Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Development of two multiplex mini-sequencing panels of ancestry informative SNPs for studies in Latin Americans: an application to populations of the State of Minas Gerais (Brazil)

M.C.F. Silva[1,2*], L.W. Zuccherato[2*], G.B. Soares-Souza[1,2], Z.M. Vieira[1], L. Cabrera[3], P. Herrera[3], J. Balqui[3,4], C. Romero[3,4], H. Jahuira[3,4], R.H. Gilman[3,5], M.L. Martins[1] and E. Tarazona-Santos[2]

[1]Fundação Hemominas, Belo Horizonte, MG, Brasil
[2]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[3]Asociación Benéfica PRISMA, Urbanización Maranga, Lima, Peru
[4]Laboratorio de Investigación en Enfermedades Infecciosas, Universidad Peruana Cayetano Heredia, San Martín de Porres, Lima, Peru
[5]Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

*These authors contributed equally to this study.
Corresponding author: E. Tarazona-Santos
E-mail: edutars@icb.ufmg.br

Genet. Mol. Res. 9 (4): 2069-2085 (2010)
Received May 21, 2010
Accepted July 23, 2010
Published October 19, 2010
DOI 10.4238/vol9-4gmr911

ABSTRACT. Admixture occurs when individuals from parental populations that have been isolated for hundreds of generations form a new hybrid population. Currently, interest in measuring biogeographic ancestry has spread from anthropology to forensic sciences, direct-to-consumers personal genomics, and civil rights issues of minorities, and it is critical for genetic epidemiology studies of admixed populations. Markers with highly differentiated frequencies among human populations are informative of ancestry and are called ancestry informative markers (AIMs). For tri-hybrid Latin American populations, ancestry information is required for Africans,

Europeans and Native Americans. We developed two multiplex panels of AIMs (for 14 SNPs) to be genotyped by two mini-sequencing reactions, suitable for investigators of medium-small laboratories to estimate admixture of Latin American populations. We tested the performance of these AIMs by comparing results obtained with our 14 AIMs with those obtained using 108 AIMs genotyped in the same individuals, for which DNA samples is available for other investigators. We emphasize that this type of comparison should be made when new admixture/population structure panels are developed. At the population level, our 14 AIMs were useful to estimate European admixture, though they overestimated African admixture and underestimated Native American admixture. Combined with more AIMs, our panel could be used to infer individual admixture. We used our panel to infer the pattern of admixture in two urban populations (Montes Claros and Manhuaçu) of the State of Minas Gerais (southeastern Brazil), obtaining a snapshot of their genetic structure in the context of their demographic history.

Key words: Admixture; Latin American; Mini-sequencing

INTRODUCTION

Admixture occurs when individuals from populations that have been isolated for hundreds of generations (i.e., parental populations) form a new hybrid population. Latin-Americans (Hispanics/Latino in the United States), African-Americans and Caribbeans in the Americas (Sans, 2000; Salzano and Bortolini, 2002; Benn-Torres et al., 2008; Herrera-Paz et al., 2010), Central Asian (Comas et al., 1998) and South African colored (Patterson et al., 2010) are examples of admixed human populations. Chromosomes of admixed individuals may be conceived as mosaics of chunks with different ancestry, which sizes reduce along time by recombination among chromosomes with different ancestry (Falush et al., 2003). The recent availability of millions of markers across the human genome for different populations has made it possible to infer admixture (also called biogeographic ancestry) not only for populations, as has been traditional, but also for individuals and for specific genomic regions along a chromosome (Via et al., 2009).

The interest in biogeographic ancestry estimations is not limited to anthropology anymore; it has spread to forensic sciences, direct-to-consumers personal genomics and civil rights issues of minorities (Lee et al., 2009). Estimating biogeographic ancestry is critical for genetic epidemiology of admixed populations (Tarazona-Santos et al., 2007). In fact, in a well-designed case-control association study, cases and controls should be sampled from the same population and thereby be ethnically homogeneous. Otherwise, spurious statistical association for any allele more common in a parental population may result if the disease is more prevalent in this population and therefore, individuals with a predominant ancestry of this population are oversampled among cases. Hence, the first step in a case-control association study in an admixed population should be to measure admixture of the participants and ascertain if cases and controls are ethnically different (i.e., if population stratification exists). In this case, it is possible to test the statistical association among genetic variants and biomedical traits controlling for population stratification at the population level using genomic control methods, or at the individual level using regression analysis or structured association methods (Tarazona-Santos et al., 2007).

In this context of broad interest in admixture studies, it is important to recognize that measuring admixture is methodologically complex (Chakraborty, 1986; Choisy et al., 2004). For instance, difficulties in estimating admixture increase from population to individual and chromosomal levels (Pritchard et al., 2000; Falush et al., 2003). Another complicating factor is the number of parental populations that have contributed to the gene pool of the admixed population/individual. Markers with frequencies that are highly differentiated among populations are particularly informative of ancestry and are called ancestry informative markers (AIMs). If enough AIMs are genotyped, they allow estimates of admixture at the levels of population, individual and chromosome regions. The use of AIMs reduces the number of markers that need to be genotyped to infer admixture at population and individual levels, compared to the genotyping of randomly selected markers (Rosenberg et al., 2003; Parra et al., 2003; Pfaff et al., 2004). For tri-hybrid Latin American populations, ancestry information is required for African, European and Native American populations. However, there is no unique and optimal set of markers (or AIMs) for all Latin American populations, because informativeness depends on the combination of allele frequencies in the parental populations and on admixture proportions (Pfaff et al., 2004). In general, for Latin American tri-hybrid populations, the best AIMs have a very different frequency in one of the three parental populations and similar frequencies in the other two.

The number of markers that are necessary to estimate population admixture or individual ancestry depends on the informativeness of the markers and the required accuracy. Currently, Affymetrix and Illumina commercial arrays allow to genotype up to ~10^6 markers scattered in the human genome, at a cost of few hundreds of US dollars per individual (Chung et al., 2010). Even if most of these single-nucleotide polymorphisms (SNPs) are not AIMs, with this resolution it is possible to estimate individual and chromosomal region ancestries with high accuracy. Although the cost of genome-wide genotyping is declining, this possibility is still limited to a few research groups, in particular for admixture studies at the population level, which require large sample sizes. For small-medium laboratories, it would be advantageous to use small-medium size panels of AIMs for studies at population and individual levels, at a cost of few dozens of dollars per individual. Kosoy et al. (2009) have shown that panels of 24 AIMs are useful to ascertain the origin of subjects from particular continents and to correct for population stratification at the population level. Some low-medium cost multiplex panels of AIMs have been published (Lins et al., 2010; Santos et al., 2010); however, these may vary in their informativeness. Therefore, the best option for an investigator performing an admixture study is to assess which combination of AIMs is most informative for the target population.

We developed two multiplex panels of AIMs that include 14 SNPs to estimate admixture in Latin American populations. We tested the performance of these 14 AIMs by comparing admixture estimates obtained with this set of markers, with estimates obtained using 108 AIMs (that we assume to be more accurate). We used our panels to infer the pattern of admixture in two populations of the State of Minas Gerais (southeastern Brazil), obtaining a snapshot of their genetic structure in the context of their demographic history.

MATERIAL AND METHODS

Selection of AIMs for the two panels

To design two panels of AIMs to be genotyped by multiplex mini-sequencing reac-

tions, we pre-selected a large set of candidate AIMs by two procedures: 1) 250 unlinked AIMs were selected based on their informativeness (index I_a of Rosenberg et al., 2003) from the admixture mapping panel of Tian et al. (2006) to assess African/European admixture. 2) 150 SNPs informative of Native American admixture were selected from the SNP500 Cancer resource (Packer et al., 2006, http://variantgps.nci.nih.gov/cgfseq/pages/snp500.do), based on differences in allele frequencies between European, African and Pima-Maya Native American populations. These SNPs were pre-selected by avoiding physical proximity in the human genome. We assessed compatibility for multiplex polymerase chain reaction (PCR) amplification using the Muplex resource (Rachlin et al., 2005), which is a convenient web-enabled system that, starting from a set of targeted sequences, automatically designs sub-sets of primers that will likely co-amplify in multiplex PCR assays under a number of conditions imposed by the investigator. After applying these criteria, we selected the following two SNP panels to be tested experimentally: AFR (Africans) (rs2697520, rs8035530, rs1372115, rs2789823, rs241679, rs7512316, rs9626698, rs1443985, rs6046024, rs735480) and AMR (Native Americans) (rs8058694, rs691968, rs2234636, rs3760657, rs2619681, rs2569190, rs800292, rs2518967, rs2088102, rs700518). We also evaluated the specificity of primers using the electronic PCR tool (http://www.ncbi.nlm.nih.gov/sutils/e-pcr/reverse.cgi). Among these SNPs, the following were excluded for further genotyping because of their high rate of missing data or because of a lack of reproducible results: rs2789823, rs7512316, rs6046024, rs2569190, rs2518967, and rs700518.

Genotyping

Genotyping by mini-sequencing consists of three steps (Carvalho and Pena, 2005): 1) amplification of regions flanking the SNPs by multiplex PCR; 2) multiplex mini-sequencing; 3) analysis of mini-sequencing products by capillary electrophoresis.

Amplification of regions flanking the SNPs by multiplex PCR

Primers designed using Muplex performed well when experimentally tested and are available as Supplementary Material (Table S1). PCR was performed in a volume of 25 μL with 100 ng genomic DNA, 0.2 μM of each primer and 1X of a commercial master-mix (Qiagen Multiplex PCR Master Mix or 1.5 U Platinum Taq DNA Polymerase from Invitrogen plus 1X STR buffer from Promega). Amplification consisted of 95°C for 5 min, followed by 30 cycles of 30 s at 94°C, 90 s at 57°C, 90 s at 72°C, and a final extension for 10 min at 72°C. After the amplification, we performed enzymatic purification of the PCR product (i.e., removal of remaining PCR primers and dNTPs before the mini-sequencing reaction, using respectively exonuclease I and shrimp alkaline phosphatase, as detailed in the Supplementary Material).

Multiplex mini-sequencing

For each locus of a multiplex panel, a mini-sequencing primer with the 3'-end adjacent to the target SNP was designed to anneal with the PCR product (Figure 1B). A mini-sequencing reaction extends this primer, producing different products for each allele, which

are visualized by capillary electrophoresis. Different loci are differentiated by the sizes of the mini-sequencing primers, which include PUC 18 plasmid sequence tails with a specific size for each locus. The mini-sequencing primers are available as Supplementary Material (Table S1). We selected the SNPs for these assays to allow genotyping both by: a) monochromatic mini-sequencing (Carvalho and Pena, 2005) and b) polychromatic mini-sequencing. Monochromatic mini-sequencing uses a homemade mix with one of the nucleotides in the form of fluorescent ddNTP (in our case ddATP-FAM) and the other three as dNTPs (Figure 1C). In this way, it is only possible to genotype SNPs with an allele complementary to the ddNTP (i.e., T in our case), in which case the primer is extended by a single ddNTP. The mini-sequencing product for the other allele will be extended until the next position containing the same nucleotide (i.e., T in our case) in the sequence, and therefore alleles are differentiated by size (Figure 1E). Polychromatic mini-sequencing does not use dNTPs; it only uses four ddNTPs with different fluorescence (and therefore, primers will always be extended by a single ddNTP; Figure 1D). Thus, alleles will be distinguished by colors (Figure 1F). The commercial kits SNaPshot (Applied Biosystems) and SNuPe (GE Healthcare) are available for polychromatic mini-sequencing. To use the same set of mini-sequencing primers in monochromatic or polychromatic protocols, we avoided combinations of A/G and C/T polymorphisms in the same multiplex reaction, since these variants cannot be co-genotyped by the monochromatic protocol, an option for investigators who prefer to use homemade reagents.

We performed the multiplex monochromatic mini-sequencing in a 13.5-µL volume with 2 µL purified PCR product and 0.37 µM of each primer, 0.46 µM ddATP labeled with fluorescein (Perkin Elmer Life Sciences), 0.46 µM of unlabeled dCTP, dTTP and dGTP (GE Healthcare), 1X Thermo Sequenase reaction buffer, and 1 U Thermo Sequenase DNA Polymerase (GE Healthcare). The thermal cycling consisted of 2 min at 80°C for denaturation, followed by 30 cycles of 30 s at 95°C, 30 s at 55°C and 20 s at 72°C. The polychromatic protocol (SnaPshot Multiplex System, Applied Biosystems) consisted of a 5-µL volume reaction containing 1 µL purified PCR product, 1 µL SNaPshot™ Kit Reaction Mix and 2 µL primer mix (at a concentration of 1 µM of each primer). The thermal cycling consisted of 2 min at 96°C for denaturation, followed by 25 cycles of 10 s at 95°C, 5 s at 55°C and 30 s at 60°C. After the mini-sequencing reaction, we performed an enzymatic purification of the reaction (see Supplementary Material for details).

A*nalysis of the mini-sequencing products by capillary electrophoresis*

For the monochromatic protocol, a mixture of 2.0 µL mini-sequencing products diluted twice, plus 7.75 µL Tween 20 at 0.1% and 0.25 µL of the size standard ET-ROX 550 (GE Heathcare) was applied in a Megabace DNA sequencer (GE Heathcare). The run parameters were: injection voltage of 3 Kv, injection time of 80 s, run voltage of 10 Kv and run time of 75 min. The analyses were done with the Fragment Profiler software (GE Heathcare). For the polychromatic option, a mixture of 1.0 µL of the SNaPshot product, 8.9 µL Hi-Di formamide and 0.1 µL Liz120 Size Standard (Applied Biosystems) was applied in an ABI 3130 DNA sequencer (Applied Biosystems). The run parameters were: injection voltage of 1.2 Kv, injection time of 18 s, run voltage of 15 Kv and run time of 800 s (capillary size of 36 cm). In this case the analyses were done with the package GeneScan Analysis 3.7 or Genotyper 3.7 software (Applied Biosystems).

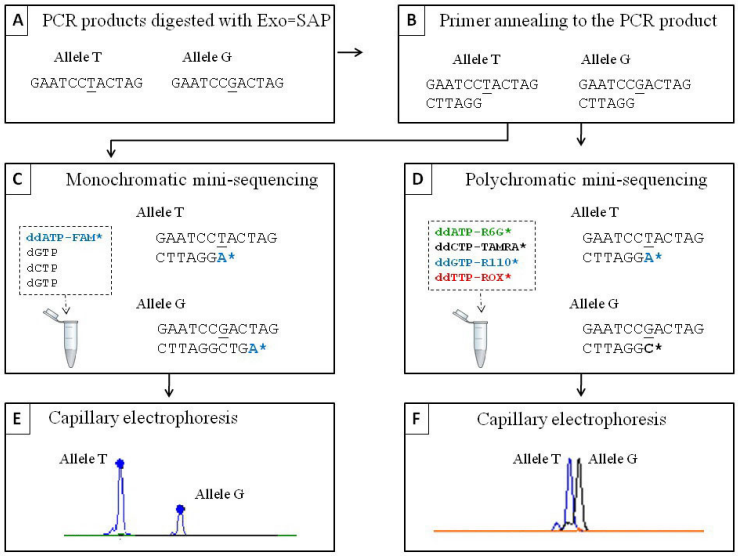

Figure 1. Representation of genotyping by monochromatic (A,B,C,E) or polychromatic (A,B,D,F) mini-sequencing. See text for details. PCR = polymerase chain reaction; Exo-SAP = a reaction containing *E. coli* exonuclease I + shrimp alkaline phosphatase.

Samples of parental and admixed populations

We used the following three sets of individuals as putative parental populations of the Latin American admixed samples: 1) 31 European ancestry and 2) 24 African ancestry from the SNP500Cancer panel (http://variantgps.nci.nih.gov/cgfseq/pages/snp500.do; Packer et al., 2006). We also used 3) 85 Peruvian Native Americans settled between the eastern slope of the Andes and the Amazon tropical forest (in the region called High Forest or "Selva Alta"). Some of these individuals are from the region of Cusco and belong to the communities of Shimaa (N = 30) and Monte Carmelo (N = 15) from the Matsiguenga linguistic group, and some of them (N = 40) reside in Ashaninka villages along the Tambo River (Region of Junin).

Admixed samples included three sets of individuals: 1) 23 Latin American catalogued as Hispanic in the SNP500Cancer initiative, 2) 24 Brazilian individuals from the city of Montes Claros, at north of the State of Minas Gerais, and 3) 30 Brazilian individuals from the city of Manhuaçu, eastern Minas Gerais. Brazilian samples were from healthy and unrelated blood donors attending centers of the Minas Gerais Blood Bank in their respective cities. The inclusion of European, African and Latin American individuals from the SNP500Cancer initiative is convenient because they have been genotyped for a large set of polymorphisms in the context of the SNP500Cancer initiative, and this information can be used to assess the informativeness of the panels of AIMs that we developed. Institutional Review Boards from the participant institutions approved this study.

Statistical and population genetics analyses

We estimated population admixture using: 1) The gene identity method developed by Chakraborty (1985), as implemented in the ADMIX95 software (developed by Bernardo Bertoni and available at http://www.genetica.fmed.edu.uy/software.htm). This method takes into account sampling error and the effect of genetic drift in the parental and admixed populations (Chakraborty, 1986) and 2) The

coalescent-based method by Dupanloup and Bertorelle (2001), which, in addition to sampling and drift errors in parental and admixed populations, considers the degree of divergence at the time of admixture.

Individual admixture was estimated using the Bayesian clustering algorithms developed by Pritchard and implemented in the STRUCTURE v2.3.2 program (Pritchard et al., 2000; Hubisz et al., 2009). We assumed that three parental populations (K = 3 clusters) contributed to the genome of the admixed individuals. STRUCTURE estimates individual admixture conditioning in Hardy-Weinberg and linkage equilibrium in each of the K = 3 clusters, which represent the parental populations. We ran the program using a burn-in period of 100,000, and 100,000 repetitions of MCMC after burning. We used prior population information for individuals from the parental populations to assist clustering (USEPOPINFO = 1) and assumed the admixture model for individuals from the admixed populations, inferring the alpha parameter for each population. We also used the parameters GENSBACK = 2 and MIGRPRIOR = 0.05. Moreover, we assumed that allele frequencies were correlated and that the different populations have different levels of differentiation (F_{ST}, with prior mean = 0.01 and standard deviation = 0.05). Based on individual admixture estimations obtained with STRUCTURE, population admixture was averaged over individuals.

RESULTS

The allele frequencies for the 14 AIMs of this study in European, African and Native American populations (for which public data are available), as well as for the parental and admixed samples, are given in Table 1. In general, there were large differences between continental groups and small differences within these groups, which confirm that the selected markers are informative for ancestry.

Table 1. Frequencies of the genotyped single-nucleotide polymorphisms (SNPs) for the different populations.

SNP	Population																
	AFR1	AFR	YRI	NiloS	Edo	AFA	EUR1	EUR	CEU	DEU	AMR	MCA	ASH	SHI	MCU	MOC	HISP
AFR panel																	
rs1443985	0.15	0.09	0.07	0.11	0.13	0.21	0.85	0.91	0.90	0.94	0.84	0.85	0.76	0.92	0.71	0.57	0.71
rs9626698	0.50	0.79	0.79	0.76	0.81	0.57	0.00	0.04	0.03	0.04	0.00	0.00	0.00	0.00	0.12	0.17	0.12
rs2416791	0.76	0.05	0.03	0.10	0.00	0.18	0.88	0.92	0.92	0.92	0.53	0.64	0.44	0.50	0.71	0.57	-
rs2697520	0.17	0.07	0.05	0.10	0.06	0.28	0.80	0.88	0.89	0.85	0.31	0.31	0.26	0.34	0.36	0.50	0.50
rs8035530	0.70	0.82	0.79	0.86	0.79	0.67	0.08	0.02	0.02	0.02	0.48	0.44	0.53	0.48	0.18	0.25	0.19
rs735480	0.09	0.03	0.00	0.06	0.02	0.21	0.90	0.95	0.95	0.96	0.88	0.82	0.99	0.84	0.82	0.58	0.86
rs1372115	1.00	0.99	1.00	0.97	1.00	0.79	0.37	0.07	0.17	0.23	0.41	0.07	0.62	0.54	0.43	0.56	0.92
AMR panel																	
rs2234636	0.20	0.00	-	-	-	-	0.28	0.28	-	-	0.78	0.77	0.73	0.83	0.35	0.21	0.37
rs3760657	0.09	0.02	-	-	-	-	0.10	0.07	-	-	0.26	0.12	0.40	0.26	0.03	0.06	0.07
rs8058694	0.59	0.23	-	-	-	-	0.88	0.61	-	-	0.70	0.62	0.75	0.72	0.37	0.54	0.39
rs800292	0.63	0.63	-	-	-	-	0.25	0.18	-	-	0.93	1.00	0.87	0.91	0.35	0.52	0.35
rs2619681	0.17	0.06	-	-	-	-	0.13	0.17	-	-	0.86	0.88	0.90	0.79	0.15	0.15	0.26
rs691968	0.41	0.42	-	-	-	-	0.04	0.00	-	-	0.01	0.00	0.00	0.02	0.10	0.09	0.17
rs2088102	0.71	0.65	-	-	-	-	0.54	0.50	-	-	0.33	0.27	0.40	0.33	0.98	0.89	0.34

AFR1 = Africans (SNP500Cancer panel, genotyped in this study); AFR = Africans (Tian et al., 2006); YRI = Yoruban (West Africans from the HapMap project); NiloS = Kanuri (Nilo-Saharan speakers from Nigeria); Edo = Bini (Niger-Congo group of Bantu speakers); AFA = African-Americans (Coriell Institute for Medical Research); EUR1 = European (SNP500Cancer panel, genotyped in this study); EUR = Europeans (Tian et al., 2006); CEU = CEPH European; DEU = European-Americans from New York City; AMR = Average for all Amerindians (genotyped in this study); MCA = Monte Carmelo Amerindians; ASH = Ashaninka Amerindians; SHI = Shimaa Amerindians; HIS = Hispanic (SNP500Cancer panel, genotyped in this study); MCU and MOC = Brazilian samples from Manhuaçu and Montes Claros, respectively. The data presented for Africans (AFR, YRI, NiloS, Edo, AFA) and Europeans (EUR, CEU, DEU) were obtained from Tian et al., 2006 (AFR panel) or SNP500Cancer database (AMR panel).

Population admixture

First, we tested our set of 13 AIMs by estimating population admixture for the Latin American sample of the SNP500Cancer project, hereafter called Hispanics to follow the SNP500Cancer nomenclature. Hughes et al. (2008), using ~350 K SNPs, have shown the predominant European ancestry of this set of individuals. Although our set of AIMs includes 14 SNPs, in the analysis including Hispanics we conservatively considered only 13 AIMs, due to the high missing rate of rs241679, specifically in the Hispanic sample. We compared our estimates with those obtained using 108 AIMs, selected for admixture estimation among thousands of SNPs genotyped in the SNP500Cancer project using the criterion that they show $F_{ST} > 0.20$ between European, African and Native American (Pima and Maya populations from CEPH; Cann et al., 2002) and $F_{ST} < 0.10$ between populations within these groups (unpublished data; see Table S2 of the Supplementary Material, for additional details). Individual genotypes and allele frequencies for these 108 SNPs in the parental populations are available in the SNP500Cancer website. We assume that these 108 AIMs provide a more accurate estimate of admixture than our 13 AIMs. Both sets of 108 and 13 AIMs confirm the predominant European ancestry of the Hispanic sample. By using the 13 AIMs in a sample with predominant European ancestry, we obtained accurate estimates of European admixture at the population level (Table 2). However, our set of 13 AIMs seems to overestimate African and underestimate Native American admixture. The methods developed by Chakraborty (1985) and Dupanloup and Bertorelle (2001) to infer population admixture, consistently estimate higher African admixture and lower European admixture than the STRUCTURE method (focused on individual admixture) (Table 1).

Table 2. Population admixture estimation obtained by three methods of analysis for Hispanic and Brazilian samples from Montes Claros and Manhuaçu, Minas Gerais.

	Parental populations					
	African		European		Native American	
	Point estimate	95% CI or SD	Point estimate	95% CI or SD	Point estimate	95% CI or SD
Dupanloup and Bertorelle (2001)						
Hispanic-13 SNPs	0.31	0.05	0.62	0.06	0.07	0.05
Hispanic-108 SNPs	0.15	0.03	0.66	0.03	0.20	0.02
Montes Claros-BR	0.41	0.05	0.54	0.07	0.05	0.06
Manhuaçu-BR	0.27	0.04	0.63	0.06	0.11	0.06
Chakraborty (1985)						
Hispanic-13 SNPs	0.34	0.02	0.58	0.02	0.08	0.02
Hispanic-108 SNPs	0.16	0.00	0.64	0.00	0.20	<0.01
Montes Claros-BR	0.41	0.00	0.54	0.00	0.05	<0.01
Manhuaçu-BR	0.27	0.01	0.63	0.01	0.09	0.01
Pritchard et al. (2000)						
Hispanic-13 SNPs	0.23	-	0.69	-	0.08	-
Hispanic-108 SNPs	0.09	-	0.75	-	0.16	-
Montes Claros-BR	0.39	-	0.52	-	0.09	-
Manhuaçu-BR	0.19	-	0.73	-	0.08	-

CI = confidence interval; SD = standard deviation; SNPs = single-nucleotide polymorphisms.

After testing the performance of our 14 AIMs to estimate population admixture, we interpreted admixture estimation in the Brazilian Minas Gerais samples of Montes Claros and Manhuaçu. Both populations have a predominant European admixture (>50%). On the basis of our test with the Hispanic sample, we consider estimates of African and Native American ancestry for the Minas Gerais samples as maximum and minimum values, respectively. In agreement with its geographical proximity to northern Bahia State, the Montes Claros population showed a higher African contribution than Manhuaçu.

Individual admixture

The 14 AIMs used in this study correctly assigned the individuals from the parental populations to their actual group, even when no information was given to the STRUCTURE algorithm about their population of origin (i.e., when we ran the program using all the parameters specified in the Material and Methods section, but specifying USEPOPINFO = 0; Figure S1 of the Supplementary Material). This implies that individuals with ancestry near 100% for any of the parental populations are correctly identified by our small panel of AIMs. However, this does not imply that admixture will be accurately estimated in more admixed individuals.

To assess the accuracy of individual admixture inferences using the 14 AIMs, we ran the STRUCTURE software in the Hispanic sample by using the set of 108 AIMs (Figure 2), and compared these results with those obtained by subsets of 80, 50 and 20 randomly selected SNPs, and also with our set of 13 AIMs (Figure 3). Using subsets of 50 or 80 AIMs, the correlation of admixture estimations was higher than 85%, while with fewer than 20 AIMs, this correlation was less than 66% (Table 3). Thus, even if our set of 13 AIMs contains information to estimate population admixture, individual admixture inferences for admixed individuals are not accurate with less than 20 AIMs. This is also evident when the Bayesian 90% probability interval (90% PI) of individual admixture generated by the STRUCTURE software is plotted against individual admixture (Figure 3); for admixed individuals, the 90% PI of individual admixture rises dramatically when reducing the number of AIMs below 50.

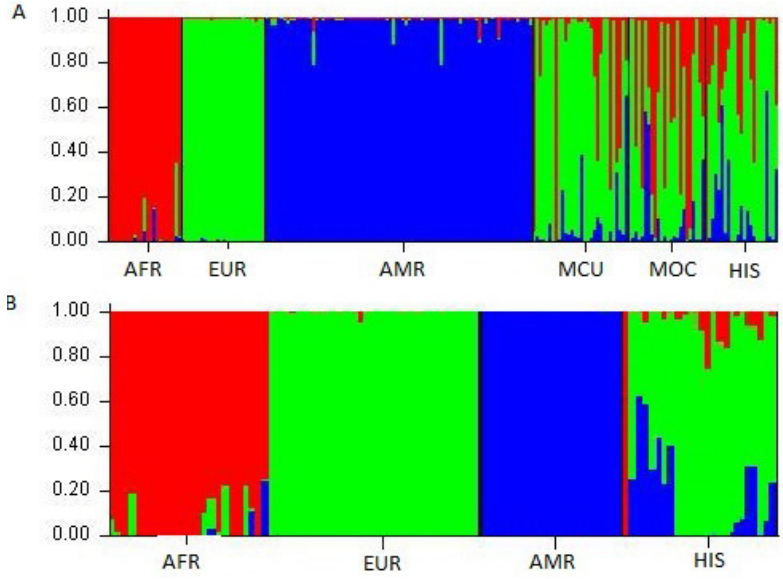

Figure 2. Individual admixture estimated by STRUCTURE with 14 ancestry informative markers (AIMs) for this study (A) and for the Hispanics using 108 AIMs (B). Each individual is represented by a vertical bar. Inferred African, European and Native American admixture is represented by red, green and blue, respectively. AFR = Africans; EUR = Europeans; AMR = Native Americans; MCU = Manhuaçu; MOC = Montes Claros; HIS = Hispanics from SNP500Cancer. Admixture estimates obtained using 108 AIMs considered only Monte Carmelo as Native American parental population.

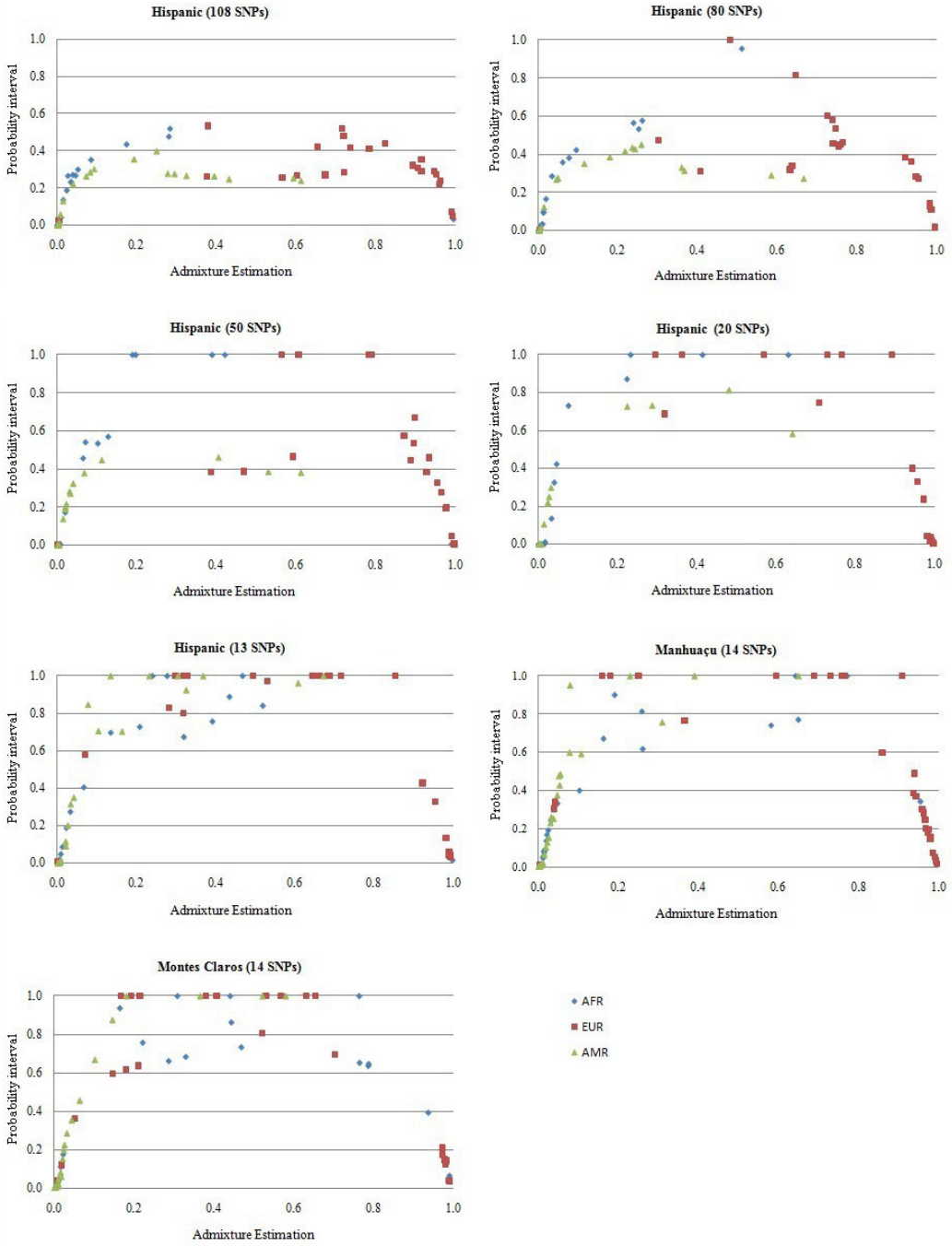

Figure 3. Variation of 90% probability intervals of individual admixture estimates as a function of point estimates of individual admixture, as inferred by the STRUCTURE algorithm with different numbers of single-nucleotide polymorphisms (SNPs) and in different populations. AFR = Africans; EUR = Europeans; AMR = Native Americans.

Table 3. Correlation coefficients between individual admixture point estimations in the Hispanic sample by using 108 ancestry informative markers (AIMs), and admixture estimations using randomly selecting subsets of 80, 50, 20 AIMs, and also by using our set of 13 AIMs.

AIM subsets	Parental population					
	African		European		Amerindian	
	Correlation coefficient	P	Correlation coefficient	P	Correlation coefficient	P
80 SNPs	0.93	>0.001	0.93	>0.001	0.96	<0.001
50 SNPs	0.96	>0.001	0.93	>0.001	0.88	<0.001
20 SNPs	0.58	0.004	0.66	0.001	0.64	0.001
13 SNPs	0.60	0.003	0.49	0.019	0.52	0.011

SNPs = single-nucleotide polymorphisms.

DISCUSSION

We presented two new multiplex panels of AIMs (for a total of 14 SNPs) developed to assist investigators of small-medium size laboratories to estimate admixture in Latin American populations. We gave methodological information in detail to allow other investigators to use these panels, to use individual genotypes of the parental populations in other admixture studies, or to follow the same steps to design additional panels of AIMs more suitable for specific populations to obtain more accurate estimates of admixture. Specifically, we recommend the use of the Muplex resource (Rachlin et al., 2005) to design primers for multiplex amplification and subsequent genotyping. Muplex may also be used to design multiplex panels of insertion-deletion, that have proven to be cost-effective markers for admixture and population structure studies (Bastos-Rodrigues et al., 2006; Santos et al., 2010).

A methodological issue when estimating admixture and population structure using few markers is to test the accuracy of the results. This may be achieved by using a reference set of samples that have been genotyped for the small number of markers (in this case the 14 AIMs) and for a large set of polymorphisms. This strategy has been followed by Bastos-Rodrigues et al. (2006), using the reference HGDP-CEPH panel of DNA samples (Cann et al., 2002) to test the performance of 48 INDELS to study the genetic structure of human populations. We tested the performance of our 14 AIMs by comparing population and individual admixture estimations obtained with this set of markers with those obtained using 108 AIMs (that we assume to be more accurate) in the Hispanic sample of the SNP500Cancer project. This sample, as well as the European and African ancestry used as parental populations, and the Pima and Maya samples used to select the 108 AIMs, is an appropriate reference because they are available as immortalized cells in the Coriell repository (www.coriell.org), which can provide unlimited good-quality DNA to reproduce or extend our results. Immortalized cells are available for reference samples, such as the CEPH-HGDP (Cann et al., 2002), the SNP500Cancer (Packer et al., 2006) and HapMap (the International HapMap Consortium 2007), for which large genome-wide genotype datasets are publicly available. It is important that new sets of markers developed to study admixture or population structure be tested using these resources. By testing the performance of our set of markers, we identified their strengths and limitations. At the population level they are appropriate to estimate European admixture, they overestimate African ancestry and underestimate Native American ancestry. As expected because of the use of few markers, they do not provide adequate estimates of individual admixture, except to identify individuals from the parental populations. However, the ability to identify individuals from the parental populations (European, African or Native American)

should not be generalized to the power to accurately estimate ancestry of admixed individuals.

A pervasive methodological issue in admixture studies is the identification of appropriate parental European, African and Native American populations (Glass and Li, 1953; Chakraborty, 1986). With the use of AIMs, this issue is mitigated by the choice of markers that have frequencies that are very different among the parental groups, but are very homogeneous within them. The selection of markers with these characteristics is facilitated by the availability of datasets of genome-wide surveys for different populations, which include the 52 worldwide populations of the CEPH-HGDP panel and the populations of Phase III of HapMap (International HapMap Consortium, 2010), although Native Americans are still under-represented in these studies. Most of the 14 AIMs that we selected reasonably fit the pattern of differentiation required for AIMs (Table 1). Thus, the effect of our suboptimal choice of parental populations (a limitation shared with most admixture studies) is partially counterbalanced by our use of AIMs.

We estimated admixture in the populations of Montes Claros and Manhuaçu of the State of Minas Gerais (southeastern Brazil), which hosted one of the largest Brazilian populations of African ancestry slaves during the Colonial period. In the geographic area of the region of Manhuaçu (eastern part of the state), slaves were 40% of the population in 1840 (Luna and Klein, 2004), but since the end of the 19th century, this geographic region received a large number of European ancestry immigrants attracted by a flourishing agricultural economy, mainly based on coffee (Botelho et al., 2007). The predominant European ancestry in the Manhuaçu sample complements these historical records, suggesting that recent European immigrants had a substantial impact in the genetic structure of this population. Montes Claros is located in the northern part of Minas Gerais. Though it is a region with one of the smallest populations of African ancestry slaves during the Colonial period (15% of the total population in 1833; Botelho, 1994; Luna and Klein, 2004), our results suggest a substantial African contribution. This may be related to its geographical proximity to Bahia, currently the Brazilian state with the largest proportion of self-identified "Black" individuals (IBGE, 2007). Our results suggest that at least in the State of Minas Gerais (one of the largest in extension and population in Brazil), historical demographic data about African ancestry slaves are not good indicators of the contribution of African ancestry to the current urban local population.

In conclusion, we developed two multiplex panels informative to estimate African (seven SNPs) and Native American (seven SNPs) ancestry, useful to assist investigators of small laboratories in studying the genetic structure of Latin American populations. Our panel of 14 AIMs allows accurate estimation of European population ancestry, but has limited power to estimate individual admixture. Thus, it is a useful tool to be used in combination with other available sets of markers to assess admixture and the genetics structure of Latin American populations (Bastos-Rodrigues et al., 2006; Kosoy et al., 2009; Lai et al., 2009; Lins et al., 2010). Flexibility in measuring admixture is important because depending on the degree of admixture of the targeted populations, the optimal set of markers to infer admixture may vary (Pfaff et al., 2004). Assessing the informativeness, strengths and limitations of new panels of AIMs is necessary to make correct inferences about admixture processes in Latin American populations.

ACKNOWLEDGMENTS

We are grateful to all blood donors and to Telma Regina Guedes Machado (Hemocentro Regional de Montes Claros) and Simone Avelino Rodrigues (Núcleo Regional de Ma-

nhuaçu) from Fundação Hemominas for sample collections, Laboratório de Biodiversidade e Evolução Molecular and Prof. Fabricio Santos, Núcleo de Análise de Genomas e Expressão Gênica (NAGE) and Prof. Santuza Teixeira, Laboratório de Genética Bioquímica and Prof. Gloria Franco for access to the Megabace sequencers, and Rinaldo Pereira and Túlio Lins for collaboration with the Applied Biosystems protocols. Research supported by the National Institutes of Health - Fogarty International Center (1R01TW007894-01 to E. Tarazona-Santos), Brazilian National Research Council (CNPq), Brazilian Ministry of Education (CAPES) and Minas Gerais State Foundation in Aid of Research (FAPEMIG).

REFERENCES

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

Benn-Torres J, Bonilla C, Robbins CM, Waterman L, et al. (2008). Admixture and population stratification in African Caribbean populations. *Ann. Hum. Genet.* 72: 90-98.

Botelho TR (1994). Famílias e Escravarias: Demografia e Família Escrava no Norte de Minas Gerais no Século XIX. Master's thesis, Universidade de São Paulo, São Paulo.

Botelho TR, Braga MP and de Andrade CV (2007). Immigration and family in Minas Gerais at the end of the 19th century. *Rev. Bras. Hist.* 27: 155-176.

Cann HM, de Toma C, Cazes L, Legrand MF, et al. (2002). A human genome diversity cell line panel. *Science* 296: 261-262.

Carvalho CM and Pena SD (2005). Optimization of a multiplex minisequencing protocol for population studies and medical genetics. *Genet. Mol. Res.* 4: 115-125.

Chakraborty R (1985). Gene Identity in Racial Hybrids and Estimation of Admixture Rates. In: Genetic Differentiation in Human and Other Animal Populations (Ahuja YR and Neel JV, eds.). Indian Anthropological Association, Delhi, 171-180.

Chakraborty R (1986). Gene admixture in human populations: models and predictions. *Am. J. Phys. Anthropol.* 29: 1-43.

Choisy M, Franck P and Cornuet JM (2004). Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Mol. Ecol.* 13: 955-968.

Chung CC, Magalhaes WC, Gonzalez-Bosquet J and Chanock SJ (2010). Genome-wide association studies in cancer - current and future directions. *Carcinogenesis* 31: 111-120.

Comas D, Calafell F, Mateu E, Perez-Lezaun A, et al. (1998). Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. *Am. J. Hum. Genet.* 63: 1824-1838.

Dupanloup I and Bertorelle G (2001). Inferring admixture proportions from molecular data: extension to any number of parental populations. *Mol. Biol. Evol.* 18: 672-675.

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.

Glass B and Li CC (1953). The dynamics of racial intermixture; an analysis based on the American Negro. *Am. J. Hum. Genet.* 5: 1-20.

Herrera-Paz EF, Matamoros M and Carracedo A (2010). The Garifuna (Black Carib) people of the Atlantic coasts of Honduras: Population dynamics, structure, and phylogenetic relations inferred from genetic data, migration matrices, and isonymy. *Am. J. Hum. Biol.* 22: 36-44.

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

Hughes AL, Welch R, Puri V, Matthews C, et al. (2008). Genome-wide SNP typing reveals signatures of population history. *Genomics* 92: 1-8.

Instituto Brasileiro de Geografia e Estatística (IBGE) (2007). Síntese de Indicadores Sociais: Uma Análise das Condições de Vida da População Brasileira. Available at [http://www.ibge.gov.br/home/estatistica/populacao/condicaodevida/indicadoresminimos/sinteseindicsociais2007/indic_sociais2007.pdf]. Accessed April 12, 2010.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-861.

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.

Lai CQ, Tucker KL, Choudhry S, Parnell LD, et al. (2009). Population admixture associated with disease prevalence in the

Boston Puerto Rican health study. *Hum. Genet.* 125: 199-209.
Lee SS, Bolnick DA, Duster T, Ossorio P, et al. (2009). Genetics. The illusive gold standard in genetic ancestry testing. *Science* 325: 38-39.
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.
Luna FV and Klein HS (2004). Economy and slave society: Minas Gerais and São Paulo in 1830. *Rev. Bras. Est. Pop.* 21: 173-193.
Packer BR, Yeager M, Burdett L, Welch R, et al. (2006). SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res.* 34: D617-D621.
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.
Patterson N, Petersen DC, van der Ross RE, Sudoyo H, et al. (2010). Genetic structure of a unique admixed population: implications for medical research. *Hum. Mol. Genet.* 19: 411-419.
Pfaff CL, Barnholtz-Sloan J, Wagner JK and Long JC (2004). Information on ancestry from genetic markers. *Genet. Epidemiol.* 26: 305-315.
Pritchard JK, Stephens M and Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.
Rachlin J, Ding C, Cantor C and Kasif S (2005). MuPlex: multi-objective multiplex PCR assay design. *Nucleic Acids Res.* 33: W544-W547.
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.
Salzano FM and Bortolini MC (2002). The Evolution and Genetics of Latin American Populations. Cambridge University Press, Cambridge.
Sans M (2000). Admixture studies in Latin America: from the 20th to the 21st century. *Hum. Biol.* 72: 155-177.
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.
Tarazona-Santos E, Raimondi S and Fuselli S (2007). Controlling the Effects of Population Stratification by Admixture in Pharmacogenetics. In: Pharmacogenomics in Admixed populations (Guilherme Suarez-Kurtz, ed.). Landes Bioscience, Austin, 1-16.
Tian C, Hinds DA, Shigeta R, Kittles R, et al. (2006). A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. *Am. J. Hum. Genet.* 79: 640-649.
Via M, Ziv E and Burchard EG (2009). Recent advances of genetic ancestry testing in biomedical research and direct to consumer testing. *Clin. Genet.* 76: 225-235.

Supplementary Material

Table S1. Primers for the polymerase chain reaction (PCR) and mini-sequencing reactions of the African and Native American panels.

SNP	Forward PCR primers	Reverse PCR primers	Mini-sequencing primers
AFR_rs2697520	TGCCTGTGTGTGTGTGTATGT	CGTATTCTTTAAATGGTGCAACA	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtcaca gcttgtctTTTGTCCTAAAAGTATCCCCTGTTA (95)
AFR_rs8035530	CCTCTTCAGTACCTCTTCCAGTG	CAACAGAAGGGAGAGCAAGAC	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtcacagc GCAATACAAGAAAAGCACCTTCAT (88)
AFR_rs137211	GCTTGCCTATTGATAGAAGGGT	CCCTCTTTCACGATACTTAGCAT	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcCACCACAGTTTTGA TTTCAGTCC (65)
AFR_rs2416791	CCCAGCTTGTGGTCATTTATTA	GATTCTGGCAACGCTGACT	tcgcgcgtttcggtgatgacggtgaaaCATCTACCATCAGCCCAATTC (48)
AFR_rs9626698	GGCAACGCTCCTAGTGATC	CAACAGCCAAGAAATATGGTG	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtcac agcttgtctgtaagcgGAGAGGGGTTTTTGCAATAACATTA (102)
AFR_rs1443985	CGGGTGCTATGTGTACTTTCAT	CAAATTCTGCTTCCTGAAATAGA	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtcacagctt gtctgtaagcggCTTGAGAAATGTCTATACAAATGCCCTATTAA (110)
AFR_rs735480	TTAGTTCGCTGCTGCTCTAAT	GATGGAGGCCTCAGAGAGA	tcgcgcgtttcggtgatgacggtgaaaacctcGCTTAAAAGTCATGCGCCTAAG (55)
AMR_rs8058694	GATTTCTTCCTGCAGCCTC	TGGAAACCTACACCACCTCTC	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatGGAAGGATTGCATCAC GATACA (62)
AMR_rs2234636	TTCCATGTCCATCCAATGTA	ATGGCTGTGGAGTTCGAA	tcgcgcgtttcggtgatgacggtgaaaacctctgaGCGACTCCAAAAAGAAGAC AA (56)
AMR_rs3760657	AGGAGCAGACAAACAGTCCTAT	GGATGGCAACAATACCTCACT	tcgcgcgtttcggtgatgacggtgaaaaGTGTCAGATGACACAGCAC (47)
AMR_rs2619681	CATGCTTGCCAACACGAT	AACATATGAACAGATGTCCAACA	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagcCAACATCATTATTA ATCAGGAAAATTAAAATCACATT (82)
AMR_rs800292	CCAAACATATCCAGAAGGCAC	AATTCAGGCATAATTGCTACAAT	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcaCCCTGGATATAGATCT CTTGGAAAT (68)
AMR_rs2088102	GGATCTTAGCCTCTTCTGCTACA	CCTTCTTCCAGCACAGGG	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtcacagc ttgtctgtaagcggatgccgggaCCCAGGTGCTCAACAGAA (105)
AMR_rs691968	TTTCACTCCTCCAAGGACC	TGATGGCTACACAAGAGCAAA	tcgcgcgtttcggtgatgacggtgaaaacctctgacacatgcagctcccggagacggtcacagc ttgtctgtGCCATCCTGGTGCCTA (88)

We designed mini-sequencing primers with the following characteristics: minimum length of 40 nucleotides, annealing temperature near 50°C, minimum difference of four nucleotides between predicted mini-sequencing product lengths (the length of the primers was modified by the addition of tails with the PUC 18 plasmid sequence to the 5' extremity of the specific primer sequences, which are represented in lowercase in Table S1), absence of predicted primer dimers or hairpin structures. The primers were commercially synthesized and ranged from 47 to 110 nucleotides.

Enzymatic purification of PCR and mini-sequencing products

The purification of 3 µL PCR products was done by a reaction containing 1 unit of the enzyme *Escherichia coli* exonuclease I (*Exo*I, 10 units/µL), 0.9 units of shrimp alkaline phosphatase (SAP, 1 unit/µL) and 0.2 µL 10X SAP reaction buffer. The Exo-SAP reaction was performed in order to eliminate the excess of PCR primers and dNTPs of the PCR products before the mini-sequencing reaction. The reaction was incubated at 37°C for 90 min, followed by inactivation of the enzymes by heating at 80°C for 20 min.

The purification of the monochromatic mini-sequencing products was performed by a reaction containing 0.3 units SAP, 0.2 µL 10X SAP reaction buffer, 1.6 µL H_2O, and 5 µL of the mini-sequencing product. The products of the polychromatic mini-sequencing were purified in a reaction with 0.5 units SAP and 0.5 µL 10X SAP buffer reaction, added directly to 5 µL of the SNaPshot product. The SAP reaction was incubated at 37°C for 60 min, followed by the inactivation of the enzymes by heating at 75°C for 15 min.

Other technical issues

We imposed on the Muplex (Rachlin et al., 2005) the condition that the minimum difference between PCR product lengths within each panel had to be 10 bp. This allows us, during the set up period of the multiplex PCR, to better evaluate if all primers are properly working, using polyacrylamide gel electrophoresis.

Table S2. Set of 108 ancestry informative markers, selected from the SNP500Cancer project database, using the criterion of F_{ST} >0.20 among Europeans, Africans and Native Americans, and F_{ST} <0.10 among populations within these groups.

SNP500Cancer ID	dbSNP ID	SNP500Cancer ID	dbSNP ID	SNP500Cancer ID	dbSNP ID	SNP500Cancer ID	dbSNP ID
ABCA1_17	rs2230808	EPHX2_04	rs1126452	IL4_03	rs2070874	PIM1_03	rs262933
AKR1C3_36	rs7921327	ERCC1_06	rs3212948	IL6R_04	rs8192284	POLB_08	rs2953983
AMACR_03	rs34689	ERCC5_01	rs1047768	IL6_04	rs1800797	POLD1_13	rs1726787
ANKK1_01	rs1800497	ESR1_17	rs2273206	IL7R_01	rs1494555	RAD52_07	rs6413436
APC_09	rs2229992	FANCA_03	rs1061646	INSR_13	rs919275	RAG1_01	rs2227973
AURKA_16	rs10485805	FASLG_01	rs929087	KRT23_03	rs2269858	RB1CC1_24	rs1129660
BCL2L1_02	rs1484994	FBXW7_44	rs2676329	LCAT_05	rs1109166	RERG_24	rs6488766
BCL6_09	rs3774309	FUT2_05	rs603985	LIPC_37	rs1968689	RGS5_01	rs15049
BIC_34	rs4817027	GATA3_25	rs520236	LRP5_01	rs312016	RNASEL_02	rs486907
BRIP1_09	rs1015771	GDF15_02	rs1059369	MATR3_01	rs11738738	SCARB1_03	rs4765621
CASP3_08	rs1049216	GHR_47	rs7712701	MBL2_46	rs10824793	SEPP1_01	rs7579
CASP8_07	rs2293554	GPX2_21	rs2737844	MSH3_07	rs3797896	SLAMF1_03	rs164283
CASR_11	rs4678045	GPX3_28	rs8177426	MTRR_19	rs8659	SLC23A1_09	rs4257763
CAT_02	rs769214	GSTM3_06	rs1537234	MX1_28	rs455599	SLC4A2_02	rs10245199
CAV1_29	rs6950798	HSD17B2_01	rs1424151	MYBL2_03	rs34771484	SLC6A3_10	rs6347
CDK5_16	rs1549760	HSD3B1_24	rs4659182	MYC_02	rs3891248	SOAT2_01	rs2280699
CDKN2A_03	rs3088440	HSD3B2_14	rs12411115	MYNN_01	rs1317082	SOD1_01	rs2070424
CGA_06	rs932742	IFNAR2_06	rs7279064	NCF2_03	rs2274064	SOD3_05	rs2855262
CYP19A1_01	rs700518	IGF1R_05	rs2137680	NCOA3_04	rs2076546	TCTA_04	rs6784820
CYP1A1_14	rs2606345	IGF2_16	rs3213221	NFKB1_02	rs3774937	TERT_02	rs2075786
CYP1B1_27	rs162556	IGFBP5_10	rs1978346	NFKBIE_01	rs483536	TLR2_06	rs4696480
CYP2E1_31	rs8192766	IGFBP6_19	rs822688	NR1H4_05	rs35724	TP73L_13	rs9840360
CYP3A7_01	rs12360	IL13_01	rs20541	OCA2_23	rs1900758	VCAM1_05	rs3176879
DHDH_02	rs4987162	IL15_02	rs10833	PAK6_13	rs2242119	WDR79_06	rs17886268
DRD2_03	rs1079597	IL1B_03	rs1143627	PCNA_10	rs17352	XRCC4_05	rs2075685
EFNB3_02	rs3744262	IL2_03	rs2069763	PCTP_01	rs2114443	XRCC5_12	rs2440
ENPP1_04	rs1044582	IL4R_07	rs1805016	PHB_02	rs4987082	-	rs1719889

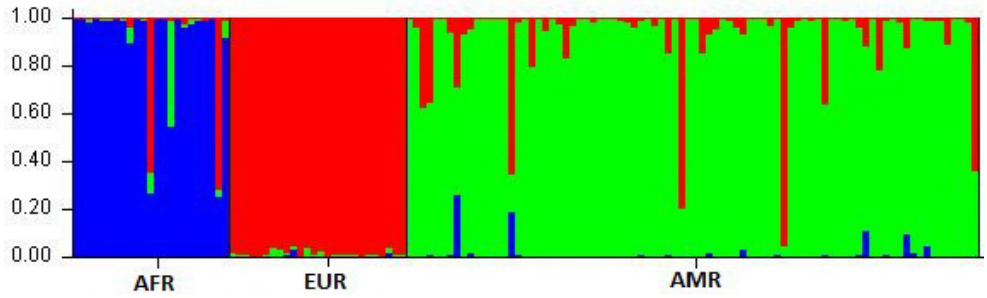

Figure S1. Analysis of parental populations (AFR = Africans; EUR = Europeans; AMR = Native Americans) with the 14 ancestry informative markers selected in our study. Each vertical bar represents an individual subject. Analyses were performed with the admixture model, K = 3 and without any prior population assignment.