Forensic Science International: Genetics 31 (2017) 29–33

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

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

ARTICLE INFO

Keywords:
Precision ID ancestry panel
Ancestry prediction
Massive parallel sequencing (MPS)
Ion torrent PGM
Kichwa
Mestizo
Afro-Ecuadorian

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.

http://dx.doi.org/10.1016/j.fsigen.2017.08.012
Received 26 April 2017; Received in revised form 8 August 2017; Accepted 11 August 2017
Available online 14 August 2017
1872-4973/

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 multicultural 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 ≥ 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 ≤ 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