Forensic Science International: Genetics 10 (2014) 12–16

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Short Communication

SNP variation with latitude: Analysis of the SNP*for*ID 52-plex markers in north, mid-region and south Chilean populations

F. Moreno [a], A. Freire-Aradas [b], C. Phillips [b,*], M. Fondevila [b], Á. Carracedo [b,c], M.V. Lareu [b]

[a] Unidad de Genética Forense, Departamento Laboratorios, Servicio Médico Legal, Santiago de Chile, Chile
[b] Unidad de Genética Forense, Instituto de Ciencias Forenses Luis Concheiro, Grupo de Medicina Xenómica, Universidade de Santiago de Compostela, Spain
[c] Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 5 November 2013
Received in revised form 20 December 2013
Accepted 24 December 2013

Keywords:
SNPs
Chile
Human identification
Allele frequency databases
Population sub-structure

ABSTRACT

Chile is a disproportionately long and narrow country defined by the southern Andes and Pacific coastline where a level of genetic sub-structure resulting from distances of several thousand kilometers might be expected across the most distantly separated regions. Although STR databases created for the Chilean Legal Medical Service indicate an absence of sub-structure, such a characteristic requires further exploration when introducing additional forensic markers. Notably, Single Nucleotide Polymorphisms (SNPs) have a much lower mutation rate than STRs and can show more stable distributions of genetic variation if population movement is restricted. In this study we evaluated 451 Chilean urban samples from the North, North-Central, Central, South-Central and South regions of Chile for the 52 SNPs of the SNP*for*ID forensic identification panel to explore the underlying genetic structure of Chilean populations. Results reveal similar genetic distances between groups suggesting a single SNP database for the whole of Chile is appropriate. To further understand the genetic composition of Chilean populations that comprise the bulk of individuals with both European and Native American ancestries, ancestral membership proportions were evaluated and pairwise comparisons to other American populations were made.

1. Introduction

Chile is a country of over 17,000,000 inhabitants at the western extreme of South America with a characteristically extensive latitudinal distance, the longest in the world, running 4300 km north to south from 17° to 56° latitude. Despite this length, the major part of the Chilean population is restricted to the central regions of Metropolitan Santiago and its neighboring areas. Chile borders Peru to the north, Bolivia to the northeast and Argentina to the east, however the southern Andes form a strong natural barrier to any population movement other than north to south along the Pacific coastal plain.

Chile has recently established national databases for use of autosomal and Y STRs plus mtDNA HVI and HVII sequence data for routine forensic application of these markers [1–4]. However, in situations of complex kinship testing or handling highly degraded DNA (typically in the Pinochet dictatorship victim identification program), use of STRs alone provides inadequate data and has prompted assessment of short amplicon markers in Chile. Since Single Nucleotide Polymorphisms (SNPs) offer considerably reduced PCR amplicon lengths, low mutation rates and simple binary polymorphism statistics, they make ideal supplements to STRs for challenging identification cases. The SNP*for*ID panel of 52 autosomal SNPs, amplified in one PCR and two single base extension reactions [5], has been comprehensively validated for forensic purposes [6] and numerous global populations have now been characterized, including several American populations (genotypes and allele frequency data available at the SNP*for*ID browser [7]). In the present study, we analyzed a total of 451 Chilean individuals across the full latitudinal range from north to south for the above 52 SNPs to build a Chilean forensic SNP database and, at the same time, to gauge population sub-structure across the complete length of the country.

2. Materials and methods

2.1. Population samples

Buccal swabs were collected from 451 unrelated males in five different locations of: Iquique (N); Santiago (central); Concepción (central); Temuco (central) and Punta Arenas (S). All participants provided written informed consent for the collection of samples and subsequent analysis. Ethical approval was granted by the Ethical Committee of the Chilean Health Ministry. Subsequently, samples

* Corresponding author. Tel.: +34 981 583 015.
E-mail address: c.phillips@mac.com (C. Phillips).

1872-4973/$ – see front matter © 2014 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.12.009

were re-grouped, according to each donor's stated birthplace, into five geographic regions: (i) North $N = 97$ (Arica, Parinacota, Tarapacá, Antofagasta, Atacama, Coquimbo, Valparaíso); (ii) North-Central $N = 59$ (Santiago); (iii) Central $N = 38$ (O'Higgins, Maule); (iv) South-Central $N = 108$ (Biobío, La Araucanía) and (v) South $N = 149$ (Los Ríos, Los Lagos, Aysén del General Carlos Ibáñez del Campo, Magallanes). A map of sampling points and locations of re-grouped population clusters is shown in Supplementary Fig. S1. Additionally, SNP genotype data from a total of 1068 samples combined into three population groups (Africa, Europe and America) was collected from the SNP*for*ID browser [7] for the 52 SNPs.

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.12.009.

2.2. DNA extraction, quantification and SNP genotyping

DNA was extracted using standard phenol/chloroform methods and quantified by a semi-quantitative method using agarose gels and ethidium bromide. Samples were genotyped with the SNP*for*ID 52plex assay using the protocols described by Sanchez et al. [5]: based on a single 52plex PCR amplification followed by parallel 23/29plex single base extension (SBE) reactions using the SNaPshot™ primer extension kit (Applied Biosystems: AB).

2.3. Quality control

The reproducibility of the SNP*for*ID 52plex assay was previously tested in a number of inter-laboratory exercises within the SNP*for*ID consortium [6] and the European DNA Profiling Group [8]. Automated calls were made with AB GeneMapper® v3.2 and exported data was manually reviewed. This study followed the guidelines for the publication of genetic population data [9] as well as ISFG recommendations on analysis of DNA polymorphisms [10] for use of nomenclature, quality control and population statistics.

2.4. Statistical analysis of data

Hardy–Weinberg equilibrium (HWE) analysis, analysis of molecular variance (AMOVA), pairwise population F_{ST} values, and allele frequencies were calculated using *Arlequin* software v. 3.11 [11]. For HWE analysis, significance levels were corrected with Bonferroni's correction and P values below 0.00096 were considered statistically significant [12]. Forensic informativeness metrics, allele frequencies and observed heterozygosity were calculated using Promega PowerStats v. 1.2 [13]. The cumulative discrimination probability and exclusion power, combining 52 SNPs, were calculated using in-house Excel calculators, available on request. Allele frequency estimates for the 52 SNPs in the study populations are available online at the SNP*for*ID frequency browser [7]. STRUCTURE software v. 2.3.3 [14] was used to analyze population characteristics applying admixture and correlated frequencies models [15]. Runs comprised 100,000 Markov Chain Monte Carlo steps after a burn-in length of 100,000. Five independent replicates were performed for each population cluster value. Graphics were constructed using *Clumpp* v1.1.2 [16] and *Distruct* v1.1 software [17]. Principal Component Analysis (PCA) was made using R software v 2.15.0 with the *SNPassoc* statistical package.

The ancestral composition of individuals from the five Chilean regions was estimated by comparison with reference samples from the three main contributing ancestral groups relevant to South America: European, Native American and African using supervised STRUCTURE runs by assigning POPFLAG = 1 to this data. Separate STRUCTURE analyses of Chilean populations alone, to assess sub-structure, were unsupervised. Although the SNPs from the SNP*for*ID 52plex were designed primarily for human identification, a degree of ancestry information can be inferred from their analysis. Genetic data for the above three reference populations are available from the SNP*for*ID browser [7], although Somalis and African Americans, previously indicated to be admixed [18,19], were not included in the African reference data used.

3. Results

3.1. Analysis of genetic structure

In order to evaluate any underlying genetic structure in the five re-grouped regions of the Chilean population (North, North-Central, Central, South-Central and South), results from a STRUCTURE analysis were assessed for indications of population sub-structure. When analysing Chilean populations alone and without using ancestry reference data, the results revealed a homogenous pattern for the five Chilean groups, evaluating different population-clustering models from $K = 2$ to $K = 5$. Cluster plots from these runs are shown in Supplementary Fig. S2. Results indicate a complete absence of genetic sub-structure across the studied regions of Chile for the SNPs genotyped.

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.12.009.

A separate STRUCTURE analysis then compared the five Chilean regions with reference data Europeans, Native Americans and Africans. The optimum number of clusters were considered to be K:3 from the $X|K$ plot (Fig. 1, lower left) and an absence of clearly

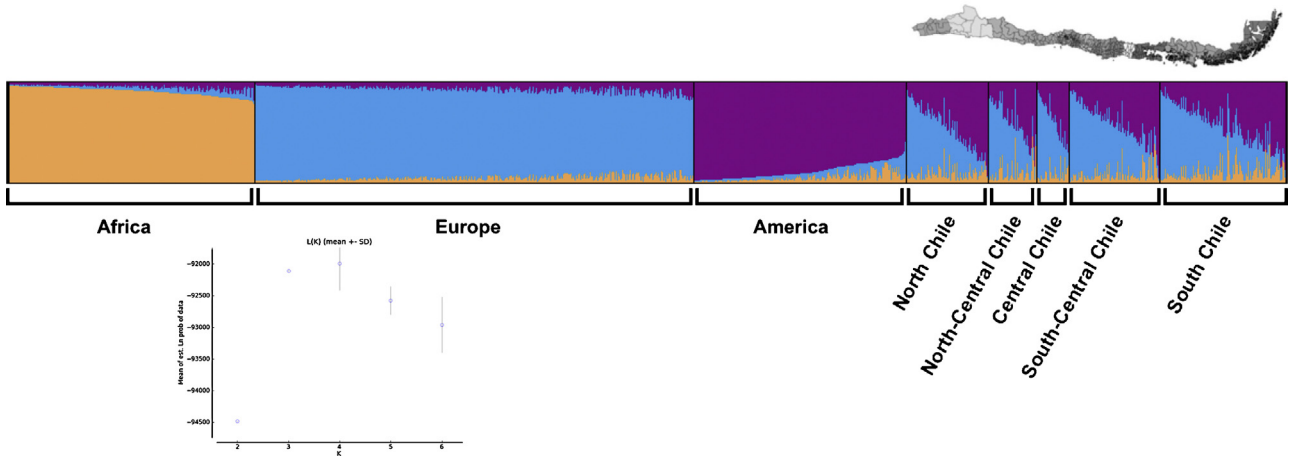

Fig. 1. STRUCTURE cluster plot for optimum K:3 (second point from left on L(K) plot lower left). Chilean population samples arranged in order of descending European co-ancestry proportions within each region.

differentiated cluster patterns at K:4 onwards (data not shown). Fig. 1 shows the corresponding STRUCTURE cluster plot for this ancestry analysis of Chileans applying a three-population model ($K = 3$), with the five tested Chile populations on the right. The cluster membership proportions are listed in Supplementary Table S1, along with summary boxplots showing their ranges in each region. Variability in ancestry proportions is high between individuals but with a common pattern revealed by the boxplots of Supplementary Table S1 of comparable European and Native American co-ancestry contributions in all five regions. Boxplots indicate a slightly higher proportion of European co-ancestry in all but the North region, but in contrast, African ancestry contributions are a much smaller minor component of co-ancestry except for a handful of individuals. This largely fits with the accepted demographic history of Chile, where a series of European immigrations predominated in the settlement of most of the country and African slavery was considerably less established than in nearly all other South American regions.

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.12.009.

3.2. Population divergences assessed with principal component analysis

PCA results analyzing 48 populations are shown in the two-dimensional principal components plot of Fig. 2A. The same reference-population Africans, Europeans and Americans previously used for STRUCTURE analysis are plotted and although the 52 identification SNPs have relatively low population heterogeneity, differences among populations are discernable. The first principal component (PC1) is 11.39% of variation and separates Africa, while PC2 = 7.11% of variation and separates Europe from America. Chilean populations are seen to mainly cluster between Europeans and Americans. Fig. 2B shows a modified PCA with Africans removed and this again indicates Europe and America are well separated (PC1: 9.04%), while Chilean individuals remain located in the middle of these clusters. In this PCA the second component (PC2: 3.05%) shows a separation of Greenlanders from other American populations. This separation was also observed in the three-dimensional plot of the first PCA (Supplementary Fig. S3). Finally, Fig. 2C shows PCA of Chilean populations alone resulting in a single homogenous cluster where all the individuals are positioned randomly and independently of their region of origin. This further underlines an absence of detectable population substructure across the full latitudinal range of Chile.

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.12.009.

3.3. Analysis of molecular variance and genetic distances

An initial AMOVA analysis of the five Chilean populations indicated that as much as 99.87% of the variation comprises the within-population component, so genetic differences between populations is almost beyond detection at just 0.13%. Subsequent genetic distance analyses took the five Chilean populations as a whole and derived pairwise F_{ST} values comparing Chileans to Africans, Europeans and Americans. The pairwise F_{ST} values in Table 1 show that the most distant population group to Chileans is the African group (F_{ST}: 0.14630) with reduced distances for Europe (F_{ST}: 0.03791) and America (F_{ST}: 0.04731). When specific American populations are compared Colombia and Argentina are closest (F_{ST}: 0.00552 and 0.00616 respectively) while the isolated Brazil

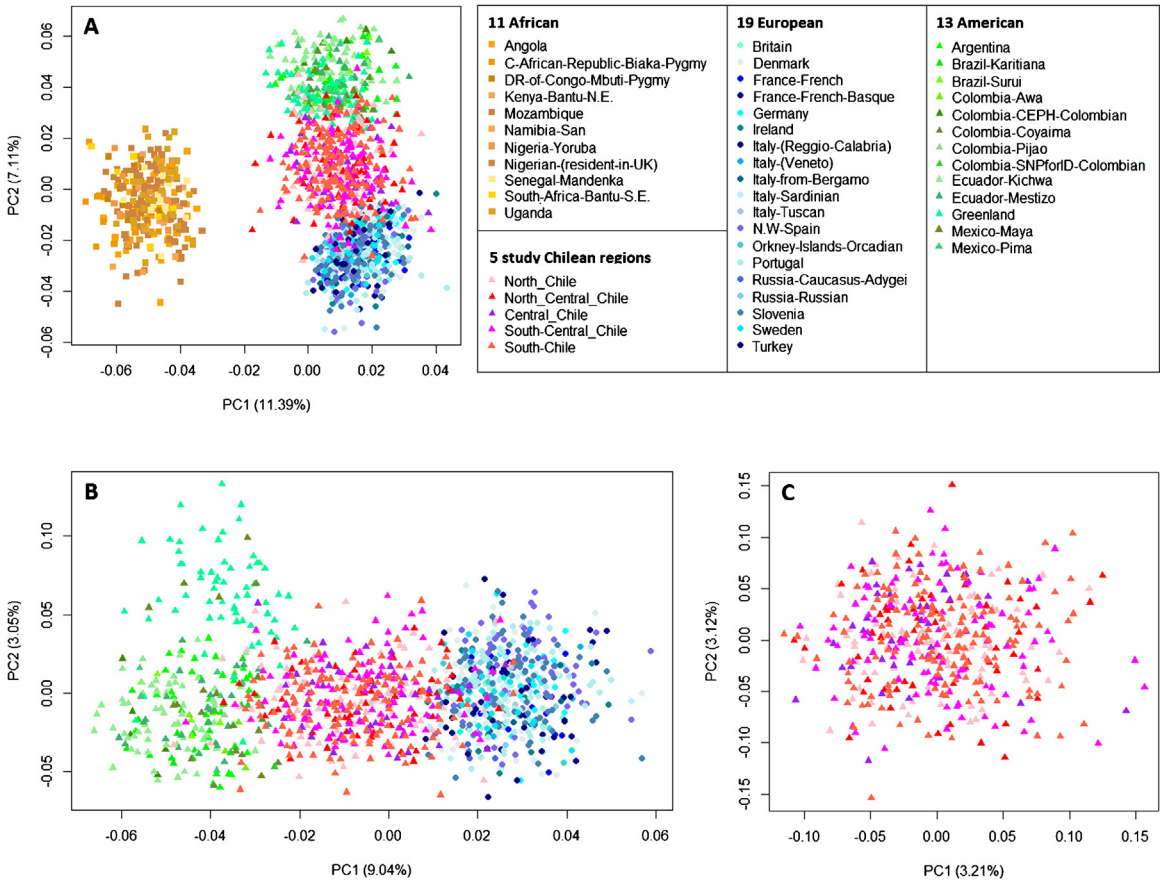

Fig. 2. Principal component analyses of study and reference populations. (A) African, European and American populations compared with five Chilean study populations. (B) With African populations removed. (C) Chilean populations alone.

Table 1
Between-population pairwise F_{ST} estimates for combined Chilean populations (all Chile) vs. three population groups or 15 American populations. All F_{ST} P values were below 0.000005 under the permutation test (10,000) with the exception of two pairwise comparisons: Chile vs. Colombian Pijao: $P = 0.0033$ and Chile vs. Colombian SNP*for*ID: $P = 0.0073$.

All Chile	Pairwise F_{ST}
Africa	0.14630
Europe	0.03791
Native America	0.04731
Argentina	0.00616
Brazil-Karitiana	0.09082
Brazil-Surui	0.07614
Colombia-Awa	0.07809
Colombia-CEPH	0.06762
Colombia-Coyaima	0.08473
Colombia-Embera	0.07327
Colombia-Mulalo	0.05505
Colombia-Pijao	0.04194
Colombia-SNP*for*ID	0.00552
Ecuador-Kichwa	0.06434
Ecuador-Mestizo	0.02898
Greenland	0.04763
Mexico-Maya	0.02178
Mexico-Pima	0.02788

Table 2
Forensic informativeness metrics.

	Cumulative random match probability	Combined discrimination probability	Cumulative exclusion probability (%)
Africa	4.8E−17	2.1E+16	99.71
Europe	3.9E−21	2.6E+20	99.98
East Asia	9.5E−19	1.1E+18	99.91
America	2.5E−19	4.0E+18	99.93
Chile	4.9E−20	3.1E+19	99.99

Amazonian Karitiana population is, unsurprisingly, the most distant Native American population.

3.4. Evaluation of forensic parameters

Examining HWE after Bonferroni correction (P-value <0.00096), all SNPs were in equilibrium except rs907100 and rs1528460: in both North and South regions, as well as analyzing all individuals together. Allele frequencies are outlined in Supplementary Table S2 and forensic informativeness metrics: random match probability (RMP), discrimination probability (Dp), exclusion probability (PE), observed heterozygosity and allele number are listed in Supplementary Table S3. Table 2 summarizes overall forensic metrics for Chile and includes equivalent values for African, European and East Asian groups.

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.12.009.

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.12.009.

4. Discussion

Although some level of population sub-structure could be expected within such a long and geographically restricted region, we found no evidence for its presence across Chile. Recently introduced STR databases that also extensively sampled population variation across Chile further indicate that the long distances from north to south have not produced any detectable levels of sub-structure. It can be argued that 52 SNPs and 15 STRs together do not represent the most powerful tools for the detection of population sub-structure but the object of this study was to confirm that a single database of SNP variation measured in Chilean populations will be sufficient to compile allele frequency data without necessary regard for differences in allele frequency distributions across Chilean regions. Furthermore, it should be noted that sampling of urban populations will overlook sub-structure where it exists between non-urban populations, but this represents a very small fraction of the Chilean demographic profile. Therefore, the present study has provided consistent results indicating absence of genetic stratification. This is shown most conclusively by principal component analysis of Chilean individuals creating a single heterogenous cluster and by the level of within-population variation accounting for 99.87% of the total 52-SNP variation in Chile. We found the smallest genetic distance was with the Colombian and Argentinian populations, underlining the dominance of European co-ancestry in the demographic history of all regions of Chile sampled, with Native American components forming slightly smaller proportion of co-ancestry amongst the individuals analyzed with STRUCTURE. The cluster plots from this analysis also show the range of European-Native American co-ancestry proportions is remarkably consistent in all five Chilean regions analyzed.

From the technical point of view, SNPs rs907100 and rs1528460 give peak imbalance that may point to uncharted SNPs in primer binding sites and with the deviations from HWE detected both SNPs are currently excluded from SNP profiles used in Chile until they can be further investigated. Introduction of degenerate primers can likely address the high peak imbalance and possible presence of null alleles in both markers.

The forensic informativeness metrics estimated from Chilean data indicates the fifty remaining SNPs will provide sufficiently informative data for routine forensic identification purposes. As a result of the reported Chilean population variability findings, a single allele frequency database for the 52 SNP*for*ID markers has been established that combines the five geographical regions of Chile as a whole, matching the framework for the STR database. This step accelerates the introduction of SNP analysis appropriate for identification of victims of the former dictatorship in Chile, or indeed any forensic application where it is necessary for Chilean laboratories to analyze highly degraded DNA or add extra genetic data to complex relationship testing.

Acknowledgements

MVL was supported by funding from Xunta de Galicia INCITE 09 208163PR. AFA was supported by a grant from Deputación da Coruña and MF by a fellowship from Fundación Pedro Barrié de la Maza.

References

[1] H. Jorquera, B. Budowle, Chilean population data on ten PCR-based loci, J. Forensic Sci. 43 (1998) 171–173.
[2] L. Cifuentes, M. Acuña, D. Sepúlveda, Allele frequencies for six STRs in a Chilean population, J. Forensic Sci. 47 (2002) 904–905.
[3] R.K. Roby, S.D. Gonzalez, N.R. Phillips, J.V. Planz, J.L. Thomas, J.A. Pantoja Astudillo, J. Ge, E. Aguirre Morales, A.J. Eisenberg, R. Chakraborty, P. Bustos, B. Budowle, Autosomal STR allele frequencies, Y-STR and mtDNA haplotypes in Chilean sample populations, Forensic Sci. Int. Genet. Suppl. Ser. 2 (2009) 532–533.
[4] I.A. Vergara, P. Villouta, S. Herrera, F. Melo, Autosomal STR allele frequencies for the CODIS system from a large random population sample in Chile, Forensic Sci. Int. Genet. 6 (2012) e83–e85.
[5] J.J. Sanchez, C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, et al., A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006) 1713–1724.
[6] E. Musgrave-Brown, D. Ballard, K. Balogh, K. Bender, B. Berger, M. Bogus, C. Børsting, M. Brion, et al., Forensic validation of the SNP*for*ID 52-plex assay, Forensic Sci. Int. Genet. 1 (2007) 186–190.
[7] http://spsmart.cesga.es/snpforid.php.

[8] J.J. Sanchez, C. Børsting, K. Balogh, B. Berger, M. Bogus, J.M. Butler, Á. Carracedo, D.S. Court, et al., Forensic typing of autosomal SNPs with a 29 SNP-multiplex-results of a collaborative EDNAP exercise, Forensic Sci. Int. Genet. 2 (2008) 176–183.
[9] Á. Carracedo, J.M. Butler, L. Gusmão, A. Linacre, W. Parson, L. Roewer, P.M. Schneider, New guidelines for the publication of genetic population data, Forensic Sci. Int. Genet. 7 (2013) 217–220.
[10] P.M. Schneider, Scientific standards for studies in forensic genetics, Forensic Sci. Int. 165 (2007) 238–243.
[11] L. Excoffier, G. Laval, S. Schneider, Arlequin (version 3.0): an integrated software package for population genetics data analysis, Evol. Bioinform. Online 1 (2007) 47–50.
[12] S.W. Guo, E.A. Thompson, Performing the exact test of Hardy–Weinberg proportion for multiple alleles, Biometrics 48 (1992) 361–372.
[13] A. Tereba, Tools for analysis of population statistics, Profiles DNA 9 (1999) 14–16.
[14] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.
[15] L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, A. Carracedo, M.V. Lareu, An overview of STRUCTURE: applications, parameter settings, and supporting software, Front. Genet. 4 (2013) 98.
[16] M. Jakobsson, N.A. Rosenberg, CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure, Bioinformatics 23 (2007) 1801–1806.
[17] N.A. Rosenberg, DISTRUCT: a program for the graphical display of population structure, Mol. Ecol. Notes 4 (2004) 137–138.
[18] Y.Y. Ruiz, M.A. Chiurillo, L. Borjas, C. Phillips, M.V. Lareu, Á. Carracedo, Analysis of the SNP*for*ID 52-plex markers in four Native American populations from Venezuela, Forensic Sci. Int. Genet. 6 (2012) e142–e145.
[19] A. Ibarra, A. Freire-Aradas, M. Martínez, M. Fondevila, G. Burgos, M. Camacho, H. Ostos, Z. Suarez, et al., Comparison of the genetic background of different Colombian populations using the SNP*for*ID 52plex identification panel, Int. J. Legal Med. (2013), http://dx.doi.org/10.1007/s00414-013-0850.