A rapid screening of ancestry for genetic association studies in an admixed population from Pernambuco, Brazil

A.V.C. Coelho[1,2], R.R. Moura[1,2], C.A.J. Cavalcanti[1,2], R.L. Guimarães[1,2], P. Sandrin-Garcia[1,2], S. Crovella[3] and L.A.C. Brandão[2,4]

[1]Departamento de Genética, Universidade Federal de Pernambuco, Recife, PE, Brasil
[2]Laboratório de Imunopatologia Keizo Asami, Universidade Federal de Pernambuco, Recife, PE, Brasil
[3]IRCCS Burlo Garofolo, Trieste, Itália
[4]Departamento de Patologia, Universidade Federal de Pernambuco, Recife, PE, Brasil

Corresponding author: R.R. Moura
E-mail: ronaldmoura1989@gmail.com

ABSTRACT. Genetic association studies determine how genes influence traits. However, non-detected population substructure may bias the analysis, resulting in spurious results. One method to detect substructure is to genotype ancestry informative markers (AIMs) besides the candidate variants, quantifying how much ancestral populations contribute to the samples' genetic background. The present study aimed to use a minimum quantity of markers, while retaining full potential to estimate ancestries. We tested the feasibility of a subset of the 12 most informative markers from a previously established study to estimate influence from three ancestral populations: European, African and Amerindian. The results showed that in a sample with a diverse ethnicity (N = 822) derived

from 1000 Genomes database, the 12 AIMs had the same capacity to estimate ancestries when compared to the original set of 128 AIMs, since estimates from the two panels were closely correlated. Thus, these 12 SNPs were used to estimate ancestry in a new sample (N = 192) from an admixed population in Recife, Northeast Brazil. The ancestry estimates from Recife subjects were in accordance with previous studies, showing that Northeastern Brazilian populations show great influence from European ancestry (59.7%), followed by African (23.0%) and Amerindian (17.3%) ancestries. Ethnicity self-classification according to skin-color was confirmed to be a poor indicator of population substructure in Brazilians, since ancestry estimates overlapped between classifications. Thus, our streamlined panel of 12 markers may substitute panels with more markers, while retaining the capacity to control for population substructure and admixture, thereby reducing sample processing time.

Key words: Brazilian genetic admixture; Population structure; Ethnicity; Ancestry informative markers; SNP; Association studies

INTRODUCTION

Genetic association studies (GAS) are conducted to determine which genetic factors underlie susceptibility to complex diseases. Single nucleotide polymorphisms (SNPs) are the main genetic variations analyzed in GAS. In fact, SNPs have a great part in inter-individual differences (Sachidanandam et al., 2001). Moreover, SNP genotyping can be performed easily and inexpensively in a wide range of high-throughput technologies.

Generally, two groups of unrelated subjects are genotyped in GAS: one carrying a determined trait, such as a disease (cases), and another not carrying it (controls). Thus, if one of the alleles is more frequent in one group than the other, it is associated with the presence/absence of the trait, meaning that its presence may contribute to risk/protection in relation to the disease (Lewis, 2002).

A positive association may represent three situations: 1) the allele has indeed a causal role in the trait; 2) the allele is not causal, but is in linkage disequilibrium with the true causal polymorphism or 3) the observed association could be a spurious one (Cordell and Clayton, 2005). The last circumstance could be interpreted as a false-positive result.

A false-positive result, due to spurious association, may arise from undetected population substructure (stratification), For example, in an admixed population, individuals heterogeneous for their genetic backgrounds result from unequal genetic contribution from ancestral populations. Thus, during sampling, a subpopulation may be overrepresented in the cases and/or in the controls. This introduces bias on allele frequencies, which could result in a false positive during statistical analysis (Balding, 2006).

Population substructure thus is regarded as a confounding factor, and as a consequence, some methods were proposed to control for this issue. They all have in common the requirement for genotyping of unlinked polymorphisms not associated with the trait of interest in all sampled subjects (both cases and controls) and have been defined as "null SNPs" (Balding, 2006).

Some authors have proposed SNP panels for use in this context. Kosoy et al. (2009),

for instance, applied the "informativeness" algorithm developed by Rosenberg et al. (2003) to select 128 SNPs, designated ancestry informative markers (AIMs). They were used to infer European, African and Amerindian ancestry from ethnically diverse populations from United States (US) cities. It is important to clarify that we understand the term "ethnicity" as the social identity of a group or population with few or no relationship with its actual genetic background (Ali-Khan et al., 2011).

Brazilian populations were also founded by admixture from these three ancestral populations. This resulted in a peculiar population composition, where self-reported ancestry (based on skin color) does not correlate or reflect real ancestry, thus being a poor predictor of population substructure, when compared to more "homogeneous" populations in other countries (Pena et al., 2009).

Thus, our research group intended to explore these AIMs to estimate ancestry proportions in populations from the Recife metropolitan region, to avoid spurious results in GAS. Recife is the capital of Pernambuco State, which is located in Northeast Brazil, a region that received influence from three ancestral populations: native Amerindian, European (16th century settlers) and African (trafficked slaves), contributing to origin of the population of present times through admixture (Ribeiro, 1995; Parra et al., 2003).

However, genotyping of 128 AIMs would be costly and labor-intensive when analyzing a substantial number (usually more than 100 individuals) of cases and controls. Therefore, in trying to meet the necessity of ancestry identification encountered by our research group in the daily activity of simple association studies, we proposed a simple panel of AIMs, suitable for low-/medium-throughput laboratories with simple genotyping technologies and small financial resources.

Since Kosoy et al. (2009) claimed that the use of smaller (48 or 24 AIMs) subsets resulted in consistent estimates when compared to the larger datasets of the original 128 markers, we thus decided to examine the informativeness of these AIMs after further streamlining this panel, and using it to estimate the genetic ancestry and control for substructure in our admixed population during GAS.

MATERIAL AND METHODS

Ethics statement

The Research Ethics Committee of the Center of Health Sciences, Federal University of Pernambuco approved the study (protocol No. 257.941). Each subject gave written consent.

Ancestry informative marker subset selection, genotyping and sample data

Kosoy et al. (2009) selected 128 AIMs through the genome using allele frequency differences between populations (>45% difference between European-derived and African-derived populations and between European-derived and Amerindian-derived populations) and informativeness criteria (the ability to estimate the proportion of individual genetic ancestries) as elaborated by Rosenberg et al. (2003).

To streamline the number of AIMs, we selected the twelve most informative markers for substructure assignment of the three subpopulations (top 12) described by Kosoy et al. (2009). Briefly, the 12 AIMs were extracted from genotyping data of 643 subjects gathered by

the authors: 128 European Americans, 42 West African subjects, 105 Amerindians, 188 East Asian Americans, and 64 South Asian Americans. In addition, 60 subjects from European [Utah residents (CEPH) with Northern and Western European ancestry] and 56 from Yoruban (Ibadan, Nigeria) populations were included (The HapMap Consortium, 2003).

To test the feasibility of this streamlined AIM subset in admixed populations, we obtained the corresponding genotypes of both the 128 original and the 12 AIMs in an independent subset of unrelated individuals from populations with available data in 1000 Genomes database (Genomes Project Consortium et al., 2012).

This subset included 822 individuals from the following populations: Finland (N = 93), England and Scotland (N = 89), Spain (N = 14), Tuscany, Italy (N = 97), Yoruba from Ibadan, Nigeria (N = 85), Luhya from Webuye, Kenya (N = 96), African-American from Southwest US (N = 64), Colombia (N = 60), Mexico (N = 64), and Puerto Rico (N = 55). The 105 Amerindian subjects were from the original Kosoy et al. (2009) subset included in this procedure.

After *in silico* analysis using the independent subset described above, we genotyped the 12 AIMs in 192 samples of individuals from the metropolitan region of Recife. Each subject answered an epidemiological questionnaire, which required self-report in five ethnic categories as defined by the Instituto Brasileiro de Geografia e Estatística (IBGE - Brazilian Geography and Statistics Institute), the Brazilian census bureau, based on skin color. The five IBGE categories are: *branco* (white or Caucasian), *preto* (black), *pardo* (multiracial), *amarelo* ("yellow", Asian descent) and *indígena* (Amerindian) (IBGE, 2010).

Table 1 summarizes information about the top 12 AIMs (marker ID, chromosome location and allele frequencies, including the Recife frequencies).

Table 1. The twelve most informative markers for assignment in three subpopulations (K = 3), TaqMan probe assays, chromosome, reference allele frequencies as presented in Kosoy et al. (2009) paper and reference allele frequencies in Recife sample.

Marker	TaqMan probe assay	Location	EURA frequencies	AFR frequencies	AMI frequencies	Recife frequencies
rs4908343	C___2494120_10	1p36.11	0.82	0.04	0.95	0.59
rs7554936	C__26139689_10	1q21.3	0.34	0.99	0.12	0.44
rs6548616	C__29071253_10	3p12.3	0.25	0.96	0.05	0.43
rs7657799	C__29422763_10	4q24	0.05	0.86	0.01	0.21
rs10007810	C___1386349_10	4p13	0.25	0.96	0.05	0.41
rs6451722	C___2938090_10	5p12	0.24	0.90	0.01	0.36
rs1040045	C___8767011_10	6p25.1	0.73	0.10	0.98	0.67
rs10108270	C__30263561_10	8p23.2	0.35	0.97	0.03	0.39
rs772262	C___8340116_10	12q13.2	0.06	0.87	0.63	0.33
rs9530435	C__27192660_10	13q22.2	0.79	0.07	0.95	0.69
rs11652805	C__31084340_10	17q24.1	014	0.98	0.13	0.35
rs4891825	C__27956007_10	18q22.2	0.89	0.09	0.90	0.68

EURA = European American populations; AFR = African populations; AMI = Amerindian populations.

Since the self-reported ethnicity was recorded (Table 2), it was possible to observe how the estimated genetic ancestry would relate to the arbitrary skin-color categories, an approach followed by previous studies (Smith et al., 2004; Lins et al., 2010), although using a different set of markers. The Mann-Whitney test was used to make comparisons between the genetic ancestry estimates between self-reported ethnicities.

Table 2. Demographic characteristics of subjects from Recife, including self-reported race according to the classification of Instituto Brasileiro de Geografia e Estatística (IBGE - Brazilian Geography and Statistics Institute), the Brazilian census bureau.

Characteristic	N = 192	%
Gender		
Female	128	66.7
Male	64	33.3
Self-reported race		
Pardo (multiracial)	93	48.4
Branco (White or Caucasian)	81	42.2
Preto (Black)	18	9.4

All SNPs were genotyped using TaqMan SNP Genotyping Assays on ABI 7500 real-time PCR platform, following manufacturer instructions (Life Technologies, USA).

Population substructure assignment and comparison between AIM sets

All population substructure analyses were performed with the STRUCTURE software version 2.3.4 (Pritchard et al., 2000; Falush et al., 2003, 2007; Hubisz et al., 2009), as recommended by Kosoy et al. (2009). For runs with the original 128 AIMs, 10,000 burn-in cycles and 50,000 replicates in the admixture model were used. For models using the 12 AIMs, 20,000 burn-in cycles and 100,000 replicates also in the admixture model were used instead. Thus, different α (Dirichlet parameter, which in this case represents the degree of admixture) were calculated for each estimated population. Apart from this, all remaining options were set to their default, and each run had three iterations. All runs were performed with two to four ancestry clusters (K).

For ancestry proportion estimation in the Recife sample, the European American, African, and Amerindian top 12 genotypes (thus excluding the EAS and SAS populations) were inputted together with the genotypes from the new samples to serve as references. This was done to help determine which cluster generated by the STRUCTURE software corresponds to each ancestry, i.e., European, African and Amerindian (K = 3) ancestries, respectively.

Ancestry cluster plots were performed through the Distruct software version 1.1 (Rosenberg, 2004). The mean estimates of the 1000 Genomes populations resulting from the three iterations with both the original 128 and the top 12 AIMs were compared using Pearson correlation. Additionally, Bland-Altman plots were produced to help visualize the degree of concordance of the correlations, an approach similar to that used by Aldrich et al. (2008).

RESULTS

When analyzing the data of the 1000 Genomes populations, the ancestry estimates obtained by the two AIM sets, i.e., the original 128 AIMs genotyped by Kosoy et al. (2009) and the top 12, were similar for all three ancestries. The high correlation indices showed that the estimates had a high degree of concordance (Figure 1). Bland-Altman plots showed in more detail the degree of concordance between these estimates (Figure 2). Since these results showed that both estimates were remarkably similar in a population dataset regardless of the subset which was used for AIMs selection, the top 12 were used as an approximation to the 128 AIMs to genotype the sample from Recife.
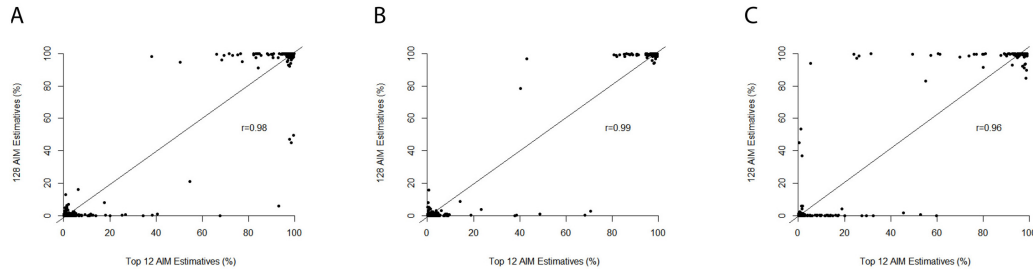
Figure 1. Correlation plots comparing estimates of 128 ancestry informative markers (AIMs) and a subset of its 12 most informative AIMs (top 12) in 822 ethnically diverse subjects derived from 1000 Genomes database. Comparison of **A.** European ancestry estimates, **B.** African ancestry estimates, and **C.** Amerindian ancestry estimates.
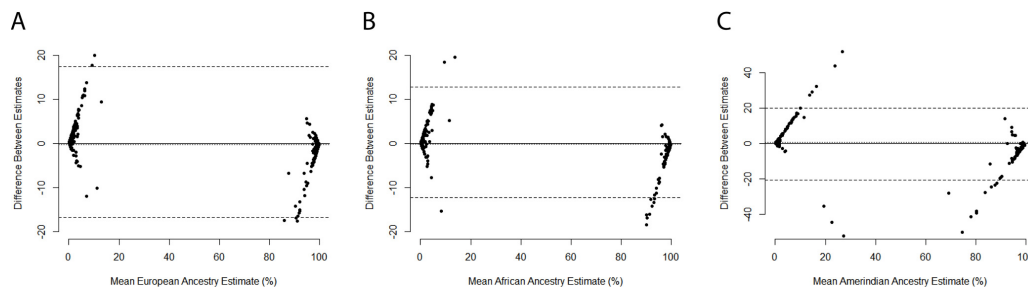

Figure 2. Bland-Altman plots showing the great concordance between estimates from the two ancestry informative marker subsets (128 and the 12 most informative markers) in 822 ethnically diverse subjects. **A.** European ancestry, **B.** African ancestry, and **C.** Amerindian ancestry comparisons. The solid line represents no difference between two estimates. The dashed lines represent 95% confidence intervals of the difference between estimates. Since the majority of data points are between these lines, it is demonstrated that the estimates did not differ significantly. The dotted line represents the mean estimates taking in consideration the two ancestry informative marker subsets (128 and the top 12).

All genotypes were in conformity with Hardy-Weinberg equilibrium. The output from the STRUCTURE software pointed out that K = 3 had the largest mean likelihood (L = -6095.2 against K = 2, L = -6311.0, and K = 4, L = -6110.6).

Considering individual Q (ancestry) estimates in the sample from Recife, Brazil, it was observed that all subjects had greater European influence when compared with African and Amerindian, without taking self-reported race into consideration (mean estimates: 59.7, 23.0 and 17.3%, respectively). However, there was considerable overlap of these estimates.

After stratifying according to self-reported ethnicity, subjects identified as black and *pardos* had marginally similar African proportions (27.9 *vs* 24.6%, Mann-Whitney U-test = 585, P = 0.04). Whites had more European ancestry proportion (61.9%) than did *pardos* (58.3%, U-test = 5237.5, P << 0.01) and blacks (56.6%, U-test = 1164, P << 0.01). Blacks and *pardos* had similar proportions of European ancestry (U-test = 998.5, P = 0.20).

Regarding Amerindian ancestry, *pardos* and whites had similar backgrounds (18.0 *vs* 17.1%, U-test = 4364, P = 0.07), with blacks having the lower proportion (15.5%) when compared with whites (U-test = 1086, P = 0.001) and *pardos* (U-test = 1107, P = 0.03). Table 3

summarizes these estimates and Figure 3 depicts the overall estimates of the reference populations (European American, African, and Amerindian) as well as Recife estimates.

Table 3. Comparisons between estimates of subjects from Recife (Northeast Brazil) according to self-reported races.

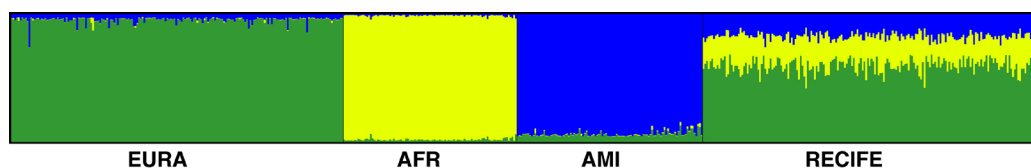| Self-reported race	European ancestry (%)		African ancestry (%)		Amerindian ancestry (%)	
	Mean ± SD	Min.-Max.	Mean ± SD	Min.-Max.	Mean ± SD	Min.-Max.
White (N = 81)	61.9 ± 5.5	41.9-68.4	20.1 ± 7.1	10.8-47.7	18.0 ± 3.0	10.4-24.6
Pardo* (N = 93)	58.3 ± 5.5	42.2-68.5	24.6 ± 7.1	11.7-46.7	17.1 ± 3.0	10.6-25.8
Black (N = 18)	56.6 ± 5.6	49.0-65.8	27.9 ± 7.2	18.0-38.2	15.5 ± 3.0	11.2-19.3
All (N = 192)	59.7 ± 5.5	41.9-68.5	23.0 ± 7.1	10.8-47.7	17.3 ± 3.0	10.4-25.8

*Multiracial.

Figure 3. Plot of the overall top 12 estimates. The European American (EURA), African (AFR) and Amerindian (AMI) populations were included together with Recife samples to help ancestry assignment. The resulting Recife estimates demonstrate Brazilian populations admixture: a large contribution of European ancestry with less influence of African and Amerindian ancestries.

DISCUSSION

Population substructure and genetic ancestry is a fundamental issue to be considered when designing and developing GAS, because they could result in spurious association detection (Balding, 2006). The present study aimed to propose a streamlined panel of AIMs that could replace larger panels of markers, thus improving cost-benefit in the analysis of highly-admixed populations such as from Brazil, more specifically in Northeast Brazil, Recife city, Pernambuco State. Our results confirm the notion that the population from Recife had origin from three ethnically ancestral groups: Amerindians, Africans and Europeans (Alves-Silva et al., 2000).

Moreover, the analyses indicated that reduced numbers of markers were still capable of adequately estimating individual ancestry proportions. We compared the 12 most informative SNPs with the original 128, published elsewhere (Kosoy et al., 2009). There was remarkable concordance between the estimates, as revealed by Pearson correlation coefficients (Figure 1).

Currently, several genotyping methods are used to determine genetic ancestry. The most common are small insertion or deletion (indels) markers, which provide accurate information about ancestry since they are scattered throughout the genome (Mills et al., 2006). The standard methods of indel genotyping could be relatively costly and time-consuming, requiring expensive reagents for PCR, PCR product purification and special hardware, such as capillary gel electrophoresis instruments. Thus, some authors have reported ancestry estimation using SNP panels with real-time PCR platforms (Smith et al., 2004; Seldin and Price, 2008; Lins et al., 2010).

Estimates from Recife (Northeast Brazil) confirmed that Brazilian populations are admixed, since each individual had substantial genetic contribution from each subpopulation cluster. In summary, the European genomic contribution was most representative in our population, followed by the African and then the Amerindian contribution (average contributions of 59.7, 23.0 and 17.3% ancestry, respectively). This means that the Recife population has a low level of genetic substructure, since self-reported race, which is mainly based on skin-color, did not relate well to genetic proportions. Subjects identified as blacks, had even more European ancestry influence than African ancestry itself. Thus, skin-color does not correctly control for population substructure in admixed populations such as the Brazilian during GAS analysis.

These results confirm the observations of Pena et al. (2009), who also detected high European influence in the Brazilian genetic background. The authors also used samples from Pernambuco State. Our average European ancestry estimates in self-declared white subjects were somewhat lower than their own (61.9 *vs* 71.1%, respectively), and our African ancestry estimates were slightly higher (20.1 *vs* 14.2%, respectively).

Moreover, when comparing the estimates from white subjects from São Paulo State (Southeast Brazil) with white subjects from the Recife sample, it is evident how different regions of Brazil received different influences from the ancestral populations. Northeast Brazil received more influence from West African populations, whereas the Southeast received large numbers of European immigrants during the 19th century, a process that shaped these inter-regional differences in genetic background. The subjects from São Paulo had 77.9% European ancestry, 11.6% African and 10.5% Amerindian ancestry, whereas the subjects from Recife had 61.9, 20.1 and 18.0%, respectively (Pena et al., 2009). Similar results were obtained in another study by Lins et al. (2010), who used a set of 28 SNPs. In their sample from Northeast Brazil, the subjects had 77.4, 13.6 and 8.9% estimated ancestries.

These discrepancies may be related to differences in sample number and the markers employed to estimate the ancestries, since in one of them, a set of indel markers was used. Despite these differences, these studies and others performed in several Brazilian populations, e.g., Santos et al. (2010), show that, qualitatively, European ancestry makes up most of the genetic background of Brazilians, followed by minor contributions from African and Amerindian ancestral populations. Thus, we again highlight the importance of considering this admixture when performing GAS in Brazilian populations.

Keeping this in mind, we decided to further explore SNP usage for ancestry estimation in GAS, to obtain a streamlined panel and report a proof of concept that ancestry estimation could be simpler using a smaller number of markers than in previous studies.

We propose a rapid streamlined SNP panel based on 12 markers, spread through the genome, to be used for genetic ancestry estimation aimed at controlling the population substructure in GAS in Brazilian admixed populations. Our SNP panel, deriving from our daily experience in small-scale association studies with relatively small financial resources, has been specifically tailored for laboratories with low-/medium-throughput genotyping instrumentation and needs.

ACKNOWLEDGMENTS

REFERENCES

Aldrich MC, Selvin S, Hansen HM, Barcellos LF, et al. (2008). Comparison of statistical methods for estimating genetic admixture in a lung cancer study of African Americans and Latinos. *Am. J. Epidemiol.* 168: 1035-1046.

Ali-Khan S, Krakowski T, Tahir R and Daar A (2011). The use of race, ethnicity and ancestry in human genetic research. *HUGO J.* 5: 47-63.

Alves-Silva J, Santos MS, Guimaraes PE, Ferreira AC, et al. (2000). The ancestry of Brazilian mtDNA lineages. *Am. J. Hum. Genet.* 67: 444-461.

Balding DJ (2006). A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* 7: 781-791.

Cordell HJ and Clayton DG (2005). Genetic association studies. *Lancet* 366: 1121-1131.

Falush D, Stephens M and Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164: 1567-1587.

Falush D, Stephens M and Pritchard JK (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* 7: 574-578.

Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56-65.

Hubisz MJ, Falush D, Stephens and Pritchard JK (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* 9: 1322-1332.

Instituto Brasileiro de Geografia e Estatística (IBGE) (2010). Atlas do Censo Demográfico 2010. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro.

Kosoy R, Nassir R, Tian C, White PA, et al. (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum. Mutat.* 30: 69-78.

Lewis CM (2002). Genetic association studies: design, analysis and interpretation. *Brief. Bioinform.* 3: 146-153.

Lins TC, Vieira RG, Abreu BS, Grattapaglia D, et al. (2010). Genetic composition of Brazilian population samples based on a set of twenty-eight ancestry informative SNPs. *Am. J. Hum. Biol.* 22: 187-192.

Mills RE, Luttig CT, Larkins CE, Beauchamp A, et al. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 16: 1182-1190.

Parra FC, Amado RC, Lambertucci JR, Rocha J, et al. (2003). Color and genomic ancestry in Brazilians. *Proc. Natl. Acad. Sci. U. S. A.* 100: 177-182.

Pena SD, Bastos-Rodrigues L, Pimenta JR and Bydlowski SP (2009). DNA tests probe the genomic ancestry of Brazilians. *Braz. J. Med. Biol. Res.* 42: 870-876.

Pritchard JK, Stephens M, Rosenberg NA and Donnelly P (2000). Association mapping in structured populations. *Am. J. Hum. Genet.* 67: 170-181.

Ribeiro D (1995). O Povo Brasileiro: A Formação e o Sentido do Brasil. Companhia das Letras, São Paulo.

Rosenberg NA (2004). Distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes* 4: 137-138

Rosenberg NA, Li LM, Ward R and Pritchard JK (2003). Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* 73: 1402-1422.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409: 928-933.

Santos NP, Ribeiro-Rodrigues EM, Ribeiro-Dos-Santos AK, Pereira R, et al. (2010). Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel. *Hum. Mutat.* 31: 184-190.

Seldin MF and Price AL (2008). Application of ancestry informative markers to association studies in European Americans. *PLoS Genet.* 4: e5.

Smith MW, Patterson N, Lautenberger JA, Truelove AL, et al. (2004). A high-density admixture map for disease gene discovery in African Americans. *Am. J. Hum. Genet.* 74: 1001-1013.

The HapMap Consortium (2003). The International HapMap Project. *Nature* 426: 789-796.