Straightforward Inference of Ancestry and Admixture Proportions through Ancestry-Informative Insertion Deletion Multiplexing

Rui Pereira1,2, Christopher Phillips2, Nádia Pinto1,3,4, Carla Santos2, Sidney Emanuel Batista dos Santos5, António Amorim1,3, Ángel Carracedo2,6, Leonor Gusmão1*

1 IPATIMUP – Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal, 2 Institute of Forensic Sciences Luis Concheiro, University of Santiago de Compostela, Santiago de Compostela, Spain, 3 Faculty of Sciences, University of Porto, Porto, Portugal, 4 Mathematics Research Centre, University of Porto, Porto, Portugal, 5 Laboratório de Genética Humana e Médica, Universidade Federal do Pará, Belém, Brazil, 6 Genomics Medicine Group, CIBERER, University of Santiago de Compostela, Santiago de Compostela, Spain

Abstract

Ancestry-informative markers (AIMs) show high allele frequency divergence between different ancestral or geographically distant populations. These genetic markers are especially useful in inferring the likely ancestral origin of an individual or estimating the apportionment of ancestry components in admixed individuals or populations. The study of AIMs is of great interest in clinical genetics research, particularly to detect and correct for population substructure effects in case-control association studies, but also in population and forensic genetics studies. This work presents a set of 46 ancestry-informative insertion deletion polymorphisms selected to efficiently measure population admixture proportions of four different origins (African, European, East Asian and Native American). All markers are analyzed in short fragments (under 230 basepairs) through a single PCR followed by capillary electrophoresis (CE) allowing a very simple one tube PCR-to-CE approach. HGDP-CEPH diversity panel samples from the four groups, together with Oceanians, were genotyped to evaluate the efficiency of the assay in clustering populations from different continental origins and to establish reference databases. In addition, other populations from diverse geographic origins were tested using the HGDP-CEPH samples as reference data. The results revealed that the AIM-INDEL set developed is highly efficient at inferring the ancestry of individuals and provides good estimates of ancestry proportions at the population level. In conclusion, we have optimized the multiplexed genotyping of 46 AIM-INDELS in a simple and informative assay, enabling a more straightforward alternative to the commonly available AIM-SNP typing methods dependent on complex, multi-step protocols or implementation of large-scale genotyping technologies.

Citation: Pereira R, Phillips C, Pinto N, Santos C, Santos SE, et al. (2012) Straightforward Inference of Ancestry and Admixture Proportions through Ancestry-Informative Insertion Deletion Multiplexing. PLOS ONE 7(1): e29684. doi:10.1371/journal.pone.0029684

Editor: Manfred Kayser, Erasmus University Medical Center, The Netherlands

Received August 22, 2011; Accepted December 2, 2011; Published January 17, 2012

Copyright: © 2012 Pereira et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was partially supported through PhD grants to RP (SFRH/BD/30039/2006), NP (SFRH/BD/37261/2007), and CS (SFRH/BD/75627/2010) awarded by the Portuguese Foundation for Science and Technology (FCT) and co-financed by the European Social Fund (Human Potential Thematic Operational Programme). This work was also partially supported by the project Xunta de Galicia INCITER09 208 163 PR (PI: Victoria Lareu). IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Education and Science and is partially funded by FCT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. No additional external funding received for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: lgusmao@ipatimup.pt

Introduction

Initial studies of human genetic variation focused on Short Tandem Repeats (STRs) and Single Nucleotide Polymorphisms (SNPs) [1,2], and only later explored Copy Number Variants (CNVs) [3–6] and Insertion Deletion Polymorphisms (INDELS) [7–9] unveiling previously unknown sources of genetic diversity that are likely to be important factors underlying inherited traits and diseases in humans. Moreover, advances in genotyping technologies have allowed progressively higher genome coverage using resources within the normal scope of most genetics laboratories. These developments have led to an increase in Genome Wide Association Studies (GWAS) in the search for genetic variants associated with a wide range of complex diseases and phenotypic traits including, for example, obesity, schizophrenia, autism, diabetes, height, eye and skin color [10–12].

These investigations have identified a large number of candidate gene variants showing strong association with specific conditions or phenotypes and subsequent replication studies and meta-analysis have strengthened or weakened these initial findings. One of the major problems in case-control association studies is the presence of undetected population structure that can lead to finding false positive associations when an excess of ancestry differentiated markers stratifies the case and the control groups. Alternatively false negative results may occur if real associations are missed if weak while greater allele frequency differentiation exists between study and control groups due to differences in ancestry [13,14]. Therefore, association studies must be accompanied by an evaluation and correction of the possible effects of population structure between both sample groups. In recent years the prevailing strategies to overcome the dangers of population stratification use genomic control to measure the possible effects of...
Ancestry-Informative INDELs

Ethics Statement
The current study was approved by the Institute of Molecular Pathology and Immunology of the University of Porto institutional review board. Besides the HGDP-CEPH diversity panel human cell line samples, all other samples involved in the study are long-lasting anonymized DNA extracts previously obtained with informed written consent from healthy individuals for research purposes.

Population samples
A total of 1002 DNA samples were used in this study comprising: i) reference samples from the HGDP-CEPH diversity panel standardized subset H952 [30,31] with origin in Africa (AFR), Europe (EUR), East Asia (EAS), America (NAM) and also Oceania (OCE), representing a total 584 individuals from 40 populations. Individuals 1219, 1339, 1344 and 1041 were not included in the study since no DNA was available for analysis; in substitution of 1041 we used 1042 who had been excluded from subset H952 due to a parent/offspring relationship with 1041 [31]; ii) samples from Angola (48), Portugal (48), Taiwan (48) and Brazilian Amazonas tribes (48) used in a preliminary evaluation of the AIM-INDEL assay and as example testing samples; iii) samples from the city of Belem [226], an admixed population in northeastern Amazonas, Brazil.

AIM-INDEL selection and development of the multiplex reaction
An initial pool of candidate INDELS was assembled by collecting previously available population data on this type of polymorphism included in the Marshfield Diallelic Insertion/Deletion Polymorphisms database website (http://www.marshfieldclinic.org/mgs/; [7]) and from later studies that also characterized some candidate INDELS in different population groups [27,28,32,33]. Considering the allele frequency data compiled from the diverse sources, all markers were sorted according to frequency differentials (δ) [34] comparing four human population groups of Africans, Europeans, East Asians and Native Americans. For this study we selected a set of 46 markers (Table 1) among the most informative INDELS for each population group (all with δ=0.40 between at least two groups) and optimized a unique multiplex reaction allowing the simulta-
| MID* | rs number  | Chromosome | Position (bp)** | Alleles described in dbSNP | References |
|------|------------|------------|----------------|----------------------------|------------|
| MID-1470 | rs2307666 | 11 | 64729920 | -/GT TAC | [7,27,33] |
| MID-777 | rs1610863 | 16 | 6551830 | -/GAA | [7] |
| MID-196 | rs16635 | 6 | 99789757 | -/CAT | [7,27,32] |
| MID-881 | rs1610965 | 5 | 79746093 | -/ACTT | [7,32] |
| MID-3122 | rs35451359 | 18 | 4510983 | -/ATCT | [7] |
| MID-548 | rs140837 | 6 | 378909 | -/CT | [7] |
| MID-659 | rs1160893 | 2 | 224794577 | -/CT | [7] |
| MID-2011 | rs2308203 | 2 | 109401291 | -/CTAGA | [7,27] |
| MID-2929 | rs33974167 | 8 | 87813725 | -/TA | [7] |
| MID-593 | rs1160852 | 6 | 137345857 | -/TT | [7] |
| MID-798 | rs1610884 | 5 | 56122323 | -/GGGAA | [7,32] |
| MID-1193 | rs2067280 | 5 | 89818959 | -/AT | [7] |
| MID-1871 | rs2308067 | 7 | 127291541 | -/TT | [7] |
| MID-17 | rs4183 | 3 | 3192524 | -/TAAC | [7,32] |
| MID-2538 | rs3054057 | 15 | 86010538 | -/AACA | [7] |
| MID-1644 | rs2307840 | 1 | 36099090 | -/GT | [7,32] |
| MID-3854 | rs60612424 | 6 | 84017514 | -/TCTA | [7] |
| MID-2275 | rs3033053 | 14 | 42554496 | -/TCACGAG | [7] |
| MID-94 | rs16384 | 22 | 42045009 | -/AAC | [7,33] |
| MID-3072 | rs34611875 | 18 | 6723917 | -/GCCCC CA | [7] |
| MID-772 | rs1610859 | 5 | 128317275 | -/TAG | [7] |
| MID-2313 | rs3045215 | 1 | 234740917 | -/ATTATAACT | [7,32] |
| MID-397 | rs25621 | 6 | 13985158 | -/TTCT | [7] |
| MID-1636 | rs2307832 | 1 | 55590789 | -/AA | [7,32] |
| MID-51 | rs16343 | 4 | 17635560 | -/TTTAT | [7,32] |
| MID-2431 | rs3031979 | 8 | 73501951 | -/ATTG | [7] |
| MID-2264 | rs34122827 | 13 | 63778778 | -/AAGT | [7] |
| MID-2256 | rs133052 | 22 | 41042634 | -/CAT | [7,32] |
| MID-128 | rs6490 | 12 | 108127168 | -/ATT | [7] |
| MID-15 | rs4181 | 2 | 42577803 | -/AAATACACAC | [7,32] |
| MID-2241 | rs3030826 | 6 | 67176774 | -/GTCCAATA | [7,32] |
| MID-419 | rs140708 | 6 | 17072016 | -/AATGGCA | [7,32] |
| MID-943 | rs1611026 | 5 | 82545545 | -/TGAT | [7] |
| MID-159 | rs16438 | 20 | 25278470 | -/CCCCA | [7] |
| MID-2005 | rs2308161 | 10 | 69800909 | -/AA AAT | [7,33] |
| MID-250 | rs16687 | 7 | 83887882 | -/CA | [7,32] |
| MID-1802 | rs2307998 | 5 | 7814345 | -/GGA | [7] |
| MID-1607 | rs2307803 | 3 | 108981031 | -/TG | [7] |
| MID-1734 | rs2307930 | 6 | 84476378 | -/CCAT | [7] |
| MID-406 | rs25630 | 6 | 14784341 | -/AG | [7] |
| MID-1386 | rs2307582 | 1 | 247768775 | -/AACTTAT CSTCTTTCACCCT | [7,27] |
| MID-1726 | rs2307922 | 1 | 39896964 | -/CAAGACTATAAT/CACACTTATAT | [7,27,32] |
| MID-3626 | rs11267926 | 15 | 45526069 | -/AATATACTTCTCCA | [7] |
| MID-360 | rs25584 | 12 | 11245217 | -/AA | [7] |
| MID-1603 | rs2307799 | 5 | 70828427 | -/TTGT | [7,27,33] |
| MID-2719 | rs34514193 | 20 | 30701405 | -/AACT | [7,28] |

*Nomenclature according to [7] and Marshfield Diallelic Insertion/Deletion Polymorphisms database; **Mapping data according to dbSNP (build 132).

doi:10.1371/journal.pone.0029684.t001

Table 1. AIM-INDELs used in the multiplex.
neous genotyping of all AIMs in a single PCR and electrophoretic run. The multiplex development followed a similar workflow as in Pereira et al. [20,35] except for the accommodation of certain longer amplicons into a broadened size window up to 230 bp in order to type an extended number of INDELs in a single reaction.

**Amplification and genotyping**

PCR amplification of the 46 AIM-INDELs used the QIAGEN Multiplex PCR kit (Qiaxen) at 1× Qiagen multiplex PCR master mix, 0.1 μM of all primers [sequence details in Table S1] and 0.3–5 ng of genomic DNA in a 10 μl final reaction volume. Thermocycling conditions were: initial step at 95°C for 15 min; 30 cycles at 94°C for 30 sec, 60°C for 90 sec, and 72°C for 45 sec; and a final extension at 72°C for 60 min. The PCR products were then prepared for capillary electrophoresis (CE) by adding 1 μl of amplified product to 10 μl Hi-Di™ Formamide (Applied Biosystems) and 0.3 μl of GeneScan™ 500 LIZ® size standard (Applied Biosystems). CE was performed using a 3130 Genetic Analyzer prepared with DS-33 matrix standard, POP-7TM polymer and applying virtual filter G5 (Applied Biosystems). The electropherograms were analyzed and genotypes were automatically assigned with GeneMapper v4.0 (Applied Biosystems). For practical reasons INDEL short alleles were coded as 1 and long alleles as 2.

**Statistical analysis**

Estimation of allele frequencies, exact tests of Hardy-Weinberg equilibrium (HWE), Fst genetic distances and linkage disequilibrium tests were assessed using Arlequin v3.5.1.2 [36]. Ancestry inferences were performed using STRUCTURE v2.3.3 [17,37] with a burn-in length of 100,000 followed by 100,000 MCMC repetitions and a variety of parameter sets were tested depending on the objective of the analysis. Initial runs were made without any prior information on the origin of samples, using the “Admixture Model” and considering either correlated or independent “Allele Frequency Models”; a minimum of 3 independent runs were performed for each testing K value, ranging from K = 1 to K = number of presumed clusters present in the dataset plus three. The estimated ln probability of data (−lnP(D)) values were plotted using Structure harvester v0.6.6. (http://taylor0.biology.uclla.edu/structureHarvester/). In a second phase, when using reference samples as training sets to test for “unknown” individuals or populations, STRUCTURE analyses were carried out using the same parameters as before or selecting the “Use Population Information” option. In these cases, allele frequencies were updated using only the reference individuals with POPFLAG = 1 data (option under the Advanced tab). Here, 3 independent runs were performed only for the appropriate number of clusters, as evaluated by the initial analysis. Unless otherwise indicated, results are presented for the default settings considering the “Admixture Model” and correlated allele frequencies. CLUMPP v1.1.2 [38] was used to obtain the average permuted individual and population Q-matrices throughout the three replicates for each K value. Those matrices were used as input to distruct v1.1 [39] to obtain bar plots where each individual is represented as a segment divided into K colors that represent the estimated membership coefficients from each cluster.

Principal component analysis (PCA) was performed as an additional and independent approach to estimate the number of populations present in the data set. We used R 2.11.1 [40] with SNPassoc package [41] to obtain two and three dimensional graphics and the information percentage values associated to each principal component.

The efficiency of the 46 AIM-INDEL set for assigning individuals to population groups was further evaluated by one-out cross-validation based on a flexible single profile analysis system very similar to STRUCTURE, calculating likelihood ratio values obtained with a Bayesian classification algorithm implemented in the “Snipper app suite” website (http://mathgene.ucsd.edu/snipper/; [23]).

**Results**

A simple and informative multiplex was developed for the simultaneous analysis of 46 AIM-INDELs reported to have high δ values between the AFR, EUR, EAS or NAM population groups. All markers were analyzed in short fragments (<230 bp) through a single PCR followed by capillary electrophoresis (Figure 1). The workflow of the INDEL assay is straightforward, reducing considerably the steps and resources needed to genotype a large set of biallelic AIMs.

After optimization of the method we created a database including HGDP-CEPH diversity panel genetic data, commonly used by the research community as reference populations for the four groups AFR, EUR, EAS, NAM and also from Oceania (complete database included in File S1).

**Genetic characterization of reference populations**

Patterns of INDEL variability observed in the HGDP-CEPH samples from the population groups AFR, EUR, EAS and NAM are detailed in Table S2 as well as δ and pairwise Fst for each marker. With few exceptions, the vast majority of the INDELs show high allele frequency differentials and genetic distances between at least two groups (39 with δ̂≥0.4 and 44 with δ̂≥0.3). No significant departures from Hardy-Weinberg equilibrium were found in the studied populations and pairwise linkage disequilibrium exact tests did not detect significant associations within the marker set.

One interesting finding was the occurrence of an unexpected third allelic state (coded as allele 3) for MID360 and MID2264. Sequencing analysis confirmed our observations as a result of additional sequence length variants within the amplicon fragments. For MID360 the third allelic state is observed due to a T insertion associated with the short allele, 8 bases downstream of the targeted polymorphism (allele 1D8Tins). Conversely for MID2264, allele 3 corresponds to a T deletion occurring in the long allele background (allele 2D6Sdel). Interestingly, the MID360 variant alleles were only found in AFR samples whereas the MID2264 variants seemed specific of EUR.

**Inferring genetic ancestry**

- The AIM-INDEL panel efficiently distinguishes four major population groups. Before implementing the HGDP-CEPH diversity panel reference database, a preliminary evaluation on the performance of a panel comprising 44 INDELs at the time (without MID94 and MID1734) had been performed using 48 samples with origin in each of the four groups under study (detailed results in Figure S1). In brief, analyses with STRUCTURE, PCA and one-out cross validation clearly supported the efficiency of the panel in clustering individuals into four population groups.

The results obtained for the complete AIM-INDEL panel with HGDP-CEPH AFR, EUR, EAS and NAM populations strongly corroborate these preliminary findings (Figure 2). STRUCTURE ancestry estimates considering K = 4 still produce an enhancement in −ln P(D) values while a plateau is reached thereafter, which
points to 4 as the smallest K number capturing the major population structure in the data and supports the inference that a four group clustering better fits the genetic data (Figure 2A and 2B).

PCA for the same dataset allows an independent non-model based view of the individual clustering. The first three PCs define approximately half of the variance in the dataset (46.1%) yet allow a clear spatial separation of four different groups (Figure 2C). Likewise, cross-validation studies (Figure 2D) revealed the INDEL panel to show a high accuracy of population assignment, with a global classification error of 1.26% (specifically 7 of 556). All AFR, EUR and NAM were correctly assigned whereas misclassified individuals were all from the Yakut population in Siberia except for one individual from Oroquen, China.

- HGDP-CEPH genetic data as reference genotypes to test individuals or populations of unknown origin. Reference HGDP-CEPH diversity panel genetic data from the four population groups (AFR, EUR, EAS and NAM) was used to estimate ancestry proportions of individuals/populations from different geographic locations. We tested samples from Angola, Portugal, Taiwan and Brazil (Amazonas Amerindian tribes and Belém, a northeastern Amazonas city). The individual and global admixture estimates obtained with genetic data only (no prior population information) correspond well with expected patterns, knowing the origin of the subjects (Figure 3; Table 2). In general, individuals from the non-admixed populations show high membership proportion in the same cluster as HGDP-CEPH representatives of the same population group. In contrast individuals from Belém show highly variable admixture patterns mainly of European, Native American and African origin (Figure 3; Table 2), resulting in average ancestry proportions of 53.5% EUR, 22.9% NAM, 14.8% AFR and 8.8% EAS. Considering the historical formation and peopling of Brazil in which there were three main contributing ancestral populations (NAM, EUR and AFR) we performed a three-group analysis for the particular case of Belém – specifically, excluding EAS and using only NAM, EUR and AFR ancestral groups with K = 3 (Figure S2). In particular the Native American proportion increased (53.7% EUR, 29.5% NAM and 16.8% AFR), having captured most of the previous East Asian component.

- Indications of population differentiations beyond four groups from inclusion of Oceanians. The AIM-INDEL panel was primarily designed as a tool for ascertaining ancestry from four major population groups. Nonetheless, as there is general interest in AIM panels able to distinguish populations at the broader continental level, we extended our study to HGDP-CEPH Oceanian samples and assessed the ability of the panel to differentiate populations with origin in all five continent regions. Following the same evaluation strategy as before, the assay proved to consistently recognize a fifth cluster corresponding to Oceanians and that K = 5 captures most of the structure in the dataset (Figure 4; Figure S3 for details). PCA plots (Figure S3C) show most HGDP-CEPH Oceanians form a distinguishable cloud lying between EUR and EAS even though the separation is not perfectly achieved. In a five-group classification, the one-out cross validation error rate increased slightly to 1.54% (9/584). The assignment of Oceanians was accurately made but two EAS (from Cambodia) were now misclassified as OCE.
Figure 2. Analysis of HGDP-CEPH diversity panel samples from four continental origins using a set of 46 AIM-INDELs. A) ancestral membership proportions (based on STRUCTURE results from 3 independent runs treated in CLUMPP and plotted with distruct; individuals were first sorted by geographic origin of population, and within those by ascending population code and HGDP individual number); B) estimated ln probability of the data (=ln P(D) obtained with STRUCTURE and plotted using Structure harvester); C) principal component analysis 3D plots. D) estimation of population assignment success (results from one-out cross validation studies using the Snipper app suite; see methods for details of the analyses). AFR: Africa; EUR: Europe; EAS: East Asia; NAM: Native America. doi:10.1371/journal.pone.0029684.g002

Figure 3. Ancestral membership proportions for testing population samples from different continental origins using the HGDP-CEPH diversity panel genetic data as training sets. Angola (Africa); Portugal (Europe); Taiwan (East Asia); Brazilian Amazonas tribes (Native America); Belém is an example of a highly admixed Brazilian city in northeastern Amazonas. doi:10.1371/journal.pone.0029684.g003
Discussion

The main objective of this study was to provide a simple tool for inferring ancestry and estimate admixture proportions from four different population origins that can be widely applied to genetic studies. We describe a new AIM assay comprising 46 INDELs that are simply analyzed in a multiplex PCR followed by CE detection. With this approach we were able to combine the ancestry informative power of biallelic markers (exemplified by AIM-SNP panels) with the simplified analysis based in fragment size separation (as in STR typing). The methodology of the assay is straightforward and can be readily and inexpensively implemented in any molecular genetics laboratory. In contrast, the majority of AIM sets published in recent years involve more complex genotyping protocols or are limited to specific platforms not available to all laboratories and therefore requiring additional resources [e.g. 23,27,46–49]. Kosoy et al. [44] have shown that small AIM sets can distinguish major population groups and correct for false positive results in association studies. Other studies have addressed ancestry prediction of the HGDP-CEPH samples using large-scale SNP datasets obtained with high-throughput microarrays, and have also evaluated the performance of small subsets of markers ascertained following different strategies such as FST, allele differentials, informativeness of assignment index I_n [51] or PCA (e.g. [42,48,52]). These studies have shown that inference of continental ancestry for the HGDP-CEPH panel is quite clear, and can be performed with a relatively small number of SNPs (10

Table 2. Ancestral membership proportions for HGDP-CEPH diversity panel samples and testing populations from four continental origins.

| 46 AIM-INDELs (this study) | 210 INDELs [32] | 48 Lr,4 AIM-SNP set [44] |
|---------------------------|-----------------|-------------------------|
| HGDP-CEPH                  | AFR  | EUR  | EAS  | NAM  | AFR  | EUR  | EAS  | NAM  | AFR  | EURA | EAS  | AMI  |
| AFR                        | 0.969 | 0.011 | 0.012 | 0.008 | 0.977 | 0.009 | 0.009 | 0.005 | 0.97  | 0.02  | 0.01 | 0.01 |
| EUR                        | 0.008 | 0.963 | 0.014 | 0.007 | 0.976 | 0.013 | 0.013 | 0.013 | 0.96  | 0.02  | 0.01 | 0.01 |
| EAS                        | 0.006 | 0.018 | 0.952 | 0.024 | 0.007 | 0.021 | 0.955 | 0.017 | EAS   | 0.01  | 0.04 | 0.91 |
| NAM                        | 0.008 | 0.041 | 0.027 | 0.924 | 0.011 | 0.028 | 0.015 | 0.946 | AMI   | 0.01  | 0.03 | 0.04 |
| Testing populations:       | AFR  | EUR  | EAS  | NAM  |
| Angola                     | 0.970 | 0.011 | 0.011 | 0.008 |
| Portugal                   | 0.018 | 0.966 | 0.008 | 0.008 |
| Taiwan                     | 0.004 | 0.003 | 0.984 | 0.009 |
| Br. Amazonas tribes        | 0.010 | 0.013 | 0.032 | 0.945 |
| Belem (4G-analysis)        | 0.148 | 0.535 | 0.088 | 0.229 |
| Belem (3G-analysis)        | 0.168 | 0.537 | -     | 0.295 |

(AFR: Africa; EUR: Europe; EAS: East Asia; NAM: Native America). doi:10.1371/journal.pone.0029684.t002

Figure 4. Ancestral membership proportions for HGDP-CEPH diversity panel samples from five continental origins using a set of 46 AIM-INDELs (based on STRUCTURE results from 3 independent runs treated in CLUMPP and plotted with distruct; individuals were first sorted by geographic origin of population, and within those by ascending population code and HGDP individual number). AFR: Africa; EUR: Europe; EAS: East Asia; NAM: Native America; OCE: Oceania. doi:10.1371/journal.pone.0029684.g004
to 50). They also showed that, when using SNPs, it is possible to predict individual ancestry down to the population level, although such approaches require an increased number of markers ranging from several hundred to thousands [40,52]. Due to the multiplexing limitation associated with the number of markers that it is possible to analyze in a single PCR, and in the same respect for other small-scale AIM sets, the 46 AIM-INDEL assay we outline is only going to be particularly useful when broad assignment to continental ancestry is desired, or when estimating admixture proportions in individuals/populations that received ancestral contributions of different continental origins. Assessing within-continent population structure requires much larger arrays of markers, well beyond the number included in our set and in most of the alternative AIM sets, and therefore it will have limited application for that purpose.

The AIM-INDEL assay allowed a rapid and cost-effective genotyping of a large number of samples including HGDP-CEPH individuals from five continental groups (AFR, EUR, EAS, NAM and OCE) and representative testing samples with different origins and admixture levels. From the genetic characterization of the reference ancestral samples we observed a high level of differentiation from the chosen INDELS, as expected from the selection criteria. Although some markers revealed lower differences than those expected from previous data, this is possibly due to the samples representing each group and allele frequency estimation strategies being different [e.g. 7]. The pairwise FST values obtained with the 46 AIM-INDELS (Table S2) are clearly above the usually found at the continental level with random markers [45,53] and similar to those obtained with other AIM panels for the same population groups [44-46].

The results from the HGDP-CEPH diversity panel and other representative populations underlined the capacity of the panel to distinguishing four continental population groups. Furthermore, the ancestry estimates obtained in a four-group analysis are very similar to those obtained in Kosoy et al. [44] with a 48 I_n4 AIM-SNP set for equivalent population groups (Table 2), as well as using a much larger number of INDELS (210) for the same HGDP-CEPH individuals (Table 2; [32]). This concordance in the ancestry estimates highlights the accuracy of the AIM-INDEL panel in inferring ancestry proportions from African, European, East Asian and Native American origin. Furthermore, in spite of the assay being primarily designed for studies considering only four major population groups, extension to five groups revealed the capacity to reliably distinguish Oceanians.

The population assignment cross validation studies based on Bayesian likelihood ratios provided additional evidence of the utility of the assay, particularly for forensic applications where single profiles are often analyzed one at a time. Here the error rates in classifications considering either four or five population groups were low (1.26% and 1.54% respectively). The AIM-INDEL panel achieves very high accuracy for population assignment in the five broad continental regions, similar to results observed by Paschou et al. when using subsets of 50 SNPs ascertained by PCA and estimation of I_n metrics [48]. In our study, the great majority of misclassified individuals were from a single population (Yakut of eastern Siberia) localized near the northeastern fringe of the Asian continent. This intermediate position between East Asia and the American continent can explain differences in patterns of divergence between individuals and their misclassification as American. Likewise, the cross validation studies with five groups revealed two misclassified Cambodians as Oceanians. Together, these results suggest a weaker performance of the panel with differentiation of East Asians. In fact, the accumulated divergence assessed for EAS is

others is slightly smaller than for the other ancestral groups, and the fact that the HGDP-CEPH EAS group analyzed is so diverse (229 individuals from 10 subpopulations) may contribute to this reduced differentiation for East Asians. Another important aspect is the proximity of the East Asian and Native American gene pools. Considering the history of modern humans these groups have diverged over the shortest time, and furthermore, the original peopling of Americas from Beringia involved a significant bottleneck effect that is still reflected in Native American variability. Despite this slightly reduced level of differentiation in the AIM-INDELS selected, STRUCTURE, PCA and cross validation studies together support the capacity of the panel to properly distinguish both groups.

AIM panels are regularly applied in population genetics studies to analyze admixed populations by estimating admixture proportions both at the individual and population level. Depending on the historical context of populations under study, there are different principal ancestral contributors to the formation of the current ancestry characteristics of the region. For example, Brazil and the majority of south-American countries underwent admixture between the pre-existing Native Americans, colonizing Europeans and later African influences resulting from the slave trade to create essentially tri-hybrid populations. In such cases, it is appropriate for genetic studies to perform three-group analyses of ancestry estimates. Our study analyzed ancestry proportions in Belém. We first considered the possibility of a fourth EAS minor ancestral contributor in initial analyses and K = 4 resulted in a low level but detectable fraction of membership of this cluster at 8.8% (Table 2). However, although not statistically significant (exact test of differentiation p value = 0.136), the three-group membership proportion estimates at K = 3 showed a noticeable increase in the Native American component to 29.5% (Table 2) which is in very close agreement with the admixture proportions previously reported for the same population but using a different set of AIM-INDELS (average NAM estimate: 28.4%; [27]). Nevertheless, a preliminary four-group analysis has persuasive arguments for considering all four potential contributors to admixture in these regions. In particular, some locations in Brazil (e.g. São Paulo, Campinas; IBGE - Instituto Brasileiro de Geografia e Estatistica, www.ibge.gov.br) include significant East Asian communities, despite having joined these populations rather recently. When using Brazilian samples from such geographic areas, particularly as case and control samples for association studies, a preliminary four-group analysis is recommended to detect the presence of East Asian ancestry amongst individuals in the study. Otherwise there is considerable risk that the *a priori* rejection of this hypothesis based on three-group analyses could lead to an over-estimation of the Native American proportion in the global admixture estimates (data not shown) due to a strong bias caused by the presence of East Asian individuals in the population under study. Conversely, when “forcing” a four-group analysis in south-American tri-hybrid populations, it is possible that the fourth East Asian component can produce a spurious fraction of membership arising from the Native American component, due to the close relationship of the East Asian and Native American population groups. In summary, we advocate adopting an approach taking due regard for the particular population under study. Consideration of the known recent population history and demographics helps make appropriate adjustment for the different principal ancestral contributors. In the special case of south-American populations, we recommend a preliminary study taking advantage of the full potential of the AIM-INDEL assay to identify and possibly exclude East Asian study subjects, and subsequently perform a comprehensive three-group analysis. The AIM-INDEL assay can be efficiently used in
Supporting Information

Figure S1 Analysis of population samples from four different continental origins using a preliminary set of 44 AIM-INDELs (without MID94 and MID1734). A) ancestral membership proportions (based on STRUCTURE results from 3 independent runs treated in CLUMPP and plotted with distruct); B) estimated ln probability of the data (−lnP(D)) obtained with STRUCTURE and plotted using Structure harvester; C) principal component analysis 3D plots; D) estimation on population assignment success (results from one-out cross validation studies using the Snipper app suite; see methods for details on the analyses). Angola (Africa); Portugal (Europe); Taiwan (East Asia); Brazilian Amazonas tribes (Native America).

Figure S2 Ancestral membership proportions in the Brazilian city of Belém using HGDP-CEPH diversity panel genetic data of three main ancestral contributors as training sets. A) bar plots based on STRUCTURE results from 3 independent runs treated in CLUMPP and plotted with distruct (AFR: Africa; EUR: Europe; NAM: Native America); B) triangular plots based on STRUCTURE results from the run with highest −lnP(D) (left: admixture model; right: using population information; red: Africa; green: Europe; blue: Native American; yellow: Belém).

Table S1 PCR primer sequences used in the multiplex.

Table S2 Allele frequencies, δ and FST values for the 46 AIM-INDELs in HGDP-CEPH diversity panel population samples from Africa (AFR), Europe (EUR), East Asia (EAS) and Native America (NAM).

File S1 Genotypic data (STRUCTURE format) for the 46 AIM-INDELs in HGDP-CEPH diversity panel population samples from Africa, Europe, East Asia and Native America.

Acknowledgments

The authors would like to acknowledge Dayse Alencar for the technical support with the samples from Belém. The assistance of Fundación Botín is also acknowledged with appreciation.

Author Contributions

Conceived and designed the experiments: RP AC LG. Performed the experiments: RP CS SEBdS. Analyzed the data: RP CP NP CS LG. Contributed reagents/materials/analysis tools: CP SEBdS AA AC LG. Performed the experiments: RP CS SEBdS. Analyzed the data: RP CP NP CS LG. Contributed reagents/materials/analysis tools: CP SEBdS AA AC LG. Wrote the paper: RP LG. Critically reviewed the manuscript: CP CS NP SEBdS AA AC.

References

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
2. International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928–933.
3. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444–444.
4. Conrad DF, Pinto D, Redon R, Feuk L, Gokcenen O, et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature 464: 704–712.
5. Sudmant PH, Kitzman JO, Antonacci F, Allkan C, Malig M, et al. (2010) Diversity of human copy number variation and multicopy genes. Science 330: 641–646.
6. Allkan C, Coe BP, Echler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12: 363–376.
7. Weber JL, David D, Hell J, Fan Y, Zhao C, et al. (2002) Human diallelic insertion/deletion polymorphisms. American Journal of Human Genetics 71: 435–862.
8. Mills RE, Luttig CT, Larrkin CE, Beauchamp A, Tsai C, et al. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Research 16: 1102–1109.
9. Mills RE, Pintard WS, Mallaney JM, Farooq U, Creasy TH, et al. (2011) Natural genetic variation caused by small insertions and deletions in the human genome. Genome Research 21: 830–839.
10. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95–108.
11. Donnelly P (2008) Progress and challenges in genome-wide association studies in humans. Nature 456: 728–731.
12. Hirschhorn JN, Gajdoue ZK (2011) Genome-wide association studies: results from the first few years and potential implications for clinical medicine. Annual Review of Medicine 62: 11–24.
13. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nature Genetics 36: 512–517.
14. Tian G, Gregersen PK, Seidlin MF (2008) Accounting for ancestry: population substructure and genome-wide association studies. Human Molecular Genetics 17: R141–150.
15. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
16. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11: 459–463.
17. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–955.
18. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67: 170–181.
19. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: 904–909.
20. Luca D, Ringquist S, Klei L, Lee AB, Giger C, et al. (2008) On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82: 453–463.

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

22. Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Ginjupalli S, et al. (2003) A classifier for the SNP-based inference of ancestry. Journal of Forensic Sciences 48: 771–782.

23. Phillips C, Salas A, Sanchez JJ, Fondevila M, Gonzalez-Tato A, et al. (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Science International: Genetics 1: 273–280.

24. Phillips C, Prieto I, Fondevila M, Salas A, Gonzalez-Tato A, et al. (2009) Ancestry analysis in the 11-M Madrid bomb attack investigation. PLoS ONE 4: e6083.

25. Kayser M, de Knijff P (2011) Improving human forensics through advances in genetics, genomics and molecular biology. Nat Rev Genet 12: 179–192.

26. Londin ER, Keller MA, Mahta C, Smith G, Mamounas LA, et al. (2010) CoAIM: a cost-effective panel of ancestry informative markers for determining continental origins. PLoS One 5: e13443.

27. Santos NP, Ribeiro-Rodrigues EM, Ribeiro-de-Santos AK, Pereira R, Guimaraes L, et al. (2010) Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INDEL) ancestry-informative marker (AIM) panel. Human Mutation 31: 184–190.

28. Pereira R, Phillips C, Alves C, Amorim A, Carracedo A, et al. (2009) A new multiplex for human identification using insertion/deletion polymorphisms. Electrophoresis 30: 3692–3699.

29. Mullane PY, Mills RE, Pittard WS, Devine SE (2010) Small insertions and deletions (INDELS) in human genomes. Human Molecular Genetics 19: R131–136.

30. Cann HM, de Tomo C, Cazes L, Legrand MF, Morel V, et al. (2002) A human genome diversity cell line panel. Science 296: 261–262.

31. Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Annals of Human Genetics 70: 841–847.

32. Rosenberg NA (2006) Assorted differences in the genetic structure and histories of human populations in Europe and the Americas. Proc Natl Acad Sci USA 103: 9476–9481.

33. Yang N, Li H, Criswell LA, Gregersen PK, Alarcon-Riquelme ME, et al. (2005) Clines, clusters, and the effect of study design on the inference of human genetic structure from DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. Human Genetics 118: 382–392.

34. Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, et al. (1997) Ethnic affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 60: 957–964.

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

36. Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 10: 564–567.

37. 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.

38. Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801–1806.

39. Rachow NA (2004) Distinction: a program for the graphical display of population structure. Molecular Ecology Notes 4: 157–158.

40. R Development Core Team (2010) R: A language and environment for statistical computing. 2.11.1 ed. Vienna, Austria: R Foundation for Statistical Computing.

41. González JR, Armgardt L, Sole X, Guino E, Mercader JM, et al. (2007) Ancestry informativeness of genome-wide SNP sets and admixture proportions in common populations in America. Human Mutation 30: 69–76.

42. Nassir R, Kosoy R, Tian C, White PA, Butler LM, et al. (2009) Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Human Mutation 30: 69–76.

43. Halder I, Shriver MD, 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. Human Mutation 29: 648–653.

44. Lao O, van Duijn K, Kreibergen P, de Knijff P, Kayser M (2006) Proportioning whole-genome single-nucleotide-polymerism diversity for the identification of geographic population structure and genetic ancestry. Am J Hum Genet 78: 680–690.

45. Kosoy R, Nasir R, Tian C, White PA, Butler LM, et al. (2009) Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Human Mutation 30: 69–76.

46. Kersbergen P, van Duijn K, Klooosterman AD, den Dunnen JT, Kayser M, et al. (2009) Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. BMC Genet 10: 67.

47. Pollack J, Vivalle FM, Cible MJ, Diegoli TM, van Oven M, et al. (2010) Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. Human Mutation 31: E1873–1893.

48. Paschou P, Lewis J, Javed A, Drineas P (2010) Ancestry informative markers for five-scale individual assignment to worldwide populations. Journal of Medical Genetics 47: 835–847.

49. Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, De La Vega FM, et al. (2011) Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. Investig Genet 2: 1.

50. Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, De La Vega FM, et al. (2011) Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. Investig Genet 2: 1.

51. Tandon A, Patterson N, Reich D (2011) Ancestry informative marker panels for African Americans based on subsets of commercially available SNP arrays. Genetic Epidemiology 35: 80–83.

52. Lao O, Vivalle FM, Cible MJ, Diegoli TM, van Oven M, et al. (2010) Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. Human Mutation 31: E1873–1893.

53. Paschou P, Lewis J, Javed A, Drineas P (2010) Ancestry informative markers for five-scale individual assignment to worldwide populations. Journal of Medical Genetics 47: 835–847.

54. Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, De La Vega FM, et al. (2011) Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. Investig Genet 2: 1.

55. Tandon A, Patterson N, Reich D (2011) Ancestry informative marker panels for African Americans based on subsets of commercially available SNP arrays. Genetic Epidemiology 35: 80–83.

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

57. Biwas S, Scheinfeldt LB, Akey JM (2009) Genome-wide insights into the genetic diversity of African Americans based on subsets of commercially available SNP arrays. Genetic Epidemiology 35: 80–83.

58. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. (2002) The Genographic Project: Human Diversity and the Human Genome. Science 298: 2381–2385.