



OPEN

## Searching for the roots of the first free African American community

Beatriz Martínez<sup>1,5</sup>, Filipa Simão<sup>2,5</sup>, Verónica Gomes<sup>3</sup>, Masinda Nguidi<sup>2</sup>, Antonio Amorim<sup>3,4</sup>, Elizeu F. Carvalho<sup>2</sup>, Javier Marrugo<sup>1</sup> & Leonor Gusmão<sup>2</sup>

San Basilio de Palenque is an Afro-descendant community near Cartagena, Colombia, founded in the sixteenth century. The recognition of the historical and cultural importance of Palenque has promoted several studies, namely concerning the African roots of its first inhabitants. To deepen the knowledge of the origin and diversity of the Palenque parental lineages, we analysed a sample of 81 individuals for the entire mtDNA Control Region as well as 92 individuals for 27 Y-STRs and 95 for 51 Y-SNPs. The results confirmed the strong isolation of the Palenque, with some degree of influx of Native American maternal lineages, and a European admixture exclusively mediated by men. Due to the high genetic drift observed, a pairwise  $F_{ST}$  analysis with available data on African populations proved to be inadequate for determining population affinities. In contrast, when a phylogenetic approach was used, it was possible to infer the phylogeographic origin of some lineages in Palenque. Contradicting previous studies indicating a single African origin, our results evidence parental genetic contributions from widely different African regions.

San Basilio de Palenque is a small town near Cartagena, Colombia, founded by runaway slaves (Supplementary Fig. S1). At the end of the sixteenth century, African slaves started to escape from the coastal city of Cartagena to take refuge in the nearby region of Montes de María, establishing the foundations of the town of San Basilio de Palenque (hereafter referred to as Palenque). Exactly when the city was founded is unknown, but there are studies indicating that this community was already established in the second half of the seventeenth century and ultimately became the first free African community in America<sup>1,2</sup>.

Due to its strategic position located on the north coast of Colombia, Cartagena city was the centre of the Spanish slave trade and one of the main South American ports of arrival for slaves brought from different regions of Africa. The paucity of historical records makes it difficult to establish the exact place of departure of the slaves from Africa. Nonetheless, it is thought that until the early seventeenth century, the Africans arriving in Cartagena would have left from the region of Upper Guinea. Later, the Congo and Angola, together with Upper Guinea, would have been the major regions from which slaves were taken to Spanish America. At the end of the eighteenth century/beginning of the nineteenth century, slaves would have come from several regions, from Senegambia to Mozambique<sup>3,4</sup>.

The village of Palenque is currently inhabited by approximately 4000 Afro-descendants who maintained a cultural and ethnic identity for more than 3 centuries, with high endogamy and little influence from neighbouring communities<sup>5</sup>. Palenque preserves the ethnic conscience and cultural traits of African roots such as the social organisation, complex funeral rituals, and traditional medical practices, among others<sup>6</sup>. Moreover, it is the only African American population speaking a Creole language with a Spanish lexical base<sup>5,7</sup>. Due to these characteristics, Palenque was declared by UNESCO as a Masterpiece of the Oral and Intangible Heritage of Humanity in 2005. During the last decade, the recognition of the historical and cultural importance of Palenque has promoted several studies with the aim of reviving its history and searching for the African roots of its first inhabitants.

Given the high diversity of slaves arriving in Cartagena, one might suppose that several ethnic groups would be behind the foundation of Palenque. However, this theory has been questioned by linguistic and anthropological evidence, pointing to the region between Congo and Angola (the ancient Kingdom of Kongo) as the origin of the first habitants of Palenque<sup>2,3</sup>, with an almost exclusive contribution from a single Bantu ethnic group, the Bakongo, speakers of Kikongo<sup>5</sup>.

However, cultural and genetic features do not always come together, and few genetic studies have been performed to confirm a single geographic source of the founders of Palenque. The first genetic studies carried out

<sup>1</sup>Molecular Genetics Laboratory, Institute for Immunological Research, University of Cartagena, Cartagena 36-100, Colombia. <sup>2</sup>DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), 20550-900 Rio de Janeiro, Brazil. <sup>3</sup>IPATIMUP/i3S, Instituto de Investigação e Inovação em Saúde, Universidade do Porto, 4200-135 Porto, Portugal. <sup>4</sup>Faculty of Sciences, University of Porto (FCUP), 4169-007 Porto, Portugal. <sup>5</sup>These authors contributed equally: Beatriz Martínez and Filipa Simão. ✉email: [bmartineza1@unicartagena.edu.co](mailto:bmartineza1@unicartagena.edu.co)

with human leukocyte antigen (HLA) markers revealed limited Native American and European gene flow, as well as close genetic distances with African populations, especially from western Africa<sup>8,9</sup>. Nevertheless, a higher than expected European input (38%) was detected when studying the pool of paternal lineages in Palenque<sup>10</sup>. Due to the strong isolation of Palenque, paternal European admixture most likely occurred before its foundation<sup>10</sup>. The results based on autosomal ancestry informative markers showed approximately 10% European ancestry, supporting a sex biased European influx<sup>11</sup>.

More recently, Ansari-Pour et al.<sup>12</sup> investigated uniparental markers from the Y chromosome and mitochondrial DNA (mtDNA). According to these authors, the Yombe from the Republic of the Congo is the most likely group from which the original male settlers of Palenque came.

Taking together all genetic evidence available, it is safe to assume that Palenque has preserved a high African ancestry and that the European background was essentially mediated by males. Nevertheless, the information available is not enough to clearly assign the continental/regional origin of all parental lineages that are present in Palenque. Concerning mtDNA, the results from Hypervariable segment I (HVSI) did not allow us to determine the continental origin in more than 25% of the studied haplotypes<sup>12</sup>.

For the Y chromosome, the available information is also fragmentary. Data available in Noguera et al.<sup>10</sup>, for a small number of samples, do not allow us to evaluate founder effects, given the lack of information on Y chromosome specific short tandem repeats (Y-STRs). Based on the results from Ansari-Pour et al.<sup>12</sup>, it is not possible to quantify non-African influx due to the low resolution of lineages outside haplogroup E1b1a-M2.

Aiming to obtain a deeper knowledge of the diversity and origin of the Palenque parental lineages, in this study, we analysed the entire mtDNA control region (CR), as well as Y chromosome-specific markers. For both maternally and paternally inherited gene pools, we intended to determine the degree of isolation from Native American and European influx. Based on comparisons with data from African populations, we also intended to contrast the hypothesis of a single against a multiple African origin.

## Materials and methods

A total of 95 male children (aged between 5 and 18) were selected for this study. The volunteers identified themselves as having Palenque descent for at least 3 generations (all parents and grandparents born in Palenque). The samples were collected under written informed consent from the guardians of the participants included in the study. The project and informed consent were approved by Act No. 40 of the ethical committee of the University of Cartagena, Colombia; and the ethical principles of the 2000 Helsinki Declaration of the World Medical Association (<http://www.uma.net/e/policy/b3.htm>) were followed. Based on genealogical information, only unrelated individuals (not sharing grandparents) for at least three generations were selected. A total of 48 children were recruited at the Benkos Biojó Rural school in Palenque (PR), and 47 children resided in the urban area of Cartagena city (PU). The children from PU were recruited in schools participating in an ethnopedagogy program created to preserve aspects of Palenque culture such as dance, music, religion, and especially the *Palenquero* language<sup>7</sup>.

DNA was extracted from blood samples using a standard salting-out protocol.

The 95 samples were genotyped for 51 Y chromosome-specific single nucleotide polymorphisms (Y-SNPs) using previously described methods (see details in Supplementary Fig. S2). Ninety-two samples were genotyped for the 27 Y-STR loci included in the Yfiler™ Plus kit, following the manufacturer's protocol (Thermo Fisher Scientific, Waltham, MA, USA). Amplified fragments were separated and detected on a 3500 XL Genetic Analyzer and genotyped using Gene Mapper IDX v.4.0 (Thermo Fisher Scientific).

A subgroup of 81 samples was sequenced for the full control region of mtDNA. The fragments between positions 16024 and 576 were amplified, sequenced, and detected as previously described in Simão et al.<sup>13</sup> using the primers listed in Supplementary Table S1. Haplotypes were classified using SeqScape v2.7 software (Thermo Fisher Scientific). Haplogroups were assigned using both EMPOP database v4/R12<sup>14</sup> and Haplogrep tool<sup>15</sup> and confirmed in Phylotree<sup>16</sup>. Data were submitted to the EMPOP database for quality control checks and are available for research purposes under the accession number EMP00749. Mitochondrial DNA sequences were deposited in GenBank: PopSet 1782793150 (<https://www.ncbi.nlm.nih.gov/popset/?term=1782793150>), accession numbers: MK930265–MK930345.

Haplogroup frequencies were calculated by direct counting. Haplotype and haplogroup diversities, pairwise genetic distances and non-differentiation probabilities were calculated using Arlequin ver. 3.5.1.2 software<sup>17</sup>. Pairwise genetic distances were visualised by multidimensional scaling (MDS) using the software STATISTICA ver.8.0 ([www.statsoft.com](http://www.statsoft.com)). Phylogenetic networks were constructed using Network v10.1.0.0 software (<http://www.fluxusengineering.com>). The number of Y-STRs used to construct the networks depended on the common set available to maximise the representation of African populations for each haplogroup. In most cases, we used a set of 11 loci, namely, DYS389I, DYS389II, DYS19, DYS390, DYS438, DYS392, DYS437, DYS385a/b, DYS393, and DYS439. However, for clades Y-MRCA\* (xM13, SRY10831.1) and E1b1b-M35, DYS391 was also included, since it was genotyped in most studies reporting samples from these haplogroups.

## Results

The mtDNA and Y chromosome haplotypes found in this study are described in Supplementary Tables S2 and S3, along with the corresponding haplogroup classifications. Differences in EMPOP and Haplogrep classifications were observed in only one sample (PU063), which was classified as L2a1 + 16189 + (16192) by Haplogrep and L2a1 by EMPOP. The classification from EMPOP is supported by Phylotree, since PU063 lacks 16189C and has a heteroplasmy at position 16192.

The two subsamples from Palenque (PR and PU) were compared by means of  $F_{ST}$  genetic distances and corresponding nondifferentiation probabilities (after 10100 permutations). Regarding mtDNA haplotypes, no

mtDNA			Y-SNPs		
Haplogroup	n	%	Haplogroup	n	%
L0a1a + 200	1	1.23	Y-MRCA*(xM13, SRY10831.1)	3	3.16
L1b1a + 189	10	12.35	B2a-M150* (xM109)	3	3.16
L1b1a1'4	7	8.64	E1a-M33	2	2.11
L1b1a18	1	1.23	E1b1a-M191	7	7.37
L1b1a7a	1	1.23	E1b1a-M2* (xM154, M191)	36	37.89
L1c3	1	1.23	E1b1b-M35* (xM78, M81, M123, V6, M293)	3	3.16
L1c3a	2	2.47	R1b-V88	4	4.21
L1c3a1b	4	4.94	<b>AFRICAN</b>		<b>61.05</b>
L2a1	2	2.47	E1b1b-M81	2	2.11
L2a1 + 143 + @16309	1	1.23	E1b1b-M123	8	8.42
L2a1 + 16189 + (16192)	20	24.69	G-M201	1	1.05
L2a1c3b2	1	1.23	I2-M26	1	1.05
L2b1a	1	1.23	J2-M172	1	1.05
L2d + 16129	1	1.23	R1a-SRY1831.2	9	9.47
L3d1a1a	11	13.58	R1b-M529	2	2.11
L3e1d	9	11.11	R1b-S116* (xU152, M529, M153, M167)	9	9.47
L3f1b + 16292	1	1.23	R1b-U152	1	1.05
<b>AFRICAN</b>		<b>91.36</b>	<b>EUROPEAN</b>		<b>35.79</b>
A2 + (64)	1	1.23	Q1a2-M3* (xM19, M194, M199)	3	3.16
A2af1a1	2	2.47	<b>NATIVE AMERICAN</b>		<b>3.16</b>
A2a1	1	1.23			
B2d	1	1.23			
C1c3	2	2.47			
<b>NATIVE AMERICAN</b>		<b>8.64</b>			

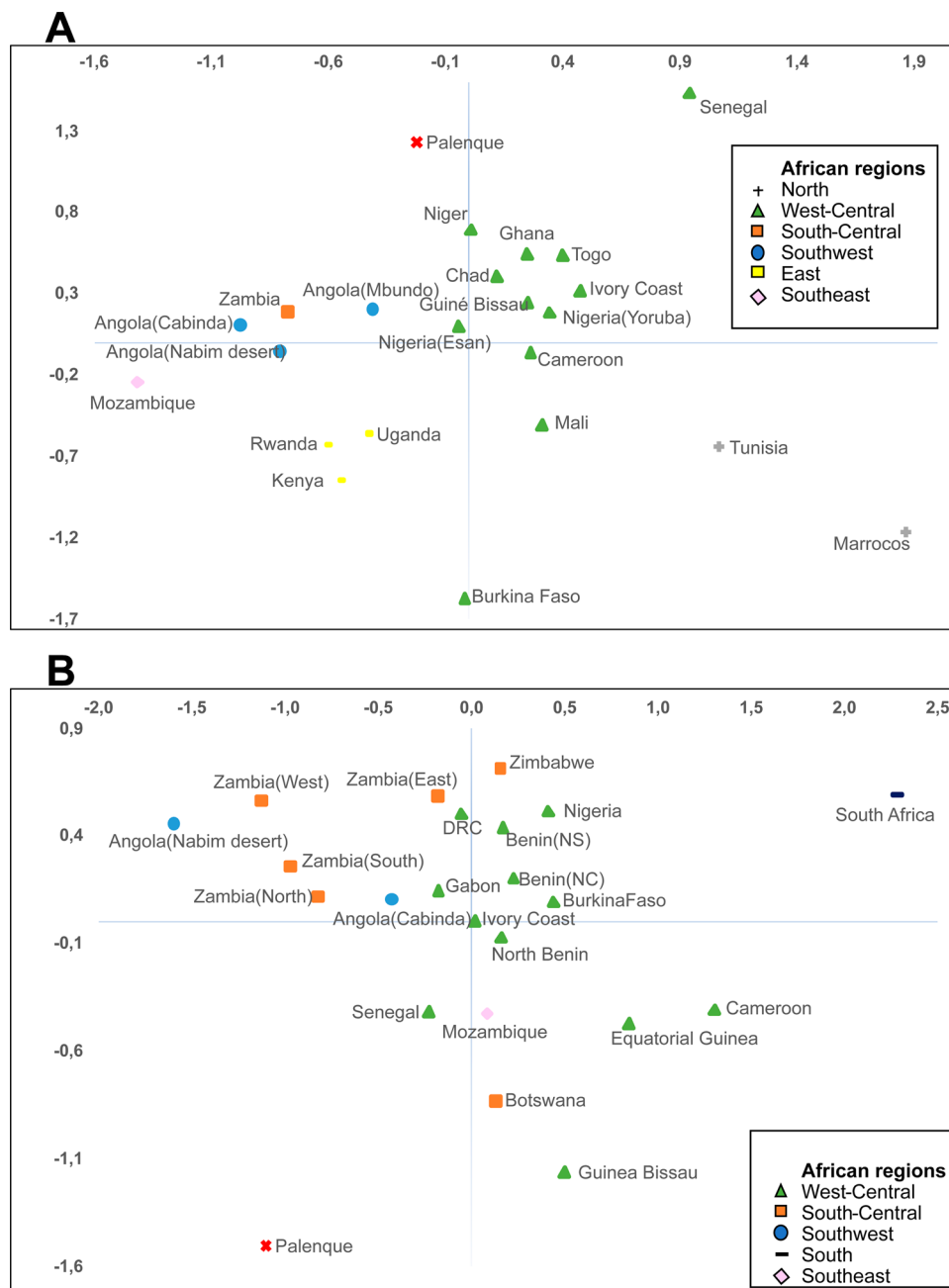
**Table 1.** Frequencies of the mtDNA and Y chromosome haplogroups detected in a population sample from Palenque. Note: The presence of E1b1b-M81 in Palenque is interpreted in this study as the result of European admixture, although it cannot be ruled out that it came from North Africa via western Africa. Although an African origin was considered to be more likely for the E1b1b-M35\* (xM78, M81, M123, V6, M293) lineage in Palenque, we cannot exclude the hypothesis that it came from Europe, since rare E1b1b-M35 subclades were reported in European populations.

statistically significant difference was found ( $F_{ST} = -0.018$ ;  $P = 0.9855 \pm 0.001$ ). For the Y chromosome, the PR subsample showed a lower proportion of African Y-SNP haplogroups than the PU (54.2% and 68.1%, respectively). However, differences between subsamples were not statistically significant for Y-STR haplotype distributions ( $F_{ST} = 0.0041$ ;  $P = 0.1119 \pm 0.0014$ ) or for Y-SNP haplogroups ( $F_{ST} = 0.0163$ ;  $P = 0.0792 \pm 0.0012$ ). Based on these results, samples were pooled for the remaining analysis.

**The genetic diversity of San Basilio de Palenque.** Regarding mtDNA, high haplogroup diversity was found ( $0.8895 \pm 0.0193$ ) for a total of 22 different haplogroups present in our sample. This high haplogroup diversity is, however, associated with a low diversity of haplotypes ( $0.9225 \pm 0.0162$ ). A large proportion of haplotypes were shared inside haplogroups, and only 33 different haplotypes were present in the studied sample. A wide separation of African haplogroups carrying few haplotypes is illustrated in a network (Supplementary Fig. S3). The great majority of the samples (91%) belong to the African macro haplogroup L (Table 1). The remaining 9% belong to Native American haplogroups (A2, A2af1a1, A2a1, B2d and C1c3). No European maternal lineages were observed.

A high diversity was also found for Y-SNP haplogroups ( $0.8185 \pm 0.0033$ ), with 17 different haplogroups being observed (Table 1). Similar to mtDNA, low Y-STR haplotype diversity was observed ( $0.9881 \pm 0.0038$ ), with many shared haplotypes inside haplogroups (Supplementary Fig. S4). In the whole sample, an African origin can be attributed to 61% of the Y-haplogroups, 36% represent European admixture, and three samples (3%) belong to a Native American haplogroup (Table 1).

**The African origins of San Basilio de Palenque lineages.** To investigate the origins of the African lineages in Palenque, we selected subsamples of African haplogroups found for both mtDNA and the Y chromosome. These subsamples were compared with populations from different African regions (see Supplementary Tables S4 and S5). A pairwise comparison was performed among all selected populations. Moreover, phylogeographic analyses of the African haplogroups in Palenque were undertaken based on haplogroup frequencies and distributions across the African continent and considering the haplotypic profile of each lineage.



**Figure 1.** MDS plot of pairwise  $F_{ST}$  genetic distances between the subset of African lineages in Palenque and different African populations, based on (A) mtDNA (HVS1 and HVS2) haplotypes and (B) 10 Y-STRs (DYS389I, DYS389II, DYS19, DYS391, DYS390, DYS438, DYS392, DYS437, DYS393 and DYS439).

**Population pairwise  $F_{ST}$  analysis.** For the mtDNA (HVS1 and HVS2),  $F_{ST}$  genetic distances were calculated between the subgroup of L-lineages in Palenque and haplotype distributions in African populations (see Supplementary Table S4). Relatively high  $F_{ST}$ s and low non differentiation probabilities were observed between Palenque and all African populations used for comparison (see Supplementary Table S6). As depicted in the MDS plot (Fig. 1A), Palenque is far from all populations, although closer to the western ones. A similar picture emerges from the comparative analysis between the Palenque Y-chromosomal African pool and reference populations from Africa (see supplementary Table S5). Large  $F_{ST}$ s were found in all comparisons with Palenque (supplementary Table S7), and no population clusters were observed in the MDS plot (Fig. 1B).

**Phylogeographic analysis of the African mtDNA lineages.** The African lineages in Palenque fall into four clades, L0, L1, L2 and L3, including 17 different haplogroups. The African origin of each haplotype in Palenque was investigated by comparison with data from African populations (Supplementary Table S4).

The southeast and east regions of Africa were identified as the likely origin for the most ancient L0 lineages (such as L0k and L0d)<sup>18,19</sup>. Previous studies show that L0 reaches the highest frequencies and sub-haplogroup diversities in these areas<sup>18</sup>. In contrast, due to a later northeast migration, more recent L0 subclades such as L0a1a show the highest frequencies in central Africa<sup>19–21</sup>. A network built using data available in the literature for L0a1a + 200 shows that none of the African haplotypes match the haplotype from Palenque (Supplementary Fig. S5). The closest haplotypes were from east (Uganda)<sup>22</sup> and west populations (Guinea Bissau and Togo)<sup>23,24</sup>. Considering historical records, an origin of this lineage in the west-central region is more likely than in East Africa.

A total of 19 samples from Palenque had haplotypes classified inside L1b (Table 1). Branches inside L1b are spread throughout Africa, although the branches are more frequent in the west-central region<sup>20,25</sup>. Both L1b1a + 189 and L1b1a18 branches were found in several regions throughout Africa, as shown in a network (Supplementary Fig. S6). Due to their broad distribution, it was not possible to point to the origin of these lineages in Palenque. L1b1a7a was not found in the collected population data, and L1b1a1'4 was found in one sample from Guinea Bissau and one from Senegal<sup>23,26</sup>. The two L1b1a1'4 haplotypes in Palenque differ by a single mutation, and the most frequent one is separated from the Guinea Bissau and Senegal haplotypes by one step (Supplementary Fig. S6). A search in EMPOP (empop.online, release v4/R13)<sup>14</sup> showed the presence of 13 and 25 samples from L1b1a1'4 and L1b1a7a, respectively, most from admixed populations in America. Apart from the presence of L1b1a1'4 in Guinea Bissau, these haplotypes could not be found in other African populations in the EMPOP.

The territory between Cameroon and Angola has been identified as the most likely origin for L1c<sup>27,28</sup>. Additionally, in this region, this haplogroup currently has the highest frequencies. L1c was observed in 7 samples from Palenque, distributed among three sub-haplogroups: one sample is L1c3, one is L1c3a and four are L1c3a1b. Additionally, one haplotype showed a point heteroplasmy at position 16355, necessary for haplogroup classification: a C at position 16355 corresponds to L1c3; a T variant puts the sample in L1c3a1b (Supplementary Table S2). A network was built for haplogroups L1c3, L1c3a and L1c3a1b (Supplementary Fig. S7). Approximately, 94% of the African samples are from south central, southwest, or west-central populations. Only a small number were from the east or southeast. In accordance with previous studies<sup>25,28</sup>, L1c revealed high internal haplotype diversity, preventing a detailed investigation of the origin of the Palenque lineages. Nonetheless, three out of five L1c3a1b samples from Palenque shared a haplotype with a sample from Ghana<sup>29</sup> (Supplementary Fig. S7). This haplotype is one mutational step apart from two Bantu samples from Angola<sup>30</sup> and two mutational steps from one sample from Cameroon<sup>18</sup> (Supplementary Fig. S7). Consequently, an origin of the Palenque lineages in the western region (central and/or south) seems more likely than in the eastern or north regions.

Haplogroup L2, which has been associated with Bantu expansion, has its most likely origin in western Africa. Currently, L2 is spread throughout the continent, presenting high frequencies in central west and southeast Africa<sup>25,31</sup>. Three (L2a, L2b and L2d) of the five main branches from L2 (L2a–L2e) were found in Palenque (Table 1). L2a is the most common and widespread clade in sub-Saharan Africa, and was described with the highest frequencies in Ghana, Sudan and Mozambique<sup>25,31</sup>. Approximately 92% (n = 24) of the L2 found in Palenque was attributed to branches inside L2a. In the networks of L2a1 (Supplementary Fig. S8) and L2a1 + 16189(16192) (Supplementary Fig. S9), it is possible to see a wide geographic dispersion, typical of L2a lineages. In both cases, a star-like distribution emerges from a central cluster of shared haplotypes from several African regions. The two L2a1 samples from Palenque are one and/or two mutational steps from the central cluster. All haplotypes from L2a1 + 16189(16192) are associated with a single founder who shares its sequence with samples from several African regions. The only L2a1 + 143 + 16309 haplotype from Palenque is separated by one position from west, southwest and east Africa haplotypes (Supplementary Fig. S10). A L2a1c3b2 sample from Palenque differed from the only African sample available for this haplogroup<sup>32</sup> in 5 polymorphic sites.

In summary, inferences concerning the African origin of Palenque haplotypes inside the L2a clade were not possible due to its wide distribution in Africa, as well as because of the uncertainty in subbranch classification associated with rapidly mutating polymorphisms<sup>25,31</sup>.

The L2b and L2d subbranches are less frequent in Africa than L1a<sup>25,27,31</sup>. Both haplogroups originated in western Africa, showing the highest frequencies in west-central and southwest populations. Although L2b1a is characterised by the 146T and 16355T polymorphisms, the L2b1a sample from Palenque lacks the 146T mutation. No samples in the literature were found to have this exact profile. In the network, the haplotype from Palenque differs from the central cluster by two mutational steps (the first one being 146) (Supplementary Fig. S11). Comparably, one sample from Cameroon<sup>24</sup> also diverges from the central cluster by a polymorphism at position 146.

L2d is the oldest L2 subclade<sup>25</sup> and shows a sporadic occurrence on the African continent. The L2d + 16129 haplotype from Palenque matches the single L2d + 16129 sample from Angola (Bantu)<sup>30</sup> (Supplementary Fig. S12). The widespread distribution previously reported for other L branches is once again observed in the L2b and L2d sub-lineages. Nevertheless, a clear southwest origin for the L2d haplotype in Palenque can be demonstrated. Regarding the L2b haplotype in Palenque, data place the most likely origin in the western region (between the Bight of Benin and Angola), while east and southeast origins can possibly be excluded.

L3 is thought to originate in east Africa, the homeland for out-of-Africa migration and diversification, where it reaches the highest frequency<sup>21,25</sup>. Nonetheless, L3 subbranches can also be found in other African regions. Approximately 28% of the L-lineages found in Palenque are distributed across three L3 branches: L3d, L3e and L3f.

The L3d branch is very common in west-central Africa<sup>25</sup>, even though it is also present in other regions such as the east and southeast. Although widely spread, L3d1a1a shows low diversity, with just two haplotypes described in the literature<sup>18,30,33</sup> (Supplementary Fig. S13). The eleven L3d1a1a samples from Palenque have the same haplotype. It was nevertheless impossible to infer the exact region of Africa where this haplotype originated, since this haplotype (i) was not found in the African samples, and (ii) differs by one step from a central cluster that includes samples from several regions (Supplementary Fig. S13).



Compared to other L3 subclades, L3e is the most frequent and widely distributed in the African continent. It is especially common in Bantu groups<sup>25</sup>. The L3e1d network shows an almost exclusive distribution in southwest<sup>18,30</sup> and south-central Africa<sup>34,35</sup> (Supplementary Fig. S13), indicating a southern origin for this lineage in Palenque. All samples from Palenque share the same haplotype (apart for one variant of a heteroplasmic sample), suggesting a single founder.

Haplogroup L3f is rare in Africa<sup>25</sup>. L3f originated in the eastern region and later expanded to central Africa. This branch is also present in west-central Africa<sup>25</sup>. Although no matching haplotypes were found with the L3f1b + 16365 sample from Palenque, the closest one was from Ghana (Supplementary Fig. S13).

**Phylogeographic analysis of the African Y chromosome lineages in Palenque.** The African paternal lineages found in Palenque fall into four different clades: Y-MRCA\* (xM13, SRY10831.1), B, E and R, comprising seven haplogroups. The most likely origins of these haplogroups were investigated by comparing the Y-STR haplotype profile of samples from Palenque and African populations (Supplementary Table S5).

In this study, we detected three chromosomes classified as Y-MRCA\* (xM13, SRY10831.1), which includes samples from haplogroup A. Haplogroup A is virtually restricted to the African continent, reaching the highest frequencies in Khoisan-speaking populations. It is also frequent in the Nilotic groups from east and northeast Africa<sup>23,36–38</sup>, and it was sporadically observed in the southeast and southwest populations<sup>39,40</sup>. Some lineages inside clade A have geographic specificity. For instance, lineages belonging to haplogroups A1-M31 were described in west Africa<sup>41</sup>, differing from those in the eastern part of the continent that belong to haplogroups A3-M13 or to other lineages inside A3.

To infer the most likely origin of the haplogroup A samples found in Palenque, a network was built, including the Y-STR profiles of two samples from this study (we had no Y-STR information for one sample) and those from 109 haplogroup A samples compiled from the literature (Supplementary Table S5). The network constructed depicts a clear separation of samples belonging to haplogroups A1-M31 (from Guinea Bissau, in west Africa), A3-M13 (including all samples from East Africa, some from northeast Africa, and one sample from Benin and one from Equatorial Guinea) and A3-M28 (including four samples from northeast Africa) (Supplementary Fig. S14). Central African Bantu and Pygmy samples from Ghana and Cameroon, not typed for Y-SNPs inside haplogroup A [Y-MRCA (xA4)], separate into two clusters, not overlapping A1-M31, A3-M28 and A3-M13 clusters. The two samples from Palenque (which share the full Yfiler Plus profile) stand between the A1-M31 and A3-M28 clusters. Although it is difficult to point to an origin of the haplogroup found in Palenque, we can at least exclude that they came from east, southeast, or south African regions. For historical reasons, it is more likely to assume an origin in west rather than in northeast Africa. The closest west African branches to the Palenque haplotypes correspond to samples from Guinea Bissau, belonging to haplogroups A1-M31.

Haplogroup B is present exclusively in sub-Saharan African populations, except for recent migration to other continents. Haplogroup B2b-M112 chromosomes are found at their highest frequency among Pygmies and are also frequent in Khoisan, with some lineages being virtually restricted to these hunter-gatherer groups<sup>42,43</sup>. In contrast, subclade B2a is widely dispersed throughout sub-Saharan Africa<sup>44,45</sup>. The presence of this haplogroup in different African regions was first attributed to dispersion of Bantu speakers. Nevertheless, recent studies showed that haplogroup B2a was already present in Khoisan groups before their contact with Bantu-speaking populations<sup>43,45</sup>. We identified three samples belonging to B2a-M150\* in Palenque (5.17% of the African lineages), all lacking the M109-derived allele usually present in southwest, southeast and south African Bantu populations<sup>40,46,47</sup>. Subclade B2a is widely dispersed throughout sub-Saharan Africa<sup>44,45</sup>. The network shows a wide dispersion of the haplotypes (Supplementary Fig. S15), which can be due to the low resolution of some lineages. Aiming at broad population coverage, we have included all possible B2a-M150\* (xM109) chromosomes, even those that were typed only for the haplogroup B diagnostic SNP. Only 33 of the 134 African reference samples included in the network were typed for M109: 3 from Eritrea, 25 from Uganda, 4 from Nigeria and 1 from Ghana. The samples from Palenque are positioned outside the two main clusters in a branch rooted in a sample from Gabon. Two samples share their haplotypes with a sample from Cabinda (Angola), and the third sample differs by a single step mutation. The results obtained are thus compatible with a single founder, who could have been brought from the ports of Loango (in modern Gabon) or Cabinda (Angola), in the ancient Kingdom of Kongo.

Clade E is the most common clade in Africa and is also present in European and western Asian populations. In accordance with the Palenque African background, this clade represents 61.05% of the Y-lineages and 82.76% of those assigned to an African origin. The haplogroup E1b1b-M81 is frequent in north Africa, particularly in Berber groups and is also present in Iberia and southern Italy as a northwest African legacy<sup>48,49</sup>. Since no direct gene flow from north Africa has been reported, the presence of this haplogroup in the Palenque population most likely resulted from European admixture. Nonetheless, a northern African origin of the E-M81 in Palenque, via western Africa, cannot be ruled out. A European origin was also attributed to the haplogroup E1b1b-M123 lineages found in Palenque. This haplogroup is absent in Sub-Saharan Africa and frequent in Arabian populations from northeast Africa and the Middle East<sup>50</sup>. This haplogroup also spread in Europe, although at low frequencies. In Iberia, it has been associated with Middle Eastern, North African, and Jewish ancestry. A direct origin in Africa can be attributed to the remaining E-haplogroups in Palenque [E1a-M33, E1b1a-M2\* (xM154,M191), E1b1a-M191 and E1b1b-M35\* (xM78,M81,M123,V6,M293)], which are not present in European populations and were more likely brought to America during the slave trade.

Apart from E1b1b-M81 and E1b1b-M123 (attributed to European influx), inside the E1b1b clade, we found three samples belonging to E1b1b-M35\* (xM78,M81,M123,V6,M293). Although an African origin of this lineage is more likely, we cannot exclude the hypothesis that it came from Europe, as rare E1b1b-M35 subclades

(harbouring V1515, V257 or V2009 mutations) were detected in populations in the western Mediterranean region, namely, Portugal, Spain, France and Italy<sup>51</sup>.

E1b1b-M35\*(xM78,M81,M123,V6,M293) is present in eastern and southern African populations<sup>37,47,48</sup>. It was not detected in central and southwest regions, namely, Niger, Nigeria, Cameroon, Equatorial Guinea, Gabon and Angola<sup>30,42,46–48,52</sup>, although it was described in west Africa at low frequencies. The E1b1b-M35 network shows many separated branches (Supplementary Fig. S16), some of which are restricted to a single African region. The high diversity and wide dispersal in African populations is explained by the low resolution of the data available and most likely represents diverse E1b1b-M35 sublineages. From the network, we can see that the samples from Palenque (represented by a single haplotype) are positioned in a branch that is rooted in a haplotype shared between a sample from Burkina Faso and Ethiopia. Apart from the samples from Palenque, this branch comprises one sample from Senegal, one from Benin, three from Guinea Bissau and two from Burkina Faso. The results obtained are thus compatible with a single founder who is more likely to have come from a region between Senegambia and the Bight of Benin than from other African regions.

Two samples from haplogroup E1a-M33 were found in Palenque, corresponding to 3.45% of the African chromosomes. This haplogroup is present exclusively in west Africa. The highest frequencies reported were in the region of Mali and Burkina Faso<sup>38,47</sup>, with a gradual decrease in frequency towards the south. It was not detected in eastern and southern populations<sup>23,41,47,53</sup>. A network was built using available information on 77 samples from haplogroups E1a-M33 (Supplementary Fig. S17). The two samples from Palenque are well apart from each other. One is in a branch together with samples from Guinea Bissau. This branch is rooted in a cluster mainly comprising samples from Burkina Faso. The other sample is positioned in the network close to a group of samples from Benin and Nigeria. The results obtained are thus compatible with two different origins in western Africa. One of the Y chromosomes is more likely to have come from the region of Upper Guinea. The origin of the other Y chromosome is more likely to be somewhere along the Bight of Benin.

The haplogroup E1b1a-M2 (and its sub-lineages) is widely spread in Africa and highly prevalent in all Bantu sub-Saharan populations, with frequencies above 80% in most populations<sup>39,40,46,47</sup>. In Palenque, 45.26% of the samples (74.14% of the African samples) belong to the haplogroup E1b1a-M2\*(xM154,M191) and to its sub-clade E1b1a-M191. The extremely high number and diversity of samples that can be included in these haplogroups did not allow us to resolve the reticulation obtained in the networks (data not shown). Therefore, for these two haplogroups, we carried out a match analysis between the haplotypes found in Palenque and those reported in African populations from the same haplogroups.

A total of 2204 samples, selected from haplogroup E1b1a-M2 (excluding those carrying M154- or M191-derived alleles), were compared with the 35 E1b1a-M2\*(xM154,M191) samples from Palenque (Supplementary Table S8). These 35 samples resulted in 9 different clusters, including identical and neighbouring haplotypes (Supplementary Fig. S4). No matches were found for five out of the 9 clusters (Supplementary Table S8); therefore, it was not possible to infer their origin. For two clusters, matching haplotypes were only found in Benin, with seven and one samples, respectively (Supplementary Table S8), placing the most likely origin of these lineages in the region of the Bight of Benin. For the two remaining clusters, identical haplotypes were found mainly in populations south of Cameroon; therefore, these lineages most likely came from the African west coast between Loango and Angola.

From E1b1a-M191, we selected 934 samples from 20 countries in sub-Saharan Africa to be compared with the seven E1b1a-M191 samples found in Palenque (Supplementary Table S9). Three of these samples shared the same haplotype (Supplementary Fig. S4). Exact matches were found for only two of the five different haplotypes. For one haplotype, a match was found in Gabon. The second haplotype has four matches, one in Gabon, two in Angola and one in Zambia, placing the most likely origin of these lineages in a region south of Cameroon, possible between Loango and Angola.

The haplogroup R1b-V88 was found in 6.15% of the African lineages in Palenque. Clade R originated outside Africa and has high frequencies in European populations. However, except for rare sub-lineages, the R1b-V88 sub-clade is essentially restricted to the African continent, with high frequencies in the central Sahel populations<sup>54</sup>. In sub-Saharan Africa, this haplogroup has the highest frequencies in the northern Cameroon Chadic groups<sup>55</sup>. In western Africa, this haplogroup was also found in Gabon and Equatorial Guinea and is absent in the southern populations<sup>42,53,55</sup>. Based on its geographic distribution and diversity levels in different regions, haplogroup R1b-V88 has been associated with the trans-Saharan spread of proto-Chadic populations during the early mid-Holocene<sup>54,55</sup>. A network was built using available information on 56 African samples from haplogroup R after excluding samples that were assigned to sublineages outside the R1b-V88 branch and believed to result from recent European migration (Supplementary Fig. S18). All four samples from Palenque share a Y-STR haplotype, which is located close to four out of the seven R1b-V88 chromosomes found in the Punu people from Gabon (two are separated by a single-step mutation and two by two-steps). In the same cluster is also one sample from Equatorial Guinea and another from Benin (both are two steps apart from the Palenque samples). The results obtained are compatible with a single founder from western Africa. Close haplotypes were found spread along the Bights of Benin and Biafra and along the Loango coast. The closest genetic proximity was found with two samples from Punu-speakers, putting their most likely origin in the current region of Gabon.

## Discussion

**Ancestry profile of Palenque.** Although previous studies have been undertaken to establish maternal and paternal ancestral contributions to the current population of Palenque, only rough estimates are available due to poor data resolution. In the present study, we analysed numerous Y chromosome-specific markers and extended the mtDNA analysis to the whole CR, enabling a clear assignment of the continental origin of the lineages in Palenque.

A high African maternal ancestry was found, in accordance with the report by Ansari-Pour et al.<sup>12</sup>. Nevertheless, it was not possible to perform a straightforward comparison between the two studies, since previous data included only HVSI information, and many samples could not be assigned to a specific haplogroup, namely, inside some African and Native American branches. Even so, a comparison between HVSI haplotypes showed no statistically significant differences between the two studies ( $F_{ST} = 0.0062$ ;  $P = 0.1278 \pm 0.004$ ).

In contrast with mtDNA, a high influx of European males was observed, and only three out of the 95 Y chromosomes have a Native American origin. Although a previous study reported an even higher European male lineage input (38.5%), a European origin was attributed to all R1b lineages<sup>10</sup>. However, in the present study, we found samples in Palenque that belong to an African R1b subhaplogroup, characterised by the V88-derived allele, not investigated in the previous study of Noguera et al.<sup>10</sup>. Due to different resolutions in the haplogroup definition, we could not make a meaningful comparison between our data and that from Ansari-Pour et al.<sup>12</sup>. In comparison with this study, Ansari-Pour et al.<sup>12</sup> performed a more detailed characterisation of the E1b1a-M2 sublineages (including the U175, U181 and U290 markers) but did not distinguish between African, European or Native American haplogroups inside BF-M213\* (xM9) and QR-92R7\* (xSRY10831.2).

The overall results show that non-African admixture was mainly mediated by European males, responsible for the introduction of at least 9 different haplogroups present in our sample, and no admixture with European women was detected. Native American influx was higher for the maternal than for the paternal lineages. The three Native American Y chromosomes detected could be explained by a single entry, since their haplotypes differ only at DYF387S1, which is known to have a higher than average Y-STR mutation rate ( $> 1 \times 10^{-2}$ )<sup>56</sup>.

**African roots of Palenque.** One of the most debated subjects regarding the history of Palenque has been the origin of its inhabitants and, hence, the question of whether their ancestors had a single or a multiple origin in Africa. The hypothesis of a single Bakongo origin of the Palenque founders, based on linguistic evidence, showed support for Y-chromosomal diversity, as reported in a previous study<sup>12</sup>. In the study from Ansari-Pour et al.<sup>12</sup>, a comparative analysis between Palenque and populations representing different sub-Saharan African groups showed a close  $F_{ST}$  genetic distance with the Yombe group from the Republic of the Congo for the Y chromosome. However, regarding mtDNA lineages, no evidence was found concerning maternal African roots<sup>12</sup>. In contrast, our results based on pairwise genetic distances did not allow us to place the origin of the Y-haplotypes from Palenque in a specific African region. An AMOVA performed between Palenque in one group and all African populations in a second group revealed a higher variation between Palenque (subsample of African lineages) and the African group (1.76%) than among populations from different African regions (0.88%). The large  $F_{ST}$  between the African substrate of Palenque and all African populations can only be explained by genetic drift, which is in accordance with the low haplotype diversity within Y chromosome haplogroups. Regarding mtDNA lineages in Palenque, similar to the observations by Ansari-Pour et al.<sup>12</sup>, it was not possible to establish maternal African roots based on pairwise genetic distance analyses.

Overall, diversity and genetic distance results from both mtDNA and Y chromosome data underline the importance of genetic drift in the separation of Palenque from its homeland population(s). Therefore, the analysis of genetic distances based on allele frequency distributions is not the best strategy to look for affinities in Africa. In this situation,  $F_{ST}$  genetic distance or principal component analysis will reflect random drift rather than distant genetic affinities. A phylogeographic approach of each of the lineages present in the Palenque is therefore more suitable for the search for their origins in different African regions.

Among the 79 samples from Palenque, at least 17 potential mtDNA founders could be identified (Supplementary Fig. S3). The comparison with data from Africa allowed us to place the most likely origin of seven of these founders (Fig. 2). For almost 70% of the African mtDNA sequences in Palenque, it was not possible to infer the origin. Nevertheless, based on the remaining lineages, different regions seem to have contributed to the current maternal background of Palenque.

Among the 58 samples from Palenque attributed to African Y-haplogroups, at least 20 potential founders could be identified based on haplotype diversity. The comparison with data from Africa allowed placing the most likely origin of 12 out of the 20 potential founders. The results pointed to multiple origins in western Africa in a territory extending from Upper Guinea to Angola (Fig. 2). Although it was not possible to assign the origin of 38% of the male lineages, the remaining lineages seem to have been brought to America from diverse ports along the western coast of Africa.


The detection of lineages originating mainly from the west coast of Africa, as well as their variety of origins, reflects the existing information on the arrival of slave vessels in Cartagena. According to the Slave Voyages database (<https://www.slavevoyages.org/>), ships from a wide variety of ports in west Africa, from Senegambia to Angola, arrived in Cartagena during the slave trade.

Although the mtDNA for the full control region was analysed in this work, in comparative analyses, we could use only HVSI and HVSII information to comprise the most relevant African regions. For the same reason, only a small set of all genotyped Y-STRs could be used. Moreover, there is little overlap among publications concerning the Y-SNPs used for haplogroup determination. While affinities were found between Palenque and some African regions, it was not possible to pinpoint the origin for many lineages, and some were traced to a vast region, preventing the estimation of the exact proportion in which different African regions contributed to Palenque. The phylogeographic approach used is strongly influenced by the knowledge of the genetic diversity of the original populations, being conditioned by the existence of data with high resolution. However, the gene pool of sub-Saharan African populations is still understudied, and some relevant areas in the history of the trans-Atlantic slave trade have been poorly investigated thus far.

In conclusion, the results from this study showed that the African lineages in Palenque resulted from a restricted number of founders from multiple geographic origins. The results do not contradict the important



	Haplogroup	n	Freq.	Most likely African origin
mtDNA	L0a1a+200	1	1.3%	Between Upper Guinea and Bight of Benin
	L1b1a1'4	7	8.9%	Between Senegambia and Upper Guinea
	L1c3a1b	5	6.3%	Between Gold Coast and Angola
	L2b1a	1	1.3%	Bight of Benin and Angola
	L2d+16129	1	1.3%	Angola
	L3e1d	10	12.7%	Angola
	L3f1b+16365	1	1.3%	Gold Coast
	Other samples	53	67.1%	Unknown
Y chromosome	Y-MRCA* (xM13,SRY10831.1)	2	3.4%	Upper Guinea
	B2a-M150* (xM109)	3	5.2%	Loango
	E1a-M33	1	1.7%	Upper Guinea
	E1a-M33	1	1.7%	Between Bight of Benin and Bight of Biafra
	E1b1a-M191	1	1.7%	Between Loango and Angola
	E1b1a-M191	1	1.7%	Loango
	E1b1a-M2* (xM154,M191)	13	22.4%	Bight of Benin
	E1b1a-M2* (xM154,M191)	7	12.1%	Between Loango and Angola
	E1b1b-M35* (xM78,M81,M123,V6,M293)	3	5.2%	Between Senegambia and Bight of Benin
	R1b-V88	4	6.9%	Between Bight of Benin and Loango
	Other samples	22	37.9%	Unknown



**Figure 2.** Most likely African origin of the mtDNA and Y chromosome lineages detected in Palenque.

influence of the Kikongo speakers in the early Palenque, as largely documented and patent in the *Palenquero* language. Nevertheless, this results show that there are still lacunae in the history of the Palenque people, who seem to carry in their genes other roots beyond the Kikongo. The results are compatible with the presence of several African substrates in the early inhabitants of Palenque, although one has been dominant in terms of culture and language. Concerning the European paternal legacy, it remains to be investigated at which point in the history of the Palenque these lineages would have been introduced.

Received: 21 August 2020; Accepted: 13 November 2020

Published online: 26 November 2020

## References

- Peláez, M. C. N. *San Basilio de Palenque: Memoria y Tradición: Surgimiento y Avatares de las Gestas Cimarronas en el Caribe Colombiano* (Programa Editorial Universidad del Valle, Valle del Cauca, 2008).
- Arrazola, R. *Palenque, Primer Pueblo Libre de America: Historia de las Sublevaciones de los Esclavos de Cartagena*. (Ediciones Hernandez, 1970).
- Friedemann, N. S. de. San Basilio en el universo Kilombo-África Y Palenque-America. Tomo VI. In *Geografía humana de Colombia—Los Afrocolombianos* 63–83 (Instituto Colombiano de Antropología e Historia, 1998).
- Borucki, A., Eltis, D. & Wheat, D. *From the Galleons to the Highlands: Slave Trade Routes in the Spanish Americas* (University of New Mexico Press, New Mexico, 2020).
- Schwegler, A., Kirschen, B. & Maglia, G. *Orality, Identity, and Resistance in Palenque (Colombia)—An interdisciplinary approach* (John Benjamins Publishing Co, Amsterdam, 2017).
- Mosquera, C., Pardo, M. & Hoffmann, O. *Afrodescendientes en las Américas: Trayectorias Sociales e Identitarias* (Universidad Nacional de Colombia, Colombia, 2002).
- Schwegler, A. Combining population genetics (DNA) with historical linguistics. In *Spanish Language and Sociolinguistic Analysis* 33–88 (2016).
- Jiménez, S. *et al.* Análisis Inmunogenético y antropológico de la población del Palenque de San Basilio (Colombia). In *Polimorfismo génico (HLA) en poblaciones hispanoamericanas* 247–269 (Real Academia de Ciencias Exactas, Físicas y Naturales, 1996).
- Arnaiz-Villena, A. *et al.* HLA genes in Afro-American Colombians (San Basilio de Palenque): The first free Africans in America. *Open Immunol. J.* **2**, 59–66 (2009).
- Noguera, M. C. *et al.* Colombia's racial crucible: Y chromosome evidence from six admixed communities in the Department of Bolívar Palenque. *Ann. Hum. Biol.* **44**, 453–459 (2014).
- Martínez, B. *et al.* Ancestry estimates in afrodescendant population from San Basilio de Palenque, Colombia. *Forensic Sci. Int. Genet. Suppl. Ser.* **6**, E224–E225 (2017).
- Ansari-Pour, N. *et al.* Palenque de San Basilio in Colombia: Genetic data support an oral history of a paternal ancestry in Congo. *Proceedings R. Soc. B* **283**, 20152980 (2016).
- Simão, F. *et al.* Defining mtDNA origins and population stratification in Rio de Janeiro. *Forensic Sci. Int. Genet.* **34**, 97–104 (2018).
- Parson, W., Brandstätter, A., Pircher, M., Steinlechner, M. & Scheithauer, R. EMPOP—The EDNAP mtDNA population database concept for a new generation, high-quality mtDNA database. *Int. Congr. Ser.* **1261**, 106–108 (2004).
- Weissensteiner, H. *et al.* HaploGrep 2: Mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–W63 (2016).
- van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386–E394 (2009).
- Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 35: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
- Cerezo, M. *et al.* Comprehensive analysis of Pan-African mitochondrial DNA variation provides new insights into continental variation and demography. *J. Genet. Genomics* **43**, 133–143 (2016).
- Chan, E. K. F. *et al.* Human origins in a southern African palaeo-wetland and first migrations. *Nature* **575**, 185–189 (2019).
- Rito, T. *et al.* The first modern human dispersals across Africa. *PLoS ONE* **8**, 1–16 (2013).

21. Soares, P., Rito, T., Pereira, L. & Richards, M. B. A genetic perspective on African prehistory. In *Africa from MIS 6–2: Population Dynamics and Paleoenvironments* 195–212 (2016).
22. Gomes, V. *et al.* Mosaic maternal ancestry in the Great Lakes region of East Africa. *Hum. Genet.* **134**, 1013–1027 (2015).
23. Carvalho, M. *et al.* Paternal and maternal lineages in Guinea-Bissau population. *Forensic Sci. Int. Genet.* **5**, 114–116 (2011).
24. Göbel, T. M. K. *et al.* Mitochondrial DNA variation in Sub-Saharan Africa: Forensic data from a mixed West African sample, Côte d'Ivoire (Ivory Coast), and Rwanda. *Forensic Sci. Int. Genet.* **44**, 102202 (2020).
25. Salas, A. *et al.* The making of the African mtDNA landscape. *Am. J. Hum. Genet.* **71**, 1082–1111 (2002).
26. Graven, L. *et al.* Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. *Mol. Biol. Evol.* **12**, 334–345 (1995).
27. Salas, A. *et al.* The African Diaspora: Mitochondrial DNA and the Atlantic Slave Trade. *Am. J. Hum. Genet.* **74**, 454–465 (2004).
28. Batini, C. *et al.* Phylogeography of the human mitochondrial L1c haplogroup: Genetic signatures of the prehistory of Central Africa. *Mol. Phylogenet. Evol.* **43**, 635–644 (2007).
29. Fendt, L. *et al.* MtDNA diversity of Ghana: A forensic and phylogeographic view. *Forensic Sci. Int. Genet.* **6**, 244–249 (2012).
30. Coelho, M., Sequeira, F., Luiselli, D., Beleza, S. & Rocha, J. On the edge of Bantu expansions : mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. *BMC Evol. Biol.* **18**, 1–18 (2009).
31. Silva, M. *et al.* 60,000 years of interactions between Central and Eastern Africa documented by major African mitochondrial haplogroup L2. *Sci. Rep.* **5**, 1–13 (2015).
32. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
33. Plaza, S. *et al.* Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. *Hum. Genet.* **115**, 439–447 (2004).
34. Barbieri, C., Butthof, A., Bostoen, K. & Pakendorf, B. Genetic perspectives on the origin of clicks in Bantu languages from south-western Zambia. *Eur. J. Hum. Genet.* **21**, 430–436 (2013).
35. Barbieri, C. *et al.* Migration and interaction in a contact zone: mtDNA variation among Bantu-speakers in Southern Africa. *PLoS ONE* **9**, e99117 (2014).
36. Semino, O. *et al.* The genetic legacy of paleolithic Homo sapiens sapiens in extant Europeans: A Y chromosome perspective. *Science* (80-) **290**, 1155–1159 (2000).
37. Iacovacci, G. *et al.* Forensic data and microvariant sequence characterization of 27 Y-STR loci analyzed in four Eastern African countries. *Forensic Sci. Int. Genet.* **27**, 123–131 (2017).
38. Cruciani, F. *et al.* A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am. J. Hum. Genet.* **70**, 1197–1214 (2002).
39. Ansari-Pour, N., Plaster, C. A. & Bradman, N. Evidence from Y-chromosome analysis for a late exclusively eastern expansion of the Bantu-speaking people. *Eur. J. Hum. Genet.* **21**, 423–429 (2012).
40. Rowold, D. *et al.* At the southeast fringe of the Bantu expansion: Genetic diversity and phylogenetic relationships to other sub-Saharan tribes. *Meta Gene* **2**, 670–685 (2014).
41. Rosa, A., Ornelas, C., Jobling, M. A., Brehm, A. & Villems, R. Y-chromosomal diversity in the population of Guinea-Bissau: A multiethnic perspective. *BMC Evol. Biol.* **7**, 1–11 (2007).
42. Berniell-Lee, G. *et al.* Genetic and demographic implications of the bantu expansion: Insights from human paternal lineages. *Mol. Biol. Evol.* **26**, 1581–1589 (2009).
43. Batini, C. *et al.* Signatures of the preagricultural peopling processes in sub-saharan africa as revealed by the phylogeography of early Y chromosome lineages. *Mol. Biol. Evol.* **28**, 2603–2613 (2011).
44. Scozzari, R. *et al.* Molecular dissection of the basal clades in the human Y chromosome phylogenetic tree. *PLoS ONE* **7**, e49170 (2012).
45. Barbieri, C. *et al.* Refining the Y chromosome phylogeny with southern African sequences. *Hum. Genet.* **135**, 541–553 (2016).
46. Beleza, S., Gusmão, L., Amorim, A., Carracedo, A. & Salas, A. The genetic legacy of western Bantu migrations. *Hum. Genet.* **117**, 366–375 (2005).
47. De Filippo, C. *et al.* Y-chromosomal variation in sub-Saharan Africa: Insights into the history of Niger-Congo groups. *Mol. Biol. Evol.* **28**, 1255–1269 (2011).
48. Cruciani, F. *et al.* Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa. *Am. J. Hum. Genet.* **74**, 1014–1022 (2004).
49. Capelli, C. *et al.* Moors and Saracens in Europe: Estimating the medieval North African male legacy in southern Europe. *Eur. J. Hum. Genet.* **17**, 848–852 (2009).
50. Luis, J. R. *et al.* The levant versus the horn of Africa: Evidence for bidirectional corridors of human migrations. *Am. J. Hum. Genet.* **74**, 532–544 (2004).
51. Trombetta, B. *et al.* Phylogeographic refinement and large scale genotyping of human Y chromosome haplogroup E provide new insights into the dispersal of early pastoralists in the african continent. *Genome Biol. Evol.* **7**, 1940–1950 (2015).
52. Oliveira, S. *et al.* The role of matrilineality in shaping patterns of Y chromosome and mtDNA sequence variation in southwestern Angola. *Eur. J. Hum. Genet.* **27**, 475–483 (2019).
53. González, M. *et al.* The genetic landscape of Equatorial Guinea and the origin and migration routes of the y chromosome haplogroup R-V88. *Eur. J. Hum. Genet.* **21**, 324–331 (2013).
54. D'Atanasio, E. *et al.* The peopling of the last Green Sahara revealed by high-coverage resequencing of trans-Saharan patrilineages. *Genome Biol.* **19**, 1–15 (2018).
55. Cruciani, F. *et al.* Human Y chromosome haplogroup R-V88: A paternal genetic record of early mid Holocene trans-Saharan connections and the spread of Chadic languages. *Eur. J. Hum. Genet.* **18**, 800–807 (2010).
56. Ballantyne, K. N. *et al.* Mutability of Y-chromosomal microsatellites: Rates, characteristics, molecular bases, and forensic implications. *Am. J. Hum. Genet.* **87**, 341–353 (2010).

## Acknowledgements

We thank all sample donors for their contribution to this work and all the people who helped with sample collection, namely, Regina Miranda from the Community of San Basilio de Palenque, the personnel of the Benkos Biojó Rural school, and Miguel Obeso from the *etnoeducación* program. This work was supported by the University of Cartagena and FPIT—Fundación para la promoción de la investigación y la tecnología del Banco de la República, Colombia (FPIT-388) and partially financed by FEDER—Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020—Operacional Programme for Competitiveness and Internationalization (POCI), Portugal 2020, and by Portuguese funds through FCT—Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Inovação in the framework of the projects “Institute for Research and Innovation in Health Sciences” (POCI-01-0145-FEDER-007274). FS was supported by Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro—FAPERJ (E-26/202.275/2019). LG was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq (ref. 306342/2019-7), and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro—FAPERJ (CNE-2018). VG was supported by FCT, under the program contract provided in Decree-Law

no.57/2016 of August 29. The IPATIMUP integrates the i3S research unit, which is partially supported by the Portuguese Foundation for Science and Technology.

### Author contributions

B.M. and L.G. conceived and supervised the study. Material preparation, data collection and analysis were performed by B.M., F.S., V.G., M.N. and J.M. The first draft of the manuscript was written by B.M., F.S. and L.G. and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-77608-8>.

**Correspondence** and requests for materials should be addressed to B.M.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020