PLOS ONE

RESEARCH ARTICLE

Outlining the Ancestry Landscape of Colombian Admixed Populations

Humberto Ossa[1,2], Juliana Aquino[3], Rui Pereira[4,5], Adriana Ibarra[6], Rafael H Ossa[2,7], Luz Adriana Pérez[8], Juan David Granda[6], Maria Claudia Lattig[8], Helena Groot[8], Elizeu Fagundes de Carvalho[3], Leonor Gusmão[3,5]*

1 Pontificia Universidad Javeriana, Facultad de Ciencias, Bogotá, Colombia, **2** Laboratório de Genética y Biología Molecular, Bogotá, Colombia, **3** DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil, **4** i3S (Instituto de Investigação e Inovação em Saúde), Universidade do Porto, Porto, Portugal, **5** IPATIMUP (Instituto de Patologia e Imunologia Molecular da Universidade do Porto), Porto, Portugal, **6** 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, **7** Universidad El Bosque, Facultad de Medicina, Bogotá, Colombia, **8** Laboratorio de genética humana, Universidad de los Andes, Bogotá, Colombia

* lgusmao@ipatimup.pt

Citation: Ossa H, Aquino J, Pereira R, Ibarra A, Ossa RH, Pérez LA, et al. (2016) Outlining the Ancestry Landscape of Colombian Admixed Populations. PLoS ONE 11(10): e0164414. doi:10.1371/journal.pone.0164414

Editor: Tzen-Yuh Chiang, National Cheng Kung University, TAIWAN

Received: June 16, 2016

Accepted: September 23, 2016

Published: October 13, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Financial support was granted by DNA Program – State University and Justice Court of Rio de Janeiro (UERJ/TJRJ/MPRJ), Brazil. IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education, and is partially supported by FCT, the Portuguese Foundation for Science and Technology. JA was supported through a Doctoral fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil. RP is

Abstract

The ancestry of the Colombian population comprises a large number of well differentiated Native communities belonging to diverse linguistic groups. In the late fifteenth century, a process of admixture was initiated with the arrival of the Europeans, and several years later, Africans also became part of the Colombian population. Therefore, the genepool of the current Colombian population results from the admixture of Native Americans, Europeans and Africans. This admixture occurred differently in each region of the country, producing a clearly stratified population. Considering the importance of population substructure in both clinical and forensic genetics, we sought to investigate and compare patterns of genetic ancestry in Colombia by studying samples from Native and non-Native populations living in its 5 continental regions: the Andes, Caribe, Amazonia, Orinoquía, and Pacific regions. For this purpose, 46 AIM-Indels were genotyped in 761 non-related individuals from current populations. Previously published genotype data from 214 Colombian Natives from five communities were used for population comparisons. Significant differences were observed between Native and non-Native populations, among non-Native populations from different regions and among Native populations from different ethnic groups. The Pacific was the region with the highest African ancestry, Amazonia harboured the highest Native ancestry and the Andean and Orinoquían regions showed the highest proportion of European ancestry. The Andean region was further sub-divided into 6 sub-regions: North East, Central West, Central East, West, South West and South East. Among these regions, the South West region showed a significantly lower European admixture than the other regions. Hardy-Weinberg equilibrium and variance values of ancestry among individuals within populations showed a potential stratification of the Pacific population.

supported by a postdoctoral fellowship from FCT (SFRH/BPD/81986/2011), Portugal. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Introduction

In a geographic framework, Colombia occupies the southern extreme of the bridge that connects the two American subcontinents. Therefore, since prehistorical times, Colombia has been subject to an intense genetic and cultural flow carried by Native American migrations, which ultimately resulted in a high diversity of ethnic groups inhabiting the country and to a noticeable heterogeneity between geographic regions [1,2] (DANE censo general 2005; http://www.dane.gov.co). This heterogeneity is still maintained in the current Colombian population, whose diversity has been further shaped by admixture with people from other continents. The Andes is the region in which the Chibcha civilization flourished, the third most developed group in the Americas after the Aztecs and the Incas. The Caribbean region is clearly differentiated from the remaining regions, and its territory is inhabited by highly diverse ethnic groups. These two regions, both before and after the European conquest, were and remain the most densely populated and economically active in the country. The other three natural regions include the Pacific region and the forest area that today corresponds to the Orinoquía and Amazonia regions, which are the largest in the territory but the least populated and least economically developed (DANE censo general 2005).

During and after the colonial era, the genetic background of the populations currently inhabiting the Colombian territory was ultimately shaped by different levels of admixture between Natives and European and African incomers [3–5]. This admixture was associated with different migration patterns, drift effects and the more or less pronounced geographic isolation of the populations.

Some studies have been conducted to genetically characterize Colombian populations and to correlate the observed diversity with history events, using genetic markers with different inheritance patterns and degrees of susceptibility to detect populations' differentiation by mutation, drift, and admixture [6–10].

The genetic ancestry of Colombia has been investigated using lineage markers. Several studies were published describing the Y chromosome profile of Colombian populations, but most of these studies were restricted to Y-STRs (e.g., [11–13]), and only a few included the Y-SNP markers that allow for a more robust analysis of the paternal ancestry of a population [7,10,14–16]. Some of the above-mentioned studies also included mtDNA data, revealing the unequal maternal *vs*. paternal ancestry of the admixed Colombian populations, which harbour a gene pool that is mainly composed of Native American mtDNA and European Y chromosome haplogroups [7,14–16]. European mtDNA haplogroups are the second-most represented in non-Native populations in Colombia, with the exception of some African descendant populations in which the African L-haplotypes are predominant [6,17,18]. Most Native groups still preserve an almost complete native maternal ancestry, and signs of European admixture can be observed in the Y chromosome gene pool that varies with the degree of cultural and geographic isolation of the group [16,19,20].

Although uniparental markers have been useful in revealing differences in populations with respect to paternal and maternal inheritance, a comprehensive description of the genetic profile of populations in terms of ancestry can only be achieved by the study of recombining markers.

Regarding recombining autosomal markers, the studies available for Colombia have been even more fragmentary than for lineage markers and have only considered a restricted number of markers and/or population groups (e.g., [3,7,15,21]). Price et al. [22] have studied a large set of ancestry-informative markers (AIMs); however, in this study, the "Colombians" were represented by individuals from a single population (Antioquia), and did not account for the large heterogeneity in the country. A much larger number of population samples from across Colombia were examined by Rojas et al. [7] and Ibarra et al. [8]. However, the study by Rojas

et al. [7] included only a small number of eleven autosomal AIMs, and the set of 52 SNPs studied by Ibarra et al. [8] does not present high levels of intercontinental diversity. Therefore, although appropriate for estimating relative differences in the ancestral composition of the populations analysed, the data from both studies are not sufficient to obtain accurate estimates of the ancestry of Colombians. So far, these studies have shown that Colombia does not contain a uniform genetic pool; rather, it has a high heterogeneity of African, European and Native American ancestries, depending on the region of the country.

Considering the complex history and the high genetic heterogeneity of the present-day populations in Colombia, more comprehensive studies are still required for a fine-scale mapping of the admixture structure of the country, which is crucial in many applied fields, including clinical and forensic genetics [23,24].

In an attempt to contribute to a better understanding of the ancestry of Colombians, in the present work we used a panel of 46 autosomal ancestry-informative insertion/deletion markers (AIM-Indels) to characterize patterns of variation across Native and non-Native Colombian populations from different geographic regions.

The AIM-Indels that we used were previously selected by Pereira et al. [25] for inference of ancestry and admixture proportions of African, European, Native American and East Asian origin, and they have been applied to the study of South American populations from Bolivia, Brazil and Colombia [16,26,27].

Genetic variation in the Native and non-Native populations occupying the five continental regions following the settlement of Colombia (the Andean, Caribbean, Pacific, Orinoquía, and Amazonia regions) was investigated with an emphasis on the Andean region, which currently represents almost 80% of the total population (DANE censo general 2005). The results revealed a highly stratified population in terms of ancestry that should be considered when delineating studies and/or interpreting genetic data in different areas of research.

Materials and Methods

Ethics Statement

All samples involved in the study were anonymised DNA extracts that had been previously obtained from healthy, unrelated individuals. Blood samples were collected by free and written informed consent for research purposes. The current study complies with the ethical principles of the 2000 Helsinki Declaration of the World Medical Association and, together with the informed consent, it was approved by the Ethics Committee of the Pontificia Universidad Javeriana of Colombia.

Sample collection and DNA extraction

A total of 761 samples were collected from unrelated individuals from 28 of the 32 departments of Colombia, as indicated in Fig 1a. These are random samples representing the general population of each region, comprising individuals with admixed ancestry (Afro-Colombians and Mestizos), mainly from urban centres in each department. In the present manuscript, the term non-Native was used to distinguish these populations from the Indigenous tribes described in previous studies, which were used for comparison.

A larger sample was collected from Departments in the Andes because this is the most densely populated region of Colombia. This region was further subdivided into 6 sub-regions. The location and number of samples included in this study for each Natural Region of Colombia and for the Andean sub-regions are indicated in Fig 1.

DNA was extracted following the salting-out protocols described by Miller et al. [28]. DNA quantification was performed by spectrophotometry using the Thermo Scientific NanoDrop

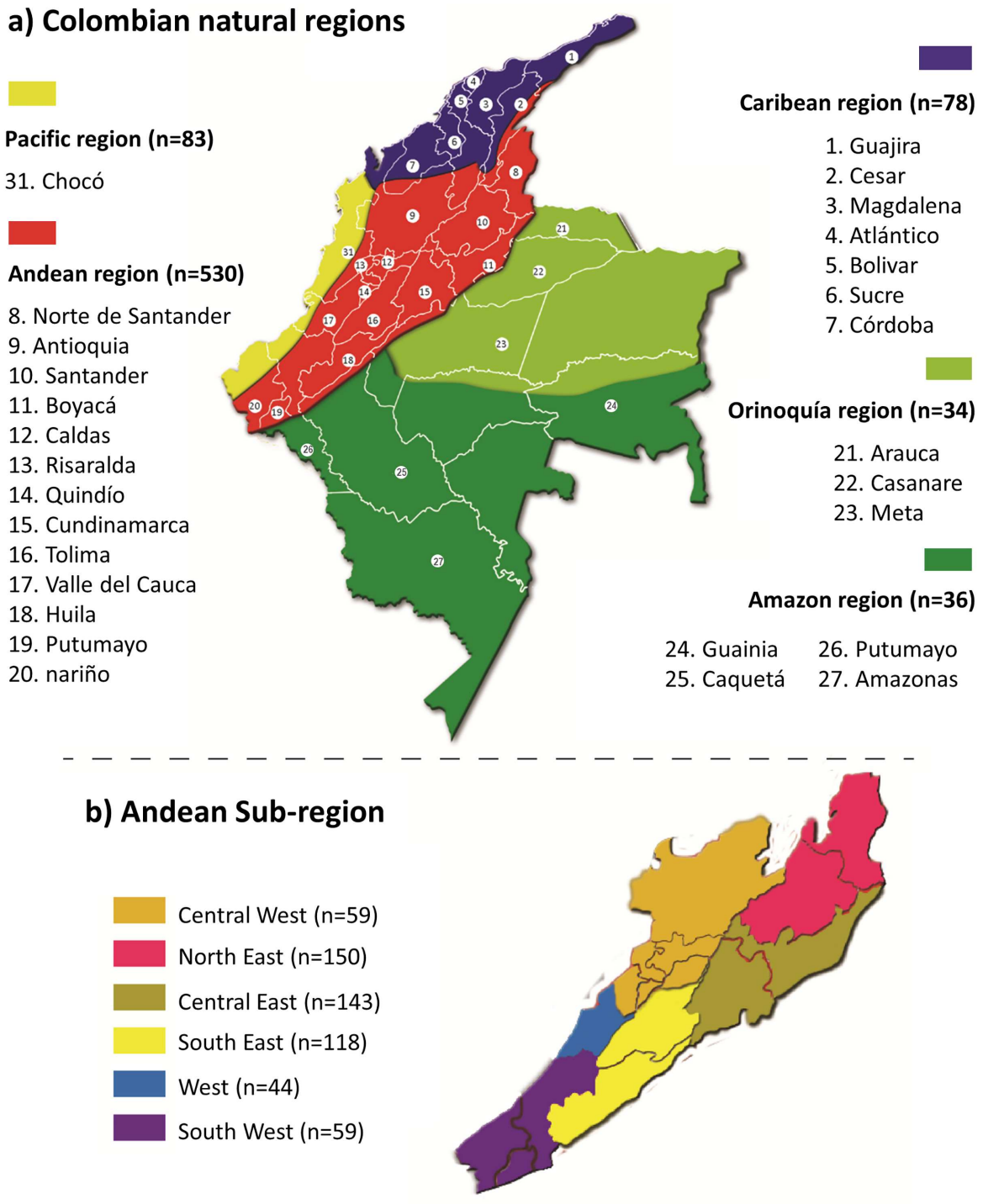


Fig 1. Map of the continental territory of Columbia showing the five main regions and geographical locations of sampling sites (a). The Andean sub-regions considered in the present work and their sample sizes are also indicated (b).

doi:10.1371/journal.pone.0164414.g001

100 apparatus (NanoDrop Technologies, Wilmington, DE, USA). Aliquots of DNA were prepared for PCR reactions at a concentration of 2.5 ng/μL.

For analysis, we included data from Native Colombian populations that were previously studied for the same AIM-Indel markers. These included (i) Natives living in the Cauca region; (ii) Emberá-Chami living in the Antioquia Department; (iii) Natives living in Guainía; (iv) Motilón-Barí living in Norte de Santander; (v) and Pijao from the Tolima Department [16,29]. These Native groups are from different departments (see Fig 1 in Ossa et al. [29]) and belong to diverse linguistic/ethnic groups: Barbacoan (Guambiano and Coconuco from Cauca), Chibchan (Motilón-Barí; Nasa from Cauca), Chocoan (Emberá-Chamí), Tucanoan (Tucano, Cubeo, Guanano and Desana from Guanía) and Arawakan (Curripaco from Guainía). Additionally, to perform ancestry analysis, we used reference data available for HGDP-CEPH samples from African, European and Native American populations [25].

Genetic markers and genotyping

All samples were genotyped for 46 AIM-Indel markers distributed across the autosomal genome. PCR reactions were carried out in a single multiplex containing all 46 primer pairs according to the protocol described by Pereira et al. [25] after adjusting the total reaction volume to 5 μL for 2.5 ng of template DNA. The following PCR thermocycling conditions were used: an initial step of 15 min at 95°C, followed by 27 cycles at 94°C for 30 s, 60°C for 1.5 min, 72°C for 45 s and a final extension at 72°C for 60 min.

Dye-labelled PCR amplified fragments were separated by capillary electrophoresis and detected using an Applied Biosystems® 3500 Series Genetic Analyzer (Thermo Fisher Scientific Inc., Waltham, Massachusetts, USA). Automated allele calls were obtained with GeneMapper v.4.1 software (Thermo Fisher Scientific, Inc.).

Statistical analyses

Genetic diversity parameters, including the estimation of allele frequency, observed and expected heterozygosity, Hardy-Weinberg values, F_{ST} genetic distances and resulting non-differentiation p-values, were assessed using Arlequin v3.5 [30]. Using this software, the significance of genetic distances is obtained by permuting individuals between populations; and the p-value of the test is the proportion of permutations leading to a F_{ST} value larger or equal to the observed one. A multidimensional scaling (MDS) plot of the pairwise F_{ST} matrix was created using the software STATISTICA v7.0. (Statsoft, Tulsa, Oklahoma; http://www.statsoft.com/).

The apportionment of genetic ancestral contributions to the different regions of Colombia was estimated as the mean of each ancestry across individuals, using Admixture v1.3 software [31]. As suggested by cross validation results of a preliminary unsupervised analysis, and considering the historical formation of the Colombian population, we assumed an essentially tri-hybrid contribution from Native Americans, Europeans and Africans (i.e., K = 3). To estimate the ancestral membership proportions in the studied populations, supervised analyses were then performed using prior information on the geographic origin of the reference samples. Ancestry analyses were initially conducted using the HGDP-CEPH populations as a reference [25]. We then investigated the possible impact of considering different Native groups (i.e., HGDP-CEPH, Emberá-Chami, Motilón-Barí and Guainía) as the American ancestral contributor when estimating genetic ancestry. Based on the coherence of the results (see the discussion below), we conducted a final supervised analysis in which the American reference comprised all samples from those four Native groups that showed individual ancestry estimates $\geq$ 90% Native in the preliminary unsupervised analysis.

Results and Discussion

The genotyping results obtained for the 761 Colombian samples from the present work and the 121 samples from the Native groups examined in Ossa et al. [29] are listed in S1 Table. The genotypes from the remaining 93 samples are available in Xavier et al. [16]. Based on the observed genotypes, allele frequencies were estimated in each population and are presented in S2 Table. Hardy-Weinberg tests showed no significant deviations in all population samples studied, with the exception of MD1636 in the Pacific region (S3 Table). A significance level of 0.001 was obtained by applying Bonferroni's correction to 46 tests performed in each population sample. It is worth mentioning that *p*-values below 0.05 were also observed for six additional markers in the Pacific population sample; in most cases (six out of seven), these markers were associated with an excess of homozygotes.

Ancestry analysis among the five natural regions of continental Colombia

Based on geographical criteria, Colombia can be divided into six natural regions, with the three Andes Mountain branches (East, Central and West Andes) separating the two coastal regions (Pacific and Caribe) from the two inner regions (Orinoquía and Amazonia). A sixth region includes the islands in both the Caribbean Sea and the Pacific Ocean (Insular region).

The apportionment of ancestral genetic contributions from Africa, Europe, and Native America were estimated in five continental regions of Colombia: Caribe, Andes, Orinoquía, Amazonia and Pacific (Fig 2). The prevalent ancestral component in Caribe, Andes and Orinoquía was the European; the Pacific and Amazonia regions showed a higher genetic contribution from Africa and Native America, respectively.

The Andes region had the highest European contribution, but also revealed a marked Native American contribution. Together with Amazonia, this region had the lowest values of African ancestry. A similar genetic ancestry profile was observed in Orinoquía, although with slightly higher Native American and African components than in the Andean region.

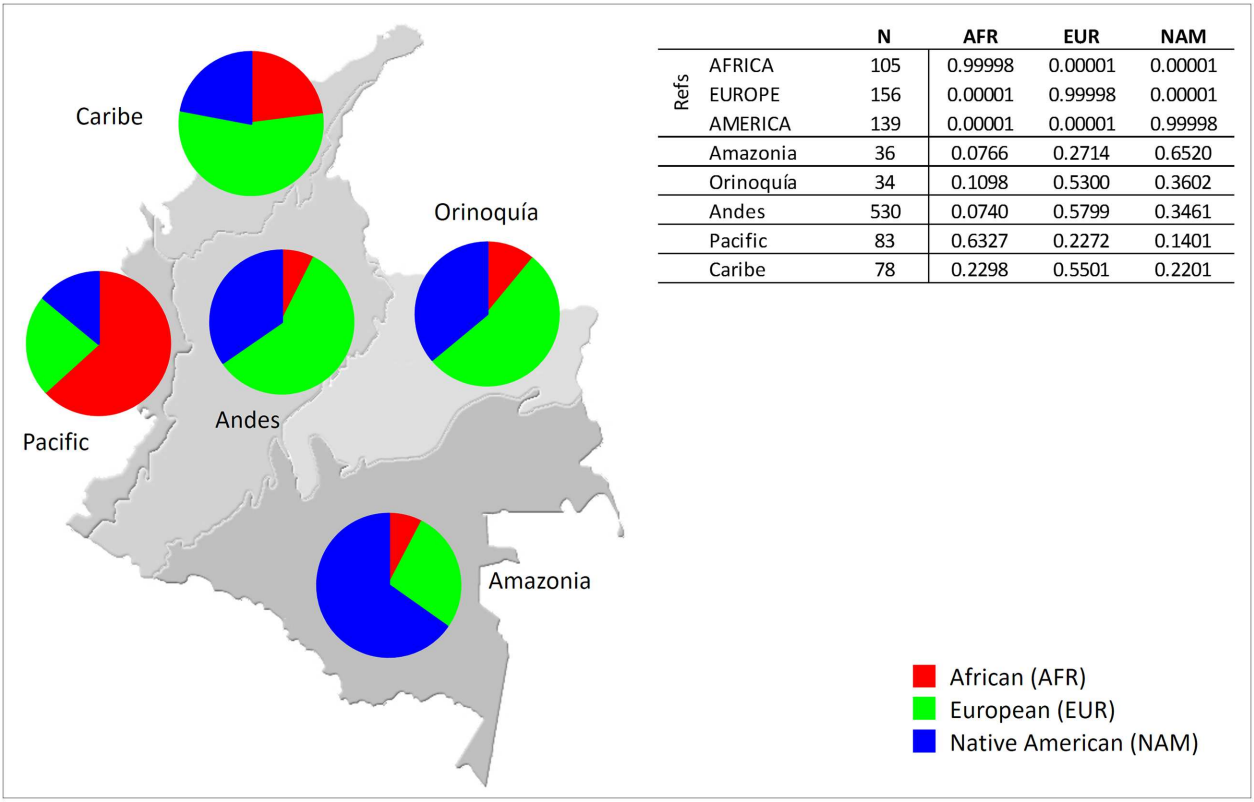
The high European contribution in Andean populations is in accordance with historical data because, during European colonization, the Magdalena River, located between the Central and Western branches of the Andes, was the main route of dispersal through the interior of the territory.

Amazonia had low values of both European and African gene flow, while more than 65% of its genes are Native American. This is probably due to the geographic isolation of this region, which has the lowest population density in the country. Although it had a higher proportion of non-Native influx than some Native groups (previously studied by Xavier et al. [16] and Ossa et al. [29] for the same markers), the sample from Amazonia showed a similar proportion of Native American ancestry as the Pijao group from Tolima [29].

The Pijao (also known as Coyaima-Natagaima) originate in different ethnic groups that share linguistic and cultural affinities. They lost their original language, and their culture and religion experienced some degree of acculturation with Hispanic influence [32].

The similarity in the proportion of Native ancestry found in the Pijao and the population from Amazonia highlights the importance of geographic isolation in the preservation of high levels of Native ancestry in non-Native communities. It also emphasizes that the self-perception of belonging to an ethnic group cannot be determined by genetics alone, as it is the case of the Pijao indigenous community in Colombia that although subjected to a high European admixture still preserves Native American ancient traditions.

As expected from its historical African background, the sample from the Pacific region has 63% African ancestry and the lowest values of both European and Native American ancestries



		N	AFR	EUR	NAM
Refs	AFRICA	105	0.99998	0.00001	0.00001
	EUROPE	156	0.00001	0.99998	0.00001
	AMERICA	139	0.00001	0.00001	0.99998
	Amazonia	36	0.0766	0.2714	0.6520
	Orinoquía	34	0.1098	0.5300	0.3602
	Andes	530	0.0740	0.5799	0.3461
	Pacific	83	0.6327	0.2272	0.1401
	Caribe	78	0.2298	0.5501	0.2201

Fig 2. African, European and Native American membership proportions in samples from the five continental regions of Colombia.

doi:10.1371/journal.pone.0164414.g002

of all regions in the country. After the arrival of the Europeans, the Pacific coast suffered different migration waves that changed its original Native American genepool. During the 16th and 17th centuries, a large number of African slaves were brought to work in gold mines and, during this period, the African population become predominant. This population pattern was reinforced with the abolition of slavery in the 19th century, when a large number of Africans came to this region, especially from the Department of Cauca.

Although much lower than in the Pacific, the Caribe region also presents a significant African ancestry (23%), which can be explained by the role of Cartagena as one of the most important ports in South America during the slave trade.

Ancestry analysis among Andean sub-regions

The Andes is the most densely populated region of Colombia. Based on cultural criteria, this region was divided into six sub-regions: Central West, West, South West, North East, Central East, and South East. When comparing the ancestry of populations across the Andes, all sub-regions are characterized by a predominant European contribution (Fig 3). Native Americans represent the second-highest contribution to the genetic background of these populations. Among the Andean sub-regions, the Central West revealed the highest European and the lowest Native American contributions (67% and 25%, respectively). This sub-region is also known as the Antioqueña colonization region because the first Europeans were established in the Antioquia Department and subsequently migrated to the neighbouring departments of Caldas, Risaralda, Quindío, Norte del Valle and Norte del Tolima.

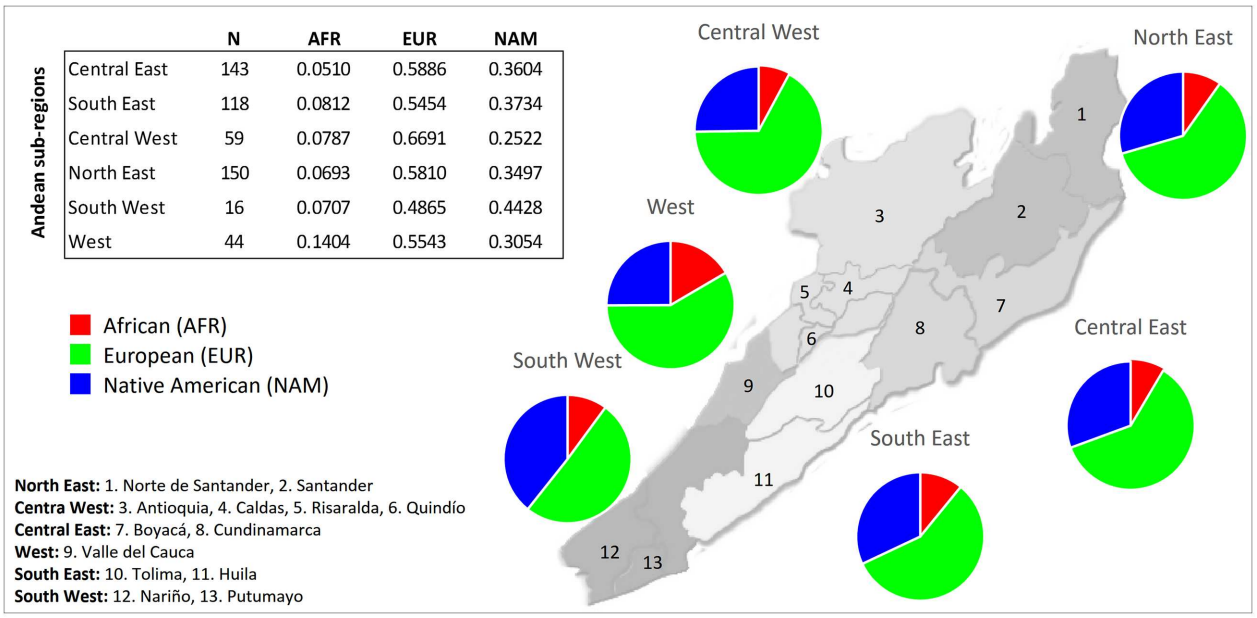


Fig 3. African, European and Native American membership proportions in samples from the Andean sub-regions of Colombia.

doi:10.1371/journal.pone.0164414.g003

The highest African contribution (14%) is found in the West sub-region (also known as Valle del Cauca), which is the Andean region that is closest to the Pacific coast. The two southern regions, including populations from the Departments of Tolima, Huila (South East), Nariño, Cauca and Putumayo (South West), have preserved a marked Native American ancestry (37–44%).

The North East (comprising the Norte de Santander and Santander Departments) and Central East (Cundinamarca and Boyacá) sub-regions presented a similar ancestry profile and, although these two Andean sub-regions had been early colonized by Europeans that largely displaced the Native populations, their samples exhibited lower European ancestry than those from the Central West region.

Comparison with previous studies

In this study, we obtained higher differences among European, African and Native American contributions than Ibarra et al. [8] due to the improved suitability of the markers selected for ancestry estimation (S4 Table). In fact, when markers with low levels of population differentiation are used to infer contributions from three source populations, they tend to overestimate values of ancestry below one third and underestimate contributions that are above this value [33]. Despite the differences observed in the absolute values of ancestry, the relative values among regions are similar in both studies: Orinoquía and most Andean populations showed higher European ancestry, followed by Native ancestry, and the Pacific populations showed a predominantly African ancestry. Differences were observed between the two samples from the South East region (S4 Table) that cannot be attributed to differences in the type of marker used. The sample from Ibarra et al. [8] (comprising individuals from Nariño) has a higher Native American contribution than that studied in this work (comprising individuals from Nariño and Putumayo). This result can be attributed to a sampling effect or may be caused by differences between Nariño and Putumayo populations. Regardless, it would be desirable to perform a larger study covering the three departments in this region (Cauca, Nariño and

Putumayo) to investigate their population substructure. The need for such an analysis is further emphasized by the high intra-population variation that was observed during this study.

The results we obtained are more difficult to reconcile with those reported by Rojas et al. [7] for eleven autosomal AIMs. In general, we obtained higher estimates of European and African ancestry in Andean populations and lower estimates of African contribution in Caribbean and Valle de Cauca (West Andean) populations. Additionally, differences among Andean sub-regions were lower in our study.

In previous studies, the ancestry of four Native groups from the Andes was investigated using the same set of AIM-Indel markers employed in this work [16,29]. When comparing the non-Native admixture, no correlations were observed between the Native and non-Native groups of each sub-Andean region. For instance, the Natives from Cauca (in the South West sub-region) and the Pijao (in the South East sub-region) both present high levels of admixture with non-Native gene pools. Conversely, although from geographic sub-regions with higher admixture with Europeans, the Emberá-Chamí (Central West) and the Motilón-Barí (North East) reveal the highest values of Native ancestry, highlighting the importance of cultural and geographic isolation in the preservation of the Native genetic heritage of these Andean Native groups.

Inter-individual ancestry variation within populations

To investigate the signals of recent admixtures or population stratifications, we compared the variations in individual ancestry estimates within the studied populations. Therefore, African, European and Native American membership proportions were assessed using the Admixture software, and the results are presented in Fig 4. A high inter-individual variation can be observed within populations, especially in the Pacific, Orinoquía, Amazonia and South West Andean regions. These populations showed higher values of average variance of African, European and Native American ancestries than the remaining populations from the Andes region and the Caribbean region (Fig 4). In S1 Fig we plotted the absolute values obtained after subtracting the individual ancestry estimates from the average value of the population. Confirming our previous observations, most individuals from the Andean (except the South West sub-region) and Caribbean populations have ancestry values that are closer to the population average, in contrast to the South West and Pacific regions, where a high proportion of the individuals show ancestry values far from the average.

The high intra-population variation found in the Pacific regions can be the result of (i) recent migration and admixture between two groups that differ in their African ancestry; (ii) the presence of population substructure in an old admixed population; or (iii) both. A recent migration of Africans is known to have occurred to the North Pacific region after the abolition of slavery. However, the new incomers joined a population that was already known to harbour an important African component. Therefore, a certain degree of population substructure can be expected in the Pacific population.

In conclusion, regions with less European admixture showed higher ancestry variation among individuals, indicating more recent admixture events and/or a stronger stratification of these populations.

Genetic distance among Colombian Native and non-Native populations

Pairwise F_{ST} genetic distances were calculated between Native and non-Native populations from Colombia and three reference samples from Africa, America and Europe.

As expected, higher values of genetic distance were obtained for the AIM-Indels included in the present work than for the SNPs previously studied by Ibarra et al. [8], which have been selected to maximize intra-population differentiation.

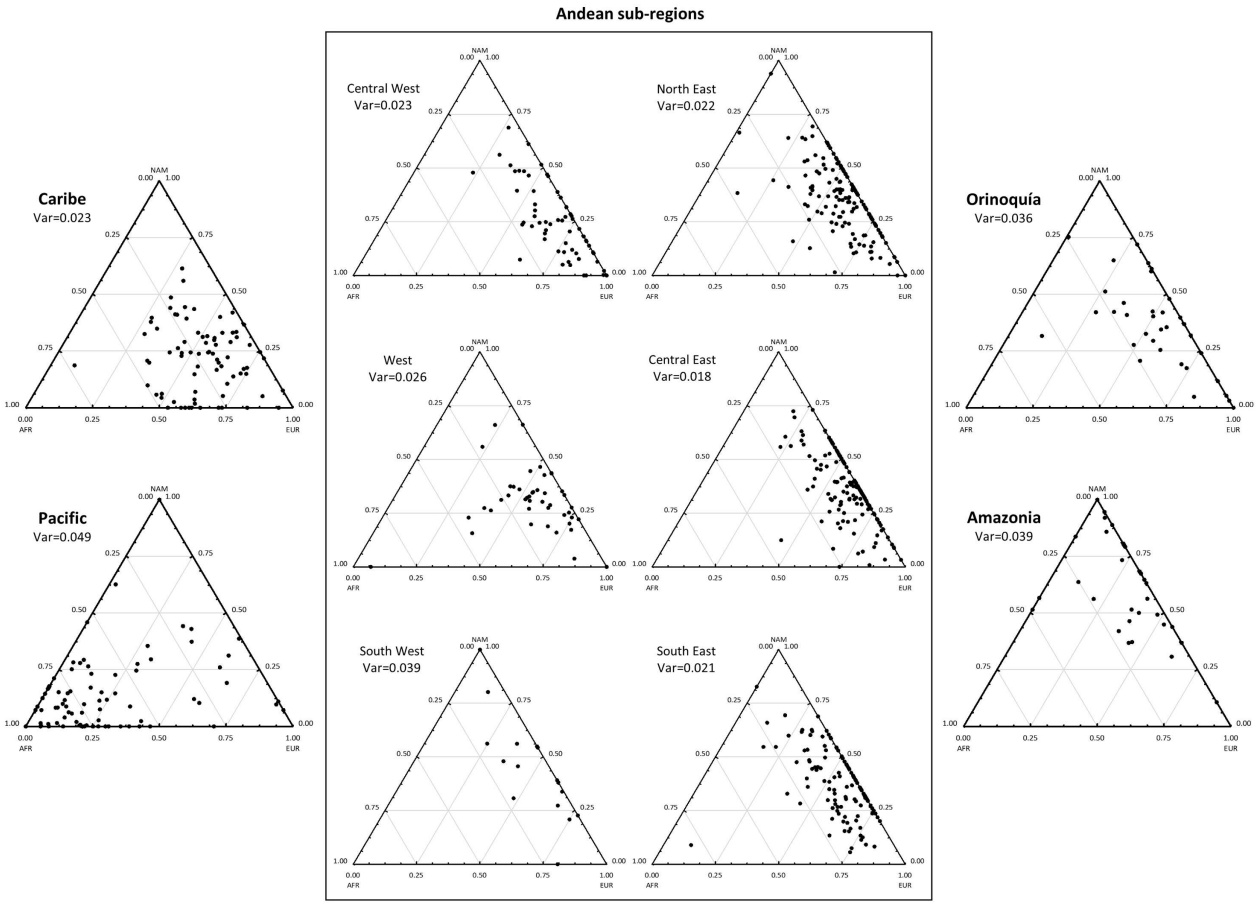


Fig 4. Individual ancestry estimates within populations obtained in samples from the five regions of Colombia. The average value of variance for the three estimates in each population (var) are also indicated.

doi:10.1371/journal.pone.0164414.g004

Significant differences were observed between Native and non-Native populations, as well as among non-Native populations from different regions and among native populations from different ethnic groups (S5 Table). When comparing the Andean sub-populations, no significant differences were found except between the Central West and North and South East. The presence of the highest European ancestry in the Central West can explain these results, although differences between the Native backgrounds of these Andean sub-populations may also exist. Indeed, significant differences were found between Natives, even for groups with similar membership proportions in ancestry estimates, including the Emberá-Chamí (Central West) and the Motilón-Barí (North East). The large genetic distance between these two Native groups and between them and the HGPD-CEPH reference sample from Native Americans reveals important genetic drift events, that are visualized in the MDS plot in Fig 5.

In the MDS plot, the Pacific sample appears closer to the African reference sample, the Amazonia sample is closer to the American Native sample, and the Andean, Orinoquía and Caribbean samples are closer to the European reference sample (Fig 5).

In summary, the position of the non-Native populations in the MDS plot closely follows the expected results based on the ancestry estimates. The high genetic distances observed between Native populations with low non-Native American influx (HGDP-CEPH reference samples from Native American, Guainía, Emberá-Chamí and Motilón-Barí groups) cannot be

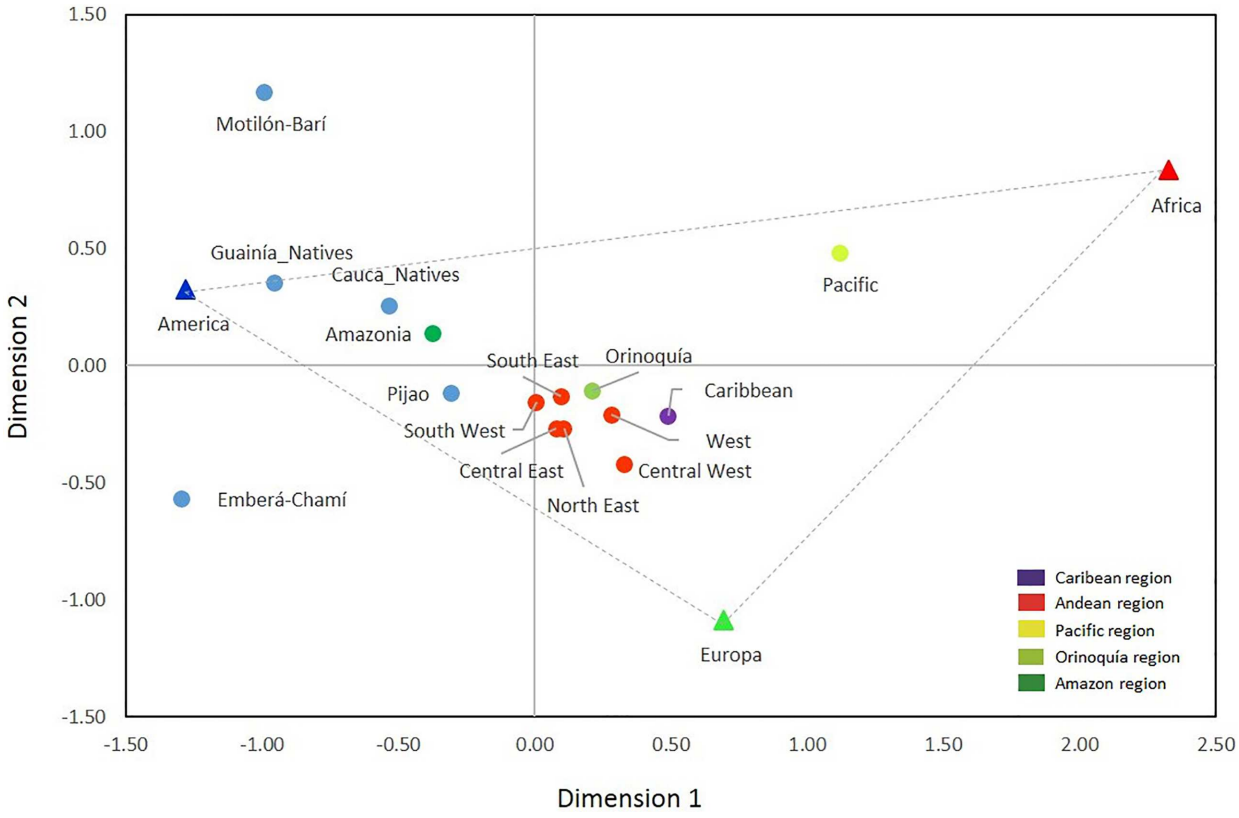


Fig 5. MDS plot of the F_{ST} pairwise genetic distances between Native and non-Native Colombian populations and the three samples used as references for Africa, Europe and Native America (stress = 0.057136).

doi:10.1371/journal.pone.0164414.g005

explained solely by differences in the ancient background of these populations; these populations must have been subject to a strong genetic drift, including founder and/or endogamy events.

Genetic differences among nearby Native American groups have been previously reported based on different types of markers, namely mtDNA, Y chromosome, HLA class II variation and ancestry informative SNPs [10,16,34–36]. The results obtained point to a subpopulation differentiation process, due to a restricted gene flow with the Andes mountains acting as a geographic barrier, or as the result of different migration routes entering South America, from Panama, during the Pleistocene [3,21,37].

The impact of Native American reference samples in ancestry estimates

Methods of inference of genetic ancestry usually model hypothetical ancestral populations or rely on the use of samples representing those putative ancestral contributors. This modelling is not always a straightforward task because, for most genetic studies, only contemporary populations are available to be used as a reference. Such populations may harbour a more or less preserved ancestral genepool. Concerning the populations under study, it is believed that since the first admixture events with Europeans or Africans, an important part of Native American genetic diversity has been gradually lost. This loss of diversity has been caused by genetic drift, creating noticeable genetic distances among groups.

Therefore, considering the perceptible differentiation among the Native American groups discussed above, we further assessed its putative impact in the inference of genetic ancestry

when, for instance, using different groups as the Native American reference population. S6 Table summaries ancestry analyses obtained for the same dataset while considering HGDP-CEPH and each of the Colombian Native groups as the reference population (i.e., Emberá-Chamí, Guainía and Motilón-Barí). The results were remarkably similar, with only small variations in the ancestry values obtained for the Colombian populations under study. This observation supports the overall robustness of the ancestry analyses performed, which seem to adequately represent the genetic diversity existing among Native American groups.

Conclusion

This study provides a general picture of the ancestry of the Colombian populations from the five continental regions of the country, complementing previous information on lineage and recombining markers. The overall results revealed a highly stratified population in terms of ancestry both between and within delimitated natural regions. This stratification should be taken into account when delineating studies and/or interpreting genetic data in different areas of research.

Supporting Information

S1 Fig. Plot of the absolute values obtained after subtracting the individual ancestry estimates from the average value of the population. Each individual is represented along the x-axis, by three points corresponding to the differences in African, European and Native American ancestries to the average.
(TIF)

S1 Table. List of 46 AIM-Indel genotypes from the Colombian samples included in the present study and in Ossa et al. (2015).
(XLSX)

S2 Table. Frequencies of the shorter alleles (allele 1) for the 46 AIM-Indels in Colombian populations. For loci with 3 alleles in a population sample, the frequency of a second allele (allele 2) was also indicated.
(XLSX)

S3 Table. Observed and expected heterozygote values and p-value for the exact test of Hardy-Weinberg equilibrium (forecasted chain length: 1,000,000; dememorization steps: 100,000), excluding the monomorphic loci (*). *p*-values above 0.05 are indicated in pink, and those below 0.001 are indicated in yellow.
(XLSX)

S4 Table. Comparison of the African, European and Native American ancestry estimates obtained in this study and in Ibarra et al. (2014).
(XLSX)

S5 Table. Pairwise F_{ST} values between Colombian and reference populations (below diagonal) and corresponding differentiation p-values (above diagonal; significant values after Bonferroni's correction [$p<0.0005$] are indicated in red).
(XLSX)

S6 Table. Comparative ancestry estimates of Colombian populations considering different Native American groups as the ancestral reference in supervised analyses.
(XLSX)

Acknowledgments

We would like to thank everyone who voluntarily donated the blood samples that made it possible to carry out this investigation. We would also like to thank Carlos Ossa Valencia and Paola Ossa Montoya for their selfless support in collecting the samples. Finally, we would like to specially thank Ashcayra Arebedora and Alvaro Azoira, who kindly opened the doors to their communities.

Author Contributions

Conceptualization: HO LG.

Data curation: RP LG.

Formal analysis: HO RP LG.

Funding acquisition: HO EFC.

Investigation: HO JA LAP.

Methodology: HO JA RP LG.

Project administration: LG.

Resources: HO AI RHO JDG MCL HG EFC.

Supervision: EFC LG.

Visualization: HO LG.

Writing – original draft: HO LG.

Writing – review & editing: HO JA RP AI RHO LAP JDG MCL HG EFC LG.

References

1. Arango Ochoa R. Los pueblos indígenas de Colombia en el umbral del nuevo milenio: población, cultura y territorio: bases para el fortalecimiento social y económico de los pueblos indígenas. Bogotá: Departamento Nacional de Planeación; 2004.
2. Gonzales De Perez MS, Rodriguez De Montes ML. Lenguas Indigenas De Colombia: Una Vision Descriptiva. Instituto Caro y Cuervo, 2000
3. Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, et al. Genetic variation and population structure in native Americans. PLoS Genet. 2007; 3:e185. doi: 10.1371/journal.pgen.0030185 PMID: 18039031
4. Roewer L, Nothnagel M, Gusmão L, Gomes V, González M, Corach D, et al. Continent-wide decoupling of Y-chromosomal genetic variation from language and geography in native South Americans. PLoS Genet. 2013; 9: e1003460. doi: 10.1371/journal.pgen.1003460 PMID: 23593040
5. Salzano FM, Sans M. Interethnic admixture and the evolution of Latin American populations. Genet Mol Biol. 2014; 37:151–170. PMID: 24764751
6. Salas A, Acosta A, Alvarez-Iglesias V, Cerezo M, Phillips C, Lareu MV, et al. The mtDNA Ancestry of Admixed Colombian Populations. Am J Hum Biol. 2008; 20:584–591. doi: 10.1002/ajhb.20783 PMID: 18442080
7. Rojas W, Parra MV, Campo O, Caro MA, Lopera JG, Arias W, et al. Genetic make up and structure of Colombian populations by means of uniparental and biparental DNA markers. Am J Phys Anthropol. 2010; 143: 13–20.
8. Ibarra A, Freire-Aradas A, Martinez M, Fondevila M, Burgos G, Camacho M, et al. Comparison of the genetic background of different Colombian populations using the SNPforID 52plex identification panel. Int J Legal Med. 2014a; 128:19–25.

9. Ibarra A, Restrepo T, Rojas W, Castillo A, Amorim A, Martinez B, et al. Evaluating the X chromosome-specific diversity of Colombian populations using insertion/deletion polymorphisms. PloS one. 2014b; 9: e87202.
10. Noguera MC, Schwegler A, Gomes V, Briceno I, Alvarez L, Uricoechea D, et al. Colombia's racial crucible: Y chromosome evidence from six admixed communities in the Department of Bolivar. Ann Hum Biol. 2014; 41: 453–459. doi: 10.3109/03014460.2013.852244 PMID: 24215508
11. Yunis JJ, Acevedo LE, Campoc DS, Yunis EJ. Population data of Y-STR minimal haplotypes in a sample of Caucasian-Mestizo and African descent individuals of Colombia. Forensic Sci Int. 2005; 151: 307–313. doi: 10.1016/j.forsciint.2005.02.005 PMID: 15939168
12. Builes JJ, Castañeda SP, Espinal C, Aguirre D, Rodríguez JR, Gómez MV, et al. Analysis of 16 Y-chromosomal STRs in a Córdoba (Colombia) population sample. International Congress Series 2006; 1288:174–176.
13. Alonso LA, Usaquén W. Y-chromosome and surname analysis of the native islanders of San Andrés and Providencia (Colombia). Hum Biol. 2013; 64: 71–84.
14. Carvajal-Carmona LG, Duque C, Alvarez VM, Soto ID, Ospina-Duque J, Bedoya G, et al. Strong Amerind/White Sex Bias and a Possible Sephardic Contribution among the Founders of a Population in Northwest Colombia. Am J Hum Genet. 2000; 67:1287–1295. doi: 10.1016/S0002-9297(07)62956-5 PMID: 11032790
15. Bedoya G, Montoya P, Garcia J, Soto I, Bourgeois S, Carvajal L, et al. Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. Proc Natl Acad Sci U S A. 2006; 103: 7234–7239. doi: 10.1073/pnas.0508716103 PMID: 16648268
16. Xavier C, Builes JJ, Gomes V, Ospino JM, Aquino J, Parson W, et al. Admixture and genetic diversity distribution patterns of non-recombining lineages of Native American ancestry in Colombian populations. PloS one. 2015; 10: e0120155. doi: 10.1371/journal.pone.0120155 PMID: 25775361
17. Rodas C, Gelvez N, Keyeux G. Mitochondrial DNA Studies Show Asymmetrical Amerindian Admixture in Afro-Colombian and Mestizo Populations. Hum Biol. 2003; 75: 13–30. PMID: 12713143
18. Yunis JJ, Yunis EJ. Mitochondrial DNA (mtDNA) haplogroups in 1526 unrelated individuals from 11 Departments of Colombia. Genet Mol Biol. 2013; 36:329–335. doi: 10.1590/S1415-47572013000300005 PMID: 24130438
19. Mesa NR, Mondragón MC, Soto ID, Parra MV, Duque C, Ortíz-Barrientos, et al. Autosomal, mtDNA, and Y chromosomal diversity in Amerinds: Pre- and post-Colombian patterns of gene flow in South America. Am J Hum Genet. 2000; 67:1277–1286. doi: 10.1016/S0002-9297(07)62955-3 PMID: 11032789
20. Usme-Romero S, Alonso M, Hernandez-Cuervo H, Yunis EJ, Yunis JJ. Genetic differences between Chibcha and Non-Chibcha speaking tribes based on mitochondrial DNA (mtDNA) haplogroups from 21 Amerindian tribes from Colombia. Genet Mol Biol. 2013; 36: 149–157. doi: 10.1590/S1415-47572013005000011 PMID: 23885195
21. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al. Reconstructing Native American population history. Nature. 2012; 488: 370–374. doi: 10.1038/nature11258 PMID: 22801491
22. Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, et al. A genome wide admixture map for Latino populations. Am J Hum Genet. 2007; 80: 1024–1036. doi: 10.1086/518313 PMID: 17503322
23. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. Hum Mol Genet. 2008; 17: R143–150. doi: 10.1093/hmg/ddn268 PMID: 18852203
24. Zhivotovsky LA, Ahmed S, Wang W, Bittles AH. The forensic DNA implications of genetic differentiation between endogamous communities. Forensic Sci Int. 2001; 119: 269–272. PMID: 11390138
25. Pereira R, Phillips C, Pinto N, Santos C, dos Santos SE, Amorim A, et al. Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. PloS one. 2012; 7: e29684. doi: 10.1371/journal.pone.0029684 PMID: 22272242
26. Manta FS, Pereira R, Caiafa A, Silva DA, Gusmão L, Carvalho EF. Analysis of genetic ancestry in the admixed Brazilian population from Rio de Janeiro using 46 autosomal ancestry-informative indel markers. Ann Hum Biol. 2013; 40: 94–98. doi: 10.3109/03014460.2012.742138 PMID: 23151124
27. Vullo C, Gomes V, Romanini C, Oliveira AM, Rocabado O, Aquino J, et al. Association between Y haplogroups and autosomal AIMs reveals intra-population substructure in Bolivian populations. Int J Legal Med. 2015; 129: 673–680. doi: 10.1007/s00414-014-1025-x PMID: 24878616
28. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988; 16:1215. PMID: 3344216

29. Ossa H, Aquino J, Sierra S, Ramírez A, Carvalho EF, Gusmão L. Analysis of admixture in Native American populations from Colombia. Forensic Sci Int Genet, Supplement Series. 2015; 5:e332–e333

30. Excoffier L and Lischer HE. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 2010; 10: 564–567. doi: 10.1111/j.1755-0998.2010.02847.x PMID: 21565059

31. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19: 1655–1664. doi: 10.1101/gr.094052.109 PMID: 19648217

32. Oliveros DE. Coyaimas y Natagaimas. In Correa F. (ed.), Geografía humana de Colombia. Región Andina Central. Santafé de Bogotá: Instituto Colombiano de Cultura Hispánica. 1996. Tomo IV, vol. 2, 153–193.

33. Aquino JG, Jannuzzi J, Carvalho EF, Gusmão L. Assessing the suitability of different sets of InDels in ancestry estimation. Forensic Sci Int Genet, Supplement Series. 2015; 5: e34–e36.

34. Briceño I, Gomez A, Bernal JE, Papiha SS.HLA-DPB1 polymorphism in seven South American Indian tribes in Colombia. Eur J Immunogenet. 1996; 23: 235–240. PMID: 8803536

35. Casas-Vargas A, Gómez A, Briceño I, Díaz-Matallana M, Bernal JE, Rodríguez JV. High genetic diversity on a sample of pre-Columbian bone remains from Guane territories in northwestern Colombia. Am J Phys Anthropol. 2011; 146: 637–649. doi: 10.1002/ajpa.21626 PMID: 21990065

36. Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, et al. Genomic Insights into the Ancestry and Demographic History of South America. PLoS Genet. 2015; 11: e1005602. doi: 10.1371/journal.pgen.1005602 PMID: 26636962

37. Bonatto S, Salzano F. A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data. Proc Natl Acad Sci USA 1997; 94: 1866–1871. PMID: 9050871