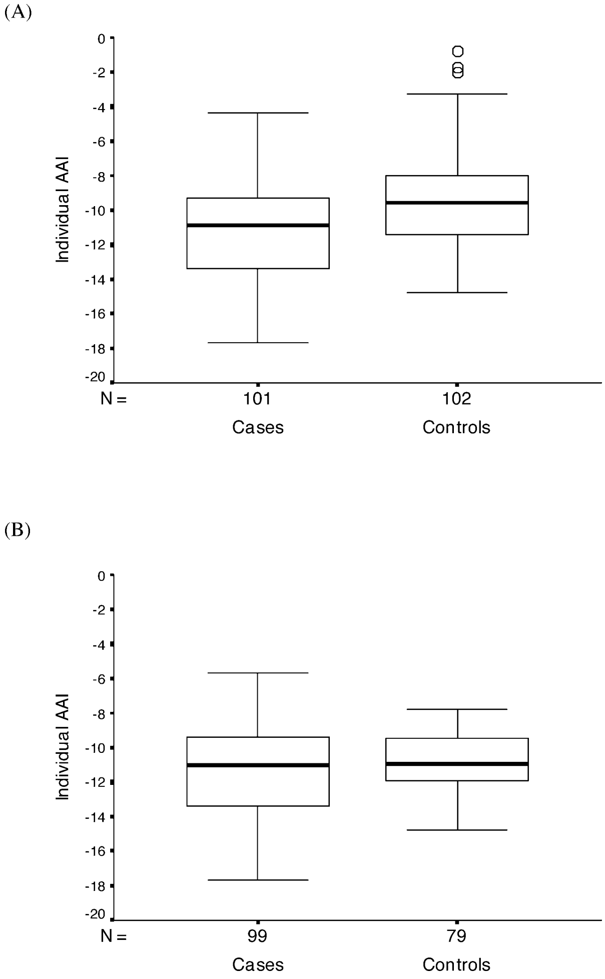

Figure 2 Distribution of individual Ancestry African Index (AAI) values in cases and in the simulated control sample with 14% African admixture. Each group is represented by a box whose top and bottom are at the lower and upper quartiles, with a line at the median; thus, the box contains the middle half (50%) of the scores of the distribution. Vertical lines outside the box extend to the largest and the smallest observations within 1.5 interquartile ranges from the box. Open circles represent extreme values. (A) Ascertained cases and a simulated control sample with 14% African admixture. (B) Only individuals with AAI lower than −4.86.

cation at the sample level. The African ancestry index used in the present study can be used at the individual level, and therefore it is possible to correct for stratification by removing some individuals from the sample without loss of statistical power due to statistical corrections.

It has been shown that, even when using proxies such as skin colour to match cases and controls, some hidden admixture could still occur, as described by Shriver *et al.* (2003) in European Americans and as shown in the present study in European Brazilians. All individuals investigated in this study were selected as being of European ancestry based on skin colour and morphological characteristics, but low levels of admixture were estimated (Table 2); therefore the simultaneous use of a second method of matching, at the individual level, will help to avoid false positives due to population substructure. On the other hand, in heterogeneous populations, such as Brazilians where there is a poor correlation between skin colour and ancestry, direct matching of cases and controls using AAI might be a more accurate and appropriate procedure, rather than using proxies such as skin colour and morphological characteristics to match individuals. Even using a population with higher levels of admixture, as shown in our hypothetical simulation, the method is still useful. If one is, for instance, dealing with a rare disease it would be possible to remove or include individuals in the control group to match them with the cases, and avoid population stratification in association analyses.

In Brazil many population studies have been carried out with the objective of assessing the degree of European, African and Amerindian contributions to their gene pools using blood groups and protein markers (for a review see Salzano & Bortolini, 2002). Most of these studies estimated the African contribution to the southern European-derived population as 8%, which is a figure close to that estimated for our control sample using DNA SNPs (6%). Parra *et al.* (2003), using the same set of SNPs as well as the same estimation method (Long, 1991), showed 13% of African admixture in southern Brazilians, which is more than twice the highest value observed in the present study. Although their sample is not strictly comparable to the one investigated herein, because it comprises individuals from all three southern states, the main difference is that Parra *et al.* (2003) used self-classification as the criterion to define European ancestry. As pointed out by Ziv & Burchard (2003) in populations from the United States and Latin America, where admixture has been ongoing for several generations, self-described ethnicity may be a less accurate predictor of genetic ancestry. Nevertheless the use of the African Ancestry Index was still useful for population controlling, as shown by the simulation performed with this level of African gene flow (Table 4).

Considering the tri-ethnic nature of the Brazilian population as a whole it would be important to define a set of markers which would be able to identify the Amerindian contribution. Although Shriver *et al.* (2003), among others, have identified specific markers for Native Americans these might not be useful for admixture estimations. As pointed out by Cavalli-Sforza *et al.* (1994) the extreme drift in many South American Native groups has generated an exceptional gene frequency variation. There is no assurance for any of the most informative markers that they were truly absent from the original American Natives and can therefore be used for inferring admixture. Even if gene frequencies are averaged over extant populations we could have no idea how these frequencies were in the past, because of the enormous intertribal and interregional drift due to the numerous processes of fission and fusion of tribes.

The limitations of this study should also be considered. The samples used in this study are small, as is the number of SNPs used. Larger samples sizes and more markers would be required for a whole–genome association study of a multigenic phenotype. Because few SNPs were used we cannot exclude the possibility that low levels of stratification remained undetected. The constraint of having a high-throughput setting severely limits the possibility to test hundreds or thousands of SNPs in several laboratories, but if this was an option it would be possible to genotype more markers, and then calculate the individual index in the same way as it was performed in this study. Nevertheless our results indicate that simple matching strategies can effectively control for population stratification, and that ten ancestry informative markers were sensitive enough to detect admixture levels as low as 2% as estimated in the case sample. Data presented herein show that if matching is done carefully, using the African Ancestry Index, genetic stratification can also be kept to a minimum. Our results are encouraging to investigators who work with more heterogeneous populations. As pointed out by Ardlie *et al.* (2002) in samples that have the same mix of genetic subgroups, stratification bias will not be a problem, nevertheless it could occur in repeated samplings from the same population, if the underlying heterogeneity in each sampling is not well matched.

In conclusion, we suggest that using markers that are particularly informative for ancestry to estimate an African ancestry index at the individual level would be a useful method for genomic control in association studies in admixed populations. More empirical data is needed to improve the feasibility of this approach in case-control association studies in admixed populations.

Acknowledgements

Thanks are due to our colleagues Fabiana M. de Andrade, Marilu Fiegenbaum, Silvana de Almeida and Domingos L. Souza Rios for help in sample collection. Financial support was provided by Programa de Apoio a Núcleos de Excelência (PRONEX, Brazil) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil).

References

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

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. (2004) Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. *Ann Hum Genet* **68**, 139–153.

Cavalli-Sforza, L. L., Menozzi, P., Piazza, A. (1994) *The history and geography of human genes*. Princeton University Press, Princeton.

Chakraborty, R. (1985) Gene identity in racial hybrids and estimation of admixture rates. In: *Genetic microdifferentiation in man and other animals* (eds J. V. Neel & Y. Ahuja), pp 171–180. Indian Anthropological Association, New Delhi.

Chen, H. S., Zhu, X., Zhao, H. & Zhang, S. (2003) Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Ann Hum Genet* **67**, 250–264.

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.

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

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

Hoggart, C. J., Parra, E. J., Shriver, M. D., Bonilla, C., Kittles, R. A., Clayton, D. G. & McKeigue, P. M. (2003) Control of confounding of genetic associations in stratified populations. *Am J Hum Genet* **72**, 1492–1504.

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

Mattevi, V. S., Zembrzuski, V. M. & Hutz, M. H. (2004) A resistin gene polymorphism is associated with body mass index in women. *Hum Genet* **115**, 208–212.

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

Parra, F. C., Amado, R. C., Lambertucci, J. R., Rocha, J., Antunes, C. M. & Pena, S. D. (2003) Color and genomic ancestry in Brazilians. *Proc Natl Acad Sci USA* **100**, 177–182.

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

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

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

Raymond, M. & Rousset, F. (1995) http://wbiomed.curtin.edu.au/genepop.

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

Rios, D. L., Vargas, A. F., Torres, M. R., Zago, A. J., Callegari-Jacques, S. M. & Hutz, M. H. (2003) Interaction between SREBP-1a and APOB polymorphisms influences total and low-density lipoprotein cholesterol levels in patients with coronary artery disease. *Clin Genet* **63**, 380–385.

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

Salzano, F. M. & Bortolini, M. C. (2002) *The evolution and genetics of Latin American populations*. Cambridge University Press, Cambridge.

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

Wang, X. L., Badenhop, R., Humphrey, K. E. & Wilcken, D. E. (1996) New MspI polymorphism at +83 bp of the human apolipoprotein AI gene: association with increased circulating high density lipoprotein cholesterol levels. *Genet Epidemiol* **13**, 1–10.

Ziv, E. & Burchard, E. G. (2003) Human population structure and genetic association studies. *Pharmacogenomics* **4**, 431–441.

Received: 16 June 2005
Accepted: 23 December 2005