RESEARCH ARTICLE

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Assessing Individual Interethnic Admixture and Population Substructure Using a 48–Insertion-Deletion (INSEL) Ancestry-Informative Marker (AIM) Panel

Ney P.C. Santos,[1] Elzemar M. Ribeiro-Rodrigues,[1] Ândrea K.C. Ribeiro-dos-Santos,[1] Rui Pereira,[2,3] Leonor Gusmão,[2] António Amorim,[2,4] João F. Guerreiro,[1] Marco A. Zago,[5] Cecília Matte,[6] Mara H. Hutz,[7] and Sidney E.B. Santos[1]*

[1]*Laboratório de Genética Humana e Médica, Universidade Federal do Pará, Belém, Pará, Brazil;* [2]*Instituto de Patologia e Imunologia Molecular, Universidade do Porto (IPATIMUP), Porto, Portugal;* [3]*Instituto de Medicina Legal, Universidade de Santiago de Compostela, Santigo de Compostela, España;* [4]*Faculdade de Ciências, Universidade do Porto, Porto, Portugal;* [5]*Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brazil;* [6]*Instituto Geral de Perícias, Rio Grande do Sul, Brazil;* [7]*Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil*

Communicated by Pui-Yan Kwok

Received 26 June 2009; accepted revised manuscript 6 November 2009.
Published online 1 December 2009 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.21159

ABSTRACT: Estimating the proportions of different ancestries in admixed populations is very important in population genetics studies, and it is particularly important for detecting population substructure effects in case-control association studies. In this work, a set of 48 ancestry-informative insertion-deletion polymorphisms (INDELs) were selected with the goal of efficiently measuring the proportions of three different ancestries (sub-Saharan African, European, and Native American) in mixed populations. All selected markers can be easily analyzed via multiplex PCR and detected with standard capillary electrophoresis. A total of 593 unrelated individuals representative of European, African, and Native American parental populations were typed, as were 380 individuals from three Brazilian populations with known admixture patterns. As expected, the interethnic admixture estimates show that individuals from southern Brazil present an almost exclusively European ancestry; Afro-descendant communities in the Amazon region, apart from the major African contribution, present some degree of admixture with Europeans and Native Americans; and a sample from Belém, in the northeastern Amazon, shows a significant contribution of the three ethnic groups, although with a greater European proportion. In summary, a panel of ancestry-informative INDELs was optimized and proven to be a valuable tool for estimating individual and global ancestry proportions in admixed populations. The ability to accurately infer interethnic admixtures highlights the usefulness of this marker set for assessing population substructure in association studies, particularly those conducted in Brazilian and other Latin American populations sharing trihybrid ancestry patterns. Hum Mutat 31:184–190, 2010. © 2009 Wiley-Liss, Inc.

KEY WORDS: ancestry-informative marker; AIM; insertion-deletion polymorphism; INDEL; admixture; population structure

Introduction

Genetic association studies are often performed in populations of unrelated individuals to identify susceptibility loci for complex human traits. If subjects are sampled from two or more subpopulations for which the frequencies of marker alleles and traits differ, spurious associations may arise due to population substructure [Pritchard et al., 2000a,b; Reiner et al., 2005; Risch et al., 2002; Schork et al., 2001].

In general, population stratification exists when a population is formed by recent mixing of subpopulations and when the admixture proportions (defined as the proportions of the genome that have ancestry from each subpopulation) vary among individuals [Shriver et al., 2003]. Population stratification is especially important when conducting association studies in admixed populations such as Latinos, African Americans [Tsai et al., 2006], and admixed populations from the Caribbean and Central and South America. Most of these populations were formed by an interethnic admixture of two parental groups (Europeans and Africans or Europeans and Native Americans) or even three parental groups (Europeans, Africans, and Native Americans) as in most Brazilian populations [Ribeiro-Rodrigues et al., 2009; Santos and Guerreiro, 1995].

The Brazilian population is one of the most heterogeneous in the world. Estimates indicate that, before the arrival of colonizers, around 2.5 million natives were living in the Brazilian territory [Cunha, 1995]. European immigration first consisted primarily of Portuguese. In the first three centuries, ~500,000 individuals came from Portugal. On the other hand, the slave trade started in the second half of the 15th century and continued until 1850. During this period, ~3.5 million Africans were introduced into Brazil by force [Curtin, 1969]. Later, Brazil also received immigrants from European countries other than Portugal. It is estimated that between 1800 and the mid-20th

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Sidney Emanuel Batista dos Santos, Instituto de Ciências Biológicas; Laboratório de Genética Humana e Médica; Cidade Universitária Prof. José da Silveira Netto; Rua Augusto Corrêa, 01, BOX: 8615; CEP: 66.075-970, Belém, PA, Brasil. E-mail: sidneysantos@ufpa.br

century, about 4 million individuals entered the country, mainly from Portugal, Italy, Spain, and Germany. This second immigration wave settled predominantly in the southeastern and southern regions. Due to the special occupation policies of such a vast territory, the admixture process occurred in different ways in different geographic regions of the country. In northeast Brazil, the African contribution is high and the Native American component is low; in the North, the contribution of Native Americans is pronounced, whereas in the South the Amerindian and African influence is reduced in comparison with all other geographic regions [Salzano and Bortolini, 2002; Santos and Guerreiro, 1995]. Furthermore, there is evidence of a so-called directed admixture process; i.e., involving predominantly European men and both Native and African women [Carvalho et al., 2008; Feio-dos-Santos et al., 2006; Marrero et al., 2005; Ribeiro-dos-Santos et al., 2007; Ribeiro-Rodrigues et al., 2009; Santos et al., 1999].

Considering the complex process that gave rise to the present Brazilian population, caution dictates that disease genetic association studies should include an assessment of population structure in order to identify and correct for possible substructure effects. Many approaches can be used, including genomic control and structured association. The structured association approach estimates individual ancestry by using a set of genetic markers and then tests for association while correcting for individual admixture. This approach is particularly favored by researchers studying admixed populations and it is often used with a set of highly informative markers for estimating ancestral proportions [Parra et al., 2001; Pritchard and Donnelly, 2001; Tsai et al., 2006]. Therefore, the identification of genetic markers capable of accurately assessing the interethnic mixture in individuals is fundamental to dealing with substructure in admixed populations.

A growing number of publications have reported the use of ancestry-informative markers (AIMs)—markers whose allele frequency varies significantly between populations of distinct geographic origins—to estimate individual admixture and to identify population substructure. In most studies, these AIMs consisted of single-nucleotide polymorphisms (SNPs) [Benn-Torres et al., 2008; Choudhry et al., 2006; Kosoy et al., 2009; Parra et al., 1998, 2001; Shriver et al., 2003], but insertion-deletion polymorphisms (INDELs) of small DNA fragments [Bedoya et al., 2006] and short tandem repeats [Pimenta et al., 2006] have also been used.

Recently, INDELs have been the focus of multiple investigations [Bastos-Rodrigues et al., 2006; Mills et al., 2006; Ribeiro-Rodrigues et al., 2009; Weber et al., 2002]. This type of polymorphism presents interesting features as genetic markers: (1) INDELs are spread throughout the human genome; (2) INDELs derive from a single event (they do not present homoplasy); (3) since the allele frequencies of many INDELs are significantly different in separated populations, they can be used as AIMs; (4) small INDELs can be analyzed using short amplicons, which improves the amplification of degraded DNA and facilitates multiplexing; and (5) INDELs can be easily genotyped with a simple dye-labeling electrophoretic approach.

The aim of this work was to develop a panel of AIMs with the following characteristics: (1) the ability to differentiate populations of three continents [Europe, Africa (sub-Saharan), and America]; (2) the ability to assess substructure in different populations; (3) the ability to accurately measure global and individual ancestry proportions in admixed populations; and (4) permissive of easy, fast, and cost-effective genotyping.

Materials and Methods

Population Samples

The study was performed with population samples of known origin consisting of 593 individuals representative of three major ancestry groups: Sub-Saharan Africans—189 individuals from Angola, Mozambique, Zaire, Cameroon, and the Ivory Coast (details in [Alves et al., 2004; Silva et al., 2006]); Europeans—161 individuals, mainly Portuguese [Alves et al., 2007]; and Native Americans—243 individuals from indigenous tribes of the Brazilian Amazon region [Santos et al., 1999].

The admixed populations used in this study were a sample of 81 individuals from southern Brazil, a sample of 196 individuals from the city of Belém in the northeast Brazilian Amazon region [Rodrigues et al., 2007], and a sample of 103 individuals from Afro-descendant communities also living in the Amazon region [Carvalho et al., 2008].

Selection of Ancestry-Informative INDELs

Sixty biallelic INDEL markers were preselected based on three main criteria: (1) great differences in allele frequencies ($\delta \geq 40\%$) between African, European, and/or Native American populations; (2) mapping to different chromosomes or to different physical regions of the same chromosome; and (3) variable size between 3 and 40 base pairs (bp) to permit simultaneous genotyping of multiple markers. The selection process was based on data from Weber et al. [2002] and from the online database at www.marshfieldclinic.org/mgs/pages/default.aspx?page = didp.

After primer design and optimization of PCR conditions for the simultaneous analysis of multiple markers, we selected 16 African ancestry markers (presenting high δ values between Africans vs. Europeans or Native Americans), 16 European ancestry markers, and 16 Native American ancestry markers.

INDEL Typing

Primers were designed using the PRIMER3 software (www.genome.wi.mit.edu/cgi-bin/primer/primer3) and tested for hairpin and primer-dimer secondary structures with the AutoDimer software [Vallone and Butler, 2004].

Information on the 48 studied INDELs is presented in Supp. Table S1, including mapping data, INDEL size, amplicon length variation, primer sequences, dye-labeling fluorochromes, and concentrations used in the multiplex assay.

DNA samples were typed for the 48 biallelic INDELs by means of three 16-plex PCR amplifications (Supp. Table S1). Each multiplex PCR was performed in a final volume of 12.5 μL containing 1 × PCR buffer with 3 mM $MgCl_2$, 125 μM of each dNTP, 2 U AmpliTaq Platinum DNA Polymerase (Invitrogen Life Technologies, Carlsbad, CA), primer concentrations according to Supp. Table S1, and 10 to 20 ng of genomic DNA. The PCR thermocycling conditions were: 11 min at 95°C; followed by 1 min at 94°C, 1 min at 60°C, and 2 min at 70°C for 10 cycles; then 1 min at 90°C, 1 min at 60°C, and 2 min at 70°C for 17 cycles; and a final extension of 60 min at 60°C.

Before capillary electrophoresis, 1 μL PCR product was added to 8.5 μL deionized formamide HI-DI (Applied Biosystems, Foster City, CA) and 0.5 μL GeneScan 500 LIZ size standard (Applied Biosystems). DNA fragments were separated using an ABI PRISM 3130 Genetic Analyzer (Applied Biosystems) and analyzed with GeneMapper ID v3.2 software (Applied Biosystems).

Statistical Analysis

Allele frequency estimates were obtained by direct gene count, and δ values were determined by subtracting allele frequency values in the studied populations. Basic parameters such as molecular diversity, Hardy-Weinberg equilibrium, and population genetic structure, as well as analysis of molecular variance (AMOVA) and F_{ST} calculations, were performed using the Arlequin 3.1 software package [Excoffier et al., 2005] with 500,000 steps in the Markov chain. The statistical significance of the F_{ST} values was estimated by permutation analysis using 500,000 permutations.

For further population structure analysis and estimation of individual ancestry proportions we utilized STRUCTURE v.2.2 software (http://pritch.bsd.uchicago.edu/software.html) [Falush et al., 2003, 2007; Pritchard et al., 2000a,b]. Individual admixture estimates (IAEs) were computed using both STRUCTURE and ADMIXMAP (http://homepages.ed.ac.uk/pmckeigu/admixmap/index.html) [Hoggart et al., 2003], and their accuracy was determined with a Pearson's correlation test using SPSS v12.0 (SPSS, Chicago, IL).

ADMIX 2.0, based on a coalescent approach, was used for estimating group admixture [Bertorelle and Excoffier, 1998; Dupanloup and Bertorelle, 2001]. The standard deviation of the group admixture estimates was calculated using 10,000 bootstraps. Global admixture proportions were also estimated by the ADMIX95 program (www.genetica.fmed.edu.uy/software.htm), which is based on the gene identity method [Chakraborty, 1985].

Results

All INDELs were successfully amplified using three 16-plex reactions followed by capillary electrophoresis, and consistent results were obtained in repeated analyses of randomly chosen DNA samples. Moreover, no discrepancies were observed in allele identification with either single or multiplex PCR.

All amplicons were relatively small; i.e., less than 300-bp long. The 48 AIMs were spaced throughout the genome (in 17 different chromosomes) and were chosen to avoid, as much as possible, bias due to unequal representation between the parental populations.

Analysis of the Parental Populations

The set of 48 AIMs was used to study parental populations from sub-Saharan Africa, Europe, and Native America. The allele frequencies observed in each group and the values of two differentiation measurements (δ and F_{ST}) between parental population pairs are presented in Table 1.

The data reveal considerable differences in allele frequencies among the parental populations. Out of the 144 possible comparisons, 44 have δ values over 50%, 20 have δ values higher than 40% and lower than 50%, and 26 have δ values between 30 and 40%. At least 21 of the studied INDELs present δ values greater than 45% in 2 out of the 3 possible comparisons among the parental population frequencies.

Considering the mean for the 48 INDEL set, the highest differentiation was observed between Africans and Native Americans, followed by Europeans and Native Americans, and finally Africans and Europeans (Table 1).

In order to test the capacity of the panel of 48 AIMs to distinguish individuals from different parental populations, the obtained genotypes were used to run the STRUCTURE software with combinations of different parameters. All runs were performed with the same run length of 100,000 burn-in and 500,000 Markov Chain Monte Carlo (MCMC) replications using an ancestry model that assumes that the analyzed individuals may have inherited fractions of the genome from K different ancestral populations.

When all individuals of the studied parental populations were grouped together, the ln likelihood values demonstrated that the population structure of the group could be best explained by three distinct clusters (K = 3; ln Prob = −25,259.3; analyses performed with K varying between 2 and 6). Figure 1 presents the results obtained assuming correlated allele frequencies either with (Fig. 1A) or without (Fig. 1B) information on the probable origin of the individuals. These results show that the panel of 48 INDELs clearly differentiates between European, African, and Amerindian populations regardless of the analysis model employed.

Population Structure

In an attempt to answer questions regarding the genetic structure of admixed populations, we genotyped 380 individuals from three Brazilian populations characterized by different levels of interethnic admixture: south Brazil, with predominantly European ancestors; Belém, a trihybrid population from the northeastern Amazon region; and a sample of Afro-descendant communities living in the Amazon region.

Multilocus genotypes analyses with STRUCTURE were first performed without any information about the origins of the individuals. Data were analyzed so as to examine the number of clusters (K) that best describes the structure of the admixed population under study.

When each admixed population was separately considered, it was not possible to detect substructure in the Afro-descendant and South Brazilian samples. By contrast, population substructuring was observed in the sample from Belém that was best explained by the existence of two clusters.

Individual Interethnic Admixture Estimates

In order to test the capacity of the INDEL panel to correctly assess individual ancestry proportions, genotyping data obtained from populations with different levels of admixture were analyzed using the two statistical programs STRUCTURE and ADMIXMAP.

The IAEs obtained using STRUCTURE were calculated assuming the Admixture Model of Ancestry with correlated or independent allele frequencies. All runs were performed with 100,000 burn-in steps followed by 500,000 MCMC iterations without any information about the population of origin of each individual in the sample and with the option K = 3. Under similar conditions, we also ran STRUCTURE assuming an Ancestry Model in which individuals of known origin (parental populations) are used to classify individuals of unknown origin (the mixed populations). The IAEs for the representatives of each population obtained using STRUCTURE are presented in Figure 2. The results obtained are concordant with those previously described for each mixed population: the Afro-descendant sample presented a high proportion of African origins (individual variation between 0.66 and 0.71), the sample from South Brazil showed a greater contribution of genes of European origin (variation 0.71–0.95), and the sample from Belém carried contributions from all ethnicities, although with the greatest contribution from Europe. In the case of Belém, the individual contributions varied substantially: between 5 and 47% for African genes, from 26 to 86% for European genes, and from 9 to 68% for Native American genes.

Table 1. Allele Frequencies for the 48 INDELs in AFR, EUR, and NAM Populations, and Differentiation Measures Observed Among Populations

	Short allele frequency			F_{ST}[a]			δ[a]		
Markers	AFR	EUR	NAM	AFR/EUR	AFR/NAM	EUR/NAM	AFR/EUR	AFR/NAM	EUR/NAM
MID1357	0.116	0.716	1.000	0.545	0.896	0.332	0.599	0.884	0.284
MID273	0.191	0.672	0.998	0.382	0.825	0.373	0.480	0.807	0.326
MID1684	0.476	0.462	0.145	0.000	0.233	0.224	0.014	0.331	0.318
MID818	0.758	0.266	0.015	0.389	0.757	0.265	0.492	0.743	0.251
MID1172	0.537	0.206	0.000	0.192	0.570	0.370	0.331	0.537	0.206
MID1176	0.913	0.309	0.242	0.562	0.619	0.009	0.603	0.671	0.068
MID1358	0.383	0.853	0.905	0.372	0.472	0.011	0.470	0.523	0.052
MID1785	0.694	0.244	0.029	0.335	0.669	0.201	0.450	0.665	0.215
MID1271	0.071	0.856	0.295	0.769	0.144	0.474	0.785	0.223	0.562
MID780	0.228	0.872	0.917	0.585	0.665	0.009	0.644	0.690	0.045
MID494	0.228	0.741	0.656	0.416	0.309	0.014	0.513	0.428	0.085
MID625	0.320	0.675	0.845	0.222	0.448	0.078	0.355	0.525	0.170
MID1379	0.620	0.902	1.000	0.190	0.414	0.118	0.282	0.380	0.098
MID2011	0.197	0.734	0.817	0.451	0.556	0.018	0.538	0.621	0.083
MID1726	0.167	0.672	0.612	0.417	0.336	0.005	0.505	0.446	0.059
MID473	0.979	0.388	0.004	0.594	0.977	0.434	0.591	0.975	0.384
MID619	0.817	0.269	0.940	0.467	0.070	0.669	0.548	0.123	0.671
MID1448	0.074	0.465	0.008	0.335	0.057	0.503	0.391	0.066	0.457
MID1923	0.275	0.109	0.355	0.080	0.012	0.144	0.166	0.080	0.246
MID856	0.437	0.849	0.254	0.305	0.070	0.516	0.413	0.182	0.595
MID99	0.786	0.439	0.990	0.226	0.206	0.596	0.347	0.204	0.551
MID93	0.206	0.772	0.000	0.485	0.227	0.812	0.566	0.206	0.772
MID1716	0.332	0.702	0.783	0.238	0.340	0.014	0.370	0.451	0.081
MID682	0.431	0.904	0.624	0.390	0.070	0.180	0.473	0.193	0.280
MID1039	0.175	0.513	0.154	0.225	0.001	0.264	0.338	0.020	0.359
MID1780	0.333	0.785	0.418	0.338	0.013	0.238	0.452	0.084	0.368
MID1470	0.135	0.593	0.008	0.376	0.124	0.632	0.458	0.127	0.585
MID132	0.198	0.606	0.549	0.296	0.226	0.004	0.407	0.351	0.056
MID1098	0.271	0.487	0.322	0.093	0.003	0.053	0.216	0.051	0.165
MID1558	0.074	0.607	0.428	0.491	0.269	0.059	0.532	0.354	0.179
MID217	0.024	0.481	0.041	0.453	0.002	0.442	0.457	0.017	0.439
MID568	0.119	0.439	0.076	0.230	0.008	0.320	0.320	0.043	0.363
MID1952	0.163	0.291	0.845	0.043	0.635	0.487	0.128	0.682	0.554
MID476	0.000	0.328	0.854	0.340	0.837	0.457	0.328	0.854	0.526
MID275	0.633	0.700	0.407	0.007	0.094	0.155	0.067	0.225	0.293
MID152	0.249	0.191	0.942	0.007	0.684	0.750	0.058	0.694	0.752
MID196	0.594	0.494	0.035	0.017	0.560	0.467	0.100	0.559	0.459
MID778	0.771	0.787	0.335	0.000	0.316	0.336	0.017	0.435	0.452
MID1603	0.088	0.391	0.979	0.226	0.896	0.617	0.302	0.891	0.589
MID1386	0.182	0.187	0.755	0.000	0.490	0.482	0.006	0.573	0.568
MID660	0.221	0.609	0.858	0.269	0.585	0.153	0.388	0.637	0.249
MID575	0.108	0.000	0.652	0.099	0.463	0.608	0.108	0.545	0.652
MID216	0.739	0.853	0.368	0.036	0.239	0.381	0.114	0.370	0.485
MID481	0.089	0.306	0.000	0.139	0.101	0.356	0.217	0.089	0.306
MID913	0.006	0.003	0.101	0.000	0.071	0.076	0.003	0.095	0.098
MID350	0.061	0.047	0.416	0.000	0.276	0.290	0.014	0.355	0.368
MID988	0.094	0.371	0.696	0.196	0.532	0.191	0.277	0.601	0.324
MID1184	0.279	0.368	0.733	0.015	0.340	0.238	0.089	0.454	0.365
Average				0.268	0.369	0.300	0.340	0.418	0.342

[a]δ and F_{ST} are differentiation measures.
AFR, African; EUR, European; NAM, Native American.

Global Interethnic Admixture Analysis

The global interethnic admixture estimates of each mixed population were obtained using the statistical programs ADMIX 2.0 and ADMIX 95 in addition to STRUCTURE v.22 and ADMIXMAP. In the two latter cases, we considered the mean individual admixture estimates to be the global interethnic admixture of the population. The results are presented in Table 2. Overall, the estimates obtained by different methodologies point towards an almost exclusive contribution of European genes in the sample from South Brazil and a predominant contribution of African genes in the sample from Amazonian Afro-descendants. Nevertheless, these estimates showed a few minor differences: the ADMIX 2.0 program was unable to identify either the contribution of African and Native American genes in the sample from South Brazil or the contribution of European genes in the sample from Afro-descendants; ADMIX 95 failed to identify the African contribution in the sample from South Brazil.

The population sample from South Brazil showed a very high level of European contribution (94%) and far fewer Native American (5%) and African (1%) genes. The Amazonian Afro-descendant communities presented a principle contribution of African genes (75%) and a smaller input from Native Americans (15%) and Europeans (10%). In the population of Belém, the highest contribution was from Europeans (60%), followed by Native Americans (28%) and a smaller contribution from Africans (12%).

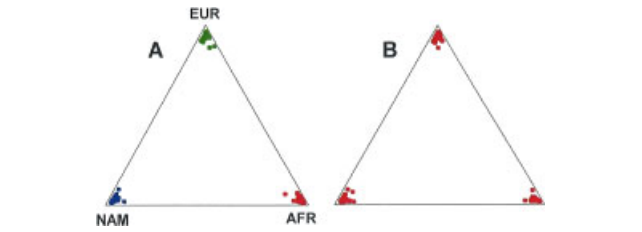

Figure 1. Schematic representation of the 48-AIM panel's ability to distinguish individuals from different parental populations, using STRUCTURE v.22 software and assuming correlated allelic frequencies. Results were obtained with (**A**) and without (**B**) prior information about the origin of the individuals.

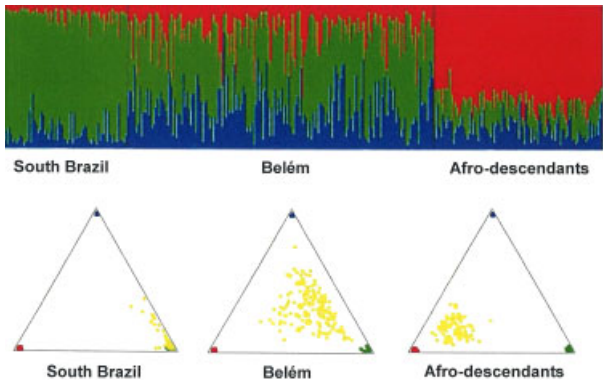

Figure 2. Schematic representation of the individual admixture estimates (IAEs) in three Brazilian mixed populations, using STRUCTURE v. 22 software for K = 3. Bar plot: each vertical line represents one individual and the correspondent European (green), African (red), and Amerindian (blue) admixture proportions. Triangle plots: each individual is represented by a colored point, and the correspondent admixture proportions are indicated by the distance to the edges of the triangle. Green, red, and blue colors correspond to individuals from the parental populations (labeled as above), and individuals from admixed populations are plotted in yellow.

Table 2. Global Interethnic Admixture Estimates in Three Mixed Populations*

	South Brazil			Afro-descendants			Belém		
Software	AFR	EUR	NAM	AFR	EUR	NAM	AFR	EUR	NAM
ADMIXMAP	3.3	89.2	7.5	72.0	14.2	13.8	14.4	57.3	28.3
STRUCTURE	1.0	95.0	4.0	80.7	9.3	10.0	11.7	61.4	26.9
ADMIX95	0.0	91.3	8.7	69.0	14.9	16.1	14.8	54.6	30.6
ADMIX2.0	0.0	100.0	0.0	79.9	0.0	20.1	6.5	65.9	27.6

*Percentages (%) of AFR, EUR, and NAM.
AFR, African; EUR, European; NAM, Native American.

Discussion

In this work we develop an AIM panel capable of identifying stratification in admixed populations and estimating both individual and global ancestry proportions.

Unlike previous studies that employed SNPs as AIMs [Benn-Torres et al., 2008; Halder et al., 2008; Reiner et al., 2005; Shriver et al., 2003], all the markers used in this panel correspond to insertions or deletions of small DNA fragments. A similar approach was used by Bastos-Rodrigues et al. [2006] who used 40 INDELs in order to make inferences about the genetic structure of human populations. However, they selected markers presenting high heterozygosity among Europeans rather than markers informative with regard to ancestry, as in the present work.

Table 3. Quantification of Statistical Bias Level Introduced in Interethnic Admixture Estimates

	AFR	EUR	NAM	Population SE
AFR	0.987	0.008	0.005	0.013
EUR	0.008	0.984	0.008	0.016
NAM	0.003	0.006	0.991	0.009
Ancestry SE	0.011	0.014	0.013	

AFR, African; EUR, European; NAM, Native American; SE, standard error.

The number of INDELs used in this panel is consistent with the estimated number of biallelic markers considered necessary to identify geographically distinct populations [Bamshad et al., 2003; Bastos-Rodrigues et al., 2006; Turakulov and Easteal, 2003].

The capacity of this panel to estimate ancestry is similar to others previously described [Benn-Torres et al., 2008; Halder et al., 2008; Reiner et al., 2005; Shriver et al., 2003; Yang et al., 2005]. We used the data generated by the STRUCTURE software to quantify the level of statistical bias (statistical error [SE]) introduced in the interethnic admixture estimates, as described by Halder et al. [2008]. SE was measured as the average proportion of outside group admixture in nonadmixed individuals (parental populations) and defined as either "population SE" (total ancestry from all noncontributing populations to a nonadmixed individual) or "ancestry SE" (the total contribution from one noncontributing population to all other populations). The results are presented in Table 3. In the three parental populations, the IAE showed over 98% affiliation within the expected group, which establishes a mean population SE of less than 2%. The ancestry SE was also very low (less than 2%), meaning that none of the parental populations can have contributed with more than 2% to the formation of the other two parental populations.

In order to obtain more reliable individual interethnic admixture estimates, we genotyped a considerable number of individuals representative of the ancestral populations. After collecting the population data, we were surprised to find that three of the selected INDELs (MID913, MID1923, and MID1098) had low δ and F_{ST} values between the parental populations. Individual and global interethnic admixture estimates and population substructure analyses performed without these markers gave results essentially equal to those obtained with their inclusion; we therefore concluded that these markers could be excluded from the panel without a significant loss of accuracy.

In the separate analyses of each admixed population, no substructuring was detected in the samples from South Brazil and Afro-descendants, and the genetic structure of Belém was best explained by the existence of two subgroups (K = 2). We then tested whether the INDEL panel was capable of identifying the expected substructure resulting from an artificial joining of the three mixed samples into a single sample. The results indicated that the structure of the group could be best explained by the existence of three clusters (K = 3), as expected when pooling such different samples. We conclude that the panel with 48 INDEL markers is capable of identifying pronounced population substructures (as in the case of the three mixed samples) and even moderate substructuring (as in the case of the population of Belém).

We estimated the individual interethnic admixture proportions in the mixed populations using different methods of analysis and two statistical programs (STRUCTURE and ADMIXMAP).

Independently of the method and program used, all estimates support the expected results: (1) in the sample from South Brazil, the individuals showed an almost exclusive contribution of European genes; (2) the individuals from the Amazonian Afro-descendants sample showed a greater contribution of African genes, but mixed with European and Native American genes; and (3) the individuals from the population of Belém showed a significant contribution of all three ethnic groups, although with a greater contribution of European genes. In order to evaluate the accuracy of the results, we compared the IAEs (for each mixed population) obtained with the two statistical programs using Pearson's correlation test. The results showed that the estimates obtained by both methodologies were statistically similar, with $P < 0.001$ for all comparisons.

We also estimated the global interethnic admixture proportions in the three mixed populations using four statistical programs, and again the results were in agreement with the descriptions and histories of the admixed populations.

In conclusion, we compiled a panel of 48 INDEL polymorphisms that are informative of ancestry and that can be genotyped using three multiplex PCR reactions and gel electrophoresis. The simplicity, accuracy, and low cost of genotyping with this panel are additional advantages over previously described AIM SNP panels, which usually require more complex techniques and workflows.

We demonstrated that this INDEL panel can be used to distinguish continental populations, specifically Europeans, Africans, and Native Americans, and it also identifies substructure in hybrid populations. In admixed populations with multiple parental groups, the marker panel permits accurate estimates of the individual and global interethnic admixture relative to the ancestry groups.

Despite these results, we understand that there is still need to extend these analyses to other continental populations (Asiatic, for instance), as well as other admixed European populations (such as in Central America). Only in this way can this ancestry marker panel's capacity to infer population substructure and interethnic mixture estimation be fully understood.

Acknowledgments

Grant sponsor: Institutos do Milênio, Programa de Apoio a Grupos de Excelência, FINEP (Financiadora de Estudos e Projetos), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), UFPA (Universidade Federal do Pará), Fundação de Amparo à Pesquisa do Estado do Pará, and Fundação para a Ciência e a Tecnologia (Grant SFRH/BD/30039/2006 and POCI 2010). There is not any financial association that might pose any potential conflict of interest to disclose.

References

Alves C, Gusmão L, Damasceno A, Soares B, Amorim A. 2004. Contribution for an African autosomic STR database (AmpF/STR Identifiler and Powerplex 16 System) and a report on genotypic variations. Forensic Sci Int 139:201–205.

Alves C, Gomes V, Prata MJ, Amorim A, Gusmão L. 2007. Population data for Y-chromosome haplotypes defined by 17 STRs (AmpFlSTR YFiler) in Portugal. Forensic Sci Int 171:250–255.

Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB. 2003. Human population genetic structure and inference of group membership. Am J Hum Genet 72:578–589.

Bastos-Rodrigues L, Pimenta JR, Pena SD. 2006. The genetic structure of human populations studied through short insertion-deletion polymorphisms. Ann Hum Genet 70:658–665.

Bedoya G, Montoya P, García J, Soto I, Bourgeois S, Carvajal L, Labuda D, Alvarez V, Ospina J, Hedrick PW, Ruiz-Linares A. 2006. Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. Proc Natl Acad Sci USA 103:7234–7239.

Benn-Torres J, Bonilla C, Robbins CM, Waterman L, Moses TY, Hernandez W, Santos ER, Bennett F, Aiken W, Tullock T, Coard K, Hennis A, Wu S, Nemesure B, Leske MC, Freeman V, Carpten J, Kittles RA. 2008. Admixture and population stratification in African Caribbean populations. Ann Hum Genet 72:90–98.

Bertorelle G, Excoffier L. 1998. Inferring admixture proportions from molecular data. Mol Biol Evol 15:1298–1311.

Carvalho BM, Bortolini MC, Santos SEB, Ribeiro-dos-Santos AKC. 2008. Mitochondrial DNA mapping of social-biological interactions in Brazilian Amazonian African-descendant populations. Genet Mol Biol 31:12–22.

Chakraborty R. 1985. Gene identity in racial hybrids and estimation of admixture rates. In: Neel JV, Ahuja Y, editors. Genetic Micro-differentiation in Man and other Animals. New Delhi: Indian Anthropological Association. p 171–180.

Choudhry S, Burchard EG, Borrell LN, Tang H, Gomez I, Naqvi M, Nazario S, Torres A, Casal J, Martinez-Cruzado JC, Ziv E, Avila PC, Rodriguez-Cintron W, Risch NJ. 2006. Ancestry-environment interactions and asthma risk among Puerto Ricans. Am J Respir Crit Care Med 174:1088–1093.

Cunha MC. 1995. História dos Índios no Brasil. São Paulo: Companhia das Letras, FAPESP, SMC-PMSP. 611 pages.

Curtin PD. 1969. The Atlantic slave trade: a census. Madison, WI: University of Wisconsin Press. 457 pages.

Dupanloup I, Bertorelle G. 2001. Inferring admixture proportions from molecular data: extension to any number of parental populations. Mol Biol Evol 18:672–675.

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50.

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

Feio-dos-Santos AC, Carvalho BM, Batista dos Santos SE, Ribeiro-dos-Santos AK. 2006. Nucleotide variability of HV-I in admixed population of the Brazilian Amazon Region. Forensic Sci Int 164:276–277.

Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T. 2008. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat 29: 648–658.

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM. 2003. Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72:1492–1504.

Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF. 2009. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat 30:69–78.

Marrero AR, Leite FPN, Carvalho BA, Peres LM, Kommers TC, Cruz IMD, Salzano FM, Ruiz-Linares A, Silva Júnior WA, Bortolini MC. 2005. Heterogeneity of the genome ancestry of individuals classified as White in the State of Rio Grande do Sul, Brazil. Am J Hum Biol 17:496–506.

Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16:1182–1190.

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD. 1998. Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63:1839–1851.

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. 2001. Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. Am J Phys Anthropol 114:18–29.

Pimenta JR, Zuccherato LW, Debes AA, Maselli L, Soares RP, Moura-Neto RS, Rocha J, Bydlowski SP, Pena SD. 2006. Color and genomic ancestry in Brazilians: a study with forensic microsatellites. Hum Hered 62:190–195.

Pritchard JK, Stephens M, Donnelly P. 2000a. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000b. Association mapping in structured populations. Am J Hum Genet 67:170–181.

Pritchard JK, Donnelly P. 2001. Case-control studies of association in structured or admixed populations. Theor Popul Biol 60:227–237.

Reiner AP, Ziv E, Lind DL, Nievergelt CM, Schork NJ, Cummings SR, Phong A, Burchard EG, Harris TB, Psaty BM, Kwok PY. 2005. Population structure, admixture, and aging-related phenotypes in African American adults: the Cardiovascular Health Study. Am J Hum Genet 76:463–477.

Ribeiro-dos-Santos AK, Carvalho BM, Feio-dos-Santos AC, Santos SE. 2007. Nucleotide variability of HV-I in Afro-descendants populations of the Brazilian Amazon Region. Forensic Sci Int 167:77–80.

Ribeiro-Rodrigues EM, dos Santos NP, dos Santos AK, Pereira R, Amorim A, Gusmão L, Zago MA, Santos SE. 2009. Assessing interethnic admixture using an X-linked insertion-deletion multiplex. Am J Hum Biol 21:707–709.

Risch N, Burchard E, Ziv E, Tang H. 2002. Categorization of humans in biomedical research: genes, race and disease. Genome Biol 3:comment2007.

Rodrigues EM, Palha TJ, Santos SE. 2007. Allele frequencies data and statistic parameters for 13 STR loci in a population of the Brazilian Amazon Region. Forensic Sci Int 168:244–247.

Salzano FM, Bortolini MC. 2002. The evolution and genetics of Latin American Population. Cambridge, UK: Cambridge University Press.

Santos SEB, Guerreiro JF. 1995. The indigenous contribution to the formation of the population of the Brazilian Amazon Region. Genet Mol Biol 18:311–315.

Santos SEB, Rodrigues JD, Ribeiro-dos-Santos AK, Zago MA. 1999. Differential contribution of indigenous men and women to the formation of an urban population in the Amazon region as revealed by mtDNA and Y-DNA. Am J Phys Anthropol 109:175–180.

Schork NJ, Fallin D, Thiel B, Xu X, Broeckel U, Jacob HJ, Cohen D. 2001. The future of genetic case-control studies. Adv Genet 42:191–212.

Shriver MD, Parra EJ, 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 CJ, McKeigue PM, Kittles RA. 2003. Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 112:387–399.

Silva WA, Bortolini MC, Schneider MP, Marrero A, Elion J, Krishnamoorthy R, Zago MA. 2006. MtDNA haplogroup analysis of black Brazilian and sub-Saharan populations: implications for the Atlantic slave trade. Hum Biol 78: 29–41.

Tsai HJ, Kho JY, Shaikh N, Choudhry S, Naqvi M, Navarro D, Matallana H, Castro R, Lilly CM, Watson HG, Meade K, Lenoir M, Thyne S, Ziv E, Burchard EG. 2006. Admixture-matched case-control study: a practical approach for genetic association studies in admixed populations. Hum Genet 118:626–639.

Turakulov R, Easteal S. 2003. Number of SNPS loci needed to detect population structure. Hum Hered 55:37–45.

Vallone PM, Butler JM. 2004. AutoDimer: a screening tool for primer-dimer and hairpin structures. Biotechniques 37:226–231.

Weber JL, David D, Heil J, Fan Y, Zhao C, Marth G. 2002. Human diallelic insertion/deletion polymorphisms. Am J Hum Genet 71:854–862.

Yang N, Li H, Criswell LA, Gregersen PK, Alarcon-Riquelme ME, Kittles R, Shigeta R, Silva G, Patel PI, Belmont JW, Seldin MF. 2005. Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. Hum Genet 118:382–392.