Title: A GHEP-ISFG collaborative study on the genetic variation of 38 autosomal indels for human identification in different continental populations

Authors: R. Pereira, C. Alves, M. Aler, A. Amorim, C. Arévalo, E. Betancor, D. Braganholi, M.L. Bravo, P. Brito, J.J. Builes, G. Burgos, E.F. Carvalho, A. Castillo, C.I. Catanesi, R.M.B. Cicarelli, P. Coufalova, P. Dario, M.E. D'Amato, S. Davison, J. Ferragut, M Fondevila, S. Furfuro, O. García, A. Gaviria, I. Gomes, E. González, A. Gonzalez-Liñan, T.E. Gross, A. Hernández, Q. Huang, S. Jiménez, L.F. Jobim, A.M. López-Parra, M. Marino, S. Marques, G. Martínez-Cortés, V. Masciovecchio, D. Parra, G. Penacino, M.F. Pinheiro, M.J. Porto, Y. Posada, C. Restrepo, T. Ribeiro, L. Rubio, A. Sala, A. Santurtún, L.S. Solís, L. Souto, E. Streitemberger, A. Torres, C. Vilela-Lamego, J.J. Yunis, I. Yurrebaso, L. Gusmão

PII: S1872-4973(17)30200-4
DOI: http://dx.doi.org/10.1016/j.fsigen.2017.09.012
Reference: FSIGEN 1786

To appear in: *Forensic Science International: Genetics*

Received date: 28-3-2017
Revised date: 9-9-2017
Accepted date: 21-9-2017

Please cite this article as: R.Pereira, C.Alves, M.Aler, A.Amorim, C.Arévalo, E.Betancor, D.Braganholi, M.L.Bravo, P.Brito, J.J.Builes, G.Burgos, E.F.Carvalho, A.Castillo, C.I.Catanesi, R.M.B.Cicarelli, P.Coufalova, P.Dario, M.E.D'Amato, S.Davison, J.Ferragut, M Fondevila, S.Furfuro, O.García, A.Gaviria, I.Gomes, E.González, A.Gonzalez-Liñan, T.E.Gross, A.Hernández, Q.Huang, S.Jiménez, L.F.Jobim, A.M.López-Parra, M.Marino, S.Marques, G.Martínez-Cortés, V.Masciovecchio, D.Parra, G.Penacino, M.F.Pinheiro, M.J.Porto, Y.Posada, C.Restrepo, T.Ribeiro, L.Rubio, A.Sala, A.Santurtún, L.S.Solís, L.Souto, E.Streitemberger, A.Torres, C.Vilela-Lamego, J.J.Yunis, I.Yurrebaso, L.Gusmão, A GHEP-ISFG collaborative study on the genetic variation of 38 autosomal indels for human identification in different continental populations, Forensic Science International: Geneticshttp://dx.doi.org/10.1016/j.fsigen.2017.09.012

A GHEP-ISFG collaborative study on the genetic variation of 38 autosomal indels for human identification in different continental populations

R. Pereira[1,2], C. Alves[1,2], M. Aler[3], A. Amorim[1,2,4], C. Arévalo[5], E. Betancor[6], D. Braganholi[7], M.L. Bravo[8], P. Brito[9], J.J. Builes[8,10], G. Burgos[11,12], E.F. Carvalho[13], A. Castillo[14], C.I. Catanesi[15,16], R.M.B. Cicarelli[7], P. Coufalova[17], P. Dario[18], M.E. D'Amato[19], S. Davison[19], J. Ferragut[20], M Fondevila[21], S. Furfuro[22], O. García[23], A. Gaviria[11], I. Gomes[24], E. González[6], A. Gonzalez-Liñan[25], T.E. Gross[24], A. Hernández[26], Q. Huang[27], S. Jiménez[28], L.F. Jobim[29], A.M. López-Parra[27], M. Marino[22], S. Marques[1,2,4], G. Martínez-Cortés[30], V. Masciovecchio[31], D. Parra[32], G. Penacino[33], M.F. Pinheiro[34], M.J. Porto[9], Y. Posada[35], C. Restrepo[36], T. Ribeiro[18], L. Rubio[37], A. Sala[38], A. Santurtún[39], L.S. Solís[40], L. Souto[41], E. Streitemberger[31], A. Torres[42], C. Vilela-Lamego[43], J.J. Yunis[44], I. Yurrebaso[23], L. Gusmão[1,2,13]

[1]IPATIMUP (Institute of Molecular Pathology and Immunology of the University of Porto), Porto, Portugal
[2]I3s (Instituto de Investigação e Inovação em Saúde, Universidade do Porto), Porto, Portugal
[3]Sección de Genética Forense y Criminalística. Servicio de Laboratorio. Instituto de Medicina Legal y Ciencias Forenses de Valencia. Spain
[4]FCUP - Faculty of Sciences of the University of Porto, Portugal
[5]Laboratorio Biología-ADN. Unidad Central de Análisis Científicos. Comisaría General Policía Científica. Madrid, Spain
[6]Laboratorio de Genética Forense del Instituto de Medicina Legal de Las Palmas. Spain
[7]Laboratório de Investigação de Paternidade - NAC-FCFAr – UNESP. Brazil
[8]GENES SAS. Medellín, Colombia
[9]Instituto Nacional de Medicina Legal e Ciências Forenses, Delegação do Centro, Portugal
[10]Instituto de Biologia, Universidad de Antioquia, Medellín, Colombia
[11]Molecular Genetics Laboratory, Cruz Vital. Ecuadorian Red Cross, Quito, Ecuador
[12] Facultad de Ciencias de la Salud, Universidad de Las Américas, Quito, Ecuador
[13]DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil
[14]Laboratorio de Genética. Universidad Industrial de Santander (UIS), Bucaramanga, Santander, Colombia
[15]Identificación Genética - IMBICE. Instituto Multidisciplinario de Biologia Celular (CONICET-UNLP-CIC). La Plata, Argentina
[16]Fac. Cs. Naturales y Museo, Universidad Nacional de La Plata (UNLP), La Plata, Argentina
[17]Department of Genetics. Institute Criminalistics in Prague. Prague, Czech Republic.
[18]Instituto Nacional de Medicina Legal e Ciências Forenses, Delegação do Sul, Portugal
[19]Department of Biotechnology, Forensic DNA Lab. University of the Western Cape. Cape Town, South Africa
[20]Laboratori de Genètica (Department de Biologia). Universitat de les Illes Balears. Palma de Mallorca, Spain
[21]Sección de Genética Forense. Instituto de Ciencias Forenses Luis Concheiro. Facultade de Medicina. Universidade de Santiago de Compostela. A Coruña, Spain
[22]Laboratorio de Análisis de ADN. Facultad de Ciencias Médicas - UNCuyo. Mendoza, Argentina
[23]Forensic Science Unit, Forensic Genetics Section, Basque Country Police, Erndio (Bizkaia), Spain
[24]Institute of Legal Medicine, Faculty of Medicine, University of Cologne, Germany
[25]Andros Day Surgery Clinic. Unilab. Forensic Genetics Laboratory. Palermo, Italy
[26]Instituto Nacional de Toxicología y Ciencias Forenses. Delegación de Canarias. Spain
[27]Laboratorio de Genética Forense y Genética de Poblaciones, Depto. de Toxicología y Legislación Sanitaria, Fac. de Medicina, UCM, Madrid, Spain
[28]Laboratorio Biología Forense. Departamento de Patología y Cirugía. Universidad Miguel Hernández. Elche, Spain
[29]Laboratório DNA referencia. Serviço de Imunologia. Hospital de Clínicas de Porto Alegre. Brazil
[30]Laboratorio de Genética Molecular. DNA Profiles. Centro Universitario de la Ciénega, Universidad de Guadalajara (CUCI-UdeG). Jalisco, Mexico
[31]IACA Laboratorios. Bahía Blanca, Argentina

[32]Departamento de Biología del Servicio de Criminalística de la Guardia Civil. Madrid, Spain
[33]Unidad de Analisis de ADN (COFyBCF). Buenos Aires, Argentina
[34]Instituto Nacional de Medicina Legal e Ciências Forenses, Delegação do Norte, Portugal
[35]IdentiGEN - Genetic Identification Laboratory and Research Group of Genetic Identification, Institute of Biology, School of Natural and Exact Sciences (FCEN), University of Antioquia. Medellin, Antioquia, Colombia
[36]Genética Molecular de Colombia Ldta. Bogotá, Colombia
[37]Departamento de Anatomia y Medicina Legal. Laboratorio de periciales Médico Legales de la Universidad de Málaga. Spain
[38]Servicio de Huellas Digitales Genéticas. Facultad de Farmacía y Bioquímica. Universidad de Buenos Aires. Argentina
[39]Unit of Legal Medicine, University of Cantabria. Santander, Cantabria, Spain
[40]Laboratorio Clínico Genetix, S.A. Panamá
[41]Laboratório de Genética Aplicada. Departamento de Biologia, Universidade de Aveiro. Portugal
[42]Laboratorio Genia Geo, Montevideo, Uruguay
[43]INTS. Paris, France
[44]Servicios Médicos Yunis Turbay y Cia. Bogotá, Colombia

Corresponding Author:
Rui Pereira
IPATIMUP/i3s
Rua Alfredo Allen, 208
4200-135 Porto, Portugal
Phone: +351 220 408 800
email: rpereira@ipatimup.pt

Highlights

- A collaborative study was performed by laboratories of the Spanish and Portuguese Speaking Working Group of the International Society for Forensic Genetics for a 38 autosomal indel multiplex.
- Allele frequency databases were created covering a comprehensive geographic area with representative samples from 21 different countries: Angola, Argentina, Brazil, Cape Verde, China, Colombia, Czech Republic, East Timor, Ecuador, France, Germany, Iraq, Italy, Mexico, Panama, Portugal, South Africa, Spain, Sudan, Uganda and Uruguay.
- Statistically significant differences were found among some admixed populations inside Latin American countries, namely Brazil, Colombia and Mexico.
- The high levels of diversity found in all populations support the usefulness of this marker set in the forensic context.

Abstract

A collaborative effort was carried out by the Spanish and Portuguese Speaking Working Group of the International Society for Forensic Genetics (GHEP-ISFG) to promote knowledge exchange between associate laboratories interested in the implementation of indel-based methodologies and build allele frequency databases of 38 indels for forensic applications. These databases include populations from different countries that are relevant for identification and kinship investigations undertaken by the participating laboratories. Before compiling population data, participants were asked to type the 38 indels in blind samples from annual GHEP-ISFG proficiency tests, using an amplification protocol previously described. Only laboratories that reported correct results contributed with population data to this study. A total of 5839 samples were genotyped from 45 different populations from Africa, America, East Asia, Europe and Middle East. Population differentiation analysis showed significant differences between most populations studied from Africa and America, as well as between two Asian populations from China and East Timor. Low F_{ST} values were detected among most European populations. Overall diversities and parameters of forensic efficiency were high in populations from all continents.

Introduction

Several PCR multiplex strategies have been optimized to increment discrimination capacity in genetic identification testing. These multiplexes consist of different types of markers, selected on the basis of their mode of transmission and suitability to produce complete genetic profiles even from degraded DNA samples [1, 2].

The most widely used markers in forensic genetics are autosomal STRs. Markers located in the autosomes allow identification since, contrarily to lineage markers (specific from Y chromosome and mtDNA), they recombine during meiosis. Moreover, autosomal markers can be used in any kinship scenario, since their transmission is not restricted to specific parent-child constellations (all alleles can be potentially transmitted from both parents, to offspring of both sexes).

The STRs have the advantage of being more polymorphic than SNPs or indels, since they have higher mutation rates and more alleles per locus. However, they present some limitations in the analysis of degraded DNA samples, due to the relatively large size of the PCR amplicons, when compared to SNPs or small indels. Therefore, in the last decade, new multiplexes comprising SNPs or indels were described to complement STR typing strategies, when partial profiles are obtained with conventional STR multiplexes, due to poor quality of the available DNA samples.

Informative SNP- or indel-based multiplexes must include a large number of markers to compensate for the lower diversity of bi-allelic markers [3, 4]. Large multiplexes are currently

available for forensic use, allowing the genotyping of many SNPs using Sanger sequencing or Massively Parallel Sequencing (MPS) [e.g. 5, 6, 7]. Indel multiplexes are advantageous over SNPs in regard to easier genotyping with conventional automated fragment size analysis, which is common technique available in most forensic laboratories worldwide [8-12].

In 2009, a multiplex of indels was described by Pereira et al. [8], combining 38 markers with high polymorphism in European, African and Asians in a single PCR. Due to the short amplicon size, ranging from 57 to 158 bp, this multiplex has proved useful in degraded samples analysis [13, 14].

In this paper, we present the results of a collaborative study performed among laboratories of the Spanish and Portuguese Speaking Working Group of the International Society for Forensic Genetics (GHEP-ISFG), for the 38 indelplex described by Pereira et al. [8]. The aim of this study was to estimate allele frequencies and forensic relevant parameters in different populations used as reference by the participating laboratories in their casework. The study was organized in two phases: the first phase consisted in the genotyping of samples included in the annual GHEP-ISFG proficiency tests; while in the second part, laboratories that successfully typed the test samples in the first phase, for the full set of markers, were invited to genotype population samples, for at least 100 unrelated individuals. In total, 54 population samples from 21 countries from Africa, America, East Asia, Europe, and Middle East, were genotyped.

Material and Methods

GHEP-ISFG working commission on autosomal indels for human identification

In 2012, a GHEP-ISFG working commission on autosomal indels for human identification was established to coordinate a collaborative exercise aiming to promote knowledge exchange between associate laboratories interested in implementing indel-based methodologies, and creating population databases for forensic use (https://ghep-isfg.org/en/working-commissions/autosomal-indels-for-identification/). This study was approved by vote at the GHEP-ISFG general meeting and complies with the ethical principles of the 2000 Helsinki Declaration of the World Medical Association.

In brief, the exercise was organized in two steps. In a first phase, four samples from the annual GHEP-ISFG proficiency tests were used for quality control purposes (samples 1 to 4 included in 2012 or 2013 kinship basic modules; see [15] for details on GHEP-ISFG proficiency tests). Written

informed consent was obtained from the donors for using these samples in the exercise. The laboratories that implemented the technique successfully and reported correct genotypes for all indel markers in the 4 control samples, qualified to enter the second phase. Participants were then invited to characterize population samples of their interest to set up allele frequency databases useful for their routine casework (minimum of 100 samples per population).

Population samples

The samples used in the second phase of the collaborative exercise were anonymised DNA extracts previously obtained from healthy unrelated individuals who consented to participate in this study under strictly confidential conditions.

In this study, 40 forensic laboratories characterized a total of 54 population samples from 4 continents. Population samples coming from the same geographic region were combined after assessing the absence of population structure (see Results and Discussion section). Figure 1 details the location and size of the final 45 different population samples characterized in this study. Population samples were coded by the name of the country, province or city they represent, depending on the sampling scheme of each laboratory.

Samples from Angola, South Africa and Uganda were named by the ethnic group that they represent (details on the Ugandan Karimojong and South African Xhosa and Zulu samples can be found in [16, 17]).

The Mexican samples from Guerrero are from natives belonging to the Nahuas, Mixtec and Tlapanecos ethnic groups. Samples from Yucatán and Chihuahua represent Mestizo admixed populations from these provinces. A subset of these two population samples was previously studied by Martínez-Cortéz et al. [18], together with other Mexican regions.

The sample from Iraq includes individuals born in Iraq but tested in Cologne, Germany, mostly obtained from immigration cases. Cape Verde includes individuals born in Cape Verde and living in Portugal.

Genotyping protocol

All samples were genotyped for a panel of 38 autosomal indels in a single PCR multiplex reaction using a previously described protocol [8]. For this collaborative exercise, a new stock of primer mix was prepared and tested by the coordinating laboratory, and then distributed in aliquots of

500 µl to all participants. Any other necessary materials for the genotyping of the samples were the responsibility of each laboratory.

To facilitate the genotyping process, dedicated files for the marker panel and allelic bins were made available for different versions of GeneMapper software (Applied Biosystems). Support to adjust markers and bins to specific conditions of electrophoretic mobility was offered upon request. The same applied for generic questions related with the method implementation in the laboratories.

Statistical Analyses

Allele frequency estimates, gene diversity values, observed and expected heterozygosities, Hardy-Weinberg equilibrium (HWE) exact tests, and pairwise F_{ST} genetic distances were all calculated using Arlequin software v3.5.2.2 [19]. A Neighbor-joining tree was built from the pairwise F_{ST} matrix with the Neighbor program implemented in the Phylip v3.695 software package [20] and plotted using TreeView v1.6.6 [21]. Principal component analysis was performed using Statistica v13.0 (Statsoft; http://www.statsoft.com/).

In the statistical analyses and interpretation of genetic data, a Bonferroni correction was considered whenever multiple testing was performed. In HWE exact tests, the significance level was adjusted for the total number of loci in the set (p=0.05/38), and in genetic distance analyses, the number of pairwise F_{ST} calculations was considered.

Results and Discussion

In a first stage, 53 participant laboratories genotyped four samples from the GHEP-ISFG proficiency tests, for quality control purposes. From these, 47 obtained identical results for all samples and markers. The results of this phase showed a good performance of the indel multiplex, similar to what was observed in a previous collaborative inter-laboratory exercise organized by the European DNA Profiling group (EDNAP) involving indel- and SNP-based ancestry informative marker panels [12]. In fact, the genotyping completeness and concordance was even higher in the present study, with only six laboratories out of 53 presenting incomplete profiles and/or genotyping errors.

The causes for inconsistencies were evaluate by the inspection of the electropherograms and protocols used, and discussed during a GHEP-ISFG meeting. Most errors concentrated in a single laboratory using an amplification master mix different from the recommended. Another

laboratory, employing an amplification buffer from a new generation forensic kit, experienced an impaired performance and reported one genotype error for the largest VIC-labelled marker (G09). Two laboratories failed to report results for the largest PET-labelled marker (R10) in all samples: (i)in one case, the weak amplification of the samples allied to the presence of noticeable dye-blobs in small scales hindered the genotyping of R10; (ii) the other laboratory did not adequately adjust the allelic bins for R10; moreover, for sample M2, an additional allele at B03 was reported as a result of pull-up from an overscaled homozygous peak at G02. Finally, one laboratory failed at R01 for sample M4 due to extreme allelic imbalance, while other incurred in a transcription error in R02.

In a second stage, the laboratories that reported correct results were asked to genotype a population sample, to create a database of allele frequencies to be used as reference in forensic casework. A total of 40 laboratories sent results for at least 100 unrelated individuals for 54 population samples from different countries: Angola, Cape Verde, South Africa, Sudan, Uganda, Iraq, China, East Timor, Czech Republic, France, Germany, Italy, Portugal, Spain, Argentina, Brazil, Colombia, Ecuador, Mexico, Panama, and Uruguay. The genotyping results obtained in these populations for a total of 5839 individuals (after removing a total of 8 samples with more than 10% missing data) are listed in Supplementary Table S1.

Comparison of samples from the same population

A first analysis was performed between samples from the same region that were typed by different laboratories, to verify if their genotype distributions were not significantly different and could be combined. Namely, two different population sample sets from Bogotá, Buenos Aires, Canarias, Chocó, Madrid, North Portugal, São Paulo and Valencian Community were genotyped by different laboratories. All sample pairs from the same populations were combined, since they showed no-significant differences ($-0.00107 \leq F_{ST} \leq 0.00157$; $0.0540 \leq P \leq 0.88199$), except for Buenos Aires samples revealing much higher F_{ST} and significant non-differentiation P-value ($F_{ST}=0.01239$; $P \leq 0.00000$). In one of these samples, an unusual excess of homozygotes was consistently observed in 6 out of the 38 studied loci associated with low P-values in the HWE exact test ($0.00000 \leq P \leq 0.00896$), indicating possible genotyping or sampling problems. For the same laboratory, a sample from another population also pointed out similar problems (an excess of homozygotes was observed in 7 loci associated with low P-values in the HWE test ($0.00000 \leq P \leq 0.00511$) and, therefore, genotyping data from this laboratory were not included in the present study.

Differentiation analysis between different samples from the same country

After the two exclusions and having pooled samples from the same geographic location, a comprehensive pairwise F_{ST} analysis was performed between the resulting 45 population samples. All F_{ST} values and corresponding non-differentiation *P*-values are presented in Supplementary Table S2.

Significant differences were not expected for closely related populations, since the multiplex under study includes 38 indels that were selected to maximize diversity within rather than between populations [8]. Therefore, whenever non-statistically significant differences were observed (Supplementary Table S2), samples from the same country were pooled, before calculating allele frequencies and other forensically relevant parameters. This was the case for two samples from South African Bantu groups (Xhosa and Zulu) (F_{ST}=0.00029; *P*=0.3499); the two samples from Ecuador (Sierra and Costa) (F_{ST}=0.00380; *P*=0.00554); the three samples from Portugal (North, Central and South regions) (F_{ST}≤0.00181; *P*≥0.02772); and the 6 samples from Argentina, including Buenos Aires, Entre Rios, Mendoza, Resistencia, San Luis and Tucuman (F_{ST}≤0.00475; *P*≥0.00079).

Small F_{ST} genetic distances were also found among samples from Spain (F_{ST}≤0.00424; *P*≥0.00208), except for the Basques that showed significant differences in 4 out of the 6 comparisons performed with other Spanish samples. Therefore, samples from Galicia, Cantabria, Madrid, Valencian Community, Malaga, Majorca and Canarias were pooled in a single database from Spain, separated from the Basque Country database.

Among four Brazilian samples, Rio de Janeiro, Espírito Santo and São Paulo showed non-significant differences, and were pooled in a single database representing Brazilian Southeast region. In the comparison of Rio de Janeiro or Espírito Santo with Porto Alegre (in South region) significant differences were observed.

Concerning Colombia and Mexico, none of their samples could be combined, since statistically significant differences were found between the four Colombian samples from Antioquia, Bogotá, Norte de Santander and Chocó, and between the three samples from Mexico (from Guerrero natives and the provinces of Chihuahua and Yucatán).

Differentiation analysis within and between countries and continents

A new pairwise F_{ST} analysis was performed in a final set of 28 population samples (after combining a total of 23 population samples in only 6 representatives, as described above). The results are presented in Supplementary Table S3, together with corresponding non-differentiation *P*-values.

The matrix of F_{ST} genetic distances was used to draw a Neighbor-joining (NJ) tree, which is represented in Fig. 2. The overall pattern of genetic distances represented in the NJ tree reflects geographic positions, with populations from different continents separated by large distances. In the extremities of the tree are four groups representing Sub-Saharan African, Asian, Native American and European populations. The Northern African population of Sudan is in between Sub-Saharan Africans and Europeans. In intermediate positions between the four continental groups are the admixed populations from South America and the African admixed population from Cape Verde.

Among African populations, significant genetic distances were observed in all pairwise comparisons (Supplementary Table S3). The smallest distance (F_{ST}=0.00619) was found between Bantu groups from Angola (Bakongo) and South Africa (Xhosa and Zulu); F_{ST}s between the remaining sample pairs were all higher than 1% ($0.01188 \leq F_{ST} \leq 0.03761$).

Europeans and Asians appear similarly distant from Sub-Saharan Africans (Fig. 2), with F_{ST} values varying between 9.7 and 12.7 % (Supplementary Table S3). The Asian samples from East Timor and Shanghai are well separated, by both geographic and genetic distances (F_{ST}=0.03679). In contrast, European populations present low genetic distances for the studied markers ($F_{ST} \leq 0.01060$), with no statistically significant differences detected between Czech Republic, Germany, France, Spain and Portugal (all F_{ST}s were below 0.00128). Such low genetic distances are in line with previous studies investigating the fine structure of European populations using high density SNP arrays [e.g. 22, 23, 24]. Iraq and Basque Country showed statistically significant differences in most comparisons with other European populations.

Concerning South and Central American populations, they are separated by large F_{ST} values, which can be attributed to different levels of African and European admixture. The samples from Argentina, Brazil, Colombia (except Chocó) and Uruguay, stand close to the European group. Interestingly, Porto Alegre (Brazil) and Uruguay show no-significant differences between them (F_{ST}=-0.00017; *P*=0.51747), or in most comparisons with Europeans. This is most probably due to the high European ancestry found both in South Brazil and Uruguay (over 70%), as documented in previous studies using autosomal ancestry informative markers [25, 26].

On the other hand, populations from Ecuador and Mexico are in a separated branch of the Neighbor-joining tree (Fig. 2; Supplementary Table S3), representing populations with an important Native American background, which is highly variable among the three Mexican populations from Guerrero Natives, Yucatán and Chihuahua.

In the NJ tree, the Colombian population from Chocó is closer to the African Bantu populations than Cape Verde. Despite the geographic disparity, this is in accordance with the higher African ancestry estimated for this Colombian region, using autosomal ancestry informative markers (63%) [27], than for Cape Verde (57%) [28].

A Principal Component Analysis (PCA) based on allele frequencies was additionally performed to evaluate the consistency of the NJ tree based on F_{ST} genetic distances. The PCA (Fig. 3) reinforced previous results showing a close relationship between African Bantu groups, and between European populations, as well as a high dispersion of American populations with different admixture levels. The three principal components capture 74.30% of the total inertia (Fig. 3). The first axis mainly separates the African from the non-African populations. The second axis separates Europeans from Asians and Native Americans, which are further separated in the third axis.

Analysis of intra-population diversity

Allele frequencies and gene diversities were calculated for the 38 indels in the 28 population samples, and are presented in Supplementary Table S4.

Average gene diversities over loci were high in all populations (Supplementary Table S4). The lowest values were found in Guerrero Native Americans (0.3688) and in Sub-Saharan African populations (varying between 0.4039 and 0.4095). Intermediate values were found in samples from Asia (0.4123 for East Timor and 0.4207 for Shanghai, China), and the highest gene diversities were obtained for Sudan, Europeans, and American admixed populations. This otherwise unexpected pattern of genetic diversity, showing lower diversity in Africans, is probably due to bias emerging during polymorphism ascertainment efforts. Ascertainment bias is reported to be higher for indel markers than for SNPs or STRs [29].

The average locus diversity over populations was high for all markers, considering the binary nature of the studied polymorphisms. A total of 32 loci showed average diversities above 0.40. The lowest value was 0.34 for marker R09, which was the less diverse indel in Asians and Native Americans. Low gene diversities were also obtained for R04 in Sub-Saharan Africans, G07 and R02 in Europeans, and G04 in populations with high Native American ancestry.

A new variant allele was found in marker B06, with an amplicon size corresponding to 1bp more than the long allele. This variant was observed in two individuals from São Paulo and two individuals from Portugal (Supplementary Table S1). When investigating the neighbour sequence on NCBI dbSNP (build149), it was not possible to find any size variant inside the amplified sequence that could justify the observed allele. Instead, a 1 bp deletion is annotated in a G homopentamer located immediately upstream the reverse primer. This variant, rs550033317, was newly reported by the 1000 Genomes project (phase 3) with a very low frequency of 0.06% (3/5008 alleles; one in "American of African ancestry in SW USA", one in "Japanese in Tokyo, Japan", and one in "Kinh in Ho Chi Minh city, Vietnam")[30].

Five cases of locus dropout were observed in this study, including one sample from Canarias, Portugal and Angola, and two samples from Timor (Supplementary Table S1). These samples showed a complete profile for 37 loci with no-amplification of marker Y09, which is compatible with the presence of silent alleles. Therefore, primer binding regions were searched for variants possibly causing this problem and a C/A transversion (rs80011419) was found, annotated 5 bp upstream the 3'end of forward primer (NCBI dbSNP build 149). The variant A allele, likely to impair primer annealing during PCR, was reported with a frequency of 0.0050 in East Asians, 0.0229 in Europeans, 0.0129 in Africans, 0.0130 in Admixed Americans, and 0.0491 in South Asians (1000 Genomes phase 3 data; [30]). The frequencies estimated in this study for silent alleles were compatible with those reported for Africans and Europeans. In East Timor, a frequency of 14% was estimated for the A allele (based on the two putative homozygotes observed in a total sample of 101 individuals), which is much higher than the reported for East Asian populations in public databases.

Exact tests of Hardy-Weinberg equilibrium revealed no-significant deviations in most populations for the 38 indels (Supplementary Table S5). The only exceptions were markers G01 and G09 in Panama that presented significant differences between expected and observed heterozygosity values (P=0.00052 and P=0.00085, respectively) associated with either an excess of heterozygous (G01) or homozygous (G09). In marker Y09, the presence of silent alleles could possibly result in an excess of homozygotes in the samples from Canarias, Portugal, Angola, and Timor, where homozygotes for silent alleles were detected. However, HWE tests showed no statistically significant deviations in this locus, indicating that a low frequency of silent alleles can be expected in all studied populations. In any case, it is important to highlight that silent alleles in Y09 can potentially produce apparent exclusions in paternity cases and therefore, its frequency should be accounted for in statistical evaluations of kinships, by considering as proxy,

the frequencies reported in dbSNP for rs80011419 allele A in East Asians, Europeans, Africans, Admixed Americans, and South Asians.

Forensically relevant parameters

Forensic efficiency parameters calculated for the 38 indels in all populations are presented in Supplementary Tables S6 and S7. As for the gene diversities, Guerrero Native Americans presented the lowest accumulated values of power of discrimination (PD) and power of exclusion (PE), followed by sub-Saharan African and Asian populations. These values are higher in European and American admixed populations. Panama and Southeast Brazil are those with the highest values (PD is 0.9999999999999995 and 0.9999999999999990, respectively; PE is 0.9987 and 0.9979, respectively). Previous reports using the same 38 indel set as in this study showed very similar results in population samples from Brazil, Portugal, and Spain [8, 31, 32].

When comparing the studied 38 indel set with the 30 indel set commercially available in the Investigator® DIPplex Kit (Qiagen, Hilden, Germany), accumulated PD values were about two orders of magnitude higher in population samples from Spain, Iraq, Brazil, and Uruguay [33-36], and three orders of magnitude higher in populations from México, South Africa, and East Asia [17, 37-39]. These results support previous findings showing higher forensic information content of the 38 indel panel compared to the Investigator® DIPplex Kit in US African American, Caucasian, East Asian, and Hispanic samples [40] (see Supplementary Table 8 for details). Moreover, the informativeness of the 38 indelplex is more uniform across worldwide population groups when compared to the Investigator® DIPplex Kit, which shows a more pronounced decrease in diversity in non-European groups [this work, 17, 39, 40], regardless of the ascertainment bias already discussed for indel genetic markers [29]. For convenience, Supplementary Table 8 presents a comparison of the forensic efficiency of this multiplex with other assays commonly used in human identification (see also Table 4 in [8]). Taking advantage of comprehensive population data available for different US groups [10, 40, 41] as an example, emphasis was given on short amplicon approaches like small indels and miniSTRs included in more recent commercial STR kits. The 38 indel set used in this study allows remarkably high *a priori* informativeness levels when compared to STRs with similar amplicon lengths, thus highlighting its utility in forensic applications, especially in challenging samples.

Conclusions

The present collaborative study allowed the successful implementation of an indel-based multiplex in the vast majority of participant laboratories and the compilation of allele frequency data from a large number of populations of different continental origin.

Population comparisons showed that differences are higher among populations within Africa, America and Asia than within Europe. American admixed populations from Mexico, Brazil and Colombia showed high variation within countries, indicating that a correction for population substructure should be applied when databases from general populations are used for these markers in forensic casework. Conversely, F_{ST} values were low within Europe, supporting the use of single databases for populations from more than one country, for example a single Iberian database, excluding Basques.

The ease of the indel genotyping procedure using standard routine platforms and the high level of informativeness found in all populations support the usefulness of this marker set in the forensic context.

Acknowledgements

RP is supported by a postdoctoral fellowship (SFRH/BPD/81986/2011) awarded by the Portuguese Foundation for Science and Technology (FCT) and co-financed by the European Social Fund (Human Potential Thematic Operational Programme – POPH.

References

1. Jobling MA, Gill P. Encoded evidence: DNA in forensic analysis. Nat Rev Genet. 2004;5(10):739-51.

2. Schneider PM. Beyond STRs: The Role of Diallelic Markers in Forensic Genetics. Transfusion Medicine and Hemotherapy. 2012;39(3):176-80.

3. Amorim A, Pereira L. Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. Forensic Sci Int. 2005;150(1):17-21.

4. Gill P. An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. Int J Legal Med. 2001;114(4 - 5):204-10.

5. Eduardoff M, Gross TE, Santos C, de la Puente M, Ballard D, Strobl C, Borsting C, Morling N, Fusco L, Hussing C, Egyed B, Souto L, Uacyisrael J, Syndercombe Court D, Carracedo A, Lareu MV, Schneider PM, Parson W, Phillips C, Consortium EU-N, Parson W, Phillips C. Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel

sequencing using the Ion PGM. Forensic Sci Int Genet. 2016;23:178-89. doi: 10.1016/j.fsigen.2016.04.008.

6. Grandell I, Samara R, Tillmar AO. A SNP panel for identity and kinship testing using massive parallel sequencing. Int J Legal Med. 2016;130(4):905-14. doi: 10.1007/s00414-016-1341-4.

7. Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D, Schneider PM, Carracedo A, Morling N. A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis. 2006;27(9):1713-24. doi: 10.1002/elps.200500671.

8. Pereira R, Phillips C, Alves C, Amorim A, Carracedo A, Gusmão L. A new multiplex for human identification using insertion/deletion polymorphisms. Electrophoresis. 2009;30(21):3682-90. Epub 2009/10/29. doi: 10.1002/elps.200900274.

9. Santos NP, Ribeiro-Rodrigues EM, Ribeiro-dos-Santos AK, Pereira R, Gusmão L, Amorim A, Guerreiro JF, Zago MA, Matte C, Hutz MH, Santos SE. Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INDEL) ancestry-informative marker (AIM) panel. Hum Mutat. 2010;31(2):184-90. Epub 2009/12/03. doi: 10.1002/humu.21159.

10. LaRue BL, Ge J, King JL, Budowle B. A validation study of the Qiagen Investigator DIPplex(R) kit; an INDEL-based assay for human identification. Int J Legal Med. 2012;126(4):533-40. doi: 10.1007/s00414-012-0667-9.

11. Zaumsegel D, Rothschild MA, Schneider PM. A 21 marker insertion deletion polymorphism panel to study biogeographic ancestry. Forensic Sci Int Genet. 2013;7(2):305-12. doi: 10.1016/j.fsigen.2012.12.007.

12. Santos C, Fondevila M, Ballard D, Banemann R, Bento AM, Borsting C, Branicki W, Brisighelli F, Burrington M, Capal T, Chaitanya L, Daniel R, Decroyer V, England R, Gettings KB, Gross TE, Haas C, Harteveld J, Hoff-Olsen P, Hoffmann A, Kayser M, Kohler P, Linacre A, Mayr-Eduardoff M, McGovern C, Morling N, O'Donnell G, Parson W, Pascali VL, Porto MJ, Roseth A, Schneider PM, Sijen T, Stenzl V, Court DS, Templeton JE, Turanska M, Vallone PM, van Oorschot RA, Zatkalikova L, Carracedo A, Phillips C, Consortium EU-N. Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (AIM) panels: Results of a collaborative EDNAP exercise. Forensic Sci Int Genet. 2015;19:56-67. doi: 10.1016/j.fsigen.2015.06.004.

13. Pereira R, Phillips C, Alves C, Amorim A, Carracedo Á, Gusmão L. Insertion/deletion polymorphisms: A multiplex assay and forensic applications. Forensic Sci Int Genet Supplement Series. 2009;2(1):513-5. doi: 10.1016/j.fsigss.2009.09.005.

14. Romanini C, Catelli ML, Borosky A, Pereira R, Romero M, Salado Puerto M, Phillips C, Fondevila M, Freire A, Santos C, Carracedo A, Lareu MV, Gusmao L, Vullo CM. Typing short amplicon binary polymorphisms: Supplementary SNP and Indel genetic information in the analysis of highly degraded skeletal remains. Forensic Sci Int Genet. 2012;6(4):469-76. Epub 2011/11/29. doi: 10.1016/j.fsigen.2011.10.006.

15. Fernández K, Gómez J, García-Hirschfeld J, Cubillo E, de la Torre CS, Vallejo G. Accreditation of the GHEP-ISFG proficiency test: One step forward to assure and improve quality. Forensic Sci Int Genet Supplement Series. 2015;5:e515-e7. doi: 10.1016/j.fsigss.2015.09.204.

16. Gomes V, Sanchez-Diz P, Alves C, Gomes I, Amorim A, Carracedo A, Gusmao L. Population data defined by 15 autosomal STR loci in Karamoja population (Uganda) using AmpF/STR Identifiler kit. Forensic Sci Int Genet. 2009;3(2):e55-8. doi: 10.1016/j.fsigen.2008.06.005.

17. Hefke G, Davison S, D'Amato ME. Forensic performance of Investigator DIPplex indels genotyping kit in native, immigrant, and admixed populations in South Africa. Electrophoresis. 2015;36(24):3018-25. doi: 10.1002/elps.201500243.

18. Martinez-Cortes G, Gusmao L, Pereira R, Salcido VH, Favela-Mendoza AF, Munoz-Valle JF, Inclan-Sanchez A, Lopez-Hernandez LB, Rangel-Villalobos H. Genetic structure and forensic parameters of 38 Indels for human identification purposes in eight Mexican populations. Forensic Sci Int Genet. 2015;17:149-52. doi: 10.1016/j.fsigen.2015.04.011.

19. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 2010;10(3):564-7. doi: 10.1111/j.1755-0998.2010.02847.x.

20. Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. 1989;5:164-6.

21. Page RD. TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci. 1996;12(4):357-8.

22. Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, Foretova L, Georges M, Janout V, Kabesch M, Krokan HE, Elvestad MB, Lissowska J, Mates D, Rudnai P, Skorpen F, Schreiber S, Soria JM, Syvanen AC, Meneton P, Hercberg S, Galan P, Szeszenia-Dabrowska N, Zaridze D, Genin E, Cardon LR, Lathrop M. Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet. 2008;16(12):1413-29. doi: 10.1038/ejhg.2008.210.

23. Nelis M, Esko T, Magi R, Zimprich F, Zimprich A, Toncheva D, Karachanak S, Piskackova T, Balascak I, Peltonen L, Jakkula E, Rehnstrom K, Lathrop M, Heath S, Galan P, Schreiber S, Meitinger T, Pfeufer A, Wichmann HE, Melegh B, Polgar N, Toniolo D, Gasparini P, D'Adamo P,

Klovins J, Nikitina-Zake L, Kucinskas V, Kasnauskiene J, Lubinski J, Debniak T, Limborska S, Khrunin A, Estivill X, Rabionet R, Marsal S, Julia A, Antonarakis SE, Deutsch S, Borel C, Attar H, Gagnebin M, Macek M, Krawczak M, Remm M, Metspalu A. Genetic structure of Europeans: a view from the North-East. PLoS One. 2009;4(5):e5472. doi: 10.1371/journal.pone.0005472.

24. Tian C, Kosoy R, Nassir R, Lee A, Villoslada P, Klareskog L, Hammarstrom L, Garchon HJ, Pulver AE, Ransom M, Gregersen PK, Seldin MF. European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. Mol Med. 2009;15(11-12):371-83. doi: 10.2119/molmed.2009.00094.

25. Manta FS, Pereira R, Vianna R, Araujo ARBd, Gitaí DLG, Silva DA, Wolfgramm EdV, Pontes IdM, Aguiar JI, Moraes MO, Carvalho EF, Gusmão L. Revisiting the genetic ancestry of Brazilians using autosomal AIM-Indels. PLoS One. 2013;8(9):e75145. doi: 10.1371/journal.pone.0075145.

26. Bonilla C, Bertoni B, Hidalgo PC, Artagaveytia N, Ackermann E, Barreto I, Cancela P, Cappetta M, Egana A, Figueiro G, Heinzen S, Hooker S, Roman E, Sans M, Kittles RA. Breast cancer risk and genetic ancestry: a case-control study in Uruguay. BMC Womens Health. 2015;15:11. doi: 10.1186/s12905-015-0171-8.

27. Ossa H, Aquino J, Pereira R, Ibarra A, Ossa RH, Perez LA, Granda JD, Lattig MC, Groot H, Fagundes de Carvalho E, Gusmao L. Outlining the Ancestry Landscape of Colombian Admixed Populations. PLoS One. 2016;11(10):e0164414. doi: 10.1371/journal.pone.0164414.

28. Beleza S, Campos J, Lopes J, Araujo, II, Hoppfer Almada A, Correia e Silva A, Parra EJ, Rocha J. The admixture structure and genetic variation of the archipelago of Cape Verde and its implications for admixture mapping studies. PLoS One. 2012;7(11):e51103. doi: 10.1371/journal.pone.0051103.

29. Romero IG, Manica A, Goudet J, Handley LL, Balloux F. How accurate is the current picture of human genetic variation? Heredity. 2009;102(2):120-6. Epub 2008/09/04. doi: hdy200889 [pii]

10.1038/hdy.2008.89.

30. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68-74. doi: 10.1038/nature15393

http://www.nature.com/nature/journal/v526/n7571/abs/nature15393.html#supplementary-information.

31. Manta F, Caiafa A, Pereira R, Silva D, Amorim A, Carvalho EF, Gusmão L. Indel markers: genetic diversity of 38 polymorphisms in Brazilian populations and application in a paternity investigation with post mortem material. Forensic Sci Int Genet. 2012;6(5):658-61. doi: 10.1016/j.fsigen.2011.12.008.

32. Cardoso S, Sevillano R, Gamarra D, Santurtún A, Martínez-Jarreta B, de Pancorbo MM. Population genetic data of 38 insertion-deletion markers in six populations of the northern fringe of the Iberian Peninsula. Forensic Sci Int Genet. 2017;27:175-9. doi: 10.1016/j.fsigen.2016.12.014.

33. Ferreira Palha Tde J, Ribeiro Rodrigues EM, Cavalcante GC, Marrero A, de Souza IR, Seki Uehara CJ, Silveira da Motta CH, Koshikene D, da Silva DA, de Carvalho EF, Chemale G, Freitas JM, Alexandre L, Paranaiba RT, Soler MP, Santos S. Population genetic analysis of insertion-deletion polymorphisms in a Brazilian population using the Investigator DIPplex kit. Forensic Sci Int Genet. 2015;19:10-4. doi: 10.1016/j.fsigen.2015.03.015.

34. Martin P, Garcia O, Heinrichs B, Yurrebaso I, Aguirre A, Alonso A. Population genetic data of 30 autosomal indels in Central Spain and the Basque Country populations. Forensic Sci Int Genet. 2013;7(2):e27-30. doi: 10.1016/j.fsigen.2012.10.003.

35. Saiz M, Andre F, Pisano N, Sandberg N, Bertoni B, Pagano S. Allelic frequencies and statistical data from 30 INDEL loci in Uruguayan population. Forensic Sci Int Genet. 2014;9:e27-9. doi: 10.1016/j.fsigen.2013.07.013.

36. Tomas C, Poulsen L, Drobnic K, Ivanova V, Jankauskiene J, Bunokiene D, Borsting C, Morling N. Thirty autosomal insertion-deletion polymorphisms analyzed using the Investigator(R) DIPplex Kit in populations from Iraq, Lithuania, Slovenia, and Turkey. Forensic Sci Int Genet. 2016;25:142-4. doi: 10.1016/j.fsigen.2016.08.006.

37. Shi M, Liu Y, Bai R, Jiang L, Lv X, Ma S. Population data of 30 insertion-deletion markers in four Chinese populations. Int J Legal Med. 2015;129(1):53-6. doi: 10.1007/s00414-014-1091-0.

38. Wang L, Lv M, Zaumsegel D, Zhang L, Liu F, Xiang J, Li J, Schneider PM, Liang W, Zhang L. A comparative study of insertion/deletion polymorphisms applied among Southwest, South and Northwest Chinese populations using Investigator® DIPplex. Forensic Sci Int Genet. 2016;21:10-4. doi: 10.1016/j.fsigen.2015.08.005.

39. Martinez-Cortes G, Garcia-Aceves M, Favela-Mendoza AF, Munoz-Valle JF, Velarde-Felix JS, Rangel-Villalobos H. Forensic parameters of the Investigator DIPplex kit (Qiagen) in six Mexican populations. Int J Legal Med. 2016;130(3):683-5. doi: 10.1007/s00414-015-1242-y.

40. Fondevila M, Phillips C, Santos C, Pereira R, Gusmao L, Carracedo A, Butler JM, Lareu MV, Vallone PM. Forensic performance of two insertion-deletion marker assays. Int J Legal Med. 2012;126(5):725-37. doi: 10.1007/s00414-012-0721-7.

41. Hill CR, Duewer DL, Kline MC, Coble MD, Butler JM. U.S. population data for 29 autosomal STR loci. Forensic Sci Int Genet. 2013;7(3):e82-3. doi: 10.1016/j.fsigen.2012.12.004.

Figure Captions

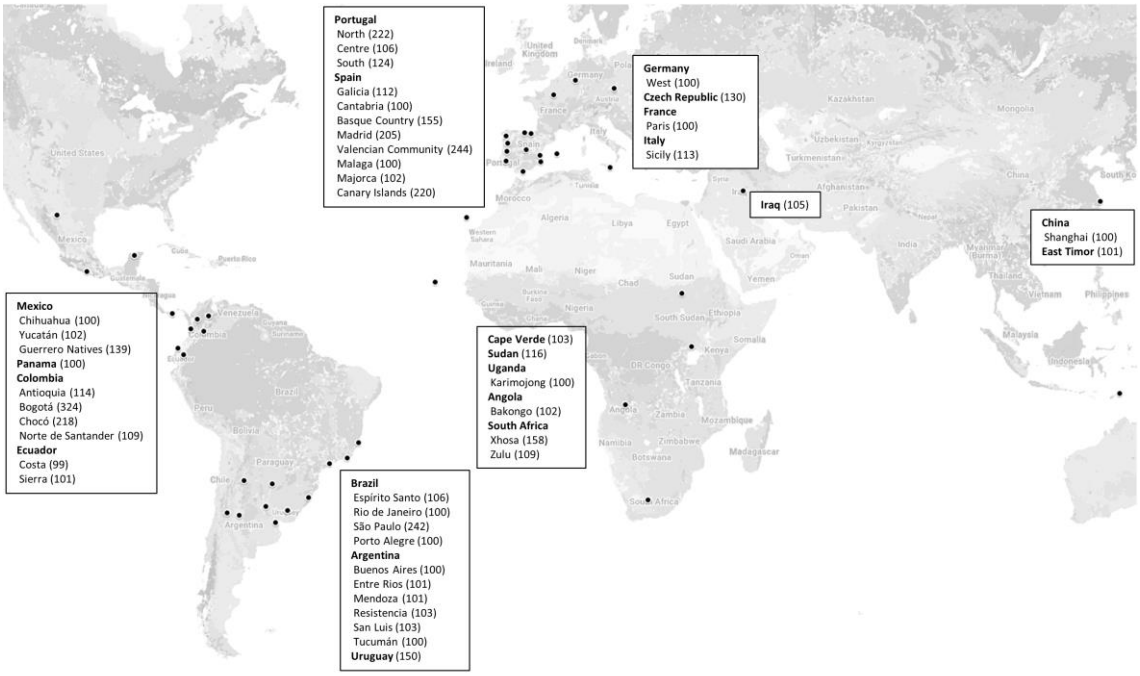

Figure 1: World map showing the 45 population samples studied in this work.

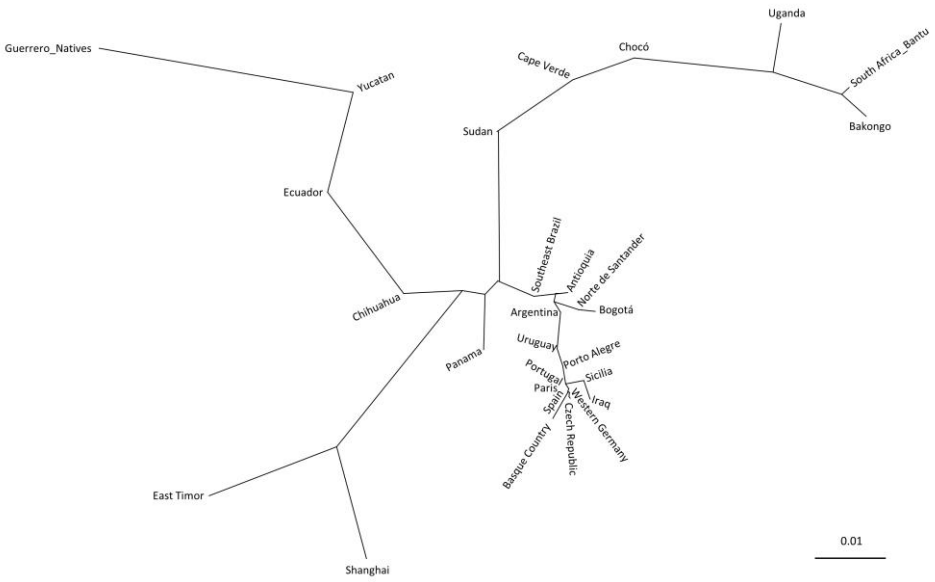

Figure 2: Neighbor-joining tree (unrooted) obtained from the matrix of pairwise F_{ST} genetic distances between 28 population samples studied in this work.

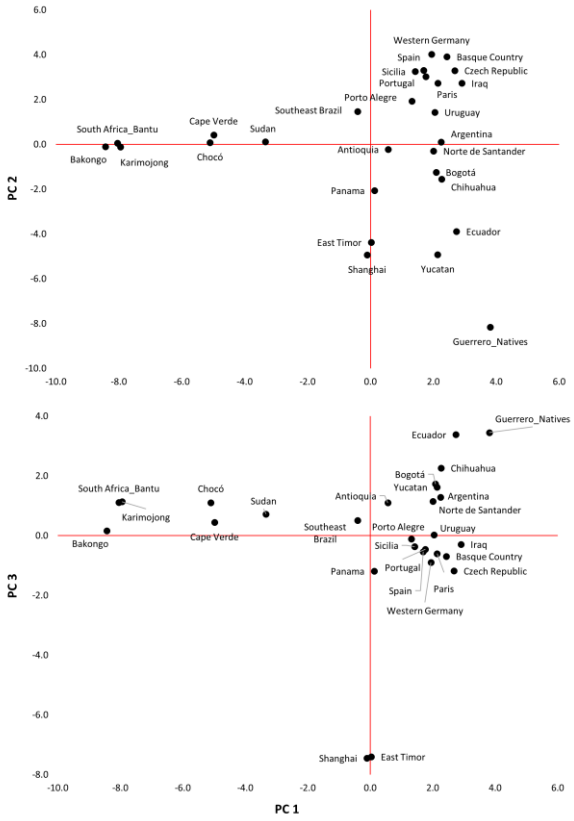

Figure 3: Principal Component Analysis (PCA) obtained from the allele frequency distributions estimated in 28 different population samples studied in this work. Upper plot represents the two Principal Components while the bellow plot details PC1 and PC3.