Research paper

Analysis of ancestry informative markers in three main ethnic groups from Ecuador supports a trihybrid origin of Ecuadorians

Roberta Santangelo[a,b,1], Fabricio González-Andrade[c,1], Claus Børsting[a], Antonio Torroni[b], Vania Pereira[a,*], Niels Morling[a]

[a] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[b] Dipartimento di Biologia e Biotecnologie "L. Spallanzani", Università di Pavia, Pavia, Italy
[c] Faculty of Medical Sciences, Central University of Ecuador, Quito, Ecuador

ABSTRACT

Ancestry inference is traditionally done using autosomal SNPs that present great allele frequency differences among populations from different geographic regions. These ancestry informative markers (AIMs) are useful for determining the most likely biogeographic ancestry or population of origin of an individual. Due to the growing interest in AIMs and their applicability in different fields, commercial companies have started to develop AIM multiplexes targeted for Massive Parallel Sequencing platforms.

This project focused on the study of three main ethnic groups from Ecuador (Kichwa, Mestizo, and Afro-Ecuadorian) using the Precision ID Ancestry panel (Thermo Fisher Scientific). In total, 162 Ecuadorian individuals were investigated. The Afro-Ecuadorian and Mestizo showed higher average genetic diversities compared to the Kichwa. These results are consistent with the highly admixed nature of the first two groups. The Kichwa showed the highest proportion of Native Amerindian (NAM) ancestry relative to the other two groups. The Mestizo had an admixed ancestry of NAM and European with a larger European component, whereas the Afro-Ecuadorian were highly admixed presenting proportions of African, Native Amerindian, and European ancestries. The comparison of our results with previous studies based on uniparental markers (i.e. Y chromosome and mtDNA) highlighted the sex-biased admixture process in the Ecuadorian Mestizo.

Overall, the data generated in this work represent one important step to assess the application of ancestry inference in admixed populations in a forensic context.

1. Introduction

The distribution of human genetic diversity has been a topic of major interest in the clinical field as well as in population and forensic genetics [1]. Genetic ancestry is inferred by comparing a sample's genetic diversity with the patterns of variation in contemporary populations [2]. The study of marker systems with different characteristics often reflects somewhat different information about population history and individual ancestry [2,3]. Lineage markers (mitochondrial DNA and Y chromosome) provide information about the maternal and paternal lineages of individuals, respectively, and can be used to study genetic events that were differentially mediated by men and women. Due to recombination, autosomal markers can provide additional information about the admixed nature of an individual.

The most common Ancestry Informative Markers (AIMs) are autosomal SNP markers that present marked allele frequency differences among populations from different geographic regions. AIMs are useful for determining the likely biogeographic ancestry or population of origin of an individual. This is particularly useful in forensic genetics, when the source individual is not known or is unable to declare his or her ancestry [4]. Due to continuous human migrations, AIM alleles are shared across all human groups. Therefore, it is not the absolute presence/absence of an allele, rather its frequency in the population that is usually analysed when inferring ancestry [4].

This study focused on the Ecuadorian population. Ecuador is located in western South America between Colombia and Peru, and bordered by the Pacific Ocean at the Equator. The country has 24 provinces distributed in 4 main regions: the Highlands (Sierra), Coast, Amazonia, and the insular region (Galapagos Islands). It has a population of 16.5 million inhabitants (April 2017), and an annual population growth rate

---

* Corresponding author at: Section of Forensic Genetics, Department of Forensic Medicine Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V's Vej 11, DK-2100, Copenhagen, Denmark.
E-mail address: Vania.pereira@sund.ku.dk (V. Pereira).
[1] These authors contributed equally to this work.

of 1.31% (2016 estimate). Similarly to other populations in South America, the Ecuadorian population is multicultural and multiethnic with a complex demographic history. The country has experienced multiple migration and admixture events in pre- and post-Columbian times, including Native Amerindian settlements, European colonization, and the import of African slaves [5]. The continuous admixture among European, Native Amerindian, and African individuals shaped the patterns of diversity in the modern Ecuadorian population, making it a very interesting case study from the population genetics point of view.

Currently, Ecuador comprises three main ethnicities:

1) Mestizo – the urban population found in all cities. The Mestizo comprises almost 85% of the Ecuadorian population, and is a Spanish-speaking admixed group with European and Native Amerindian (NAM) ancestry.

2) Native Amerindian – includes several multi-ethnic and multi-cultural groups that live in the Highlands, representing around 7% of the total Ecuadorian population [6]. Native Amerindian comprises 14 Indigenous nationalities scattered throughout the country: Kichwa, Waorani, Secoya, Siona, A'I Cofan, Shuar, Achuar, Shiwiar, Zapara, Andoans, Chachi, Awà, Tsa'Chila and Epera [6]. All native Amerindians speak two to four languages, including Spanish. The most numerous group is the Kichwa that represents around 92% of the Native Amerindian population.

3) Afro-Ecuadorian – individuals that are descendants of African slaves [7]. Afro-Ecuadorian inhabit specific regions in the North of the country (mainly the provinces of Esmeraldas and Imbabura), and along the Ecuadorian coastline. They represent almost 7% of the Ecuadorian population.

The present work focused on the genetic diversity and ancestry of the three main groups of Ecuador and was performed with the Precision ID Ancestry Panel and the Ion Torrent PGM™ platform (Thermo Fisher Scientific, Waltham, USA). The results produced in this study help to characterize the ability of this forensically-relevant AIM panel in determining ancestry in highly-admixed groups.

2. Materials and methods

2.1. Samples and DNA quantification

A total of 162 unrelated individuals were analysed in this study (29 Afro-Ecuadorian, 66 Kichwa, and 67 Mestizo). Samples were collected by the Faculty of Medical Sciences of Central University of Ecuador, and by the Laboratory of Human Genetics, Department of Biology and Biotechnology 'L. Spallanzani', University of Pavia, Italy. Informed Consent was signed by every sample donor, and all samples were anonymised. The study was approved by the ethical committee of the Capital Region of Denmark (H-1-2011-081) and complies with the ethical principles of the 2000 Helsinki Declaration of the 2006 World Medical Association [8]. Samples were only used for research purposes.

DNA was extracted from blood on FTA cards (Whatman Inc., Clifton, NJ) with a BioRobot EZ1 Workstation (Qiagen, Hilden, Germany) using the manufacturer's recommendations. The samples from the University of Pavia were extracted from blood or buccal swabs with a standard phenol/chloroform method. DNA extracts were quantified using the Qubit® dsDNA High Sensitivity (HS) assay kit and the Qubit® Fluorometer (Thermo Fisher Scientific).

2.2. Library building and sequencing

DNA was amplified with the Precision ID Ancestry Panel following the manufacturer's protocol. The panel includes 165 autosomal markers that combine a 55 AIM-assay [9] and a 123-markers assay [10]. There are 13 overlapping SNPs between the two sets.

Sample libraries were constructed using the Ion Ampliseq™ Library Kit 2.0 (Thermo Fisher Scientific) according to the manufacturer's

recommendations. The template preparation was performed with the Ion Chef™ instrument (Thermo Fisher Scientific) and the Ion PGM™ IC 200 Kit (Thermo Fisher Scientific). Sequencing was carried out on the Ion PGM™ instrument (Thermo Fisher Scientific) using the Ion PGM™ Sequencing 200 Kit v2 reagents and Ion 318™ chips v2 (Thermo Fisher Scientific). Twenty-five libraries were loaded in each chip.

2.3. Data analyses

The primary sequencing analysis of DAT files was performed on the Torrent Suite Software v.4.6 (Thermo Fisher Scientific). BAM files were generated using the HID_SNP Genotyper v.4.3.1 plug-in (Thermo Fisher Scientific). In the current work, the analyses were carried out using the following criteria: minimum coverage per marker $\geq$ 45 reads, and $0.3 \leq$ Hb $\leq$ 3. No criterion for noise was used. All genotypes that did not fulfil the two criteria were manually inspected and a decision was made whether to accept the result or not. Furthermore, only samples with $\leq$ 16 locus drop-outs were used for further analyses.

To evaluate the performance of the Precision ID Ancestry Panel assay on the Ecuadorian samples, the locus balance (Lb), heterozygote balance (Hb), and noise level were calculated. The Lb was calculated as the total number of reads for a locus divided by the average number of reads per locus per sample. The Hb was calculated as the number of reads for one nucleotide divided by the number of reads for the other nucleotide in the following alphabetical order A, C, G, and T. The noise was calculated as the number of reads that were different from the genotype call divided by the total number of reads per locus.

Allele frequencies, Hardy-Weinberg Equilibrium (HWE), and genetic diversities were calculated using the Arlequin v3.5.2.2 software [11]. HWE analysis was carried out using 1,000,000 Markov Chain Monte Carlo (MCMC) steps and 1,000,000 dememorization steps. Correction for multiple testing was done according to Bonferroni [12].

The distribution of genetic ancestry in each sample was investigated using the STRUCTURE v.2.3.4.21 software. Analyses were carried using 100,000 steps of burn-in followed by 100,000 repetitions for the MCMC. The 'admixture' and the 'correlated allele frequencies' models were considered [13,14]. Three assumed populations or clusters (K) were considered in the analyses and three independent runs were performed in order to verify the consistency of the results. The results of STRUCTURE were visualized using CLUMPP v.1.1.222 [15] and Distruct v.1.1.23 software [16].

Principal Component Analyses (PCA) of the sample genotypes were carried out using an in-house script written in Python. Reference population data were kindly provided by Professor Kenneth Kidd (Supplementary Table S1). The SNP rs10954737 was not considered in the PCA and STRUCTURE analyses due to the lack of genotype data in the reference populations.

3. Results and discussion

3.1. Evaluation of the precision ID ancestry panel performance

The threshold criteria applied in this work helped to identify the loci where the predicted genotype calls were less certain, and to guarantee that only reliable data were used in the subsequent analyses.

The Lb varied between markers, with median values per marker ranging from 0.07 to 2.70 (Supplementary Fig. S1). The majority of the markers had median Lb values close to 1.0. Some of the more imbalanced loci (e.g. rs1296819, rs12439433), were also the ones with the lowest coverages. Regarding the Hb, most of the loci were well balanced with median values close to 1.0. The median values of Hb ranged from 0.47 for rs10007810 to 1.62 for rs6990312 (Supplementary Fig. S2). For most of the loci, the median noise level was below 1%, and outliers were below 6%. However, one locus, rs7722456, had high levels of noise. The median noise level was 6%, and up to 15% of the reads were noise in some individuals

(Supplementary Fig. S3). The results are in agreement with those previously observed in a dataset of Danes and Somalis [17].

Besides calling the alleles for the Precision ID Ancestry Panel, the HID_SNP Genotyper v.4.3.1 plug-in can also provide additional information on the level of admixture of an individual and calculate the population likelihood of a genotype.

In this work, these two features have not been explored. Based on several studies carried out by our group with samples from different genetic histories, it is our opinion that the HID-SNP Genotyper plug-in still has some limitations that need to be addressed before it can be used as a tool when investigating ancestry in a forensic context. For now, it does not allow the users to integrate their own genetic data, which can considerably limit ancestry inference of admixed samples. A more thorough discussion of the plug-in and its limitations can be found at [17].

3.2. Intrapopulation analyses

The allele frequencies for the 165 AIMs in the Kichwa, Mestizo, and Afro-Ecuadorian are presented in Supplementary Table S2. Three loci were monomorphic in the three groups – rs1800414, rs3811801, and rs671. Besides these, two other loci (rs1229984 and rs3916235) were monomorphic in the Kichwa, and one (rs2196051) in the Afro-Ecuadorian. No HWE analysis was performed on monomorphic markers. The three markers that were monomorphic in all three groups were also monomorphic for the same allele in Danes, Somalis [17], and Greenlanders [18]. Further inquiries at the Ensembl genome browser http://www.ensembl.org/index.html (release 85) [19] showed that these three markers are fixed in all the European, Native Amerindian, and African samples reported to date, but they are polymorphic in East Asian populations.

HWE deviation was assessed for the remaining polymorphic markers. After correction for multiple testing (*p*-value after Bonferroni correction: 0.0003) only rs3943253 presented a statistically significant *p*-value in the Kichwa population (p = 0.0002).

Average genetic diversities in the three groups were calculated at three different levels considering: 1) all the 165 markers of the Precision ID Ancestry Panel; 2) the 55 AIM panel reported in [8]; 3) the set of 123 AIMs [10] (Table 1).

The Afro-Ecuadorian and Mestizo consistently showed higher average genetic diversity relative to the Kichwa which is consistent with admixed origin of the first two groups (that will be described in more detail in the next section).

3.3. Interpopulation analyses

Ancestry inference analyses were carried out using the STRUCTURE v.2.3.4.21 software (Fig. 1). Supplemental Fig. S4 shows a triangle plot for the same analysis. The triangle plot is an alternative way to represent the estimated membership coefficients of the analysed individuals in STRUCTURE. More information on the proportion of membership of each pre-defined population in each of the three clusters is presented in Supplementary Table S3.

The Kichwa had a high proportion of NAM ancestry (91.5% in

Table 1
Average genetic diversity in Kichwa, Mestizo, and Afro-Ecuadorian populations considering different sets of loci.

Sets of loci	Kichwa (N = 66)	Mestizo (N = 67)	Afro-Ecuadorian (N = 29)
165 loci	0.306 ± 0.148	0.386 ± 0.186	0.385 ± 0.187
55 loci (Kidd panel)	0.214 ± 0.109	0.302 ± 0.150	0.335 ± 0.168
123 loci (Seldin panel)	0.340 ± 0.164	0.421 ± 0.203	0.410 ± 0.200

average) and about 7% of European ancestry. Within the Kichwa, the majority of the individuals had ancestry proportions close to the average. A few individuals had a lower NAM component, and a European component ranging from 10% to 30%. The Mestizo had on average 65.8% of NAM ancestry, 30.1% of European ancestry, and 4% of African ancestry. A few individuals had ancestry compositions with NAM proportion of around 85%, which was comparable to the values found among Kichwa. Furthermore, some Mestizo presented a higher proportion of European ancestry (around 45%). The Afro-Ecuadorian had all three ancestry components. The African component was the most represented one (58.8%), followed by NAM (28%) and European (13%).

Given that the NAM reference populations (namely PEL) have an important EUR component, it is important to ascertain if the results may reflect a possible overestimation of the European component in our samples. For this reason, STRUCTURE analyses were also carried out excluding the Peruvian population. The admixture levels present in our samples were comparable to the previous results, indicating that the apportionment of ancestry was unbiased towards European ancestry (data not shown).

The results are consistent with what is known about the history of these groups: the Mestizo represent an admixed population that descends from Europeans (mostly Spanish) and Native Amerindians [7]; the Afro-Ecuadorian are considered a highly admixed trihybrid group with descendants from Native Amerindians, Europeans, and African slaves brought to Ecuador during the Spanish colonization [20].

To get a more visual and intuitive interpretation of the ancestry of each sample, PCA of the individual genotypes was carried out (Fig. 2). Overall, the PCA result was in agreement with the pattern observed in the STRUCTURE analysis. The combined percentage of the variance of the data set explained by the first two PCs was 52.4% (28.2% + 24.2%). The Kichwa were located closer to the NAM group, while Mestizo had an intermediate position between NAM and European populations. Afro-Ecuadorian were located between NAM and African populations.

A previous study based on Y-chromosomal STRs [7] showed that about 70% of the Y-chromosomal lineages in the Mestizo from Ecuador are of European origin. Preliminary results from the University of Pavia based on sequencing of entire mitogenomes from Mestizo individuals showed that all founding pan-American mtDNA haplogroups (A2, B2, C1, D1, and D4h3a) are present in agreement with a published study based on the sequencing of the mtDNA control region [21].

The apportionment of ancestry in Afro-Ecuadorian detected in this work is in line with the levels reported for the Y chromosome [7], where around 44% of the lineages had an African origin, 31% were Native Amerindian, and 15% European. More studies are needed on mtDNA lineages in order to ascertain the full extent of sex-biased admixture in this group.

4. Conclusions

The present work focused on the study of biogeographic ancestry of Ecuadorian populations using the Precision ID Ancestry panel. Three main ethnic groups − Kichwa, Mestizo, and Afro-Ecuadorian − were investigated. The genetic characterisation of these groups supports previous studies of the demography and genetics of Ecuador [5,7,21].

From the anthropological point of view it is important to understand the genetic origin of the Ecuadorian ethnic groups. This provides a possible explanation for the current social and cultural differences that mark current major socioeconomic differences. As shown here, the three main ethnic groups are trihybrid populations with a significant Native Amerindian genetic load. This could explain the variety of faces, features and skin colors found in present-day Ecuadorians. In addition, it is important to note that the current populations are a product of 525 years of admixture and various migratory waves to and from the country. Admixture represents the essence of the current Ecuadorians,
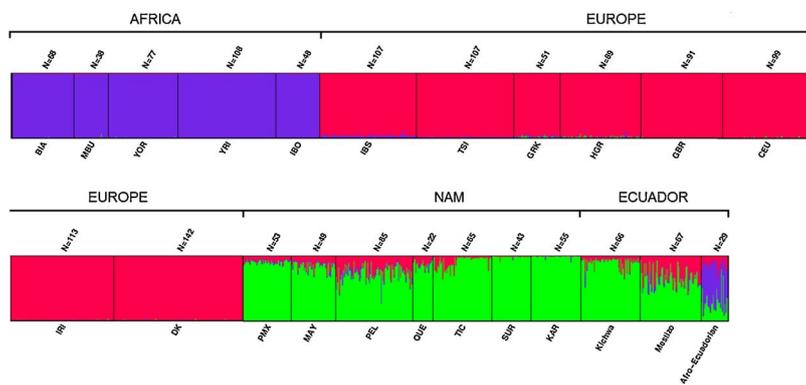
Fig. 1. STRUCTURE bar plot for K = 3 using 164 AIM SNPs. The SNP rs10954737 was not considered due to the lack of genotype data in the reference populations. Each individual is represented by a vertical line partitioned into up to K coloured segments, whose lengths are proportional to the estimated membership in each of the inferred K groups. Admixture and correlated allele frequency models were taken into account.
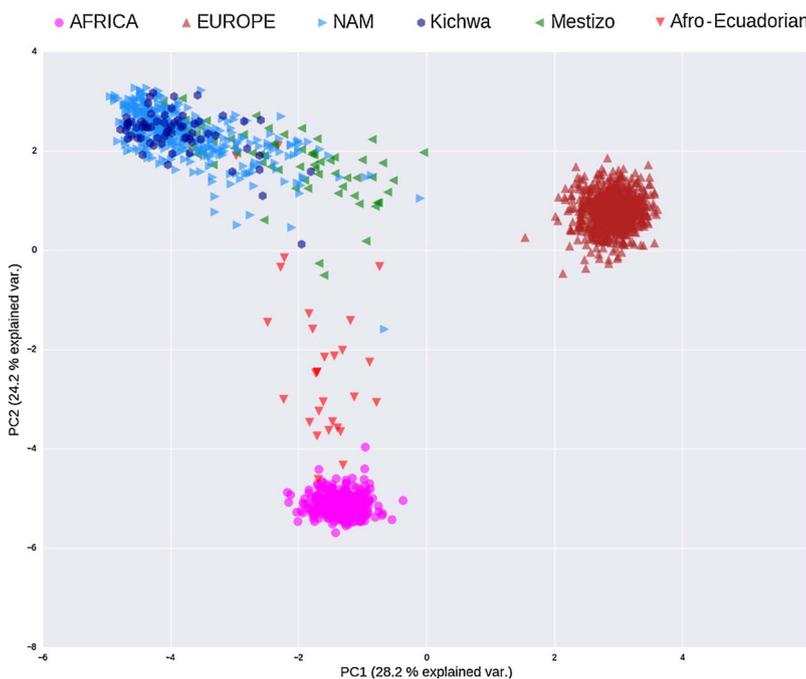
Fig. 2. PCA of Kichwa, Mestizo and Afro-Ecuadorian analysed with 164 AIM SNPs. The SNP rs10954737 was not considered due to the lack of genotype data in the reference populations. Each individual is represented as a coloured point according to the population of origin.

as reported through the study of molecular markers [7,21].

In the forensic field, the interest in ancestry inference is increasing with commercial companies providing standardized AIM panels. Before these panels can be effectively applied in casework, data for reference populations are needed together with a better understanding of how these markers perform in admixed populations. Ancestry inference is done by comparing the data generated to a group of known reference datasets. Therefore, high quality data and well-characterized and appropriate population reference databases are of extreme importance when assessing ancestry.

Regarding the Ecuadorian population, and South American populations in general, the Precision ID Ancestry Panel could still be improved by testing a greater amount of individuals as well as incorporating other ethnic groups.

Acnkowlegements

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2017.08.012.

References

[1] L.B. Jorde, W.S. Watkins, M.J. Bamshad, M.E. Dixon, C.E. Ricker, M.T. Seielstad, M.A. Batzer, The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data, Am. J. Hum. Genet. 66 (3) (2000) 979–988.

[2] C.D. Royal, J. Novembre, S.M. Fullerton, D.B. Goldstein, J.C. Long, M.J. Bamshad, A.G. Clark, Inferring genetic ancestry: opportunities, challenges, and implications, Am. J. Hum. Genet. 86 (5) (2010) 661–673.

[3] C.L. Pfaff, J. Barnholtz-Sloan, J.K. Wagner, J.C. Long, Information on ancestry from genetic markers, Genet. Epidemiol. 26 (4) (2004) 305–315.

[4] C. Phillips, A. Salas, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez-Dios, M. Calaza, M. Casares de Cal, D. Ballard, M.V. Lareu, A. Carracedo, Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic. Sci. Int. Genet. 1 (3-4) (2007) 273–280.

[5] J.R. Homburger, A. Moreno-Estrada, C.R. Gignoux, D. Nelson, E. Sanchez, P. Ortiz-Tello, B.A. Pons-Estel, E. Acevedo-Vasquez, P. Miranda, C.D. Langefeld, S. Gravel, M.E. Alarcón-Riquelme, C.D. Bustamante, Genomic insights into the ancestry and demographic history of South America, PLoS Genet. 11 (12) (2015) e1005602.

[6] F. González-Andrade, The genetic ethnography of the Native Amerindians of Ecuador, Int. J. Med. Biol. Front. 17 (2011) 117–136.

[7] F. González-Andrade, D. Sanchez, J. González-Solórzano, S. Gascon, B. Martinez-Jarreta, Sex-specific genetic admixture of Mestizos, Amerindian Kichwas, and afro-Ecuadorians from Ecuador, Hum. Biol. 79 (1) (2007) 51–77.

[8] World Medical Association, Declaration of Helsinki: ethical principles for medical

research involving human subjects, JAMA 284 (23) (2000) 3043–3045.

[9] K.K. Kidd, W.C. Speed, A.J. Pakstis, M.R. Furtado, R. Fang, A. Madbouly, M. Maiers, M. Middha, F.R. Friedlaender, J.R. Kidd, Progress toward an efficient panel of SNPs for ancestry inference, Forensic. Sci. Int. Genet. 10 (2014) 23–32.

[10] R. Nassir, R. Kosoy, C. Tian, P.A. White, L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, P.K. Gregersen, J.W. Belmont, F.M. De La Vega, M.F. Seldin, An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels, BMC Genet. 10 (2009) 39.

[11] L. Excoffier, H.E. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (3) (2010) 564–567.

[12] C.E. Bonferroni, Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8 (1936) 3–62.

[13] D. Falush, M. Stephens, J.K. Pritchard, Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, Genetics 164 (4) (2003) 1567–1587.

[14] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2) (2000) 945–959.

[15] M. Jakobsson, N.A. Rosenberg, CLUMPP: a Cluster Matching and Permutation Program for Dealing with Label Switching and Multimodality in Analysis of Population Structure, (2017).

[16] N.A. Rosenberg, DISTRUCT: a program for the graphical display of population structure, Mol. Ecol. Notes 4 (2004) 137–138.

[17] V. Pereira, H.S. Mogensen, C. Børsting, N. Morling, Evaluation of the Precision ID Ancestry Panel for crime case work: a SNP typing assay developed for typing of 165 ancestral informative markers, Forensic Sci. Int. Genet. 28 (2017) 138–145.

[18] G. Espregueira Themudo, H. Smidt Mogensen, C. Børsting, N. Morling, Frequencies of HID-ion ampliseq ancestry panel markers among Greenlanders, Forensic Sci. Int. Genet. 24 (2016) 60–64.

[19] A. Yates, W. Akanni, M.R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C.G. Girón, L. Gordon, T. Hourlier, S.E. Hunt, S.H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F.J. Martin, T. Maurel, W. McLaren, D.N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H.S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S.P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G. Spudich, S.J. Trevanion, F. Cunningham, B.L. Aken, D.R. Zerbino, P. Flicek, Ensembl 2016, Nucleic Acids Res. D1 (2016) D710–D716.

[20] F. González-Andrade, R. López-Pulles, Ecuador: public health genomics, Publ. Health Genom. 13 (3) (2010) 171–180.

[21] M. Baeta, C. Núñez, C. Sosa, M. Bolea, Y. Casalod, F. González-Andrade, L. Roewer, B. Martínez-Jarreta, Mitochondrial diversity in Amerindian Kichwa and Mestizo populations from Ecuador, Int. J. Legal Med. 126 (2) (2012) 299–302.