High-resolution HLA allele and haplotype frequencies in majority and minority populations of Costa Rica and Nicaragua: differential admixture proportions in neighboring countries

Esteban Arrieta-Bolaños[1,2,3*], Juan José Madrigal-Sánchez[4], Jeremy E. Stein[2], Priscilla Órlich-Pérez[3,5], María José Moreira-Espinoza[6], Edel Paredes-Carias[6], Yondra Vanegas-Padilla[6], Lizbeth Salazar-Sánchez[4], J. Alejandro Madrigal[2,7], Steven G.E. Marsh[2,7], & Bronwen E. Shaw[2,8]

1 Institute for Experimental Cellular Therapy, University Hospital, Essen, Germany

2 Anthony Nolan Research Institute, Royal Free Hospital, London, UK

3 Centro de Investigaciones en Hematología y Trastornos Afines (CIHATA), Universidad de Costa Rica, San José, Costa Rica

4 Escuela de Medicina, Universidad de Costa Rica, San José, Costa Rica

5 División de Banco de Células Madre, Laboratorio Clínico, Hospital San Juan de Dios, San José, Costa Rica

6 Departamento de Ciencias Morfológicas, Universidad Nacional Autónoma de Nicaragua, León, Nicaragua

7 UCL Cancer Institute, Royal Free Campus, London, UK

8 Center for International Blood and Marrow Transplant Research, Department of Medicine, Medical College of Wisconsin, Milwaukee, USA

Short title: High-resolution HLA in Costa Rica and Nicaragua

Corresponding author:

Esteban Arrieta-Bolaños
Present address:
Institute for Experimental Cellular Therapy
Universitätsklinikum Essen
Virchowstraße 171
Essen 45147, Germany
esteban.arrieta-bolanos@uk-essen.de

Conflict of interest: the authors have no conflict of interest to declare.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/tan.13280

Abstract

The HLA system shows the most extensive polymorphism in the human genome. Allelic and haplotypic frequencies of HLA genes vary dramatically across human populations. Due to a complex history of migration, populations in Latin America show a broad variety of admixture proportions, usually varying not only between countries, but also within countries. Knowledge of HLA allele and haplotype frequencies is essential for medical fields such as transplantation, but also serves as a means to assess genetic diversity and ancestry in human populations. Here, we have determined high-resolution HLA-A, -B, -C, and –DRB1 allele and haplotype frequencies in a sample of 713 healthy subjects from three Mestizo populations, one population of African descent, and Amerindians of five different groups from Costa Rica and Nicaragua and compared their profiles to a large set of indigenous populations from Iberia, Sub-Saharan Africa, and the Americas. Our results show a great degree of allelic and haplotypic diversity within and across these populations, with most extended haplotypes being private. Mestizo populations show alleles and haplotypes of putative European, Amerindian, and Sub-Saharan African origin, albeit with differential proportions. Despite some degree of gene flow, Amerindians and Afro-descendants show great similarity to other Amerindian and West African populations, respectively. This is the first comprehensive study reporting high-resolution HLA diversity in Central America, and its results will shed light into the genetic history of this region while also supporting the development of medical programs for organ and stem cell transplantation.

Keywords: admixture, ancestry, Costa Rica, ethnicity, frequencies, human leukocyte antigen, Nicaragua, population.

1. Introduction

The HLA system shows the most extensive polymorphism in the human genome. Currently, according to the IPD-IMGT/HLA website (release 3.31.0, January 2018)[1], almost 18,000 alleles have been discovered. HLA molecules owe this vast polymorphism to their central role in the immune system presenting endogenous or exogenous peptides to T cells and other immune effector cells that scan for pathogens or malignant cells. The necessity of being able to present a large and diverse repertoire of peptides has acted through selection of non-synonymous substitutions, recombination, and allele and gene conversion processes, and resulted in the concentration of this huge polymorphism in the molecules' peptide-binding grooves, and the codominant expression of several HLA loci with balancing selection at the population level[2]. Importantly, HLA loci show strong linkage disequilibrium, and allelic and haplotypic frequencies of HLA genes vary dramatically across human populations[3,4].

Due to a complex history of admixture, populations in Latin America show an extensive variety of ancestry proportions, usually varying not only between countries, but also within countries[5,6]. Majority Mestizo populations usually show evidence of varying proportions of European, American, and Sub-Saharan African genetic components. In addition, the extant Amerindian populations, the Afro-descendant populations, as well as more recent migrations from other parts of the world make this region truly genetically complex, as well as interesting. Knowledge of HLA allele and haplotype frequencies is essential for medical fields such as transplantation of organs and stem cells[7,8], disease association studies and genetic epidemiology[9], forensics, and pharmacogenetics[10], but also serves as a means to assess genetic diversity and ancestry in human populations[11]. Central America is a sub-region where there is still a lack of studies addressing this.

Costa Rica and Nicaragua cover an area of 182,000 km^2 and together have a population of approximately 11 million inhabitants[12,13]. Both neighboring countries have majority Mestizo populations, in addition to sizeable Amerindian and Afro-descendant minorities. Regional genetic variation in Costa Rica has already been suggested by other genetic markers[14-17]. Moreover, the major population of Costa Rica, that of its Central Valley, has been the subject of genetic scientific interest[18-26] due to its interesting demographic history including claims for relative isolation[27,28], while that of its northwestern province of Guanacaste has been a hub for large-scale human papilloma virus vaccine testing clinical trials[29-33]. All these characteristics make the study of HLA allele and haplotype frequencies in these countries of special interest.

We have previously performed a first study of HLA variation in Costa Rica's Central Valley population at low-resolution[34]. We now extend our study to a larger Costa Rican sample including minority and regional populations, and a sample of the Mestizo population of Nicaragua, and report for the first time high-resolution HLA-A, -B, -C, and –DRB1 allele and haplotype frequencies in 713 healthy subjects.

2. Materials and Methods

2.1. Samples

A total of 713 peripheral blood or saliva samples from unrelated healthy volunteer donors were included in this study. All samples were collected as part of the DNA biobank at the University of Costa Rica's Centre for Research in Hematology and Related Disorders (Centro de Investigaciones en Hematología y Trastornos Afines, CIHATA). DNA was extracted from blood or saliva by routine methods. All participants gave informed written consent as per CIHATA's DNA biobank standard procedures. Sample collection under CIHATA's biobank and this study were approved by the local ethics committee at the University of Costa Rica.

The samples included in the study were collected from five different populations: (1) Mestizo Costa Ricans from the Central Valley (CRCV, n=221, majority population in this country, collected in major cities of this region), (2) Mestizo Costa Ricans from the north-western Costa Rican province of Guanacaste (CRGU, n=110, regional population collected in cantons of Liberia, Santa Cruz, Nicoya, Bagaces, Cañas, Filadelfia, La Cruz, Hojancha, Las Juntas, and Puntarenas), two Costa Rican minorities: (3) Costa Ricans of Afro-Caribbean descent (CRAC, descendants from migrant workers from Jamaica and other Caribbean islands who arrived to Costa Rica's Caribbean coast at the end of the 19th century, n=102, collected in the city of Limón), and (4) Costa Rican Amerindians from five evolutionarily related groups[35,36], some of which retain their Chibchan languages (CRAI, n=125, collected in indigenous reserves throughout the country), and (5) Mestizo Nicaraguans (NICA, n=155, collected in 14 of the 17 Nicaraguan departments), the majority population in this country. The two Mestizo populations from Costa Rica were independently sampled in view of previous evidence of genetic differences suggested by other genetic markers[14-17]. Detailed information on the selection criteria, sampling regions, number of samples per region for all populations sampled, as well as Amerindian ethnicities sampled, can be found in **Supplementary Figures 1 and 2**.

2.2. HLA typing

High-resolution HLA typing was performed using Anthony Nolan's in-house sequence-based typing (SBT) methods. Briefly, HLA class I typing was done by generic amplification of exons 2, 3 and 4 of each gene. Addition of exon 4 implies a resolution higher than standard 'high resolution'[37]. For HLA class II, exon 2 of the HLA-DRB1 gene was amplified using allele group-specific primer pairs. Amplicons were purified and each exon was sequenced with specific forward and reverse primers using the Big Dye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). Sequencing products were purified by ethanol precipitation and run on an ABI 3730xL DNA Analyzer. Sequence analysis was done with Assign-SBT software (version 3.6+, Conexio Genomics, Freemantle, Australia) using IPD-IMGT/HLA database release 3.9.0[1]. Ambiguities were further solved using allele-group- (HLA class I) or codon-86-specific (HLA class II) primer combinations and sequencing of the allelic products. For HLA class I, alleles with identical sequences at codons 2, 3, and 4 could not be distinguished, and were assigned the third-field name of the allele with the lowest numerically ordered name (which is usually the more common one). Allele groups (G) were assigned to HLA class II alleles according to IPD-IMGT/HLA database specifications. All homozygous samples were confirmed by at least two determinations and using different techniques. All populations were typed for HLA-A, HLA-B, HLA-C, and HLA-DRB1.

2.3. Frequency estimation and population genetics analyses

HLA typing data for each locus for each population were tabulated and analyzed in order to determine allele and 4-locus haplotype frequencies. Allele frequencies were computed and haplotype frequencies were estimated by expectation-maximization (EM) methods[38]. Based on haplotype estimations, we generated phenotype (diplotype) estimations for each individual sample included in each population. Subsequent analyses were based on the haplotype set forming the diplotypes for each individual.

Compliance with Hardy-Weinberg equilibrium (HWE), estimated heterozygosity measures, and selective neutrality (Ewens-Watterson-Slatkin) tests were applied to each locus. Linkage disequilibrium (LD) for two-locus allele pairs was determined based on standardized residuals. Global LD for each pair of HLA loci was assessed by a likelihood-ratio test on the frequency estimations. The former assesses the difference between the observed haplotype frequency and the product of the frequencies of the alleles defining the haplotype, whereas global LD provides a summary of the situation taking into account all possible haplotypes for two given loci[39,40]. Frequency estimation and population genetics analyses were performed with Anthony Nolan's implementation of the EM algorithm[41,42] with optimizations as described by Gragert *et al.*[43] (haplotype and diplotype estimations), and by using the online tools available from the HLA-net platform (allele frequency, HWE, estimated heterozygosity, selective neutrality, LD, http://hla-net.eu/tools/)[39,40].

Slatkin's genetic distance[44] derived from pairwise F_{ST} between the five populations presented in this study was calculated on HLA-B allele frequencies using Arlequin[45] (version 3.5.2.2, University of Bern). Significance testing for inter-population distance was performed with 100 permutations, and a significance level of 0.01.

2.4. Allele and haplotype sharing

HLA alleles and extended haplotypes (HLA-A~C~B~DRB1) calculated for each population were compared and the extent of sharing in terms of number of individual alleles and haplotypes and their cumulative frequency was determined by the generation of Venn diagrams (Bioinformatics & Evolutionary Genomics, University of Gent, webtool available at http://bioinformatics.psb.ugent.be/webtools/Venn/).

2.5. Population comparisons

In order to compare the HLA profiles of the five populations analyzed in this study between each other and with other populations, a selection of 95 populations from Iberia, America and Sub-Saharan Africa with available DNA-based typing data for HLA-A and HLA-B allele groups was made and their details are shown in **Supplementary Table 1**. These 95 populations are native population samples from the three ancestral regions that have contributed most to the genetic pool of the populations of Costa Rica and Nicaragua: Iberian Europeans (35 populations), Amerindians, including also Inuit and Alaska natives (30 populations), and Sub-Saharan Africans (SSA, 30 populations). HLA frequency data were extracted from journal articles and/or the Allele Frequencies database[46]. In order to homogenize the dataset, frequencies from population samples with higher-resolution HLA-A and –B data were reduced to the first-field by adding up the frequencies belonging to each allele group (this makes these data no longer EM estimates, and hence should be taken with caution). HLA-A and –B allele group data were used to perform principal coordinates analysis (PCoA) and clustering analyses. These two loci were selected since they show the highest polymorphism of the HLA genes[1], as well the strongest geographic and ethnic-specific variation[47], summarizing well the global picture of intercontinental variation[48] while maximizing also the number of populations available for analysis[46]. In total, the population array included 142,894 chromosomes.

PCoA and clustering analysis based on Euclidean distances calculated from 50 HLA-A and HLA-B allele group frequencies from the 95 ancestral populations and the five populations presented in this study were carried out with the Multi-Variate Statistical Package (MVSP, Kovach Computing Services, Anglesey, Wales). The Eigenanalysis for the PCoA was performed at an accuracy of 1E-10 and axes were extracted according to Kaiser's rule[49]. A clustering method based on minimum variance analysis using squared Euclidean distances with randomized input order was used to generate a dendrogram.

2.6. Putative continental origin of extended haplotypes and admixture proportion approximation in Mestizo populations

Putative continental origin of extended HLA haplotypes was evaluated based on known allele and haplotype frequency distributions in worldwide populations[43,46,47,50], ethnic origin of cells used to describe specific alleles[1], and highly-conserved ethnic-specific linkages for HLA-B~C[1,43,46,47,50]. We performed our analyses on 4-locus extended (i.e. HLA-A~C~B~DRB1) haplotypes in order to work with the maximum integrated information (i.e. alleles at each of the four loci and specific B~C linkages), which is synthesized in the extended haplotype. For this, each locus in the extended haplotype was evaluated for the presence of signature ethnic-specific alleles (i.e. alleles which are exclusively found in a specific non recently-admixed human continental group or that are very rare in other continental human groups (e.g. an A*68:30, found among Amerindians; an A*02:02, only common among Sub-Saharan Africans; an A*25:01, frequent in Europeans)). In addition, the specific HLA-B~C conserved linkage in the extended haplotype was assessed for ethnic-specific allele combinations (e.g. B*40:02:01~C*03:05 seen in Amerindians, B*15:03:01~C*02:10 seen in Sub-Saharan Africans, B*44:03:01~C*16:01:01 seen commonly among Europeans). This information was then integrated to assign a putative origin to the extended haplotype.

In order to use these data as a proxy for admixture proportions, putative continental origin of extended haplotypes was applied to the estimated extended diplotypes for each individual sample in each of the Mestizo populations (i.e. CRCV, CRGU, NICA) included in this study. The proportion of extended haplotypes of putative European, Amerindian, and Sub-Saharan African origin present in each of the Mestizo population samples included in this study was then quantified. In a minority of cases where extended haplotypes could not unequivocally be assigned a most likely continental group, their putative continental origin was divided between the two putative continental groups and quantified accordingly. For comparison purposes, the same approach was also applied to the non-Mestizo populations (i.e. CRAC, CRAI)**.** A detailed explanation of this approach with examples can be found in **Supplementary Material 1.**

2.7. Statistical analyses

Euclidean distances calculated from the PCoA were compared using ANOVA and Tukey's method for multiple comparisons. The proportions of extended HLA haplotypes of putative European, Amerindian, and Sub-Saharan African origin were compared between Mestizo populations with pair-wise Chi-squared tests. For these analyses, a p-value of <0.01 was considered statistically significant. Statistical analyses were performed with Prism (version 6.05, GraphPad Software Inc., La Jolla, USA).

3. Results

3.1. Allele frequencies, heterozygosity, HWE and selective neutrality

High-resolution HLA typing was successful in 99.6% of the cases with only a handful of samples failing repeatedly for one or more loci. A summary on sample size, number of alleles per locus, heterozygosity, adherence to HWE, and neutrality tests in all populations is presented in **Table 1**. An average of 32 (range 24-37), 51 (34-64), 25 (18-30), and 35 (26-40) alleles were identified for HLA-A, -B, -C, and –DRB1 in each of the five populations, respectively. Of note, the lowest number of alleles for each locus was found in the sample of Costa Rican Amerindians (CRAI).

Compliance with HWE was confirmed for all populations and loci except for HLA-DRB1 among CRAI ($p=0.0099$). Estimated heterozygosity across loci and populations was high (overall average: 0.9227, range: 0.8289-0.9654), with lowest values among CRAI (average 0.8289) and comparable average levels in the non-Amerindian populations (≥0.9). In terms of the loci analyzed, heterozygosity was lowest for HLA-C (average for all populations: 0.9096), and highest for HLA-B (average across populations: 0.9552).

Selective neutrality was rejected for HLA-A among CRAC, HLA-DRB1 among CRCV, HLA-B and HLA-DRB1 among Nicaraguans, and all loci tested for CRGU, all showing excess of heterozygosity (min $p<0.05$), while HLA-B among CRAI showed an excess of homozygosity (min $p>0.95$).

Overall, alleles from different putative continental origins (Europe, Americas, Africa) could be found in Mestizo populations (CRCV, CRGU, NICA). There is a sizeable proportion of alleles shared between populations: 14/46 HLA-A, 19/91 HLA-B, 13/32 HLA-C, and 17/50 HLA-DRB1 alleles were found in all 5 populations, while only 7/46 HLA-A, 31/91 HLA-B, 5/32 HLA-C, and 10/50 HLA-DRB1 alleles were found in only one population. Of note, a novel HLA-A allele, HLA-A*74:23, was discovered in a CRGU sample[51]. The complete lists of alleles for each locus and population are given in **Supplementary Tables 2-5**.

A comparison of the allele frequency distributions for these populations is shown in **Supplementary Figure 3**. As expected, there is almost no correlation between the allelic frequency distributions of CRAC and CRAI (r^2=0.0007), and very weak correlations (r^2=0.1181-0.2843, and r^2=0.2264-0.4545, respectively) between the distributions in these two groups and those found in the three Mestizo populations. Stronger correlations are seen between the Mestizo samples (r^2=0.6187-0.7178).

In line with this, pairwise F_{ST} between the five populations presented in this study calculated on HLA-B allele frequencies confirm the dissimilarities among them. As presented in **Table 2**, while CRAC and CRAI showed significant pairwise F_{ST} values against the remaining 4 populations (p<0.009), the Mestizo populations did not show statistically significant differences in their HLA-B allele distribution among themselves (p=0.045-0.63). Of note, among the Mestizo populations CRGU showed a notably low F_{ST} to CRAC (0.004, p=0.009).

3.2. Haplotype frequencies

Based on the results for HLA typing on each locus we generated 4-locus haplotype frequency estimations based on the expectation-maximization algorithm. The number of the estimated extended haplotypes (frequency > 1/2N) was 173, 94, 273, 177, and 249 for CRAC, CRAI, CRCV, CRGU, and NICA, respectively. These haplotype estimations had cumulative frequencies of 99.7-100% for each population. The minimum numbers of haplotypes needed to account for 50% of the distribution were 71, 12, 62, 68, and 94 for CRAC, CRAI, CRCV, CRGU, and NICA, respectively. The CRAI sample was the only one to show extended haplotypes with a frequency of >5% (3 haplotypes). Singleton extended haplotype numbers were 151 (87.3%), 57 (60.6%), 202 (74.0%), 152 (85.9%), and 208 (83.5%) for CRAC, CRAI, CRCV, CRGU, and NICA, respectively.

Haplotype frequency distributions were found to be dissimilar between these 5 populations. As shown in **Figure 1**, a smaller extended haplotype pool size correlated with a steeper curve of cumulative frequency for CRAI, whereas NICA showed a more diverse pool. CRAC, CRCV and CRGU followed a similar distribution and only started to differentiate after the 80 most frequent extended haplotypes. After this threshold, the CRCV showed a more diverse extended haplotype pool correlating with the higher sample size for this population.

Table 3 lists the most frequent extended haplotypes estimated for each of the populations with their respective frequencies and putative continental origin. A complete list of estimated extended haplotypes for each population is given in **Supplementary Table 6**. As shown in **Table 3**, putative continental origin of the 10 most frequent extended haplotypes shows mixed origins in Mestizo populations (CRCV, CRGU, NICA), and more homogeneous ancestral putative origins for CRAI and CRAC. CRAI show a higher number of frequent extended haplotypes in comparison to the other populations.

3.3. Linkage disequilibrium

LD was assessed for allele pairs HLA-A~C, B~C, and B~DRB1 in all populations. Significant LD was determined as standardized residuals > 2. The complete lists of all allele pairs with significant LD for each pair of loci and population are given in **Supplementary Tables 7-9,** and blocks with significant LD are indicated for the extended haplotypes in **Supplementary Table 6**. We focus our analysis on the B~C pair, since these loci show the strongest linkage. The number of significantly linked HLA-B~C allele pairs found were 54/69, 38/51, 71/88, 60/76, and 83/115 for CRAC, CRAI, CRCV, CRGU, and NICA, respectively. The average standardized residual was highest for CRCV HLA-B~C allele pairs (8, range 2.0-20.98), and lowest for NICA HLA-B~C linkage (6, range 2.01-17.55). **Table 4** shows the LD data for the 10 most common significantly linked HLA-B~C allele pairs for each population. Overall, LD between these two loci followed known conserved associations, but also showed population-specific linkage related to ethnicity (e.g. B*40:02:01~C*03:05). Global LD was found to be significant only for HLA-B~C allele pairs in CRAI (p=3.34E-6).

3.4. Haplotype sharing between populations

HLA extended haplotype content for each population was compared in order to determine haplotype sharing. In contrast to allele sharing, overall sharing of haplotypes was very low (**Figure 2**), with most haplotypes being private for each population: 94.8, 62.8, 77.7, 71.8, and 83.1% private for CRAC, CRAI, CRCV, CRGU, and NICA, respectively. No extended haplotype could be found in all five populations. For Mestizo populations, only 8 extended haplotypes could be found in all three populations, and their cumulative frequency accounted for only 8.8, 8.2, and 5.3% in CRCV, CRGU, and NICA, respectively.

3.5. Population comparisons

3.5.1. Principal Coordinates Analysis

The first 2 axes for the PCoA analysis are shown in **Figure 3**. These two first components explain 57.5% of the variation of the data. As seen in the figure, native populations from Iberia, Sub-Saharan Africa, and the Americas form discrete clusters. CRAC are found among populations from Sub-Saharan African ancestry, whereas CRAI are found within the disperse cloud of Amerindian populations. Mestizo populations are located between the three ancestral clusters, with CRCV closer to the Iberians, CRGU closer to the Sub-Saharan Africans, and NICA closer to the Amerindian populations.

As shown in **Table 5,** the average distance calculated between the Mestizo populations and the ancestral clusters is lowest to Iberians, intermediate to Sub-Saharan Africans, and highest to Amerindians. The ancestral populations with the closest distance to each Mestizo population are all Iberian: NCab (0.181), Sp-Can (0.190), and AzoTI (0.243) for CRCV, CRGU, and NICA, respectively. In all cases, the distance between the Mestizo populations (CRCV vs CRGU, 0.161; CRCV vs NICA, 0.162; CRGU vs NICA, 0.148) is lower than that to any ancestral population. The closest populations to CRAC and CRAI were USAAO (0.103) and USAN (0.395), respectively.

Pairwise inter-population distances are plotted in **Supplementary Figure 4**. Results show that distances between ancestral groups are significantly larger than their intra-group distances ($p<0.0001$), albeit with a large dispersion within the Amerindian group. In addition, distances from CRAC and CRAI to the SSA and Amerindian populations, respectively, are not statistically significantly different from those within those groups ($p>0.01$). Mestizo populations, on the other hand, show statistically significantly larger distances to the Iberian ($p<0.01$, CRCV, CRGU, NICA) and SSA

($p<0.0001$, CRCV, NICA) clusters when compared to intra-group distances, but not to the Amerindians, possibly due to the large dispersion within this group. Overall, the distances of each Mestizo population to each ancestral population group were not significantly different from each other ($p>0.2588$).

3.5.2. Clustering analysis

The dendrogram generated with the clustering analysis of the HLA frequencies in populations of Costa Rica and Nicaragua is shown in **Supplementary Figure 5**. The population set splits in three main branches corresponding to Amerindian, Sub-Saharan African, and Iberian populations. The Sub-Saharan African cluster is further divided into two subclusters corresponding to (1) West African, and (2) Southern and East African populations, while the Amerindian cluster is also divided into two subclusters, one populated mainly by North American populations, and the other by groups of Mesoamerican, Andean, and South American lowlander populations. Mestizo populations described in this study appear together as a subcluster within the Iberian populations with other outliers of this group. This correlates with the fact that the distances of Mestizo populations to the Iberians in the PCoA are significantly larger than those within Iberian populations (**Supplementary Figure 4**, $p<0.0015$) and not significantly different between each other ($p>0.36$). The Amerindians from Costa Rica appear to form an outlier cluster within Amerindians, whereas the Afro-Caribbean population from Costa Rica locates to West African populations forming a close cluster with other Afro-descendants from the Americas.

3.6. Putative origin of extended haplotypes in Mestizo populations

Using our extended haplotype estimation data, we were able to also estimate extended diplotypes for each individual sample in each of the five populations included in this study. In order to use this data as a proxy for admixture proportions in Mestizo populations, we used these diplotype estimations and characterized the putative continental origin of each of the extended haplotypes forming each diplotype found in each individual sample as explained above (see section 2.6 and Supplementary Material 1). We then quantified the proportion of extended haplotypes of putative

European, Amerindian, and Sub-Saharan African origin present in each of the Mestizo samples included in this study (CRCV, CRGU, NICA), as well as in the CRAC and CRAI. The results from this analysis are shown in **Figure 4**, and show that, despite of the presence of haplotypes likely having other ethnic origins, the majority of extended haplotypes forming the diplotypes in the non-Mestizo populations, that is CRAC and CRAI, come from putative Sub-Saharan African or Amerindian origin, respectively. For the Mestizo populations (i.e. CRCV, CRGU, NICA), despite all three having extended haplotypes from the three ancestral continental regions, their proportions vary widely, with CRCV having a significantly higher proportion of extended haplotypes of likely European origin (66%, p<0.0001 vs both CRGU and NICA), CRGU showing a significantly stronger Sub-Saharan African component (28% of extended haplotypes, p<0.0001 vs CRCV, p=0.0008 vs NICA), and Nicaraguan Mestizos having the highest proportion of extended haplotypes of putative Amerindian origin (41%, p<0.0001 vs CRCV, p=0.0046 vs CRGU) among the three populations.

4. Discussion

In this study we report high-resolution HLA alleles and haplotypes in five populations from Costa Rica and Nicaragua, including the majority populations of both countries, one regional and two minority populations from Costa Rica. Our results show diverse HLA profiles between these geographically close populations, corresponding to different continental ancestries that reflect on our population genetic comparisons with their parental populations. For Mestizo populations, which show alleles and haplotypes from their three main parental continental regions, different admixture proportions can be identified. Remarkably, despite higher levels of individual HLA allele sharing between the populations analysed in this study, sharing of HLA extended haplotypes was very low, illustrating the different levels of variability of this genetic system.

Genetic diversity for the HLA loci in terms of number of alleles per locus and estimated heterozygosity was high (≥0.9) in all non-Amerindian populations, with CRAI showing slightly lower values. Moreover, selective neutrality was rejected in favor of an excess of heterozygosity in various loci, something seen commonly for HLA[40,47]. Despite this, extended haplotype frequency distributions showed differences between populations. Of note, despite its sample size being the largest in the study, the CRCV showed a relatively restricted haplotype pool in comparison to the other populations apart from Amerindians. This was also reflected on the number of alleles identified, with the sample of Nicaraguan Mestizos showing larger numbers of alleles at HLA-C and HLA-DRB1 (and comparable numbers at HLA-A and HLA-B) in comparison to the CRCV. This could relate to a relative isolation of this population in its early formation stages, something that has been a matter of debate[27,28,52].

It must be noted that due to the samples sizes included in this study, we cannot exclude an effect of sampling error on our estimates. Likewise, multi-locus EM haplotype estimations should be interpreted with caution, especially among singletons. The absence of statistically significant LD among allele blocks forming some of these extended haplotypes despite involving well-known, frequent haplotypes (e.g. A*02:01:01~C*07:02:01~B*07:02:01~DRB1*15:01:01G) or well-known, ethnic-specific B~C linkages (e.g. B*45:01~C*16:01:01 common in populations of African descent) could be reflecting a lack of statistical power to allow for this detection due to the small sample sizes in our study, especially in cases involving LD between two high-frequency alleles (e.g. A*02:01:01~C*07:02:01). In several cases, homozygosity, as well as excluded family data for some individuals also confirm the existence of the estimated extended haplotypes in their diplotypes despite the lack of full LD across the allele pairs forming them. Larger sample sizes would be needed to improve the capacity to detect these LD blocks. Of note, positive LD has been found to be more common among frequent HLA haplotypes, with rarer haplotypes not necessarily having full LD[53]. Finally, although our results do not suggest the presence of population structure, the fact that we analyzed a nationwide sample for Nicaragua as opposed to regionalized samples, and that we analyzed a sample of Amerindians from Costa Rica composed of different ethnicities could have masked regional or ethnicity-specific differences. Further studies with sufficient sample sizes for these analyses should be performed in the future.

Our previous study[34] was the first thorough characterization of HLA in the Central Valley of Costa Rica. The present study is the first to report high-resolution frequencies for these genes. A few publications have indirectly reported HLA allele or haplotype frequencies in the population of the province of Guanacaste[54-57]. These disease association-based reports give only partial data that impair thorough comparison. However, inferred allele frequencies for most common class I alleles[55] are similar to those reported in this study, and some common extended low-resolution haplotypes[57] are found among those common in this study (e.g. A*24~B*35~DRB1*04, A*02~B*39~DRB1*04, A*68~B*40~DRB1*04). Another previous study reported HLA class II alleles and haplotypes for a sample of Amerindians of Bribri, Cabécar, and Ngöbe ethnicity from Costa Rica and Panama[58] (n=50). Overall, the results in this report resemble those found by us, with reduced diversity and most individuals carrying few high-frequency HLA-DRB1 alleles (e.g. DRB1*04:07, *14:02, *16:02). A recent short population report describes second-field HLA allele and haplotype frequencies in blood donors from the general population in Managua, Nicaragua (n=339)[59]. Despite differences in sample size and resolution, frequency data in that report and the nation-wide data presented here correlate well (**Supplementary Figure 3**), with 147/197 alleles at the relevant loci found in both studies and representing 94-100% of the allele frequency distributions, and a mean allele frequency ratio of 1.24±0.84 between studies, suggesting some genetic homogeneity across this country.

The HLA profiles of CRAC and CRAI show great similarity to their ancestral continental counterparts, although gene flow from other groups can be appreciated. This is demonstrated in CRAI by the presence of HLA alleles of mainly European (B*38:01:01) or SSA putative origin (A*36:01), and in CRAC by the presence of alleles common in Amerindians (B*35:43:01, C*03:05), as well as the 8.1 haplotype (A1~B8~DR3). Gene flow in Amerindians[58,60-63] and Afro-Caribbeans[64] from Costa Rica, evident from our extended haplotype analyses, has also been reported by others. Despite this and in comparison to the other populations analyzed, the CRAI show a more restricted HLA polymorphism in terms of allele numbers per locus, estimated heterozygosity, and a smaller extended haplotype repertoire with a significant excess of homozygosity at HLA-B. This reduced diversity could be explained by the ancient high levels of population differentiation in Amerindians[65,66], but also by current low effective population size, their relative isolation[62], and endogamy and polygyny documented among some Amerindian groups in Costa Rica[67]. The significant deviation of HWE observed only for the DRB1 locus in this population, something that has been observed frequently in Amerindian populations[68,69] and even in Mestizo populations with heavy Amerindian admixture[70], could however be a signature of selective pressures acting on this locus to counteract reduced ancestral diversity[71,72]. In the case of Mestizo populations (CRCV, CRGU, NICA), their HLA gene profiles show clear evidence of the presence of tri-ethnic admixture of their parental populations. Alleles from essentially putative European (e.g. A*25:01:01, B*37:01:01, B*35:08:01), Amerindian (e.g. A*02:22, B*35:43:01, C*03:05), and Sub-Saharan African (e.g. A*02:02, B*15:03:01, DRB1*08:04:01) origin can be found in all three populations.

Our approach to approximation of admixture in Mestizo populations based on putative continental origin of extended HLA haplotypes shows differential proportions in populations of Costa Rica and

Nicaragua. Although not without limitations arising from the need for systematically assessing each haplotype forming the diplotypes in each sample, and being based on a most probable, although not necessarily unique, continental origin for extended haplotypes derived from EM estimations, our approximation integrates the various sources of information contained in extended HLA haplotypes, and proves to provide results that resemble those obtained using other strategies. It is a well-known fact that many HLA alleles can be considered region-specific[73]. This is especially relevant for Sub-Saharan African and Amerindian populations, where they represent a high proportion of the total allele pool[73]. The use of this knowledge applied to 4-locus haplotypes and coupled with assessment of B~C linkages, gives robustness to our strategy for admixture proportion approximation. Indeed, our admixture proportion estimations resemble the major patterns obtained with other genetic markers for the Central Valley of Costa Rica[6,14-16,74-76], Guanacaste[14-16,77,78], and Nicaragua[79,80]. Another study using HLA-based admixture proportion estimation applied to Mexican Mestizos also showed congruent results between HLA-B allele frequency-based and non-HLA short tandem repeat-based admixture estimates[70]. Similarly, putative HLA haplotype origin has been observed to correlate well with the mean proportion of continental genetic ancestry in subpopulations in the United States[81].

The strong SSA component found for the CRGU population (28% of the extended haplotypes), exemplified by the high frequency of essentially African alleles (e.g. B*15:03:01) and haplotypes (e.g. A*23:01:01~C*02:10~B*15:03:01~DRB1*13:05:01, A*24:02:01~C*02:10~B*15:03:01~DRB1*11:01:02), as well as their results on the population comparison analyses, correlates with previous knowledge on the higher frequency of traits associated with African ancestry, such as glucose-6-phosphate dehydrogenase deficiency[82,83] and hemoglobinopathies[84-87], in this region in comparison to other Costa Rican regions. A report on population structure in Guanacaste[77] found similar admixture proportions for this population (40% EUR, 40% AME and 20% SSA). Interestingly, the inclusion in that study of individuals from the canton of Tilarán in Guanacaste (absent in this study and known to have been founded by settlers from the Central Valley of Costa Rica[88]), which showed a different admixture profile (60% EUR, 30% AME and 10% SSA), also reveals the differences in ancestry proportions between these two Costa Rican regions. Of note, the strong European component in the CRCV correlates with a strong presence of the pharmacogenetic risk allele HLA-B*57:01[89], in turn much more rare in CRGU and not detected in our sample of Nicaraguans.

Apart from its anthropological interest, the description of HLA allele and haplotype frequencies is essential for medical fields such as transplantation. Costa Rica has recently opened the first public cord blood bank in Central America and the results from this study will undoubtedly be fundamental for guiding the rational development of this and other transplantation resources. The great diversity of HLA polymorphism within and across populations in these countries represents thus a challenge. This is even more relevant in view of the increasing gene flow between them: Guanacaste has a long history of emigration to other Costa Rican regions[88]; there are currently six times more Mulattoes (6.7%) than non-admixed Costa Rican Afro-Caribbeans (1.1%) in the Costa Rican population[90]; and, due to intense migration in the last 30 years, Nicaraguans now represent 7% of the Costa Rican population and 1 in 7 births of Costa Ricans between 2000-2016 were from Nicaraguan women[91]. Future studies should extend the analysis of HLA frequencies to other regional populations from Costa Rica and to regional and minority populations of Nicaragua as a step toward ensuring access to therapies to the majority of their inhabitants.

5. Acknowledgements

The authors would like to thank José Manuel Nunes (Department of Genetics and Evolution, University of Geneva) for his expert advice on population genetics analyses for this study. This work was supported by grants from University College London, the University of Costa Rica, and the Costa Rican National Council for Scientific and Technologic Research (CONICIT) to EAB. Molecular typing was performed at Anthony Nolan Research Institute.

6. References

1. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* 2015;43(Database issue):D423-431.
2. Meyer D, Thomson G. How selection shapes variation of the human major histocompatibility complex: a review. *Ann Hum Genet.* 2001;65(Pt 1):1-26.
3. Askar M, Daghstani J, Thomas D, et al. 16(th) IHIW: global distribution of extended HLA haplotypes. *International journal of immunogenetics.* 2013;40(1):31-38.
4. Fernandez Vina MA, Hollenbach JA, Lyke KE, et al. Tracking human migrations by the analysis of the distribution of HLA alleles, lineages and haplotypes in closed and open populations. *Philos Trans R Soc Lond B Biol Sci.* 2012;367(1590):820-829.
5. Adhikari K, Mendoza-Revilla J, Chacon-Duque JC, Fuentes-Guajardo M, Ruiz-Linares A. Admixture in Latin America. *Curr Opin Genet Dev.* 2016;41:106-114.
6. Wang S, Ray N, Rojas W, et al. Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet.* 2008;4(3):e1000037.
7. Pidala J, Lee SJ, Ahn KW, et al. Nonpermissive HLA-DPB1 mismatch increases mortality after myeloablative unrelated allogeneic hematopoietic cell transplantation. *Blood.* 2014;124(16):2596-2606.
8. Tiercy JM, Claas F. Impact of HLA diversity on donor selection in organ and stem cell transplantation. *Hum Hered.* 2013;76(3-4):178-186.
9. Ghodke Y, Joshi K, Chopra A, Patwardhan B. HLA and disease. *Eur J Epidemiol.* 2005;20(6):475-488.
10. Pirmohamed M, Ostrov DA, Park BK. New genetic findings lead the way to a better understanding of fundamental mechanisms of drug hypersensitivity. *J Allergy Clin Immunol.* 2015;136(2):236-244.
11. Sanchez-Mazas A, Fernandez-Vina M, Middleton D, et al. Immunogenetics as a tool in anthropological studies. *Immunology.* 2011;133(2):143-164.
12. National Nicaraguan Institute of Development Information (INIDE). Población Total, estimada al 30 de Junio del año 2012. *Sistema Nacional de Estadìsticas Vitales.* 2012.
13. Costa Rican National Institute for Statistics and Censuses (INEC). 2011-2025. Proyecciones distritales. Población total proyectada por grupos de edades, según región de planificación y sexo. 2015.
14. Campos-Sanchez R, Raventos H, Barrantes R. Ancestry informative markers clarify the regional admixture variation in the Costa Rican population. *Hum Biol.* 2013;85(5):721-740.
15. Morera B, Barrantes R, Marin-Rojas R. Gene admixture in the Costa Rican population. *Ann Hum Genet.* 2003;67(Pt 1):71-80.
16. Segura-Wang M, Raventos H, Escamilla M, Barrantes R. Assessment of genetic ancestry and population substructure in Costa Rica by analysis of individuals with a familial history of mental disorder. *Ann Hum Genet.* 2010;74(6):516-524.
17. Morera B, Marin-Rojas R, Barrantes R. [Analysis of various classical genetic markers in the Costa Rica population]. *Rev Biol Trop.* 2001;49(3-4):1237-1252.
18. Amorim CE, Wang S, Marrero AR, Salzano FM, Ruiz-Linares A, Bortolini MC. X-chromosomal genetic diversity and linkage disequilibrium patterns in Amerindians and non-Amerindian populations. *Am J Hum Biol.* 2011;23(3):299-304.
19. Bare LA, Ruiz-Narvaez EA, Tong CH, et al. Investigation of KIF6 Trp719Arg in a case-control study of myocardial infarction: a Costa Rican population. *PloS one.* 2010;5(9).
20. Bertisch H, Mesen-Fainardi A, Martin MV, et al. Neuropsychological performance as endophenotypes in extended schizophrenia families from the Central Valley of Costa Rica. *Psychiatr Genet.* 2009;19(1):45-52.
21. Carmiol N, Peralta JM, Almasy L, et al. Shared genetic factors influence risk for bipolar disorder and alcohol use disorders. *Eur Psychiatry.* 2014;29(5):282-287.
22. Fears SC, Service SK, Kremeyer B, et al. Multisystem component phenotypes of bipolar disorder for genetic investigations of extended pedigrees. *JAMA Psychiatry.* 2014;71(4):375-387.
23. Greenwood TA, Beeri MS, Schmeidler J, et al. Heritability of cognitive functions in families of successful cognitive aging probands from the Central Valley of Costa Rica. *J Alzheimers Dis.* 2011;27(4):897-907.
24. Hersh CP, Raby BA, Soto-Quiros ME, et al. Comprehensive testing of positionally cloned asthma genes in two populations. *Am J Respir Crit Care Med.* 2007;176(9):849-857.
25. Mathews CA, Reus VI, Bejarano J, et al. Genetic studies of neuropsychiatric disorders in Costa Rica: a model for the use of isolated populations. *Psychiatr Genet.* 2004;14(1):13-23.

26. Schuler J, Weiss NT, Chavira DA, et al. Characteristics and comorbidity of ADHD sib pairs in the Central Valley of Costa Rica. *Compr Psychiatry.* 2012;53(4):379-386.

27. Carvajal-Carmona LG, Ophoff R, Service S, et al. Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica. *Hum Genet.* 2003;112(5-6):534-541.

28. Morera B, Barrantes R. Is the Central Valley of Costa Rica a genetic isolate? *Rev Biol Trop.* 2004;52(3):629-644.

29. Cortes B, Schiffman M, Herrero R, et al. Establishment and operation of a biorepository for molecular epidemiologic studies in Costa Rica. *Cancer Epidemiol Biomarkers Prev.* 2010;19(4):916-922.

30. Herrero R, Hildesheim A, Rodriguez AC, et al. Rationale and design of a community-based double-blind randomized clinical trial of an HPV 16 and 18 vaccine in Guanacaste, Costa Rica. *Vaccine.* 2008;26(37):4795-4808.

31. Herrero R, Wacholder S, Rodriguez AC, et al. Prevention of persistent human papillomavirus infection by an HPV16/18 vaccine: a community-based randomized clinical trial in Guanacaste, Costa Rica. *Cancer Discov.* 2011;1(5):408-419.

32. Rodriguez AC, Avila C, Herrero R, et al. Cervical cancer incidence after screening with HPV, cytology, and visual methods: 18-Year follow-up of the Guanacaste cohort. *Int J Cancer.* 2017;140(8):1926-1934.

33. Tota JE, Struyf F, Merikukka M, et al. Evaluation of Type Replacement Following HPV16/18 Vaccination: Pooled Analysis of Two Randomized Trials. *J Natl Cancer Inst.* 2017;109(7).

34. Arrieta-Bolanos E, Maldonado-Torres H, Dimitriu O, et al. HLA-A, -B, -C, -DQB1, and -DRB1,3,4,5 allele and haplotype frequencies in the Costa Rica Central Valley Population and its relationship to worldwide populations. *Hum Immunol.* 2011;72(1):80-86.

35. Barrantes R, Smouse PE, Neel JV, Mohrenweiser HW, Gershowitz H. Migration and genetic infrastructure of the Central American Guaymi and their affinities with other tribal groups. *American journal of physical anthropology.* 1982;58(2):201-214.

36. Barrantes R, Smouse PE, Mohrenweiser HW, et al. Microevolution in lower Central America: genetic characterization of the Chibcha-speaking groups of Costa Rica and Panama, and a consensus taxonomy based on genetic and linguistic affinity. *Am J Hum Genet.* 1990;46(1):63-84.

37. Nunes E, Heslop H, Fernandez-Vina M, et al. Definitions of histocompatibility typing terms: Harmonization of Histocompatibility Typing Terms Working Group. *Hum Immunol.* 2011;72(12):1214-1216.

38. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol.* 1995;12(5):921-927.

39. Nunes JM. Using uniformat and gene[rate] to Analyze Data with Ambiguities in Population Genetics. *Evol Bioinform Online.* 2015;11(Suppl 2):19-26.

40. Nunes JM, Buhler S, Roessli D, Sanchez-Mazas A, collaboration HL-n. The HLA-net GENE[RATE] pipeline for effective HLA data analysis and its application to 145 population samples from Europe and neighbouring areas. *Tissue Antigens.* 2014;83(5):307-323.

41. Eberhard HP, Madbouly AS, Gourraud PA, et al. Comparative validation of computer programs for haplotype frequency estimation from donor registry data. *Tissue Antigens.* 2013;82(2):93-105.

42. Maldonado-Torres H, Robinson J, Madrigal JA, Marsh SGE. Cactus, a population genetics analysis environment. 'Genetics and The Immune Response' Abstracts of the 35th Annual Scientific Meeting of the Australasian Society for Immunology and 14th International HLA & Immunogenetics Workshop. *Tissue Antigens.* 2005;66(5):486.

43. Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum Immunol.* 2013;74(10):1313-1320.

44. Slatkin M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics.* 1995;139(1):457-462.

45. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour.* 2010;10(3):564-567.

46. Gonzalez-Galarza FF, Takeshita LY, Santos EJ, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res.* 2015;43(Database issue):D784-788.

47. Solberg OD, Mack SJ, Lancaster AK, et al. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum Immunol.* 2008;69(7):443-464.

48. Di D, Nunes JM, Sanchez-Mazas A. The influence of HLA resolution level on population comparisons strongly depends on loci and geographic ranges. *HLA.* 2017;89(6):371.

49. Kaiser H. The application of electronic computers to factor analysis. *Ed Psychol Meas.* 1960;20:141-151.

50. Maiers M, Gragert L, Klitz W. High-resolution HLA alleles and haplotypes in the United States population. *Hum Immunol.* 2007;68(9):779-788.

51. Arrieta-Bolanos E, McWhinnie AJ, Madrigal-Sanchez JJ, et al. A novel HLA-A allele, A*74:23, identified in an individual from Costa Rica. *Tissue Antigens.* 2014;84(6):583-584.

52. Service SK, Ophoff RA, Freimer NB. The genome-wide distribution of background linkage disequilibrium in a population isolate. *Hum Mol Genet.* 2001;10(5):545-551.

53. Alter I, Gragert L, Fingerson S, Maiers M, Louzoun Y. HLA class I haplotype diversity is consistent with selection for frequent existing haplotypes. *PLoS Comput Biol.* 2017;13(8):e1005693.

54. Wang SS, Wheeler CM, Hildesheim A, et al. Human leukocyte antigen class I and II alleles and risk of cervical neoplasia: results from a population-based study in Costa Rica. *J Infect Dis.* 2001;184(10):1310-1314.

55. Wang SS, Hildesheim A, Gao X, et al. Comprehensive analysis of human leukocyte antigen class I alleles and cervical neoplasia in 3 epidemiologic studies. *J Infect Dis.* 2002;186(5):598-605.

56. Carrington M, Wang S, Martin MP, et al. Hierarchy of resistance to cervical neoplasia mediated by combinations of killer immunoglobulin-like receptor and human leukocyte antigen loci. *J Exp Med.* 2005;201(7):1069-1075.

57. Carreon JD, Martin MP, Hildesheim A, et al. Human leukocyte antigen class I and II haplotypes and risk of cervical cancer. *Tissue Antigens.* 2005;66(4):321-324.

58. Mack S, Tsai, Y., Sanchez-Mazas, A., Erlich, HA. Chapter 3: Anthropology/ human genetic diversity population reports. 13th International Histocompatibility Workshop Anthropology/Human Genetic Diversity Joint Report. In: Hansen JA, ed. *Immunobiology of the human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference.* Vol I. Seattle, WA: IHWG Press; 2007:580-652.

59. Weiskopf D, Grifoni A, Arlehamn CSL, et al. Sequence-based HLA-A, B, C, DP, DQ, and DR typing of 339 adults from Managua, Nicaragua. *Hum Immunol.* 2017.

60. Melton PE, Baldi NF, Barrantes R, Crawford MH. Microevolution, migration, and the population structure of five Amerindian populations from Nicaragua and Costa Rica. *Am J Hum Biol.* 2013;25(4):480-490.

61. Ruiz-Narvaez EA, Santos FR, Carvalho-Silva DR, Azofeifa J, Barrantes R, Pena SD. Genetic variation of the Y chromosome in Chibcha-speaking Amerindians of Costa Rica and Panama. *Hum Biol.* 2005;77(1):71-91.

62. Barrantes R. [Genetic diversity and racial mixture in Amerindians from Costa Rica and Panama]. *Rev Biol Trop.* 1993;41(3A):379-384.

63. Azofeifa J, Barrantes R. Genetic variation in the Bribri and Cabecar Amerindians from Talamanca, Costa Rica. *Rev Biol Trop.* 1991;39(2):249-253.

64. Madrigal L, Ware B, Miller R, Saenz G, Chavez M, Dykes D. Ethnicity, gene flow, and population subdivision in Limon, Costa Rica. *American journal of physical anthropology.* 2001;114(2):99-108.

65. Wang S, Lewis CM, Jakobsson M, et al. Genetic variation and population structure in native Americans. *PLoS Genet.* 2007;3(11):e185.

66. Reich D, Patterson N, Campbell D, et al. Reconstructing Native American population history. *Nature.* 2012;488(7411):370-374.

67. Barrantes R, Azofeifa J. [Demography and genetics of a Guaymi Amerindian population of Limoncito, Costa Rica]. *Rev Biol Trop.* 1981;29(1):123-131.

68. Chen JJ, Hollenbach JA, Trachtenberg EA, et al. Hardy-Weinberg testing for HLA class II (DRB1, DQA1, DQB1, and DPB1) loci in 26 human ethnic groups. *Tissue Antigens.* 1999;54(6):533-542.

69. Hollenbach JA, Thomson G, Cao K, et al. HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives. *Hum Immunol.* 2001;62(4):378-390.

70. Zuniga J, Yu N, Barquera R, et al. HLA class I and class II conserved extended haplotypes and their fragments or blocks in Mexicans: implications for the study of genetic diversity in admixed populations. *PloS one.* 2013;8(9):e74442.

71. Titus-Trachtenberg EA, Rickards O, De Stefano GF, Erlich HA. Analysis of HLA class II haplotypes in the Cayapa Indians of Ecuador: a novel DRB1 allele reveals evidence for convergent evolution and balancing selection at position 86. *Am J Hum Genet.* 1994;55(1):160-167.

72. Erlich HA, Mack SJ, Bergstrom T, Gyllensten UB. HLA class II alleles in Amerindian populations: implications for the evolution of HLA polymorphism and the colonization of the Americas. *Hereditas.* 1997;127(1-2):19-24.

73. Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G. Signatures of demographic history and natural selection in the human major histocompatibility complex Loci. *Genetics.* 2006;173(4):2121-2142.

74. Chen W, Brehm JM, Boutaoui N, et al. Native American ancestry, lung function, and COPD in Costa Ricans. *Chest.* 2014;145(4):704-710.

75. Ruiz-Narvaez EA, Bare L, Arellano A, Catanese J, Campos H. West African and Amerindian ancestry and risk of myocardial infarction and metabolic syndrome in the Central Valley population of Costa Rica. *Hum Genet.* 2010;127(6):629-638.

76. Koehl AJ, Long JC. The contributions of admixture and genetic drift to diversity among post-contact populations in the Americas. *American journal of physical anthropology.* 2017.

77. Wang Z, Hildesheim A, Wang SS, et al. Genetic admixture and population substructure in Guanacaste Costa Rica. *PloS one.* 2010;5(10):e13336.

78. Azofeifa J, Ruiz-Narvaez EA, Leal A, Gerlovin H, Rosero-Bixby L. Amerindian ancestry and extended longevity in Nicoya, Costa Rica. *Am J Hum Biol.* 2017.

79. Nunez C, Baeta M, Sosa C, et al. Reconstructing the population history of Nicaragua by means of mtDNA, Y-chromosome STRs, and autosomal STR markers. *American journal of physical anthropology.* 2010;143(4):591-600.

80. Morera B. Estimacion de la mezcla genética en la población de Nicaragua. *Cuadernos de Antropología.* 2006;16:39-46.

81. Hollenbach JA, Saperstein A, Albrecht M, et al. Race, Ethnicity and Ancestry in Unrelated Transplant Matching for the National Marrow Donor Program: A Comparison of Multiple Forms of Self-Identification with Genetics. *PloS one.* 2015;10(8):e0135960.

82. Beutler E, Kuhl W, Saenz GF, Rodriguez W. Mutation analysis of glucose-6-phosphate dehydrogenase (G6PD) variants in Costa Rica. *Hum Genet.* 1991;87(4):462-464.

83. Chaves M, Saenz GF, Quintana E, Montero A, Jimenez J. [Polymorphism of erythrocytic glucose-6-phosphate dehydrogenase in Costa Rica]. *Sangre (Barc).* 1988;33(1):12-14.

84. Sáenz G, Chaves M, Quintana E. Las hemoglobinopatías en Costa Rica. Aspectos históricos, culturales y epidemiológicos. *Rev Cost Cienc Méd.* 1986;7(4):95-106.

85. Rodriguez Romero WE, Saenz Renauld GF, Chaves Villalobos MA. [Hemoglobin S haplotypes: their epidemiologic, anthropologic and clinical importance]. *Rev Panam Salud Publica.* 1998;3(1):1-8.

86. Saenz Renauld GF. [Hemoglobinopathies in Caribbean Basin countries]. *Rev Biol Trop.* 1988;36(2B):361-372.

87. Saenz GF, Altafulla M, Sancho G, Salgado M. Abnormal hemoglobins and thalassemias in Costa Rica, other countries of Central America, and Panama. *Bull Pan Am Health Organ.* 1988;22(1):42-59.

88. Jiménez-Castro W. *Migraciones internas en Costa Rica.* Washington: Unión Panamericana. Available at http://ccp.ucr.ac.cr/bvp/pdf/migracion/migracion_internaCR/index.htm; 1956.

89. Arrieta-Bolanos E, Madrigal JA, Marsh SG, Shaw BE, Salazar-Sanchez L. The frequency of HLA-B*57:01 and the risk of abacavir hypersensitivity reactions in the majority population of Costa Rica. *Hum Immunol.* 2014;75(11):1092-1096.

90. Costa Rican National Institute for Statistics and Censuses (INEC). Censo 2011. Indicadores étnico-raciales según cantón. *Available at: http://wwwinecgocr/censos/censos-2011.* 2011.

91. Costa Rican National Institute for Statistics and Censuses (INEC). Total de nacimientos por nacionalidad de la madre, según país de origen de la madre. *Available at: http://wwwinecgocr/poblacion/nacimientos.* 2000-2016.

Table 1. Heterozygosity, Hardy-Weinberg equilibrium, and neutrality testing for each locus and population

Population	Locus	N typed	Alleles	Estimated Heterozygosity[1]	HWE (p value)	Neutrality (p value)	
						Min	Max
CRAC	A	102	31	0.9361	1.0000	**0.020**	0.446
	B	102	43	0.9445	1.0000	0.164	0.703
	C	102	25	0.8981	1.0000	0.199	0.778
	DRB1	102	34	0.9299	1.0000	0.206	0.769
CRAI	A	124	24	0.8242	0.2347	0.754	0.925
	B	123	34	0.8238	1.0000	**0.997**	**0.997**
	C	122	18	0.8127	1.0000	0.501	0.774
	DRB1	120	26	0.8289	**0.0099**	na	na
CRCV	A	221	37	0.9165	1.0000	0.195	0.597
	B	221	64	0.9511	1.0000	0.429	0.878
	C	221	25	0.9038	1.0000	0.080	0.337
	DRB1	221	38	0.9461	0.5322	**0.003**	0.131
CRGU	A	110	33	0.9275	1.0000	**0.013**	0.409
	B	110	52	0.9654	1.0000	**0.002**	0.280
	C	110	25	0.9167	1.0000	**0.018**	0.263
	DRB1	110	36	0.9434	0.3113	**0.002**	0.234
NICA	A	155	33	0.9170	1.0000	0.098	0.576
	B	155	61	0.9597	1.0000	**0.013**	0.495
	C	155	30	0.9197	1.0000	0.055	0.440
	DRB1	154	40	0.9442	0.3742	**0.000**	0.043

[1] Estimated heterozygosity refers to the population parameter calculated from the allele frequencies at HWE equilibrium. CRAC, Costa Ricans of Afro-Caribbean descent; CRAI, Amerindians from Costa Rica; CRCV, Mestizo Costa Ricans from the Central Valley; CRGU, Mestizo Costa Ricans from Guanacaste province; HWE, Hardy-Weinberg equilibrium; na, not applicable; NICA, Nicaraguan Mestizos.

Table 2. Slatkin's pairwise F_{ST} distances calculated on HLA-B allele frequencies among the five populations presented in this study and their significance

F_{ST} \ significance	CRAC	CRAI	CRCV	CRGU	NICA
CRAC		+	+	+	+
CRAI	0.0996		+	+	+
CRCV	0.0185	0.0586		ns	ns
CRGU	0.0043	0.0660	0.0030		ns
NICA	0.0163	0.0446	0.0009	-0.0006	

Upper half of the matrix: statistical significance. Lower half of the matrix: F_{ST} value. CRAC, Costa Ricans of Afro-Caribbean descent; CRAI, Amerindians from Costa Rica; CRCV, Mestizo Costa Ricans from the Central Valley; CRGU, Mestizo Costa Ricans from Guanacaste province; HLA, human leukocyte antigen; NICA, Mestizo Nicaraguans; ns, not significant; +, statistically significant ($p<0.01$) F_{ST} value.

Table 3. Frequent (counts ≥ 3*) extended HLA haplotypes for each population

Population	Extended haplotype	Frequency	Counts	Putative origin
CRAC	A*01:01:01~C*07:01:01~B*08:01:01~DRB1*03:01:01G	0.0196	4	**EUR**
	A*02:02~C*04:01:01~B*53:01:01~DRB1*15:03:01G	0.0196	4	**SSA**
	A*23:01:01~C*04:01:01~B*53:01:01~DRB1*11:01:02	0.0196	4	**SSA**
	A*01:01:01~C*04:01:01~B*35:01:01~DRB1*03:01:01G	0.0147	3	**SSA, EUR**
	A*30:01:01~C*17:01:01~B*42:01:01~DRB1*03:02:01	0.0147	3	**SSA**
	A*74:01~C*04:01:01~B*44:03:01~DRB1*15:03:01G	0.0147	3	**SSA**
CRAI	A*24:02:01~C*03:05~B*40:02:01~DRB1*04:07:01G	0.0827	21	**AME**
	A*68:01:02~C*03:05~B*40:02:01~DRB1*04:07:01G	0.0679	17	**AME**
	A*68:30~C*03:05~B*40:02:01~DRB1*16:02:01	0.0685	17	**AME**
	A*68:01:02~C*04:01:01~B*35:01:01~DRB1*04:07:01G	0.0496	12	**AME**
	A*24:02:01~C*01:02:01~B*35:43:01~DRB1*08:02:01	0.0432	11	**AME**
	A*24:02:01~C*01:02:01~B*35:43:01~DRB1*16:02:01	0.0343	9	**AME**
	A*24:02:01~C*03:04:01~B*35:01:01~DRB1*14:02	0.0367	9	**AME**
	A*24:03:02~C*01:02:01~B*35:43:01~DRB1*04:07:01G	0.0323	8	**AME**
	A*24:02:01~C*03:03:01~B*15:01:01~DRB1*01:01:01G	0.0245	6	**EUR**
	A*29:02:01~C*16:01:01~B*44:03:01~DRB1*07:01:01G	0.0227	6	**EUR**
	A*68:01:02~C*03:05~B*40:02:01~DRB1*14:02	0.0226	6	**AME**
	A*02:01:01~C*07:02:01~B*39:08~DRB1*04:07:01G	0.0200	5	**AME**
	A*02:01:01~C*03:05~B*40:02:01~DRB1*04:07:01G	0.0169	4	**AME**
	A*24:02:01~C*01:02:01~B*35:43:01~DRB1*04:07:01G	0.0163	4	**AME**
	A*24:02:01~C*03:05~B*40:02:01~DRB1*16:02:01	0.0146	4	**AME**
	A*01:01:01~C*07:01:01~B*08:01:01~DRB1*03:01:01G	0.0121	3	**EUR**
	A*02:01:01~C*06:02:01~B*13:02:01~DRB1*07:01:01G	0.0126	3	**EUR**
	A*02:22:01~C*03:05~B*40:02:01~DRB1*16:02:01	0.0121	3	**AME**
	A*24:02:01~C*02:02:02~B*51:01:01~DRB1*11:01:01G	0.0122	3	**EUR**
	A*24:02:01~C*03:05~B*40:02:01~DRB1*08:02:01	0.0128	3	**AME**
	A*24:02:01~C*04:01:01~B*35:01:01~DRB1*04:05:04	0.0122	3	**AME**
	A*30:04:01~C*06:02:01~B*45:01~DRB1*11:02:01	0.0121	3	**EUR**
	A*31:01:02~C*03:05~B*40:02:01~DRB1*04:07:01G	0.0108	3	**AME**
CRCV	A*03:01:01~C*04:01:01~B*35:01:01~DRB1*01:01:01G	0.0204	9	**EUR**
	A*24:02:01~C*03:05~B*40:02:01~DRB1*08:02:01	0.0204	9	**AME**
	A*02:01:01~C*07:02:01~B*07:02:01~DRB1*15:01:01G	0.0181	8	**EUR**
	A*03:01:01~C*07:02:01~B*07:02:01~DRB1*11:01:01G	0.0158	7	**EUR**
	A*30:04:01~C*06:02:01~B*45:01~DRB1*11:02:01	0.0158	7	**EUR**
	A*02:06:01~C*07:02:01~B*35:01:01~DRB1*04:11:01	0.0136	6	**AME**
	A*25:01:01~C*12:03:01~B*18:01:01~DRB1*15:01:01G	0.0136	6	**EUR**
	A*26:01:01~C*05:01:01~B*44:02:01~DRB1*04:02:01	0.0136	6	**EUR**
	A*29:02:01~C*16:01:01~B*44:03:01~DRB1*07:01:01G	0.0136	6	**EUR**
	A*68:01:02~C*03:05~B*40:02:01~DRB1*04:07:01G	0.0136	6	**AME**
	A*01:01:01~C*06:02:01~B*57:01:01~DRB1*04:04:01	0.0113	5	**EUR**
	A*02:01:01~C*05:01:01~B*44:02:01~DRB1*13:01:01G	0.0113	5	**EUR**
	A*03:01:01~C*07:02:01~B*07:02:01~DRB1*15:01:01G	0.0113	5	**EUR**
	A*24:02:01~C*03:05~B*40:02:01~DRB1*04:07:01G	0.0113	5	**AME**
	A*25:01:01~C*03:03:01~B*15:01:01~DRB1*01:01:01G	0.0113	5	**EUR**
	A*33:01:01~C*08:02:01~B*14:02:01~DRB1*13:01:01G	0.0113	5	**EUR**
CRGU	A*24:02:01~C*04:01:01~B*35:12:01~DRB1*04:07:01G	0.0273	6	**AME**
	A*24:02:01~C*01:02:01~B*35:43:01~DRB1*04:07:01G	0.0227	5	**AME**
	A*24:02:01~C*03:03:01~B*15:01:01~DRB1*01:01:01G	0.0182	4	**EUR**
	A*30:02:01~C*05:01:01~B*18:01:01~DRB1*03:01:01G	0.0182	4	**EUR**
	A*34:02:01~C*04:01:01~B*44:03:01~DRB1*13:01:01G	0.0182	4	**SSA**
	A*02:01:01~C*03:04:01~B*39:02:02~DRB1*04:11:01	0.0136	3	**AME**
	A*02:01:01~C*07:01:01~B*08:01:01~DRB1*03:01:01G	0.0136	3	**EUR**
	A*23:01:01~C*02:10~B*15:03:01~DRB1*13:05:01	0.0136	3	**SSA**
	A*24:02:01~C*02:10~B*15:03:01~DRB1*11:01:02	0.0136	3	**SSA**
NICA	A*24:02:01~C*03:05~B*40:02:01~DRB1*04:07:01G	0.0226	7	**AME**
	A*68:01:02~C*03:05~B*40:02:01~DRB1*04:03:01	0.0159	5	**AME**

A*68:03:01~C*01:02:01~B*35:43:01~DRB1*04:07:01G	0.0159	5	**AME**
A*02:01:01~C*07:01:01~B*08:01:01~DRB1*03:01:01G	0.0097	3	**EUR**
A*02:01:01~C*07:02:01~B*07:02:01~DRB1*15:01:01G	0.0108	3	**EUR**
A*02:01:01~C*16:01:01~B*44:03:01~DRB1*07:01:01G	0.0097	3	**EUR**
A*02:01:01~C*16:01:01~B*45:01~DRB1*13:02:01	0.0097	3	**SSA**
A*31:01:02~C*03:05~B*40:02:01~DRB1*04:07:01G	0.0097	3	**AME**
A*31:01:02~C*04:01:01~B*35:17:01~DRB1*04:07:01G	0.0097	3	**AME**
A*68:03:01~C*07:02:01~B*35:01:01~DRB1*04:07:01G	0.0097	3	**AME**

*For CRCV only haplotypes with counts ≥ 5 are shown due to a larger number of haplotypes with counts ≥ 3 (n=36). See **Supplementary Table 6** for a complete list of extended haplotypes. AME, Americas; CRAC, Costa Ricans of Afro-Caribbean descent; CRAI, Amerindians from Costa Rica; CRCV, Mestizo Costa Ricans from the Central Valley; CRGU, Mestizo Costa Ricans from Guanacaste province; EUR, Europe; HLA, human leukocyte antigen; NICA, Nicaraguan Mestizos; SSA, Sub-Saharan Africa.

Table 4. HLA-B~C linkage disequilibrium data for the 10 most common alleles pairs with significant linkage (i.e. standardized residual > 2) for each population

Population	Allele pair	Frequency			stdres
		Observed	Expected	Difference	
CRAC	B*53:01:01~C*04:01:01	0.1127	0.0323	0.0804	6.28
	B*15:03:01~C*02:10	0.0784	0.0087	0.0697	10.65
	B*42:01:01~C*17:01:01	0.0637	0.0056	0.0581	11.04
	B*35:01:01~C*04:01:01	0.0539	0.0161	0.0378	4.21
	B*58:02~C*06:02:01	0.0490	0.0041	0.0449	10.02
	B*44:03:01~C*04:01:01	0.0441	0.0127	0.0314	3.96
	B*07:02:01~C*07:02:01	0.0392	0.0037	0.0355	8.26
	B*15:10:01~C*03:04:02	0.0343	0.0017	0.0326	11.35
	B*08:01:01~C*07:01:01	0.0294	0.0027	0.0267	7.35
	B*45:01~C*16:01:01	0.0294	0.0024	0.0270	7.95
CRAI	B*40:02:01~C*03:05	0.3566	0.1301	0.2265	9.14
	B*35:43:01~C*01:02:01	0.1352	0.0205	0.1147	12.38
	B*35:01:01~C*04:01:01	0.1025	0.0211	0.0814	8.65
	B*35:01:01~C*03:04:01	0.0492	0.0124	0.0368	5.11
	B*15:01:01~C*03:03:01	0.0287	0.0015	0.0272	10.91
	B*38:01:01~C*12:03:01	0.0287	0.0009	0.0278	14.13
	B*39:08~C*07:02:01	0.0246	0.0011	0.0235	11.01
	B*44:03:01~C*16:01:01	0.0205	0.0007	0.0198	11.94
	B*49:01:01~C*07:01:01	0.0164	0.0005	0.0159	11.47
	B*51:01:01~C*02:02:02	0.0164	0.0005	0.0159	11.47
CRCV	B*07:02:01~C*07:02:01	0.1063	0.0197	0.0866	12.85
	B*40:02:01~C*03:05	0.0905	0.0088	0.0817	18.21
	B*35:01:01~C*04:01:01	0.0584	0.0154	0.0430	7.24
	B*44:02:01~C*05:01:01	0.0475	0.0033	0.0442	16.10
	B*14:02:01~C*08:02:01	0.0475	0.0026	0.0449	18.57
	B*53:01:01~C*04:01:01	0.0339	0.0063	0.0276	7.28
	B*08:01:01~C*07:01:01	0.0294	0.0027	0.0267	10.78
	B*37:01:01~C*06:02:01	0.0271	0.0025	0.0246	10.45
	B*39:08~C*07:02:01	0.0247	0.0045	0.0202	6.28
	B*57:01:01~C*06:02:01	0.0226	0.0020	0.0206	9.55
CRGU	B*40:02:01~C*03:05	0.0591	0.0040	0.0551	12.84
	B*15:03:01~C*02:10	0.0545	0.0032	0.0513	13.38
	B*35:01:01~C*04:01:01	0.0455	0.0102	0.0353	5.15
	B*53:01:01~C*04:01:01	0.0455	0.0093	0.0362	5.54
	B*35:43:01~C*01:02:01	0.0455	0.0030	0.0425	11.59
	B*35:12:01~C*04:01:01	0.0364	0.0074	0.0290	4.96
	B*07:02:01~C*07:02:01	0.0364	0.0037	0.0327	7.94
	B*08:01:01~C*07:01:01	0.0318	0.0031	0.0287	7.57
	B*44:03:01~C*04:01:01	0.0318	0.0102	0.0216	3.15
	B*44:02:01~C*05:01:01	0.0318	0.0022	0.0296	9.43
NICA	B*40:02:01~C*03:05	0.0806	0.0096	0.0710	12.68
	B*07:02:01~C*07:02:01	0.0548	0.0087	0.0461	8.67
	B*35:01:01~C*04:01:01	0.0548	0.0146	0.0402	5.83
	B*35:43:01~C*01:02:01	0.0355	0.0031	0.0324	10.25
	B*35:17:01~C*04:01:01	0.0290	0.0062	0.0228	5.06
	B*08:01:01~C*07:01:01	0.0290	0.0023	0.0267	9.82
	B*44:02:01~C*05:01:01	0.0258	0.0009	0.0249	14.47
	B*14:02:01~C*08:02:01	0.0258	0.0013	0.0245	12.22
	B*52:01:01~C*12:02:02	0.0226	0.0005	0.0221	17.20
	B*44:03:01~C*16:01:01	0.0225	0.0020	0.0205	8.07

CRAC, Costa Ricans of Afro-Caribbean descent; CRAI, Amerindians from Costa Rica; CRCV, Mestizo Costa Ricans from the Central Valley; CRGU, Mestizo Costa Ricans from Guanacaste province; HLA, human leukocyte antigen; NICA, Mestizo Nicaraguans; stdres, standardized residual.
A complete list of haplotypes with significant LD (stdres>2) for each population is given in Supplementary Table 8.

Table 5. Mean distance values from each population presented in this study to the ancestral populations in the principal coordinates analysis and closest population from each ancestral group (and its distance)

	CRAC		CRAI		CRCV		CRGU		NICA	
	Mean	**Closest population**	**Mean**	**Closest population**	**Mean**	**Closest population**	**Mean**	**Closest population**	**Mean**	**Closest population**
Amerindians	0.647	MexMT (0.519)	0.558	USAN (0.395)	0.515	CanC (0.384)	0.480	CanC (0.326)	0.445	MexMT (0.287)
Sub-Saharan Africans	0.212	USAAO (0.103)	0.694	CapVNW (0.612)	0.369	CapVSE (0.278)	0.329	CapVSE (0.250)	0.392	CapVNW (0.301)
Iberians	0.366	Mur (0.319)	0.620	MajJD (0.561)	0.232	NCab (0.181)	0.243	Sp-Can (0.190)	0.294	AzoTI (0.243)

AzoTI, Azoreans from Terceira Island; CanC, Canadian Cree; CapVNW, Cape Verdeans from NW island; CapVSE, Cape Verdeans from SE island; CRAC, Costa Ricans of Afro-Caribbean descent; CRAI, Amerindians from Costa Rica; CRCV, Mestizo Costa Ricans from the Central Valley; CRGU, Mestizo Costa Ricans from Guanacaste province; MajJD, Majorcans of Jewish descent; MexMT, Mexican Tarasco from Michoacán; Mur, Spanish from Murcia; NCab, Spanish North Cabuernigo; NICA, Mestizo Nicaraguans; Sp-Can, Spanish-BMD-Canarias; USAAO, African Americans, USAN, Alaska Yupik. More details on the ancestral populations can be found in Supplementary Table 1.

Figure 1. HLA extended haplotype frequency distributions for populations from Costa Rica and Nicaragua. The cumulative frequency of HLA extended (HLA-A~C~B~DRB1) haplotypes is shown with relation to the number of haplotypes from the most frequent to the least frequent as estimated for each population. CRAC, Costa Ricans of Afro-Caribbean descent; CRAI, Amerindians from Costa Rica; CRCV, Mestizo Costa Ricans from the Central Valley; CRGU, Mestizo Costa Ricans from Guanacaste province; NICA, Nicaraguan Mestizos.

Figure 2. HLA extended haplotype sharing between populations of Costa Rica and Nicaragua. The presence of each of the extended (HLA-A~C~B~DRB1) haplotypes in all populations was compared by means of Venn diagrams. Extended haplotypes correspond to those forming the estimated diplotypes for each individual in each population. Numbers refer to the number of individual extended haplotypes in that population intersection. A total of 173, 94, 273, 177, and 249 extended haplotypes (estimated frequency > 1/2n) were found for Costa Ricans of Afro-Caribbean descent (CRAC), Amerindians from Costa Rica (CRAI), Costa Rican Mestizos from the Central Valley (CRCV), Costa Rican Mestizos from Guanacaste province (CRGU), and Nicaraguan Mestizos (NICA), respectively. A-C, Population intersections for each Mestizo population and CRAI and CRAC. D, Population intersection for the three Mestizo populations.

Figure 3. Principal coordinates analysis (PCoA) shows the relationship between the populations of Costa Rica and Nicaragua and indigenous populations from Iberia, America, and Sub-Saharan Africa (SSA). (A) PCoA based on HLA-A and HLA-B allele group frequencies was performed with 95 populations from Spain and Portugal (Iberian, 35 populations), Amerindians (30 populations), and Sub-Saharan Africans (30 populations), as well as the five populations presented in this study. Data were reduced to first-field (allele group) in order to homogenize the dataset. Shown are the first two components, which account for 57.5%

of the variation in the data. Populations from Iberia, SSA, and Amerindians form discrete clusters. CRAC and CRAI locate among other African and African-derived, and Amerindian populations, respectively. Mestizo populations locate between the major clusters, corresponding to their tri-ethnic admixture. SSA and Amerindian populations in boxes correspond to those forming West-African and mainly North American sub-clusters in the dendrogram presented in Figure 4, respectively. (B) Inset detailing the Iberian cluster. CRAC, Costa Ricans of Afro-Caribbean descent; CRAI, Amerindians from Costa Rica; CRCV, Mestizo Costa Ricans from the Central Valley; CRGU, Mestizo Costa Ricans from Guanacaste province; NICA, Nicaraguan Mestizos.

Figure 4. Putative origin of extended HLA haplotypes in admixed populations of Costa Rica and Nicaragua suggests differential admixture proportions. Extended HLA haplotypes forming the diplotypes assigned to each individual in each population sample were analyzed and assigned a putative ancestral origin according to known allele and haplotype frequencies, as well as ethnic-specific HLA-B-C associations in ancestral populations (i.e. Europe, Americas, Sub-Saharan Africa). The proportion of haplotypes with putative European, American, and Sub-Saharan African origin for each population is shown. While non-Mestizo populations (CRAC, CRAI) show most of their haplotypes having a putative origin corresponding to their ancestry, Mestizo populations (CRCV, CRGU, NICA) show evidence of differential admixture proportions. CRAC, Costa Ricans of Afro-Caribbean descent; CRAI, Amerindians from Costa Rica; CRCV, Mestizo Costa Ricans from the Central Valley; CRGU, Mestizo Costa Ricans from Guanacaste province; NICA, Nicaraguan Mestizos.

Supplementary Figure 1. Map of Costa Rica indicating the geographic location and sample sizes for the Costa Rican populations sampled and typed in this study. Mestizo populations from Costa Rica (CRCV and CRGU) were sample from the majority Mestizo population resident in these regions, excluding individuals from minorities (e.g. Afro-descendants, Amerindians, Costa Ricans of Chinese descent) based on self-identification. Mestizos from the northwestern province of Guanacaste, CRGU, were collected in cantons of Liberia, Santa Cruz, Nicoya, Bagaces, Cañas, Filadelfia, La Cruz, Hojancha, Las Juntas, and Puntarenas. Costa Rican Mestizos from the Central Valley, an intermontaneous region in the center of the country (approximate area shown in ellipse) that concentrates the majority (60%) of the population in Costa Rica, were sampled in the major cities of this central region. Costa Ricans of Afro-Caribbean descent (CRAC), the majority of which reside on the Caribbean coast on the east of Costa Rica (area shown in ellipse), were sampled in the provincial capital city of Limón. Individuals selected self-identified as Afro-Costa Ricans and had two parents who were also of Afro-Caribbean descent carrying English or Scottish surnames, in stark contrast to the Mestizo populations of this country. Costa Rican Amerindians (CRAI) were sampled exclusively in indigenous reservations, where, according to Costa Rican law, only people of Amerindian descent can reside. The locations of these indigenous reservations and the ethnicities of the five indigenous communities that are included as the CRAI are also indicated.

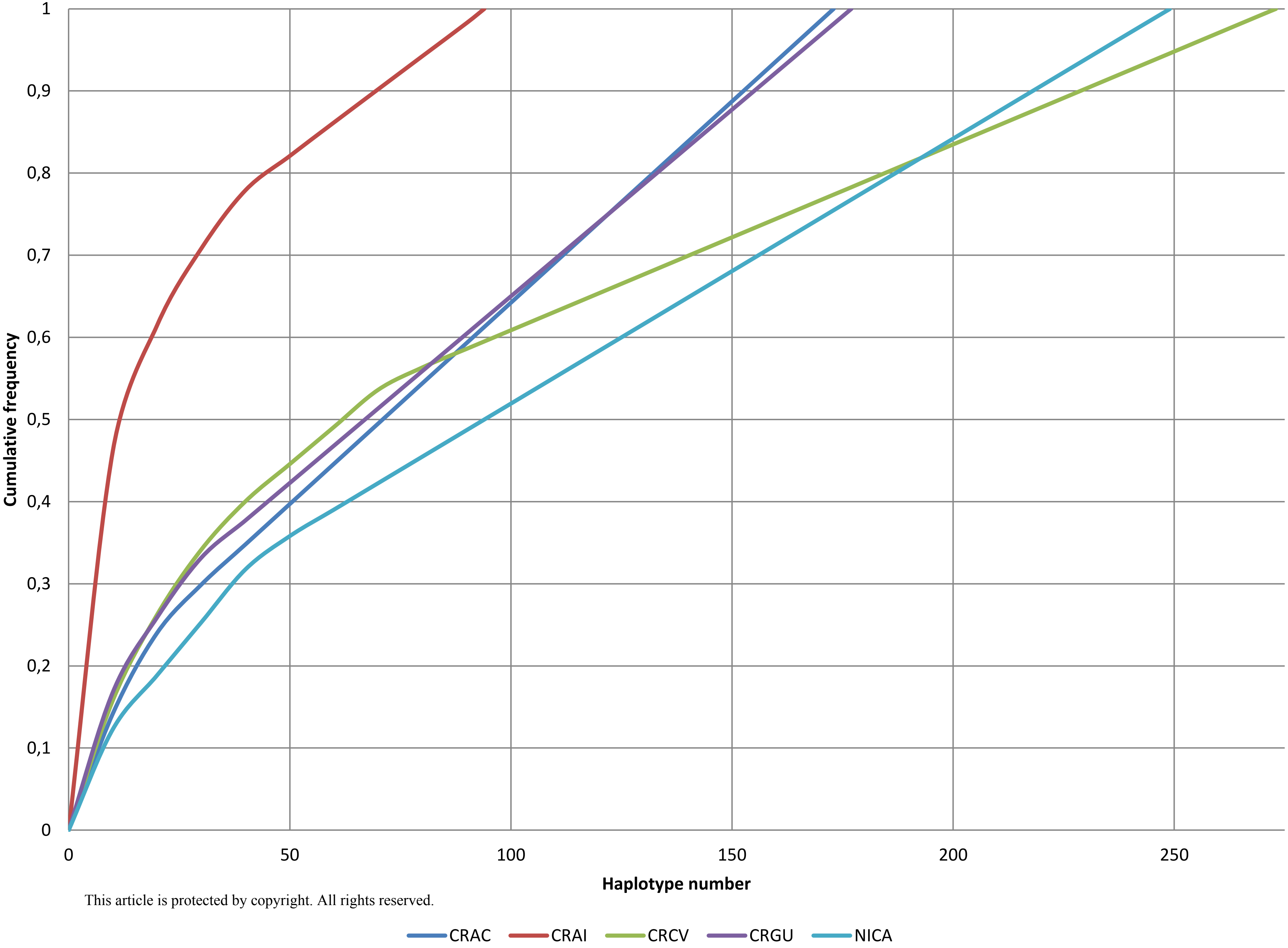
Supplementary Figure 2. Map of Nicaragua indicating the geographic origin of the Nicaraguan Mestizo (NICA) samples collected. Nicaraguan Mestizos (n=155) were sampled from the majority population of this country, excluding individuals from minorities (e.g. Afro-descendants, Amerindians) based on self-identification. Mestizo individuals originated in 14 Departments and were collected by the National Autonomous University of León. The Departments from which the samples originated are listed, including the number of samples for each of them.

Supplementary Figure 3. Correlation between allele frequency distributions in the samples from Costa Rica and Nicaragua. Pair-wise plots showing the correlation between allelic frequencies at HLA-A, -B, -C, and –DRB1 at each of the populations presented in this study. Axes for each plot show the frequency of each allele (diamonds) for each of the two populations compared (horizontal x vertical). A linear trend line and its equation and correlation coefficient (r^2) are shown for each comparison. Top panels: comparisons between CRAC and the other 4 populations; middle panels: comparison between CRAI and the three Mestizo populations; bottom panels: comparisons between the Mestizo populations. For comparison, data from a recent report from samples from blood donors from the general population in Managua, Nicaragua (MANA, see Weiskopf *et al*. ref. 59) are plotted against the NICA sample at the second field. CRAC, Costa Ricans of Afro-Caribbean descent; CRAI, Amerindians from Costa Rica; CRCV, Mestizo Costa Ricans from the Central Valley; CRGU, Mestizo Costa Ricans from Guanacaste province; NICA, Nicaraguan Mestizos.

Supplementary Figure 4. Comparison of interpopulation distances generated with the PCoA. All pair-wise interpopulation Euclidean distances based on the PCoA analysis performed with 50 HLA-A and HLA-B allele group frequencies of 95 ancestral populations and the 5 populations presented in this study are plotted and mean and standard deviations shown. Intra-ancestral group distances (i.e. AME vs AME, IBE vs IBE, SSA vs SSA) are compared to inter-ancestral group distances and the distances of each of the populations presented in this study to those of each ancestral group. Results of the ANOVA tests are presented in relation to each ancestral intra-group distance distribution: (A) Comparison of distances within and to the Amerindian populations (AME). (B) Comparison of distances within and to the Iberian (IBE) populations. (C) Comparison of distances within and to the Sub-Saharan African (SSA) populations. CRAC, Costa Ricans of Afro-Caribbean descent; CRAI, Amerindians from Costa Rica; CRCV, Mestizo Costa Ricans from the Central Valley; CRGU, Mestizo Costa Ricans from Guanacaste province; NICA, Nicaraguan Mestizos; NS, not significant.

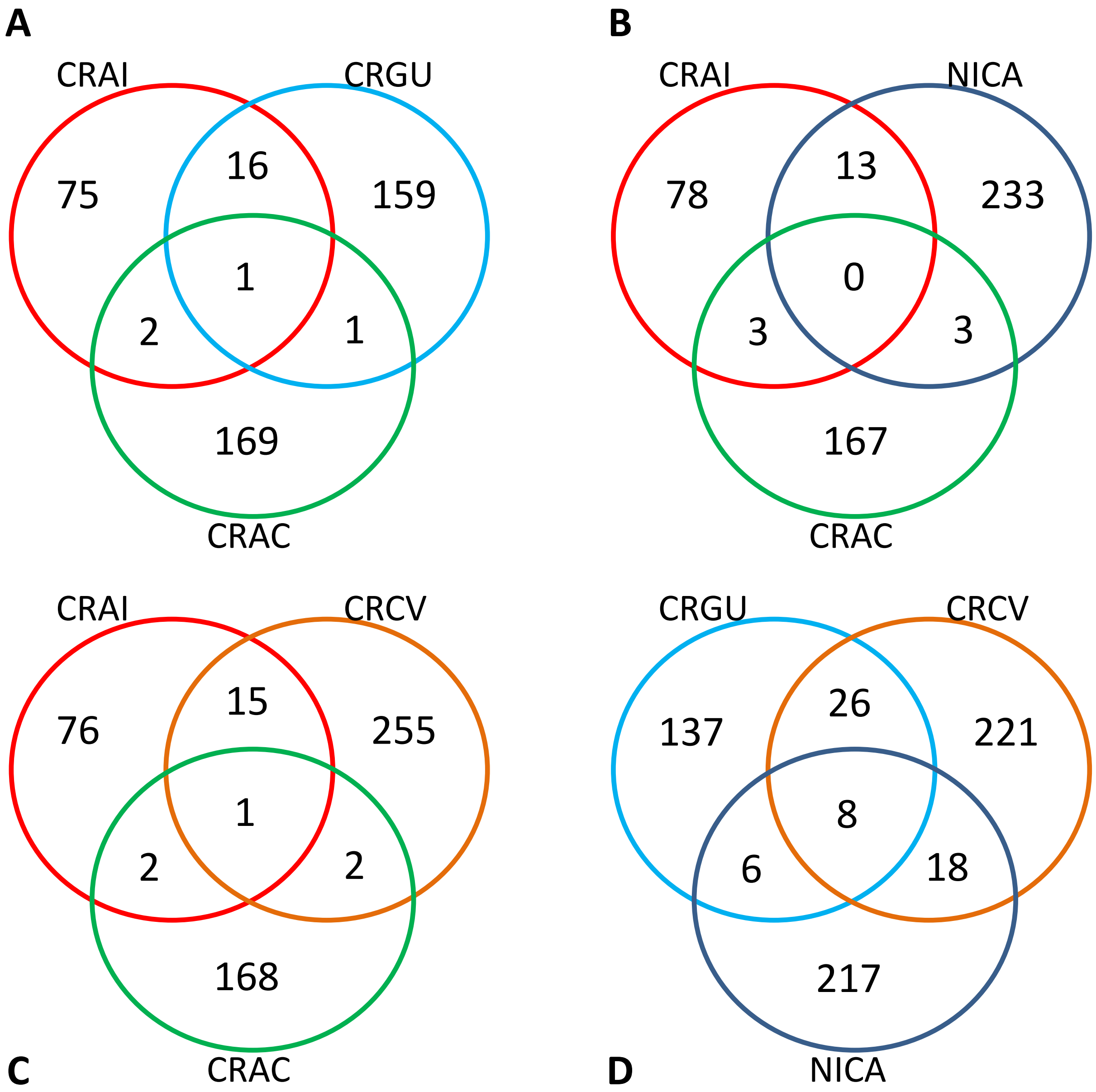
Supplementary Figure 5. **Clustering analysis defines the relationship between the populations of Costa Rica and Nicaragua and indigenous populations from Iberia, America, and Sub-Saharan Africa (SSA).** Clustering analysis based on HLA-A and HLA-B allele group frequencies was performed with 95 populations from Spain and Portugal (Iberians, 35 populations), Amerindians (30 populations), and Sub-Saharan Africans (SSA, 30 populations), as well as the five populations presented in this study. Details for the populations included in this analysis are given in Supplementary Table 1. Data were reduced to first-field (allele group) in order to homogenize the dataset. The results show the split between Amerindians (dark blue box), SSA (red box), and Iberians (light blue box). Populations presented in this study (green boxes) distribute among these main clusters, with Amerindians from Costa Rica (CRAI) and Costa Ricans from Afro-Caribbean descent (CRAC) clustering with their continental counterparts, and Mestizo populations from Costa Rica and Nicaragua (CRCV, CRGU, NICA) forming an outlier of the Iberian cluster. A-C detail of the three major population clusters. Population origin is represented by red diamonds (SSA), light blue squares (Iberians), dark blue circles (Amerindians), and green triangles (populations reported here). The dendrogram is based on minimum variance of squared Euclidean distances, with randomized input order, and unrooted hierarchical agglomerative clustering. X axis has been truncated between the first and second nodes in order to reduce the size of the graph.

[Figure 1]

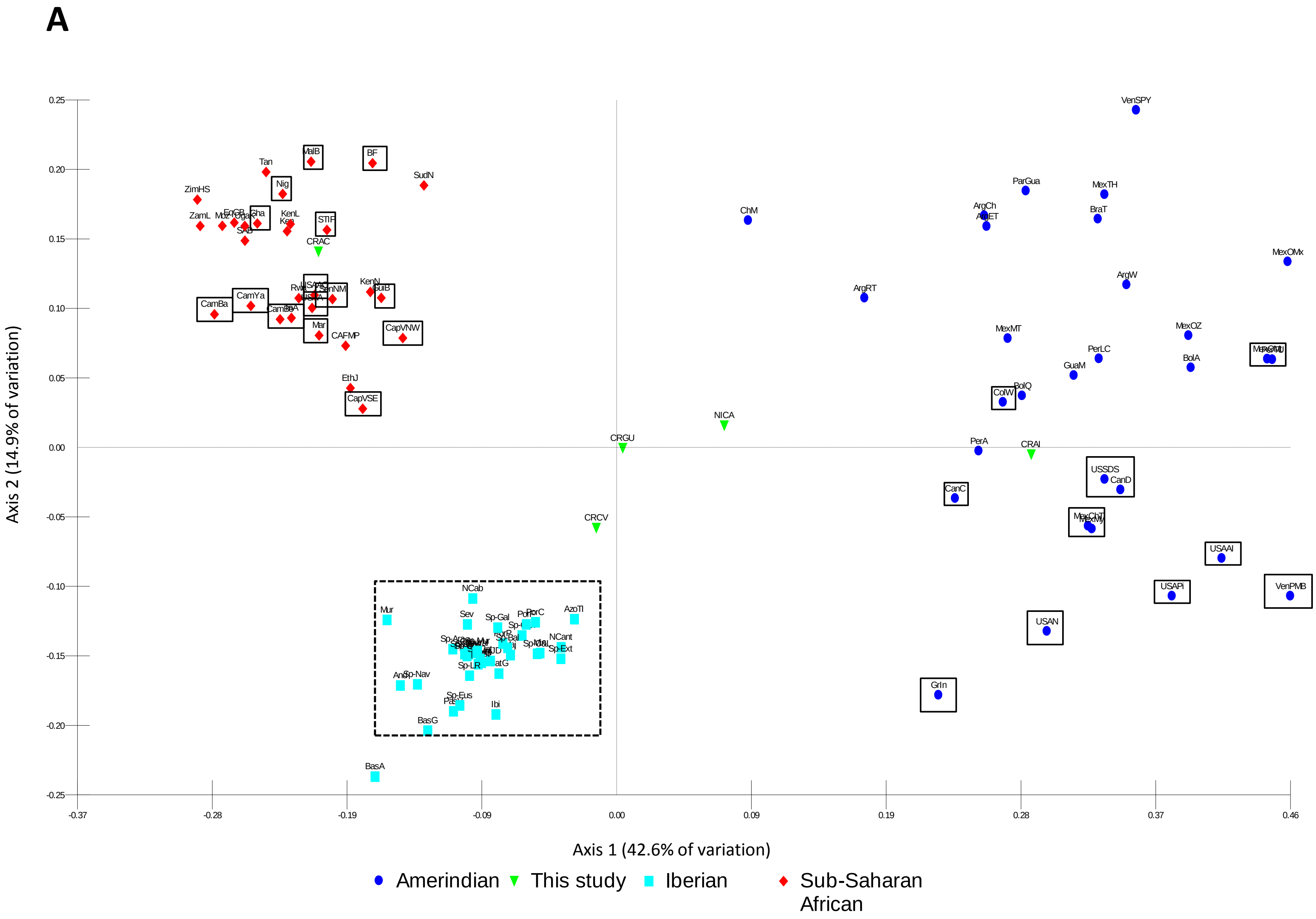


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Cumulative frequency
0
50
100
150
200
250
Haplotype number
CRAC
CRAI
CRCV
CRGU
NICA

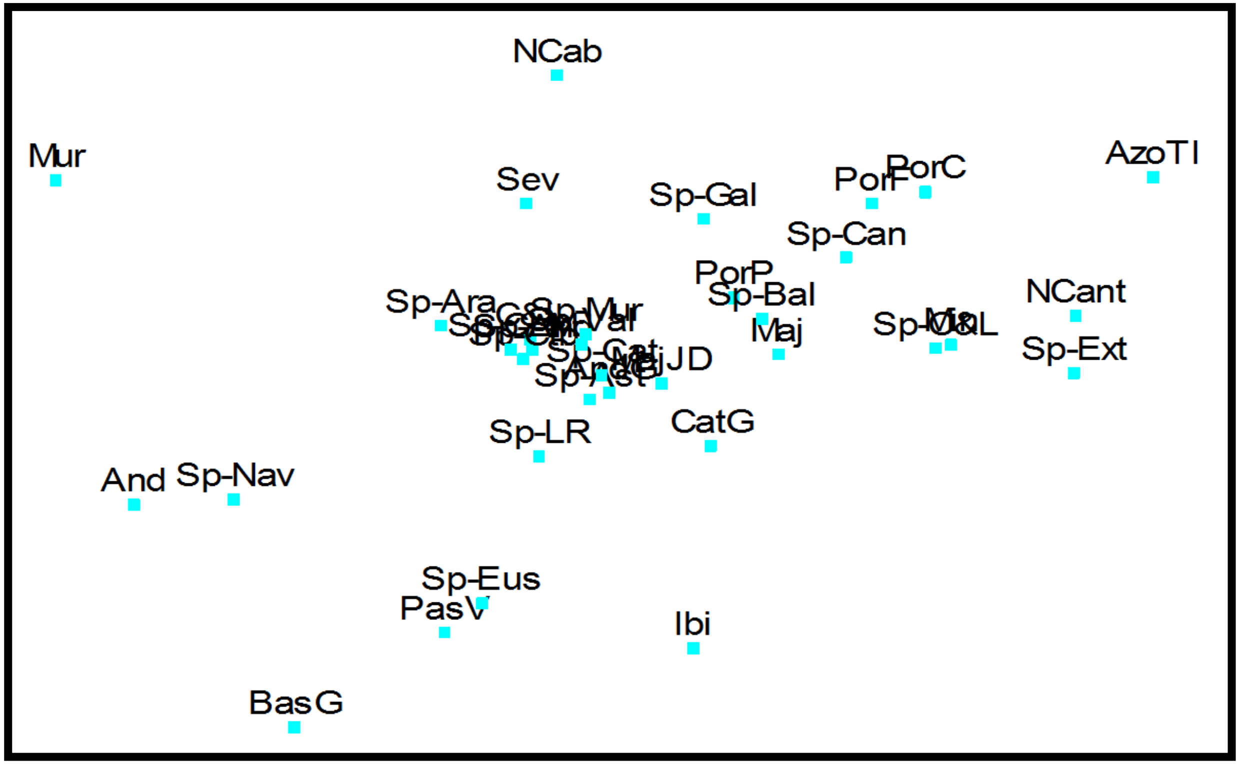
[Figure 2]


A
CRAI
CRGU
75
16
159
1
2
1
169
CRAC
B
CRAI
NICA
78
13
233
0
3
3
167
CRAC
CRAI
CRCV
76
15
255
1
2
2
168
C
CRAC
CRGU
CRCV
137
26
221
8
6
18
217
D
NICA

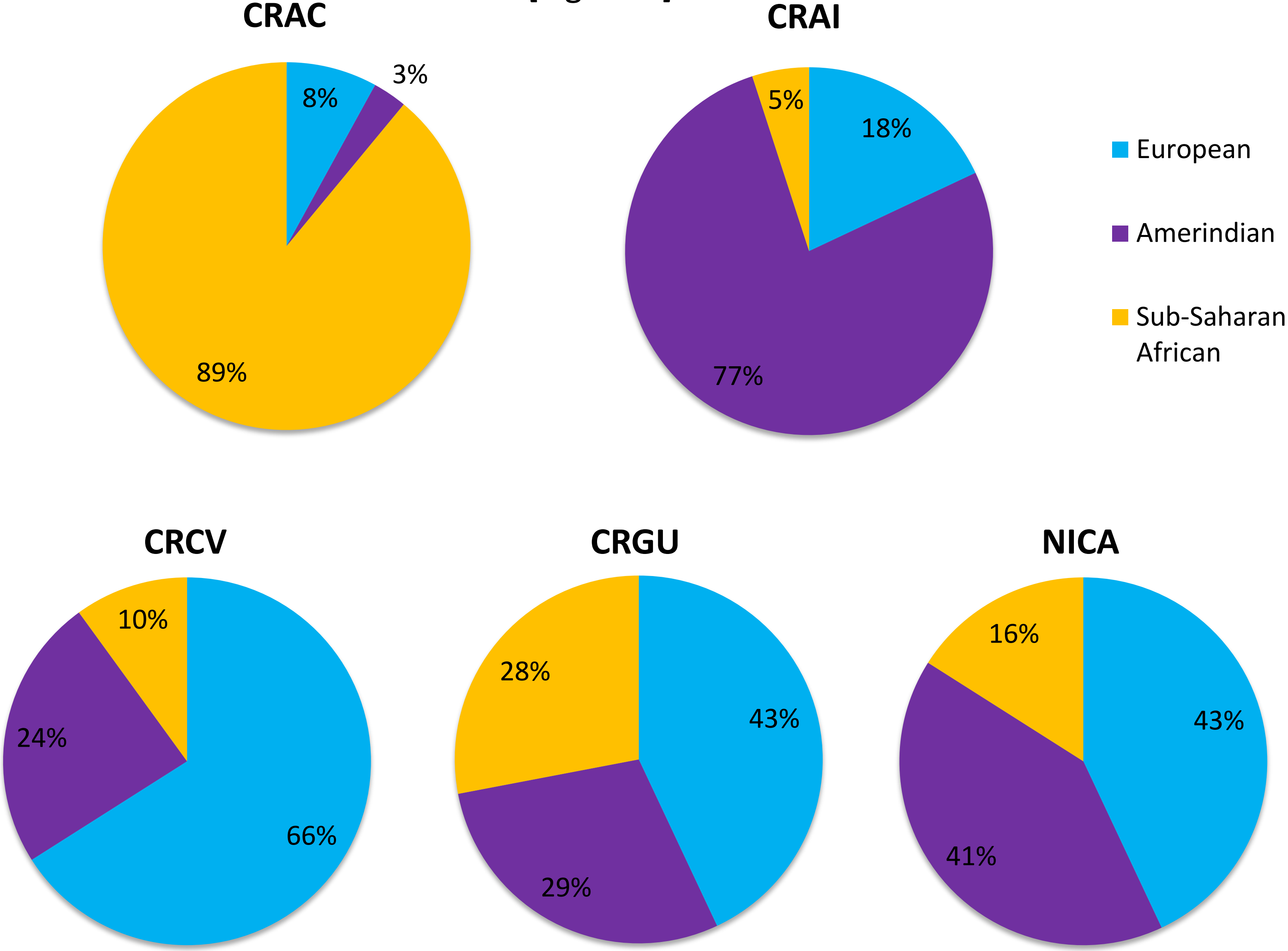
[Figure 3]


A
Axis 2 (14.9% of variation)
Axis 1 (42.6% of variation)
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.37
-0.28
-0.19
-0.09
0.00
0.09
0.19
0.28
0.37
0.46
MalB
BF
Tan
SudN
ZimHS
Nig
ZamL
Moz
EqGB
UgaK
Gha
KenL
Ken
STIF
SAB
CRAC
KenN
Rwa
USAAC
SenNM
GuiB
USAA
CamBa
CamYa
CamBe
BriA
Mar
CAFMP
CapVNW
EthJ
CapVSE
CRGU
NICA
CRCV
CRAI
NCab
Mur
Sev
Sp-Gal
Por
PorC
AzoTI
PorP
Sp-Bal
Sp-Arag
Sp-Mur
NCant
Sp-Ext
Sp-LR
atG
And
Sp-Nav
Sp-Eus
PasV
Ibi
BasG
BasA
VenSPY
ParGua
MexTH
ArgCh
ArgET
BraT
ChM
MexOMx
ArgW
ArgRT
MexMT
MexOZ
PerLC
BolA
GuaM
BolQ
ColW
PerA
USSDS
CanD
CanC
MexChT
MexMy
USAAI
USAPi
VenPMB
USAN
GrIn
● Amerindian
▼ This study
■ Iberian
♦ Sub-Saharan African

B


NCab
Mur
Sev
Sp-Gal
PorF
PorC
AzoTI
Sp-Can
PorP
Sp-Bal
NCant
Sp-Ara
Sp-Mur
Maj
Sp-Ext
Sp-Cat
JD
Sp-Ast
Sp-LR
CatG
And
Sp-Nav
Sp-Eus
PasV
Ibi
BasG

[Figure 4]


CRAC
8%
3%
89%
CRAI
5%
18%
77%
European
Amerindian
Sub-Saharan African
CRCV
10%
24%
66%
CRGU
28%
43%
29%
NICA
16%
43%
41%