HIGH RESOLUTION MITOCHONDRIAL DNA ANALYSIS SHEDS LIGHT ON HUMAN DIVERSITY, CULTURAL INTERACTIONS AND POPULATION MOBILITY IN NORTHWESTERN AMAZONIA

Leonardo Arias[1], Chiara Barbieri[2], Guillermo Barreto[3], Mark Stoneking[1], Brigitte Pakendorf[4]

[1] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D-04103, Leipzig, Germany

[2] Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History D-07745 Jena, Germany

[3] Laboratorio de Genética Molecular Humana, Universidad del Valle, Cali, Colombia,

[4] Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 69363 Lyon Cedex 07, France

Number of text pages: 31; number of figures: 15, number tables: 6

Key words: Haplogroup, South America, Language, Exogamy, Amazonia

Leonardo Arias, Max-Planck-Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103, Leipzig, Germany. Telephone number: +49 341 3550 505, Fax: +49 341 3550 555. E-mail: leoarias2@gmail.com

Grant sponsorship: Max Planck Society and COLCIENCIAS

ABSTRACT

Objectives

Northwestern Amazonia (NWA) is a center of high linguistic and cultural diversity. Several language families and linguistic isolates occur in this region, as well as different subsistence patterns: some groups are foragers while others are agriculturalists. In addition, speakers of Eastern Tukanoan languages are known for practicing linguistic exogamy, a marriage system in which partners must come from different language groups. In this study, we use high resolution mitochondrial DNA sequencing to investigate the impact of this linguistic and cultural diversity on the genetic relationships and structure of NWA groups.

Methods

We collected saliva samples from individuals representing 40 different NWA ethnolinguistic groups and sequenced 439 complete mitochondrial genomes to an average coverage of 1030x.

Results

The mtDNA data revealed that NWA populations have high genetic diversity with extensive sharing of haplotypes among groups. Moreover, groups who practice linguistic exogamy have higher mtDNA diversity, while the foraging Nukak have lower diversity. We also find that rivers play a more important role than either geography or language affiliation in structuring the genetic relationships of populations.

Discussion

Contrary to the view of NWA as a pristine area inhabited by small human populations living in isolation, our data support a view of high diversity and contact among different ethnolinguistic groups; movement along rivers has probably facilitated this contact. Additionally, we provide evidence for the impact of cultural practices, such as linguistic exogamy, on patterns of genetic variation. Overall, this study provides new data and insights into a remote and little-studied region of the world.

Northwestern Amazonia (NWA) contains tremendous biological, linguistic, and cultural diversity, which likely reflects the heterogeneity of the landscape, especially the complex and extensive network of rivers found in this area. The region (Figure 1) extends from the Andean foothills in the west to the area between the Orinoco River and the Rio Negro in the east, and extends south until the confluence between the Rio Negro and the Amazon River. The northern border is defined by the Eastern Andean Cordillera and the Colombian-Venezuelan llanos, and in the south by the full length of the Putumayo River (Eriksen 2011).

In terms of linguistic diversity, NWA harbors ethnolinguistic groups belonging to the main South American language families accepted by specialists (Campbell 1997; Chacon 2014; Dixon and Aikhenvald 1999), namely Arawakan, Carib, Tupi, and Quechua. Additionally, several local families are also present, such as Tukanoan, Guahiban, Huitotoan, Boran, Peba-Yaguan, Piaroa-Saliban and Maku-Puinave, as well as various isolate languages like Tikuna, Cofan, and Kamentsa (Landaburu 2000). Furthermore, several indigenous groups live in voluntary isolation, and almost nothing is known about their linguistic affiliation (Franco 2002). The area has been proposed as the place of origin of the Arawakan family, since it contains the highest linguistic diversity within the family (Aikhenvald 1999; Heckenberger 2002; Zucchi 2002). In addition, all 20 languages of the Tukanoan family are found in the area; these are classified into two branches: The Western Tukanoan branch (WT) distributed along the Putumayo, Caquetá, and Napo rivers, and the Eastern Tukanoan branch (ET) along the Vaupés, Rio Negro, and Apaporis rivers and their tributaries (Chacon 2014). The language families Carib, Tupi, and Quechua are probably recent immigrants in NWA, since only one language per family is present in the area. In addition, the Tupi language Nheengatú or Lingua Geral is found in the region; however, this is a very recent introduction spread by missionaries during the 17th and 18th centuries and by traders during the rubber boom in the 19th century, when it was used as a trade language (Sorensen 1967; Stenzel 2005).

[Figure 1 here]

In terms of cultural diversity, while NWA has often been viewed as a pristine area inhabited only by small, isolated, seminomadic tribes with an economy based on hunting and gathering (Denevan 1992; Meggers 1954), in fact there is considerable variation in subsistence and marriage practices. While some groups are traditional foragers, others engage in agriculture, and instead of isolated groups, archaeological and anthropological evidence now shows that NWA was indeed part of a continent-wide network of exchange and trade. Complex societies organized in chiefdoms and multiethnic confederations arose in the region, and multilingualism and extensive interactions among ethnolinguistic groups were the norm (Heckenberger 2002; Hornborg 2005; Santos-Granero 2002; Vidal 1997). In particular, the groups speaking Eastern Tukanoan languages and some of their Arawakan neighbors living in the basin of the Vaupés River and Rio Negro engage in an exceptional marital practice known as linguistic exogamy (Aikhenvald 1996; Chacon and Cayón 2013; Sorensen 1967; Stenzel 2005). According to this cultural norm, individuals are required to marry someone from a different language group, with each individual's linguistic affiliation determined by the language of the father. Linguistic exogamy thus creates a situation of multilingualism and movement of people (especially women, since it is accompanied by patrilocality and patrilineality) among the groups participating in the system (Sorensen 1967).

Historical linguistics, cultural anthropology, and archaeology are the main disciplines that have traditionally addressed questions regarding the origins, prehistory, and genetic relationships among the NWA ethnolinguistic groups (Campbell 1997; Chacon 2014; Heckenberger 2008; Lathrap 1970; Meggers 1948; Nettle 1999). However, due to the scarcity of the archaeological record, the time-depth limitations of linguistic methods based on lexical cognates to establish deep relationships (Dediu and Levinson 2012; Hock and Joseph 2009), and the insufficiency of documentation and description of a large number of the NWA societies, many of these questions remain to be fully answered. The oldest archaeological evidence of human occupation in NWA comes from a single site on the middle Caquetá River dated between 9250 and 8100 BP, containing a great variety of stone artifacts, carbonized seeds and other botanical remains from different palm species, as well as phytoliths of bottle gourd, leren, and pumpkin (Aceituno et al.

2013; Gnecco and Mora 1997), indicating that these early human groups relied on vegetable resources that are still exploited by contemporary societies in NWA.

One hypothesis about the peopling of NWA was proposed by Nimuendajú (1950), who suggested that the region was first inhabited by hunter-gatherer populations, perhaps the ancestors of the Maku-Puinave groups, most of whom still practice a foraging lifestyle. Proto-Arawakan groups then started expanding in the region from their place of origin located between the Orinoco river and the Rio Negro (Heckenberger 2002; Lathrap 1970), and finally, the Tukanoans are assumed to have arrived in the area and displaced peoples speaking Arawakan and Maku-Puinave languages from the Vaupés (the Tukanoans probably came from the Napo-Putumayo, where Western Tukanoans still live). However, this scenario does not account for the presence of groups belonging to the Carib, Guahiban, Huitotoan, and Boran language families and the various language isolates in the region.

Genetic studies can provide insights into population history, and indeed studies of mitochondrial DNA (mtDNA) genetic variation in Native American populations have contributed greatly to our knowledge about the peopling of the Americas. Early studies using restriction fragment length polymorphisms (RFLP) and sequencing of the hypervariable region one (HVS-I) identified five founder lineages or haplogroups, designated as A to D and X (Bailliet et al. 1994; Barbieri et al. 2011; Gaya-Vidal et al. 2011; Keyeux et al. 2002; Lewis et al. 2007; Schurr 2004; Torroni et al. 1993). Whereas haplogroups A-D are widely distributed in the Americas, haplogroup X is restricted to North America (Bolnick and Smith 2003; Malhi et al. 2001). The analysis of HVS-I in several Native American populations showed that haplogroups A-D exhibit similar levels of diversity (Bonatto and Salzano 1997), supporting the hypothesis of a single origin of all Native American populations from a Northeast Asian source. Additionally, HVS-I data have been used to determine the genetic relationships among indigenous populations in South America and to test hypotheses concerning how genetic variation is structured at the regional and continental levels (Barbieri et al. 2011; Gaya-Vidal et al. 2011; Lewis et al. 2007; Marrero et al. 2007; Melton et al. 2007). These studies revealed that Andean (or western) populations show higher

120 levels of diversity and low genetic distances in contrast to the Eastern populations, who show the opposite
121 pattern. However, in previous studies NWA populations have been generally underrepresented, and hence
122 the inferences about the genetic structure of the entire Amazonian region are based on a small number of
123 populations.

124 Recent developments in sequencing technology allow the determination of complete mtDNA genomes at
125 the population level and thus enable unbiased insights into the maternal history of human populations
126 (Delfin et al. 2014; Gunnarsdottir et al. 2011; Kivisild 2015). At present, no such studies are reported for
127 South American indigenous populations. Available studies of complete mtDNA genomes from Native
128 Americans have been restricted to a limited number of individuals carrying particular haplogroups,
129 usually selected based on their HVS-I sequences (Achilli et al. 2013; Bodner et al. 2012; de Saint Pierre et
130 al. 2012; Fagundes et al. 2008; Lee and Merriwether 2015; Perego et al. 2009; Perego et al. 2010), or to
131 archaeological remains from different time periods (Fehren-Schmitz et al. 2015; Llamas et al. 2016).
132 These studies have primarily focused on inferences about the peopling of the continent, the number of
133 migrations, the divergence times, and changes in the effective population size through time. Nevertheless,
134 several problems and biases are associated with this sampling strategy. First, the overall diversity might
135 be underestimated, since individuals carrying the same HVS-I sequence can exhibit considerable variation
136 in the coding region (Gunnarsdottir et al. 2011). Secondly, the reconstruction of demographic trends can
137 be skewed, since the estimation of effective population sizes through time using Bayesian coalescent
138 methods (i.e. Bayesian skyline plots in BEAST) can generate spurious signals of population growth when
139 based on samples selected by haplogroup (Gunnarsdottir et al. 2011). Lastly, the histories and origins of
140 specific populations cannot be investigated, since the coalescent age of a particular lineage does not
141 correspond to the age of the population, especially when the diversity within each lineage is unknown
142 (Schurr 2004). In this study, we use complete mtDNA sequencing in a large and representative sample of
143 populations covering the extant ethnolinguistic diversity from NWA to reconstruct their maternal history,
144 as well as to determine their genetic diversity and to make inferences about the origins of this diversity.

Finally, we aim to investigate the impact of prehistoric population dynamics and cultural interactions on the structure of the genetic variation observed among present-day NWA populations.

MATERIALS AND METHODS

Sample collection

Samples from unrelated individuals belonging to 40 ethnolinguistic groups were collected during several expeditions carried out by one of the authors (L.A.) in five departments (administrative divisions) of NWA, namely: Amazonas, Guainía, Guaviare, Meta and Putumayo (Table 1, Figure 1). The samples consisted of either saliva, collected as 3 mL of saliva in 3 mL of lysis buffer (Quinque et al. 2006), or blood samples stabilized with EDTA. Written informed consent was obtained from each participant, and from the community leader and/or local/regional indigenous organizations, after giving a full description of the aims of the study. Local translators and fieldwork assistants helped to explain and translate into the local languages when individuals or communities were not proficient in Spanish. Additionally, each participant answered a short questionnaire soliciting information regarding their birthplace, language, ethnic affiliation and that of their parents and grandparents. The study was approved by the ethics committee of the Universidad del Valle in Cali, Colombia and the Ethics Commission of the University of Leipzig Medical Faculty. All procedures were undertaken in accordance with the Declaration of Helsinki on ethical principles and an export permit was issued by the Colombian Ministry of Health and Social Protection.

[Table 1 here]

DNA sequencing and sequence processing

The DNA was extracted from blood samples with the “salting out” method (Miller et al. 1988) and from the saliva samples with the QIAamp DNA Midi kit (Qiagen), starting from 2.0 mL of the saliva/buffer

mixture. The concentration of DNA was quantified with the NanoDrop 8000 spectrophotometer (Thermo Scientific). We prepared genomic libraries with double indices and enriched for full mtDNA genomes using a hybridization-capture method described previously (Kircher et al. 2012; Maricic et al. 2010). From the enriched libraries, paired-end sequences of 100 bp length were generated on the Illumina Hiseq 2500 platform. Base-calling was performed using freeIbis (Renaud et al. 2013), and Illumina adapters were trimmed and completely overlapping paired sequences were merged using leeHOM (Renaud et al. 2014a). The sequencing data were de-multiplexed using deML (Renaud et al. 2014b) and the sequences aligned against the human reference genome 19 using BWA's *aln* algorithm (Li and Durbin 2009). After duplicate removal using PicardTools v2.1.1 (https://github.com/broadinstitute/picard), we performed an iterative alignment for each library individually to obtain mtDNA consensus sequences. In the first step, we extracted all sequencing reads of a library that aligned either to the mitochondrial genome or to a list of nuclear copies of mtDNA (NUMTs) (Li et al. 2012). We subsequently aligned these reads to the revised Cambridge Reference Sequence (rCRS; Andrews et al. 1999) using BowTie2's *very-sensitive* algorithm (Langmead and Salzberg 2012) and called a consensus sequence. In the second step, the reads were re-aligned to the library's respective consensus sequence generated in the first step, using the same BowTie2 settings. After the second alignment step, we called a final consensus sequence that was used throughout the rest of the analysis. Final sequences in fasta format were aligned to the Revised Cambridge Reference Sequence (rCRS (Andrews et al. 1999)) with the multiple sequence alignment software Mafft (Katoh and Standley 2013), and manually inspected for alignment errors with Bioedit ver. 7.2.5 (Hall 1999). The two poly-C regions (np 303–315 and 16,183–16,194) were excluded from the subsequent analysis; although one position (16,189) diagnostic for haplogroup B2 is therefore not considered in the haplogroup-calling analysis, the additional substitutions defining this haplogroup that occur elsewhere in the mitochondrial genome enable unambiguous assignment of sequences to lineage B2.

193 **Data analysis**

194 We considered populations with a sample size of 10 individuals or more, and merged populations with
195 sample sizes smaller than 10 based on linguistic criteria when our initial analyses did not show significant
196 genetic differences, as follows (Table 1). The Arawakan groups Achagua (n=6) and Piapoco (n=18) were
197 merged into a single population, since their indigenous reservations are adjacent and individuals often
198 intermarry (data available on request); one Bare (n=1) individual was added to the Curripaco (n=16)
199 sample among whom he was living when sampled on the Atabapo River; Yucuna (n=31) and Matapi
200 (n=8) were merged into a single population, since they both speak Yucuna, live along the same river, and
201 intermarry (data available on request); and Murui (n=18) and Uitoto (n=8) were merged, as these two
202 groups belong to the same language family, which is composed of several dialects that are mutually
203 intelligible (http://glottolog.org/resource/languoid/id/huit1251, accessed on 31.05.2017). Finally,
204 following the latest classification of the Tukanoan family (Chacon 2014), the Eastern Tukanoan groups
205 Piratapuyo (n=8) and Wanano (n=5) were merged as Pira-Wanano; Tukano (n=8) and Tatuyo (n=2) were
206 merged as Tuka-Tatuyo; and Tuyuca (n=7), Yuruti (n=1), Pisamira (n=1), and Karapana (n=1) were
207 merged as Other-ET. The only group with a sample size smaller than 10 that we retained as a separate
208 group in the analyses were the Carijona (n=8), since this is the only Carib-speaking group living in NWA.
209 Moreover, they are at risk of disappearing both physically and culturally, with less than 30 active speakers
210 of Carijona scattered in two communities, and they occupy an important place in the ethno-history of the
211 region (Franco 2002). We excluded Barasano (n=4), Kubeo (n=5), Cofan (n=6), Cabiyari (n=1),
212 Guambiano (n=1), and Nasa (n=1) from all the analyses except the haplotype networks, since this analysis
213 represents the evolutionary relationships among individual sequences. We furthermore excluded nine
214 individuals with maternal ancestry outside of NWA (labeled 'Mestizo' in Table 1) from all analyses.
215 After merging and filtering as described above, 412 sequences from 24 groups were kept in the
216 population-based analyses.

Based on information from D-PLACE (Kirby et al. 2016) and HG database (https://huntergatherer.la.utexas.edu/home, accessed on 06.06.2017), we divided the populations into agriculturalists (AG) and hunter-gatherers (HGP). In the latter category we included the Nukak, who currently still practice a foraging way of life, as well as the Puinave, Sikuani, and Guayabero, who have all adopted agriculture only recently (Kondo 2002; Uribe Tobón and Instituto Colombiano de Cultura 1992).

The haplogroup affiliation of the individual sequences was determined with Haplogrep (Kloss-Brandstatter et al. 2011), based on Phylotree build 16 (van Oven and Kayser 2009). Haplogroup frequencies by population were estimated by simple counting, and a correspondence analysis (CA) based on the frequency of sub-haplogroups (e.g. A2a) was performed and visualized with the R-packages FactoMineR (Le et al. 2008) and factoextra (Kassambara and Mundt 2016), respectively.

Population-based statistical analyses were performed with Arlequin v3.5 (Excoffier and Lischer 2010); these include the analysis of molecular variance (AMOVA), estimation of molecular diversity indices, the estimation of pairwise genetic distances based on Φ_{ST}, and Tajima's D test of selective neutrality. A Multidimensional Scaling analysis (MDS) was performed on the matrix of pairwise Φ_{ST} values to visualize the distances between populations. Additionally, we performed a Mantel test to evaluate the correlations between genetic distances and geographic distances. The matrix of geographic distances was built using the geographic coordinates of the location where the majority of samples for each ethnolinguistic group were collected and then calculating the great circle distances between locations via the R packages ade4 and geosphere (Dray and Dufour 2007; Hijmans 2016). Furthermore, a multiple regression analysis on distance matrices (MRM) (Goslee and Urban 2007) with the form: MRM(as.dist(gen.dist) ~ as.dist(geo.dist) + as.dist(rivers.dist)) was performed; the analysis takes into consideration a matrix of geographic distances and a matrix of proximity along rivers as predictor variables of the genetic distances (pairwise Φ_{ST} values) between populations (Pugach et al. 2016; Yunusbayev et al. 2012). For the matrix of river distances a value of zero was given to populations living

242 along the same river or on rivers that are closely connected, and a value of one was given to populations
243 living on different rivers.

244 The sharing of haplotypes between populations was estimated with in-house R scripts as the proportion of
245 pairs of identical sequences shared between populations. Additionally, networks of haplotypes were
246 constructed with the software Network ver. 4.6.1.3 and visualized with Network Publisher ver. 2.0.0.1
247 (http://www.fluxus-engineering.com). Finally, Bayesian skyline plots (BSP) were constructed with
248 BEAST ver. 1.8.2 (Drummond et al. 2012). For this analysis, the best substitution model was estimated
249 with jModeltest 2.1.7 (Darriba et al. 2012), and BEAST was used to estimate whether a strict or a relaxed
250 clock model best fits the data. This analysis was performed on both the complete sequences and the
251 sequences partitioned into coding (577-16023) and non-coding (16024-576) regions, applying the
252 corresponding substitution rates reported previously (Soares et al. 2009).

253

254 **RESULTS**

255 We generated 439 complete mitochondrial sequences to an average coverage per sample of 1030x, which
256 were deposited in GenBank with accession numbers: XXXXXXXX-XXXXXXXX and XXXXXXXX-
257 XXXXXXXX. All sequences belonged to one of the main Native American haplogroups, namely A2, B2,
258 C1 and D1. Haplogroups A2 and C1 were the most frequent lineages in the NWA populations (excluding
259 the so-called 'Mestizos'), with more than half of all sequences belonging to A2 (90 haplotypes in 138
260 sequences) and C1 (95 haplotypes in 181 sequences) together; in contrast, B2 (49 haplotypes in 73
261 sequences) and D1 (32 haplotypes in 38 sequences) were less frequent. Table 2 provides a breakdown of
262 the haplogroup frequencies for the ethnolinguistic groups included in the population analyses.

263 [Table 2 here]

264 A Correspondence Analysis (CA) (Figure 2) shows the clustering of populations based on the frequency
265 of sub-haplogroups. We observed differences among populations without a clear clustering by language

family, with the exception of the Eastern Tukanoan groups Siriano, Desano, Pira-Wanano, Tuka-Tatuyo and Other-ET, which were near one another in the left side of the plot. However, Tanimuka, who also speak an Eastern Tukanoan language, were far apart from their linguistic relatives. Additionally, Guayabero and Sikuani (who speak languages belonging to the Guahiban family) were located close to each other in the lower left pane of the plot. In addition to language affiliation, a few cases of proximity in the CA plot could be attributed to geographic proximity, as in the case of Kamentsa, Pasto and Inga, who all live close to one another in the Andean foothills. In other cases, the relatively close proximity of populations could be attributed to their being settled along the same river or on rivers that are part of the same basin (Supporting information Figure S1), as in the case of Curripaco and Puinave, who live on the Inírida and Atabapo rivers.

[Figure 2 here]

Molecular diversity indices

The genetic variation in these communities was assessed through different molecular diversity indices (Figure 3, Supporting information Table S1). On average, the gene diversity in these groups was high (0.9), but there were differences amongst them. For example, Eastern Tukanoan groups showed consistently high values of gene diversity, with the exception of the Tanimuka, who had one of the lowest values (0.73). The Western Tukanoan groups Coreguaje (0.92) and Siona (0.82) showed lower values than Eastern Tukanoan groups. Among Arawakans, the Ach-Piapoco had the lowest value (0.77). The hunter-gatherer group Nukak showed the lowest gene diversity of all groups (0.64): only four haplotypes were observed among the 16 individuals analyzed. Additionally, we observed that agriculturalist groups tended to have higher gene diversities (average = 0.92) than hunter-gatherer groups (average = 0.80) (Mann-Whitney U test, P-value = 0.03).

[Figure 3 here]

289 The mean number of pairwise differences (MPD) per population showed less variation, with an average
290 of 41.07 +/- 17.86 differences. The smallest values were found in Sikuani (24.08 +/- 11.18) and Nukak
291 (31.07 +/- 14.33), and the largest values were observed in Cocama (44.64 +/- 20.37), Carijona (42.71 +/-
292 20.8), and Siriano (41.38 +/- 19.66) The D values of Tajima's test of neutrality (Tajima 1989) ranged
293 from -0.735 to 2.318. Under neutrality, Tajima's D is expected to be equal to zero and significant
294 departures are interpreted as a result of selection or changes in population size. Although none of the D
295 values were significant (all P-values > 0.2), positive D values >1.2 were obtained for Guayabero, Nukak,
296 Sikuani and Tanimuka, which may reflect recent reductions in the size of these populations. This
297 hypothesis was supported both by the distribution of pairwise differences by population (Supporting
298 information Figure S2), which showed increased frequencies for the category of small differences (0 and
299 1 differences) and for the category of large differences (50 or more), as well as by the Bayesian
300 reconstruction of population size changes through time (BSP plots, Supporting information Figure S3 and
301 below). Furthermore, the Tanimuka and Nukak had the lowest gene diversity values.

302

303 **Shared haplotypes**

304 A total of 216 different haplotypes were observed among the 412 sequences included in this analysis,
305 pointing to a considerable number of shared sequences. Of these, 146 were unique haplotypes and 70
306 haplotypes were shared among 266 sequences: 52 within populations, 31 between populations, and 13
307 both within and between populations. The shared haplotypes accounted for 64.6% of all the sequences
308 analyzed. This amount of haplotype sharing between populations is considerably high when compared to
309 other population-based studies of complete mitochondrial genomes (Table 3). In other studies, the
310 majority of shared haplotypes were generally observed within populations, with the exception of two
311 African datasets from Burkina Faso and Zambia (Barbieri et al. 2013; Barbieri et al. 2012), which showed
312 low levels of sharing both within and between populations. The highest level of sharing between
313 populations was observed for Siberian populations spread over a large geographic area (Duggan et al.

2013); the NWA populations analyzed in this study showed the second highest value of sharing between populations.

[Table 3 here]

Figure 4 shows the proportion of pairs of sequences shared between and within NWA populations. Siriano, Other-ET, and Pasto were the only groups without shared haplotypes within the populations. The majority of between-group haplotype sharing involved Arawakan and Eastern Tukanoan groups. The Arawakan groups share mostly with groups living in close proximity (Figure S4), e.g. Yucu-Matapi shared with Tanimuka; Curripaco with Puinave and Nukak, and Ach-Piapoco with Saliba and with the Guahiban groups Sikuani and Guayabero. In contrast, most Eastern Tukanoan groups, who practice linguistic exogamy, shared haplotypes among each other (except for Tanimuka, who shared only with Yucu-Matapi). In contrast, the Western Tukanoan groups Siona and Coreguaje shared primarily within their populations and did not share haplotypes with the Eastern Tukanoan groups.

The groups from the Andean foothills--Inga, Kamentsa, and Pasto--showed different patterns of shared haplotypes, despite the fact that they live in close geographic proximity. The Pasto, a group that has lost its native language and that is highly incorporated into the admixed local population, shared no haplotypes with any population. The Kamentsa shared haplotypes only with the Inga, while the Inga also shared haplotypes with three other groups located further inside the Amazonian area: Carijona, Coreguaje and Mur-Uitoto. Finally, of the three groups living on the banks of the Amazon River close to the town of Leticia, the Cocama shared with both the Yagua and Tikuna, whereas the latter two groups lacked common haplotypes.

[Figure 4 here]

Haplotype networks

The networks of haplotypes (Supporting information Figure S5 A-D) complement the patterns of sequence sharing, but in addition allow us to discern clusters of related (not just identical) haplotypes. We

observed that some of these clusters were common among different language families and others were restricted to specific language families or to groups living in close geographic proximity; these are highlighted in Figures S5 (A-D). For instance, Arawakan and Eastern Tukanoan groups exhibited several haplotypes within haplogroups A2 (Cluster I, Figure S5A), B2 (Cluster I and II, Figure S5B), and C1 (Cluster III, IV, V and VI, Figure S5C) that were either shared or separated by only a few mutational steps. Notably, several of these clusters also included individuals speaking Maku-Puinave languages (cf. Cluster I and II, Figure S5B, and Cluster III, IV, and V, Figure S5C). Clusters of haplotypes restricted to specific groups are represented by clusters I and II in Figure S5D, exclusive to Eastern Tukanoan and Huitotoan populations, respectively. Furthermore, the haplotypes of the Inga (Quechuan) and the Kamentsa, who live in close proximity in the Andean foothills, were either shared between them or closely related (e.g. cluster II in Figure S5C). Finally, the haplotypes of the Guayabero and Sikuani (Guahiban) were mostly differentiated from those of other populations and generally shared by several individuals within the family (clusters II and III, Figure S5A; cluster I, Figure S5C). The sequences belonging to cluster I in haplogroup C (cluster I, Figure S5C) lack the diagnostic mutation A13263G for haplogroup C, but contain other diagnostic mutations that allow unambiguous assignment to haplogroup C. MtDNAs with this variant were previously identified in eastern Colombia by RFLP typing (Torres et al. 2006), where they occurred at high frequency in Guahibo, Piapoco, and Saliba groups. Given their high frequencies in the Guahiban groups, these haplotypes appear to belong to an autochthonous lineage that has then diffused into other groups living in the Orinoco basin.

Genetic structure and genetic distances

The AMOVA analysis (Table 4) allows us to test different hypotheses concerning how genetic variation is structured in NWA. We defined groups *a priori* based on language affiliation, geographic proximity, and distribution along major rivers or their tributaries to evaluate how much of the observed variation is explained by each grouping strategy. We observed that of the three grouping strategies, grouping

363 populations by their distribution along rivers resulted in the largest among-group component of the
364 genetic variance. In contrast, language was a poor predictor of the genetic structure, showing negative and
365 nonsignificant values for the component of variance due to differences among groups, indicating larger
366 genetic differences among groups of populations speaking related languages than among linguistically
367 different groups. Finally, geographic proximity was also a poor predictor; although the among-group
368 component was larger than for language, it was not significantly different from zero. An important aspect
369 to note is that although grouping by rivers performed better than grouping by geography or language, it
370 still did not provide a very good description of the genetic structure, since the percentage of variance due
371 to differences among populations within groups was still higher than the among groups component,
372 suggesting the existence of other levels of substructure within populations.

373 [Table 4 here]

374 The matrix of genetic distances between populations based on pairwise Φ_{ST} values (Supporting
375 information Figure S6) was used to construct an MDS plot (Figure 5). The populations do not form any
376 clear clustering: the majority of populations are grouped together in the center of the plot (indicated by
377 the inner circle in Figure 5) with an average pairwise $\Phi_{ST} = 0.03$, while around the main cluster a second
378 group of populations showed higher differentiation (external circle, average $\Phi_{ST} = 0.07$). Finally, Sikuani,
379 Siona, and the hunter-gatherer Nukak appeared as outliers with high genetic differentiation (average Φ_{ST}
380 = 0.22). This picture did not change after adding an additional dimension to the MDS plot (Supporting
381 information Figure S7). Particularly striking were the small genetic distances between the Eastern
382 Tukanoan groups, who clustered together in the center of the MDS plot. Although Tanimuka appeared
383 more distant from the main cluster of Eastern Tukanoan groups, their pairwise Φ_{ST} values were not
384 significantly different (Supporting information Figure S6) and the average Φ_{ST} (0.02) indicated low
385 genetic differentiation among all Eastern Tukanoan groups. In contrast, the Coreguaje and the Siona, who
386 speak languages of the Western Tukanoan branch, showed larger genetic distances, both with the Eastern
387 Tukanoan groups and with each other. Populations from each of the other language families did not form

clusters with their linguistic relatives. For example, Arawakan groups occupied different positions in the plot and their Φ_{ST} values were significantly different.

[Figure 5 here]

The results of the Mantel test showed a lack of significant correlation between geographic distances, estimated as great-circle distances, and the matrix of pairwise Φ_{ST} values ($r = 0.07$, p-value = 0.28). However, since rivers emerged as an important factor explaining the structure of genetic variation in the AMOVA results (Table 4), we also performed a multiple regression analysis on distance matrices (MRM), where we added rivers as an additional predictor variable. Adding rivers to the regression model resulted in an increase in the amount of variation explained by the model (Table 5), with rivers being a significant predictor (p-value = 0.01). We then jack-knifed over populations (Pugach et al. 2016; Ramachandran et al. 2005) and identified three populations as outliers: Sikuani, Siona, and Nukak, who appeared as outliers in the MDS plot as well (Figure 5). We performed the multiple regression analysis excluding the outliers; this resulted in an increase of 3.4 % in the R square value, a better correlation between genetic and geographic distances, and geography becoming a significant predictor factor (p-value < 0.05) (Table 5, supporting information Figure S8), although rivers were no longer a significant predictor of genetic subdivision.

[Table 5 here]

Bayesian demographic reconstruction

Bayesian skyline plots (BSP) were generated by haplogroup (A2, B2, C1 and D1) and by population. All four haplogroups showed an increase in effective population size between 17,500 – 25,000 years before present. This signal was more evident for haplogroups A2 and C1, which have the highest number of sequences (Supporting information Figure S9). In contrast, the BSP plots by population showed different outcomes. We observed four main trajectories (Supporting information Figure S3): first, a signal of population size increase shown by Yucu-Matapi, Curripaco, Desano, Siriano, Inga, Pasto, Mur-Uitoto,

412 Tikuna, and Cocama (exemplified by Yucu-Matapi in Figure S3A); second, population stability through
413 time shown by Ach-Piapoco, Tanimuka, Coreguaje, Siona, Kamentsa, Puinave, and Yagua (exemplified
414 by Coreguaje in Figure S3B); and third, population contraction shown by Sikuani, Guayabero, and Nukak
415 (exemplified by Nukak in Figure S3C), which is particularly striking for Sikuani (Supporting information
416 Figure S3D). These differences in the effective population size through time suggest that these
417 populations have followed independent demographic histories.

418

419 **DISCUSSION**

420 We have investigated the genetic diversity of ethnolinguistic groups from NWA at the level of complete
421 mitochondrial genomes. This area is underrepresented in previous studies, and our data contribute to fill a
422 gap in our knowledge about the genetic diversity of modern human populations. We have found that
423 NWA harbors a considerable amount of genetic diversity, with evidence for contact among different
424 ethnolinguistic groups, contrary to the common picture of Amazonian populations as small and isolated
425 with low genetic diversity (Fuselli et al. 2003; Wang et al. 2007). NWA populations show values of
426 nucleotide diversity as high as or higher than those observed in most other non-African populations
427 (Supporting information Figure S10), and they display the second-highest amount of sequence sharing in
428 a world-wide comparison (Table 3). The complete mitochondrial genome is the maximum level of
429 resolution one can achieve to differentiate individuals and populations at the maternal level, so the
430 presence of identical sequences among populations living in distant geographic areas indicates recent
431 contact and/or common ancestry.

432

433 **Lack of genetic structure along linguistic lines**

434 Although our dataset includes populations speaking languages belonging to different language families,
435 we found that linguistic affiliation is a poor predictor of genetic structure, as shown by the AMOVA

436 analysis (Table 4) and by the Correspondence Analysis based on sub-haplogroup frequencies (Figure 2).
437 This indicates that language does not constitute a barrier to gene flow, and that groups have been
438 interacting with other neighboring groups, especially along rivers, which in our analyses performed better
439 in explaining the patterns of genetic diversity. Archaeological and linguistic evidence demonstrates that
440 NWA has been an area of intense contact and movement of peoples of different cultural traditions,
441 evidenced by the diffusion of ceramic styles (Heckenberger 2002; Lathrap 1970; Zucchi 2002) and shared
442 subsistence strategies, by the existence of language areas and contact-induced linguistic change
443 (Aikhenvald 1999), and the generalized multilingualism among groups (Sorensen 1967; Stenzel 2005).
444 Likewise, cultural anthropology provides additional evidence of contact among groups. For example, both
445 Arawakans and Eastern Tukanoans share a ceremonial complex for male initiation known as Yurupari, in
446 which sacred flutes and trumpets are only played by males, as well as sharing myths concerning the hero
447 Kúwai (Hugh-Jones 1979; Jackson 1983; Vidal 2002). In addition, the Eastern Tukanoan groups from the
448 Pira-Parana and Apaporis rivers (Barasano, Makuna, and Tanimuka) reveal Arawakan influence, since
449 they also practice dances with masks during the season of high abundance of the palm tree fruit pupunha
450 (*Bactris gasipaes*) (Hugh-Jones 1979). The genetic distances among populations provide additional
451 evidence in this regard: although the global Φ_{ST} value of 0.11 indicates moderate differentiation (Hartl
452 and Clark 2007), this value is driven by three populations, namely the Siona, Sikuani, and Nukak. These
453 are highly differentiated from the other populations, likely reflecting the effects of genetic drift due to
454 bottlenecks, as indicated by the positive Tajima's D values (Figure 3) and the distribution of pairwise
455 differences (Supporting information Figure S2). When we exclude these populations, we observe an
456 average pairwise Φ_{ST} of 0.07, and populations appear close together in the MDS plot (Figure 5),
457 indicating low genetic differentiation among NWA populations.

458 In this general picture the Eastern Tukanoan groups stand apart, since they cluster together in the CA and
459 MDS plots (Figure 2 and Figure 5), and their pairwise genetic distances are small and non-significant
460 (Figure S6). Linguists have proposed a time depth for the Tukanoan family of 2000-2500 years, based on

461 a comparison of the diversity in Tukanoan languages with the diversity in Romance and Germanic
462 languages (Chacon 2014). The time depth of the Eastern Tukanoan branch (and thus the time to the most
463 recent common ancestor of the Eastern Tukanoan languages) would be even more recent, which might
464 indicate that the peoples speaking these languages share recent common genetic ancestry as well (at least
465 on the maternal side). However, the Eastern Tukanoan groups practice linguistic exogamy, and the close
466 genetic relationships among these populations might be the result of this marital system in which women
467 move among different ethnolinguistic groups. The consequences of the linguistic exogamy are also
468 evident in the gene diversity values and the patterns of shared haplotypes. Eastern Tukanoans are the
469 groups with the highest values of gene diversity, and they share more haplotypes among themselves than
470 with other non-Eastern Tukanoan groups. In addition, their haplotypes tend to be closely related, as can
471 be seen in the phylogenetic networks (Supporting information Figure S5). Analyses of the Y-chromosome
472 as well as nuclear markers will help to disentangle the effects of linguistic exogamy vs. recent common
473 ancestry on the patterns of genetic variation among Eastern Tukanoan groups.

474 The Tanimuka stand apart from the other Eastern Tukanoan groups in the analyses, which may reflect
475 their settlement further south, along the Apaporis and Miritíi-Paraná rivers. Moreover, they do not
476 participate in the exogamic system with other Eastern Tukanoan groups, but interact mainly with the
477 Arawakan groups Yucuna and Matapi. This is reflected in the patterns of haplotype sharing (Figure 4) as
478 well as in their language, which shows evidence of Arawakan influence (Barnes 1999; Chacon 2014).

479

480 **The role of rivers in structuring genetic variation**

481 Besides language, geography is another important factor in structuring the patterns of genetic variation in
482 human populations (Ramachandran et al. 2005; Schonberg et al. 2011; Wang et al. 2007). One of the most
483 salient characteristics of the physical landscape of NWA is the high density of rivers that drain the area,
484 and their importance for human populations was earlier recognized by explorers and ethnographers that

485 traveled through the region (Koch-Grünberg 1995; Wallace 1853). We found that the distribution along
486 rivers is an additional important factor influencing the genetic structure of NWA populations: our
487 AMOVA analyses (Table 4) show that clustering populations according to the rivers where they are
488 distributed explains more of the genetic variation that is due to differences among groups than does
489 grouping them by linguistic affiliation, i.e. populations living on the same river basin or in closely
490 connected rivers are genetically more similar than those living on different rivers. This pattern is also
491 observed in the distribution of sub-haplogroups among populations (Figure S1). For example, Curripaco
492 and Puinave, who live on the Inírida and Atabapo rivers, are located close together in the plot, and the
493 presence of Coreguaje, Yucu-Matapi, and Mur-Uitoto in the center of the plot could reflect their presence
494 in a region where the Putumayo and Caquetá rivers are separated by their shortest distance, therefore
495 facilitating contact among people inhabiting the basins and tributaries of these two rivers. Indeed, one
496 Murui individual was sampled in a Coreguaje community, and two Uitoto individuals were sampled in
497 the Miritί-Parana region, thus providing evidence for the movement of people among these groups. The
498 results of the MRM analysis provide additional evidence in this regard: even though no correlation
499 between genetic distances and geographic distances was observed via the Mantel test, we observed that
500 adding river distances as an additional predictor variable resulted in an increase of around 3% of the R-
501 square value (Table 5), indicating that rivers contribute to explaining a slightly higher percentage of the
502 variation observed in the genetic distances.

503 Rivers in Amazonia serve a double function in providing a means of communication as well as
504 subsistence, and the wide distribution of certain cultural traits (e.g., the production of Saladoid-
505 Barrancoid ceramics and circular plaza village settlement patterns), has been associated with the
506 expansion of Arawakan-speaking populations along the extensive system of NWA waterways
507 (Heckenberger 2002; Hornborg 2005; Lathrap 1970; Lowie 1948). They also mark a distinction in
508 subsistence strategies between the more numerous “river people” that build canoes, settle along rivers,
509 and rely on horticulture and fishing, and the “forest people” that inhabit the interfluvial areas, settle away

510 from the major rivers, and base their subsistence on foraging (Epps and Stenzel 2013). Additionally, the
511 rivers have profound meanings and are embedded in the cosmogonies of several NWA indigenous
512 groups. The Eastern Tukanoan creation myths describe the journeys that the ancestors of the people made
513 to settle this world on board an anaconda canoe that travelled up the Vaupes River; from the anaconda's
514 body all the Tukanoan siblings emerged (Chernela 2010; Jackson 1983). Arawakan groups also describe a
515 series of ever returning voyages from the sacred center of the world and the place of emergence of the
516 first ancestors at the rapids of Hípana on the Aiary River, covering the major arteries of the Rio Negro,
517 Orinoco, and Amazon (Wright 2002; Zucchi 2002). Therefore, our findings about the role of rivers in
518 structuring the genetic variation are in keeping with the pivotal role that rivers play for NWA populations.

519 The lack of fit between genetic and simple geographic distances may be the result of relatively recent
520 movements and the displacement of ethnolinguistic groups from their traditional territories. Population
521 dynamics and population sizes were drastically altered during the last five centuries, starting with early
522 colonial times (16th and 17th centuries), when many groups were decimated by newly introduced
523 epidemics and moved away from the accessible margins of the major rivers to avoid the slave raids of the
524 Spanish, Portuguese, and Dutch colonizers. Similar perturbations happened during the time of the
525 Christian missions in the 18th century, when many groups were forced to relocate to multiethnic mission
526 settlements, and finally during the rubber boom between the 19th and beginning of the 20th centuries,
527 when the groups who managed to escape the mercenaries exploiting the rubber fields resettled in remote
528 areas in the headwaters of small rivers (Dixon and Aikhenvald 1999; Hill and Santos-Granero 2002;
529 Stenzel 2005). The inferred reduction in population size of the Tanimuka, Sikuani, Guayabero, and
530 Nukak, as indicated by their low diversity values, the positive Tajima's D values (Figure 3), the
531 distribution of pairwise differences (Figure S2), and the reconstruction of effective population sizes
532 (Figure S3C, D), might be a result of these social upheavals.

533

534

The impact of subsistence strategies on the genetic diversity

NWA contains groups with different subsistence strategies, with manioc (*Manihot esculenta*) as the main staple among horticulturalist groups, who are best described as riverine horticultural societies, given their close association with rivers. The Nukak, in contrast, are traditionally foragers, who still rely on hunting and gathering and move throughout the extensive area between the Guaviare and Inírida rivers. Furthermore, the Guayabero, Sikuani, and Puinave are traditional foragers who have only recently undergone the transition to agriculture, and are therefore considered as HGP together with Nukak in our analyses (Table 1). Our data show that agricultural societies (AG) have higher levels of diversity on average than forager groups (HGP) as indicated by the Mann-Whitney U test (P-value = 0.03), while the HGP groups have larger values of Tajima's D statistic (Figure 3) and do not show signals of population expansion (Figure S3). These findings agree with patterns reported for other hunter-gatherer populations around the world (Aime et al. 2013; Excoffier and Schneider 1999; Oota et al. 2005) and contrast with the genetic signature of an agricultural way of life, namely higher effective population size (Patin et al. 2014), higher levels of diversity, and significantly negative values of Tajima's D test (Aime et al. 2013).

However, subsistence strategies are flexible and diverse among NWA populations. Horticulturalists complement their diet with occasional hunting and/or gathering of several kinds of palm fruit, and extensive exchanges between AG and HGP groups have been reported. In this system, HGP populations usually provide meat and several products from the forest, such as the poison curare for the tips of darts and arrows, in exchange for different cultivated products, such as manioc and other trade goods (Epps and Stenzel 2013; Jackson 1983; Milton 1984). Nonetheless, this exchange seems to be exclusively restricted to goods and labor, with little or no intermarriage documented between AG and HGP groups (Aikhenvald 1996). In contrast, we observed shared haplotypes between AG and HGP groups, which likely reflects intermarriage or recent common ancestry. For example, the most frequent haplotype in the Arawakan AG group Curripaco (Haplotype H_84 in Figure S4) is observed at high frequency in the HGP Nukak (and in the Eastern Tukanoan AG group Siriano). Moreover, the HGP Puinave share several haplotypes with the

AG group Curripaco (H_219, H_161, H_117 in Figure S4), a likely result of intermarriage between these groups, since there are communities on the Inírida River where one finds individuals from both groups. Similarly, the Guahiban HGP groups Sikuani and Guayabero exhibit a haplotype at high frequency (H_43 in Figure S4) that is shared with the AG Ach-Piapoco as well as further haplotypes related to haplotypes found in AG Arawakan groups (cluster I Figure S5C and cluster II S5B). This may reflect contact among them, since there are Piapoco communities on the lower Guaviare River as well as Sikuani communities on the Meta River, places where these groups overlap. However, it is difficult to determine the direction of the gene flow or to distinguish between contact and common ancestry as explanations for shared mtDNA haplotypes. Nevertheless, it is plausible that where haplotypes are shared the source population is the one in which the haplotype is present at higher frequency. For instance, the shared haplotype between the HGP Puinave and the AG Curripaco (H_219 in Figure S4) has a likely origin in Puinave, because of its higher frequency and the presence of related haplotypes in Puinave (cluster I Figure S5B). The source of the shared haplotype among the HGP Nukak and the AG Curripaco and Siriano (H_84 in Figure S4) is more difficult to infer, since its frequency is similar in the Nukak and in the Curripaco; furthermore, three other haplotypes present in the HGP Nukak and Guayabero are only one mutation apart from it (cluster II Figure S5B). Therefore, it is likely that this haplotype, too, moved from the HGP populations into the AG Curripaco. A similar explanation could be given for H_43 in Figure S4, which is part of the cluster I in Figure S5C, moving from the HGP Guayabero and Sikuani into the AG Ach-Piapoco. Thus, these observations seem to fit a scenario of asymmetric gene flow in which women move from HGP to AG, a pattern that has been reported for populations in Central and Southern Africa (Barbieri et al. 2014; Destro-Bisol et al. 2004; Verdu et al. 2013). However, this scenario will be further refined by analyses of Y-chromosome and genome-wide data, which will allow us to determine whether the gene flow among groups was sex-biased (i.e. involving the movement of only females or only males among groups) and to make inferences about the time and magnitude of these events.

584 In conclusion, this study provides new data from this remote and little-studied part of the world, which
585 allow insights into the impact of cultural practices on the patterns of genetic variation and on the
586 population dynamics of NWA groups. Although our current data do not allow us to distinguish whether
587 the population movements took place prior to European contact or only later, analyses of Y-chromosome
588 variation and genome-wide data will shed further light on the genetic history of NWA. Furthermore,
589 historical genetic studies will benefit from more archaeological work in NWA, since huge areas remain
590 completely unexplored.

591

592 **Acknowledgments**

593 We are grateful to all sample donors, communities, community leaders, and regional indigenous
594 organizations. L.A. especially gives thanks to Consejo Regional Indígena del Guaviare, Asociación de
595 Autoridades Tradicionales y Cabildos de los Pueblos Indígenas del Municipio de Leguízamo y Alto
596 Resguardo Predio Putumayo, Asociación de Cabildos Indígenas del Trapecio Amazónico, Organización
597 Zonal Indígena Del Putumayo, the staff of Parques Nacionales Naturales in Puerto Leguízamo, the office
598 of Indigenous Affairs in Puerto Inírida, Rafael Rodriguez, William Yucuna, the late Gustavo Arias, and
599 all people who helped during the fieldtrips for their valuable collaboration and warm welcome during our
600 stay in their communities. We also acknowledge Roland Schroeder for laboratory technical assistance,
601 and Enrico Macholdt, Alexander Huebner, Irina Pugach, and Michael Dannemann for advice with the
602 data analyses. B.P. acknowledges the LABEX ASLAN (ANR-10-LABX-0081) of Université de Lyon for
603 its financial support within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) of the French
604 government operated by the National Research Agency (ANR). L.A. was supported by a graduate grant
605 from COLCIENCIAS; research was supported by funds from the Max Planck Society.

606

607

608 **LITERATURE CITED**

609

610 Aceituno FJ, Loaiza N, Delgado-Burbano ME, and Barrientos G. 2013. The initial human settlement of
611 Northwest South America during the Pleistocene/Holocene transition: Synthesis and perspectives.
612 Quaternary International 301:23-33.
613 Achilli A, Perego UA, Lancioni H, Olivieri A, Gandini F, Kashani BH, Battaglia V, Grugni V,
614 Angerhofer N, Rogers MP et al. . 2013. Reconciling migration models to the Americas with the
615 variation of North American native mitogenomes. Proc Natl Acad Sci U S A.
616 Aikhenvald AY. 1996. Areal diffusion in Northwest Amazonia: The case of Tariana. Anthropological
617 Linguistics 38(1).
618 Aikhenvald AY. 1999. The Arawak language family. In: Dixon RMW, and Aikhenvald AY, editors. The
619 Amazonian Languages. Cambridge ; New York: Cambridge University Press.
620 Aime C, Laval G, Patin E, Verdu P, Segurel L, Chaix R, Hegay T, Quintana-Murci L, Heyer E, and
621 Austerlitz F. 2013. Human genetic data reveal contrasting demographic patterns between
622 sedentary and nomadic populations that predate the emergence of farming. Mol Biol Evol
623 30(12):2629-2644.
624 Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, and Howell N. 1999. Reanalysis
625 and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet
626 23(2):147.
627 Bailliet G, Rothhammer F, Carnese FR, Bravi CM, and Bianchi NO. 1994. Founder mitochondrial
628 haplotypes in Amerindian populations. Am J Hum Genet 55(1):27-33.
629 Barbieri C, Butthof A, Bostoen K, and Pakendorf B. 2013. Genetic perspectives on the origin of clicks in
630 Bantu languages from southwestern Zambia. Eur J Hum Genet 21(4):430-436.
631 Barbieri C, Guldemann T, Naumann C, Gerlach L, Berthold F, Nakagawa H, Mpoloka SW, Stoneking M,
632 and Pakendorf B. 2014. Unraveling the complex maternal history of Southern African Khoisan
633 populations. Am J Phys Anthropol 153(3):435-448.
634 Barbieri C, Heggarty P, Castri L, Luiselli D, and Pettener D. 2011. Mitochondrial DNA variability in the
635 Titicaca basin: Matches and mismatches with linguistics and ethnohistory. Am J Hum Biol
636 23(1):89-99.
637 Barbieri C, Whitten M, Beyer K, Schreiber H, Li M, and Pakendorf B. 2012. Contrasting maternal and
638 paternal histories in the linguistic context of Burkina Faso. Mol Biol Evol 29(4):1213-1223.
639 Barnes J. 1999. Tucano. In: Dixon RMW, and Aikhenvald AY, editors. The Amazonian Languages. New
640 York: Cambridge University Press.
641 Bodner M, Perego UA, Huber G, Fendt L, Rock AW, Zimmermann B, Olivieri A, Gomez-Carballa A,
642 Lancioni H, Angerhofer N et al. 2012. Rapid coastal spread of First Americans: novel insights
643 from South America's Southern Cone mitochondrial genomes. Genome Res 22(5):811-820.
644 Bolnick DA, and Smith DG. 2003. Unexpected patterns of mitochondrial DNA variation among Native
645 Americans from the southeastern United States. Am J Phys Anthropol 122(4):336-354.
646 Bonatto SL, and Salzano FM. 1997. Diversity and age of the four major mtDNA haplogroups, and their
647 implications for the peopling of the New World. Am J Hum Genet 61(6):1413-1423.
648 Campbell L. 1997. American Indian Languages: The Historical Linguistics of Native America: Oxford
649 University Press.
650 Chacon T. 2014. A Revised Proposal of Proto-Tukanoan Consonants and Tukanoan Family
651 Classification. Int J Am Linguist 80(3):275-322.
652 Chacon T, and Cayón L. 2013. Considerações sobre a exogamia linguística no Noroeste Amazônico.
653 Revista de Letras da Universidade Católica de Brasília 6.
654 Chernela JM. 2010. The Wanano Indians of the Brazilian Amazon: A Sense of Space: University of
655 Texas Press.

Darriba D, Taboada GL, Doallo R, and Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9(8):772.

de Saint Pierre M, Gandini F, Perego UA, Bodner M, Gomez-Carballa A, Corach D, Angerhofer N, Woodward SR, Semino O, Salas A et al. 2012. Arrival of Paleo-Indians to the southern cone of South America: new clues from mitogenomes. PLoS One 7(12):e51311.

Dediu D, and Levinson SC. 2012. Abstract Profiles of Structural Stability Point to Universal Tendencies, Family-Specific Factors, and Ancient Connections between Languages. Plos One 7(9).

Delfin F, Min-Shan Ko A, Li M, Gunnarsdottir ED, Tabbada KA, Salvador JM, Calacal GC, Sagum MS, Datar FA, Padilla SG et al. 2014. Complete mtDNA genomes of Filipino ethnolinguistic groups: a melting pot of recent and ancient lineages in the Asia-Pacific region. Eur J Hum Genet 22(2):228-237.

Denevan WM. 1992. The Pristine Myth: The Landscape of the Americas in 1492. Annals of the Association of American Geographers 82(3):369-385.

Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Caglia A, Tofanelli S, Spedini G, and Capelli C. 2004. Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. Mol Biol Evol 21(9):1673-1682.

Dixon RMW, and Aikhenvald AY. 1999. The Amazonian Languages: Cambridge University Press.

Dray S, and Dufour AB. 2007. The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software 22(4):1-20.

Drummond AJ, Suchard MA, Xie D, and Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29(8):1969-1973.

Duggan AT, Evans B, Friedlaender FR, Friedlaender JS, Koki G, Merriwether DA, Kayser M, and Stoneking M. 2014. Maternal history of Oceania from complete mtDNA genomes: contrasting ancient diversity with recent homogenization due to the Austronesian expansion. Am J Hum Genet 94(5):721-733.

Duggan AT, Whitten M, Wiebe V, Crawford M, Butthof A, Spitsyn V, Makarov S, Novgorodov I, Osakovsky V, and Pakendorf B. 2013. Investigating the prehistory of Tungusic peoples of Siberia and the Amur-Ussuri region with complete mtDNA genome sequences and Y-chromosomal markers. PLoS One 8(12):e83570.

Epps P, and Stenzel K. 2013. Upper Rio Negro: cultural and linguistic interaction in Northwestern Amazonia. Rio de Janeiro: Museu do Índio – FUNAI, Museu Nacional.

Eriksen L. 2011. Nature and culture in prehistoric Amazonia [Ph.D.]. Lund: Lund University.

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

Excoffier L, and Schneider S. 1999. Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. Proc Natl Acad Sci U S A 96(19):10597-10602.

Fagundes NJ, Kanitz R, Eckert R, Valls AC, Bogo MR, Salzano FM, Smith DG, Silva WA, Jr., Zago MA, Ribeiro-dos-Santos AK et al. 2008. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. Am J Hum Genet 82(3):583-592.

Fehren-Schmitz L, Llamas B, Lindauer S, Tomasto-Cagigao E, Kuzminsky S, Rohland N, Santos FR, Kaulicke P, Valverde G, Richards SM et al. 2015. A Re-Appraisal of the Early Andean Human Remains from Lauricocha in Peru. PLoS One 10(6):e0127141.

Franco R. 2002. Los carijonas de Chiribiquete Bogotá: Fundación Puerto Rastrojo.

Fuselli S, Tarazona-Santos E, Dupanloup I, Soto A, Luiselli D, and Pettener D. 2003. Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. Mol Biol Evol 20(10):1682-1691.

Gaya-Vidal M, Moral P, Saenz-Ruales N, Gerbault P, Tonasso L, Villena M, Vasquez R, Bravi CM, and Dugoujon JM. 2011. mtDNA and Y-chromosome diversity in Aymaras and Quechuas from Bolivia: different stories and special genetic traits of the Andean Altiplano populations. Am J Phys Anthropol 145(2):215-230.

Gnecco C, and Mora S. 1997. Late Pleistocene early Holocene tropical forest occupations at San Isidro and Pena Roja, Colombia. Antiquity 71(273):683-690.

Goslee SC, and Urban DL. 2007. The ecodist package for dissimilarity-based analysis of ecological data. Journal of Statistical Software 22(7):1-19.

Gunnarsdottir ED, Li M, Bauchet M, Finstermeier K, and Stoneking M. 2011. High-throughput sequencing of complete human mtDNA genomes from the Philippines. Genome Res 21(1):1-11.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:4.

Hartl DL, and Clark AG. 2007. Principles of population genetics. Sunderland, Massachusetts: Sinauer Associates, Inc.

Heckenberger MJ. 2002. Rethinking the Arawakan Diaspora: Hierarchy, Regionality, and the Amazonian Formative. In: Hill JD, and Santos-Granero F, editors. Comparative Arawakan Histories: Rethinking Language Family and Culture Area in Amazonia: University of Illinois Press. p 99-122.

Heckenberger MJ. 2008. Amazonian Mosaics: Identity, Interaction, and Integration in the Tropical Forest. In: Silverman H, and Isbell WH, editors. The Handbook of South American Archaeology. New York, NY: Springer New York. p 941-961.

Hijmans RJ. 2016. geosphere: Spherical Trigonometry. R package version 1.5-5 ed.

Hill JD, and Santos-Granero F. 2002. Comparative Arawakan Histories: Rethinking Language Family and Culture Area in Amazonia. Urbana: University of Illinois Press.

Hock HH, and Joseph BD. 2009. Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics, Second Edition. Trends Linguist-Stud 218:1-588.

Hornborg A. 2005. Ethnogenesis, regional integration, and ecology in prehistoric Amazonia - Toward a system perspective. Curr Anthropol 46(4):589-620.

Hugh-Jones S. 1979. The palm and the Pleiades : initiation and cosmology in northwest Amazonia. Cambridge ; New York: Cambridge University Press. xvi, 332 p., 332 leaves of plates p.

Jackson JE. 1983. The Fish People: Linguistic Exogamy And Tukanoan Identity In Northwest Amazonia. Cambridge: Cambridge University Press. xix, 287 p.

Kassambara A, and Mundt F. 2016. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.3 ed.

Katoh K, and Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30(4):772-780.

Keyeux G, Rodas C, Gelvez N, and Carter D. 2002. Possible migration routes into South America deduced from mitochondrial DNA studies in Colombian Amerindian populations. Hum Biol 74(2):211-233.

Kirby KR, Gray RD, Greenhill SJ, Jordan FM, Gomes-Ng S, Bibiko HJ, Blasi DE, Botero CA, Bowern C, Ember CR et al. 2016. D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity. Plos One 11(7).

Kircher M, Sawyer S, and Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 40(1):e3.

Kivisild T. 2015. Maternal ancestry and population history from whole mitochondrial genomes. Investig Genet 6:3.

Kloss-Brandstatter A, Pacher D, Schonherr S, Weissensteiner H, Binna R, Specht G, and Kronenberg F. 2011. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat 32(1):25-32.

Ko AM, Chen CY, Fu Q, Delfin F, Li M, Chiu HL, Stoneking M, and Ko YC. 2014. Early Austronesians: into and out of Taiwan. Am J Hum Genet 94(3):426-436.

Koch-Grünberg T. 1995. Dos años entre los indios: viajes por el noroeste brasileño, 1903-1905: Editorial Universidad Nacional.

Kondo Rd. 2002. En pos de los guahibos : prehistóricos, históricos y actuales : con pistas lingüísticas. Bogotá, Colombia: Editorial Alberto Lleras Camargo.

Landaburu J. 2000. Clasificación de las lenguas indígenas de Colombia. In: González PMS, Rodríguez, M. M. L., & Instituto Caro y Cuervo., editor. Lenguas indígenas de Colombia, una visión descriptiva. Santafé de Bogotá: Instituto Caro y Cuervo.

Langmead B, and Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357--359.

Lathrap DW. 1970. The Upper Amazon. London: Thames & Hudson.

Le S, Josse J, and Husson F. 2008. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software 25(1):1-18.

Lee EJ, and Merriwether DA. 2015. Identification of Whole Mitochondrial Genomes from Venezuela and Implications on Regional Phylogenies in South America. Hum Biol 87(1):29-38.

Lewis CM, Jr., Lizarraga B, Tito RY, Lopez PW, Iannacone GC, Medina A, Martinez R, Polo SI, De La Cruz AF, Caceres AM et al. 2007. Mitochondrial DNA and the peopling of South America. Hum Biol 79(2):159-178.

Li H, and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754--1760.

Li M, Schroeder R, Ko A, and Stoneking M. 2012. Fidelity of capture-enrichment for mtDNA genome sequencing : influence of NUMTs. Nucleic acids research:1--8.

Llamas B, Fehren-Schmitz L, Valverde G, Soubrier J, Mallick S, Rohland N, Nordenfelt S, Valdiosera C, Richards SM, Rohrlach A et al. 2016. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. Sci Adv 2(4):e1501385.

Lowie RH. 1948. The Tropical Forests: An introduction. In: Steward JH, editor. Handbook of South American Indians. Washington: U.S. G.P.O.

Malhi RS, Schultz BA, and Smith DG. 2001. Distribution of mitochondrial DNA lineages among Native American tribes of Northeastern North America. Hum Biol 73(1):17-55.

Maricic T, Whitten M, and Paabo S. 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One 5(11):e14004.

Marrero AR, Silva-Junior WA, Bravi CM, Hutz MH, Petzl-Erler ML, Ruiz-Linares A, Salzano FM, and Bortolini MC. 2007. Demographic and evolutionary trajectories of the Guarani and Kaingang natives of Brazil. Am J Phys Anthropol 132(2):301-310.

Meggers BJ. 1954. Environmental Limitation on the Development of Culture. American Anthropologist 56:801-824.

Meggers J. 1948. The archeology of the Amazon Basin. In: Steward JH, editor. Handbook of South American Indians: Washington : U.S. G.P.O.

Melton PE, Briceno I, Gomez A, Devor EJ, Bernal JE, and Crawford MH. 2007. Biological relationship between Central and South American Chibchan speaking populations: evidence from mtDNA. Am J Phys Anthropol 133(1):753-770.

Miller SA, Dykes DD, and Polesky HF. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16(3):1215.

Milton K. 1984. Protein and Carbohydrate Resources of the Maku Indians of Northwestern Amazonia. American Anthropologist 86(1):7-27.

Mizuno F, Gojobori J, Wang L, Onishi K, Sugiyama S, Granados J, Gomez-Trejo C, Acuna-Alonzo V, and Ueda S. 2014. Complete mitogenome analysis of indigenous populations in Mexico: its relevance for the origin of Mesoamericans. J Hum Genet 59(7):359-367.

Nettle D. 1999. Linguistic diversity of the Americas can be reconciled with a recent colonization. Proc Natl Acad Sci U S A 96(6):3325-3329.

Nimuendajú C. 1950. Reconhecimento dos rios Içána, Ayarí e Uaupés. Journal de la Société des Américanistes 39.

Oota H, Pakendorf B, Weiss G, von Haeseler A, Pookajorn S, Settheetham-Ishida W, Tiwawech D, Ishida T, and Stoneking M. 2005. Recent origin and cultural reversion of a hunter-gatherer group. PLoS Biol 3(3):e71.

Patin E, Siddle KJ, Laval G, Quach H, Harmant C, Becker N, Froment A, Regnault B, Lemee L, Gravel S et al. 2014. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat Commun 5:3163.

Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A, Hooshiar Kashani B, Ritchie KH, Scozzari R, Kong QP et al. 2009. Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Curr Biol 19(1):1-8.

Perego UA, Angerhofer N, Pala M, Olivieri A, Lancioni H, Hooshiar Kashani B, Carossa V, Ekins JE, Gomez-Carballa A, Huber G et al. 2010. The initial peopling of the Americas: a growing number of founding mitochondrial genomes from Beringia. Genome Res 20(9):1174-1179.

Pugach I, Matveev R, Spitsyn V, Makarov S, Novgorodov I, Osakovsky V, Stoneking M, and Pakendorf B. 2016. The Complex Admixture History and Recent Southern Origins of Siberian Populations. Mol Biol Evol 33(7):1777-1795.

Quinque D, Kittler R, Kayser M, Stoneking M, and Nasidze I. 2006. Evaluation of saliva as a source of human DNA for population and association studies. Anal Biochem 353(2):272-277.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, and Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A 102(44):15942-15947.

Renaud G, Kircher M, Stenzel U, and Kelso J. 2013. freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers. Bioinformatics 29(9):1208-1209.

Renaud G, Stenzel U, and Kelso J. 2014a. LeeHom: Adaptor trimming and merging for Illumina sequencing reads. Nucleic Acids Research 42(18):e141.

Renaud G, Stenzel U, Maricic T, Wiebe V, and Kelso J. 2014b. deML: Robust demultiplexing of Illumina sequences using a likelihood-based approach. Bioinformatics(October):1--3.

Santos-Granero F. 2002. The Arawakan Matrix: Ethos, Language, and History in Native South America. In: Hill JD, and Santos-Granero F, editors. Comparative Arawakan Histories: Rethinking Language Family and Culture Area in Amazonia. Urbana: University of Illinois Press. p 25-50.

Schonberg A, Theunert C, Li M, Stoneking M, and Nasidze I. 2011. High-throughput sequencing of complete human mtDNA genomes from the Caucasus and West Asia: high diversity and demographic inferences. Eur J Hum Genet 19(9):988-994.

Schurr TG. 2004. The peopling of the New World: Perspectives from molecular anthropology. Annu Rev Anthropol 33:551-583.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, and Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84(6):740-759.

Sorensen AP. 1967. Multilingualism in the Northwest Amazon. American Anthropologist 69(6):670-684.

Stenzel K. 2005. Multilingualism in the Northwest Amazon, Revisited. Idiomas Indígenas de Latinoamérica-II. Austin: University of Texas.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585-595.

Torres MM, Bravi CM, Bortolini MC, Duque C, Callegari-Jacques S, Ortiz D, Bedoya G, Groot de Restrepo H, and Ruiz-Linares A. 2006. A revertant of the major founder Native American haplogroup C common in populations from northern South America. Am J Hum Biol 18(1):59-65.

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, Vullo CM, and Wallace DC. 1993. Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53(3):563-590.

Uribe Tobón CA, and Instituto Colombiano de Cultura H. 1992. Geografía humana de Colombia. Santa Fe de Bogotá: Instituto Colombiano de Cultura Hispánica.

van Oven M, and Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30(2):E386-394.

Verdu P, Becker NS, Froment A, Georges M, Grugni V, Quintana-Murci L, Hombert JM, Van der Veen L, Le Bomin S, Bahuchet S et al. 2013. Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. Mol Biol Evol 30(4):918-937.

Vidal SM. 1997. Liderazgo y confederaciones multiétnicas amerindias en la amazonia luso-hispana del Siglo XVIII REVISTA ANTROPOLÓGICA 87:19-46.

Vidal SM. 2002. Secret Religious Cults and Political Leadership: Multiethnic Confederacies from Northwestern Amazonia. Comparative Arawakan Histories: University of Illinois Press. p 248-268.

Wallace AR. 1853. A narrative of travels on the Amazon and Rio Negro: with an account of the native tribes, and observations on the climate, geology, and natural history of the Amazon valley. London: Reeve and Co.

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C et al. 2007. Genetic variation and population structure in native Americans. PLoS Genet 3(11):e185.

Wright RM. 2002. Prophetic Traditions among the Baniwa and Other Arawakan Peoples of the Northwest Amazon. Comparative Arawakan Histories: University of Illinois Press. p 269-294.

Yunusbayev B, Metspalu M, Jarve M, Kutuev I, Rootsi S, Metspalu E, Behar DM, Varendi K, Sahakyan H, Khusainova R et al. 2012. The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Mol Biol Evol 29(1):359-365.

Zucchi A. 2002. A New Model of the Northern Arawakan Expansion. In: Hill JD, and Santos-Granero F, editors. Comparative Arawakan Histories: Rethinking Language Family and Culture Area in Amazonia. Urbana: University of Illinois Press. p 199-222.

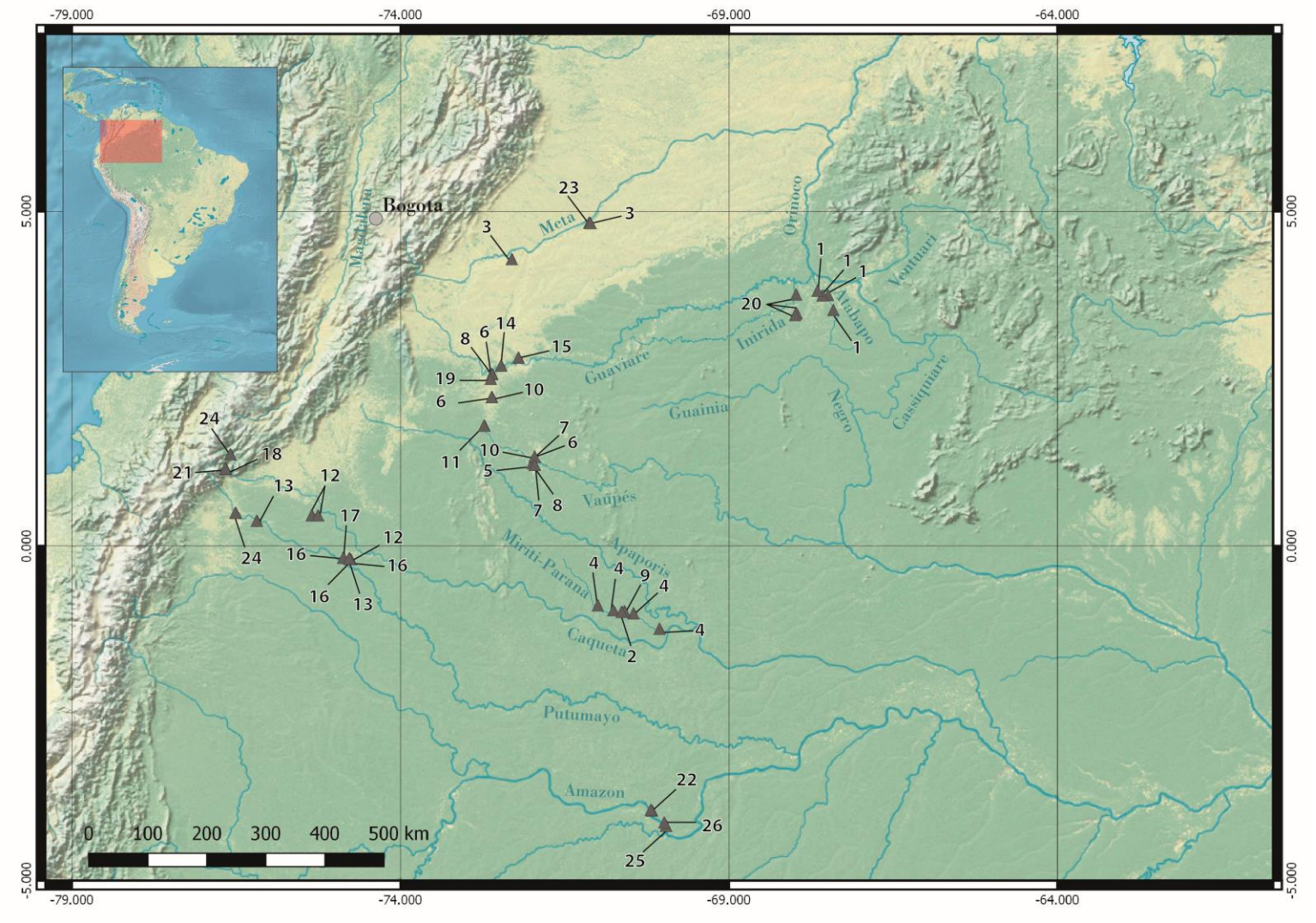

Figure 1. Geographic location of the sampling sites. Every triangle corresponds to a single community, which may contain more than one ethnolinguistic group. 1. Curripaco and Bare, 2. Matapi, 3. Ach-Piapoco, 4. Yucuna, 5. Carijona, 6. Desano, Yuruti, Pisamira, and Karapana, 7. Pira-Wanano, 8. Siriano, 9. Tanimuka, 10. Tukano, 11. Tuyuca and Tatuyo, 12. Coreguaje, 13. Siona, 14. Guayabero, 15. Sikuani, 16. Murui, 17.Uitoto, 18. Kamentsa, 19. Nukak, 20. Puinave, 21. Pasto, 22. Yagua, 23. Saliba, 24. Inga, 25. Tikuna, 26. Cocama.

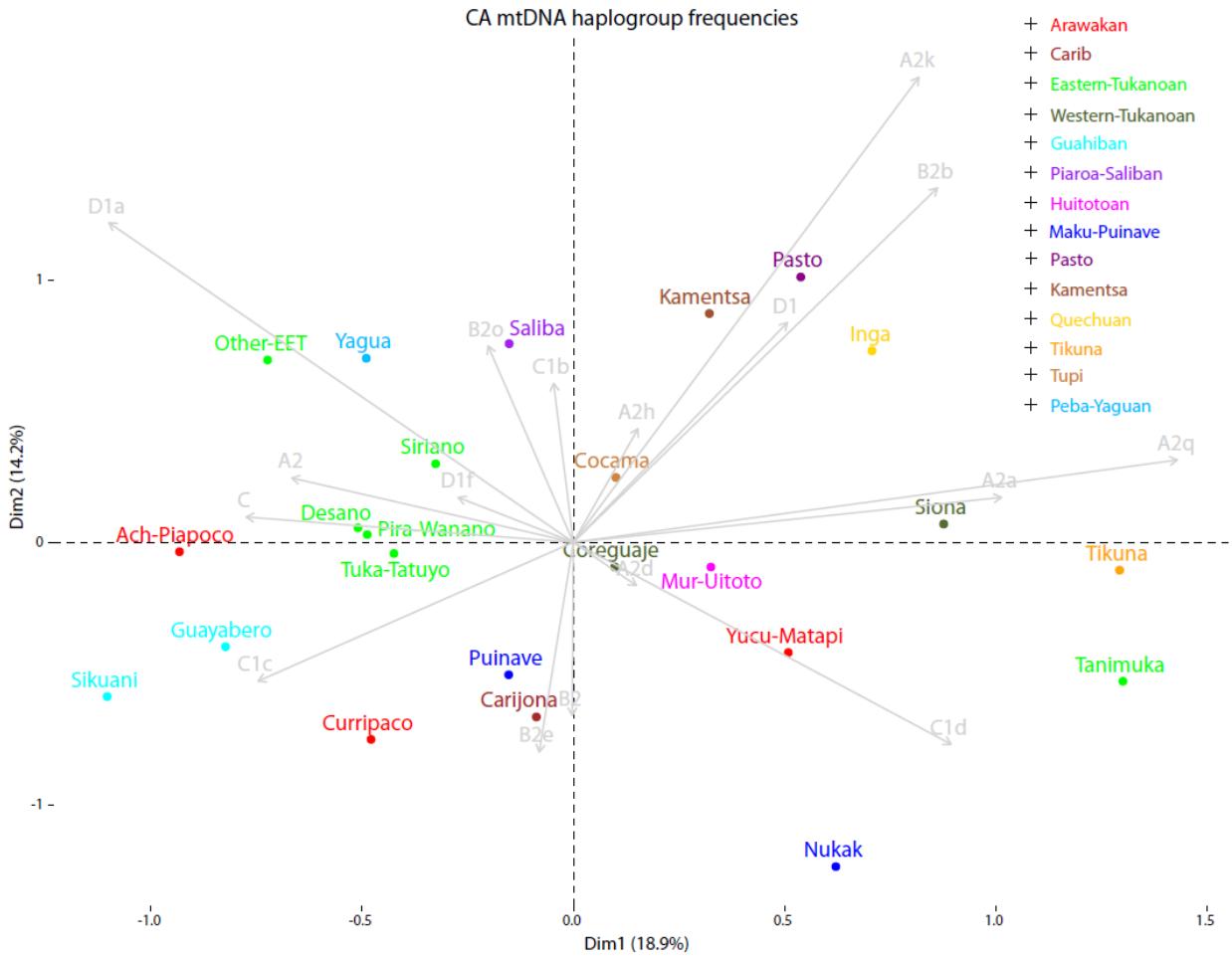

Figure 2. Correspondence analysis based on the sub-haplogroup frequencies by population. Populations are color-coded by linguistic affiliation.

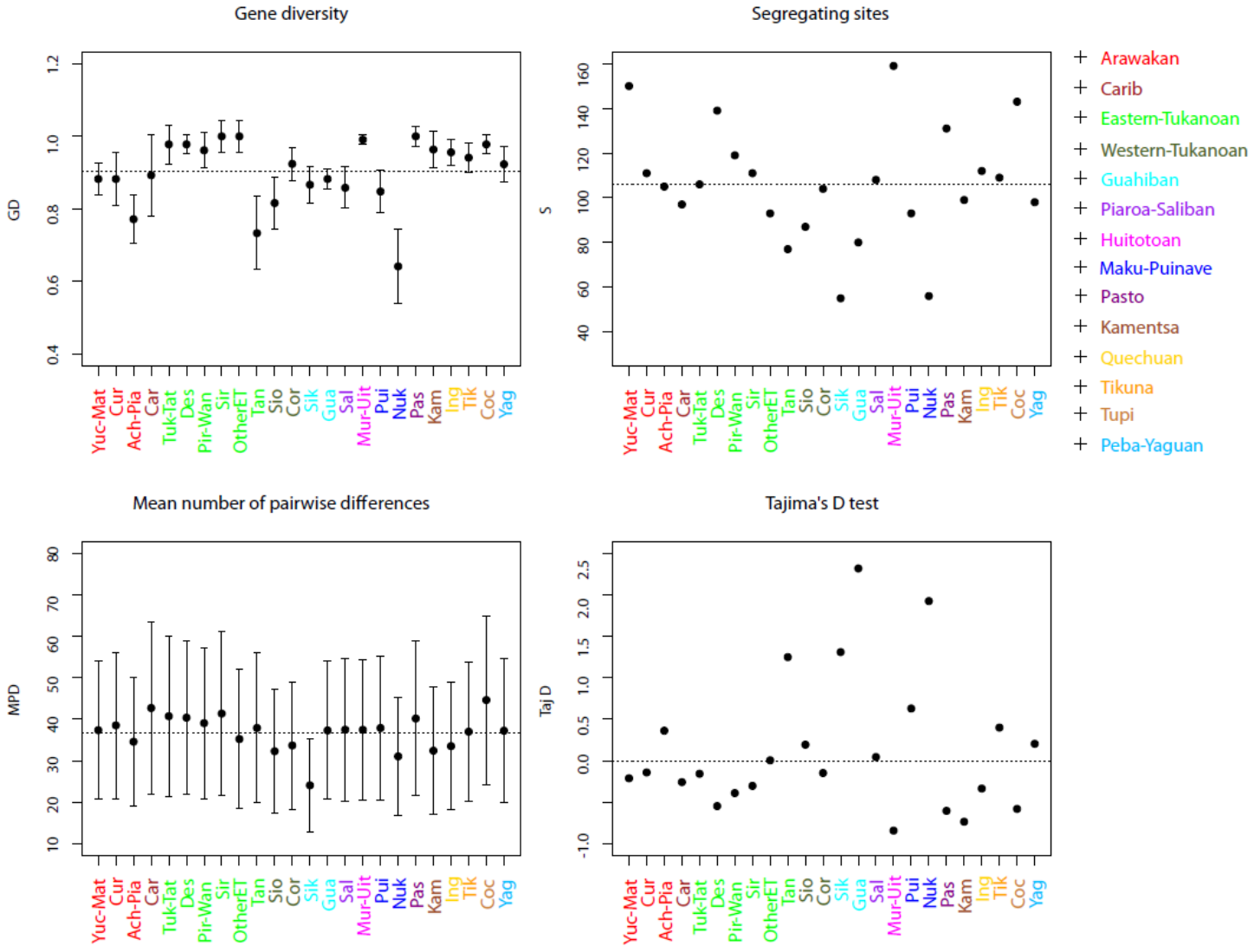

Figure 3. Molecular diversity indices by population. Dashed lines correspond to average values, except for Tajima's D test which corresponds to zero. Populations are color-coded by linguistic affiliation as in Figure 2.

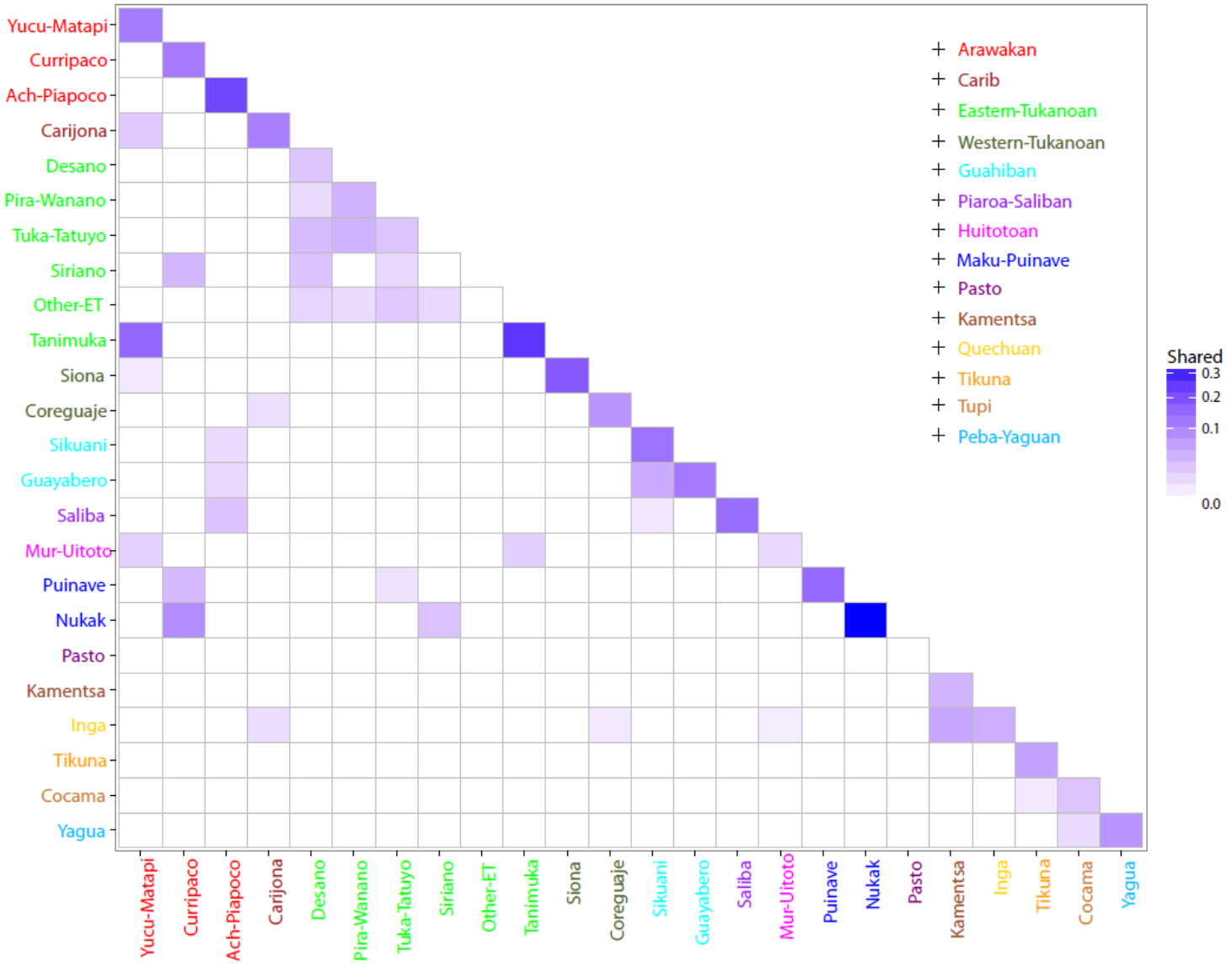

Figure 4. Matrix of shared haplotypes between populations. The color scale indicates the proportion of the total haplotypes that are shared within (on the diagonal) or between (below the diagonal) populations.

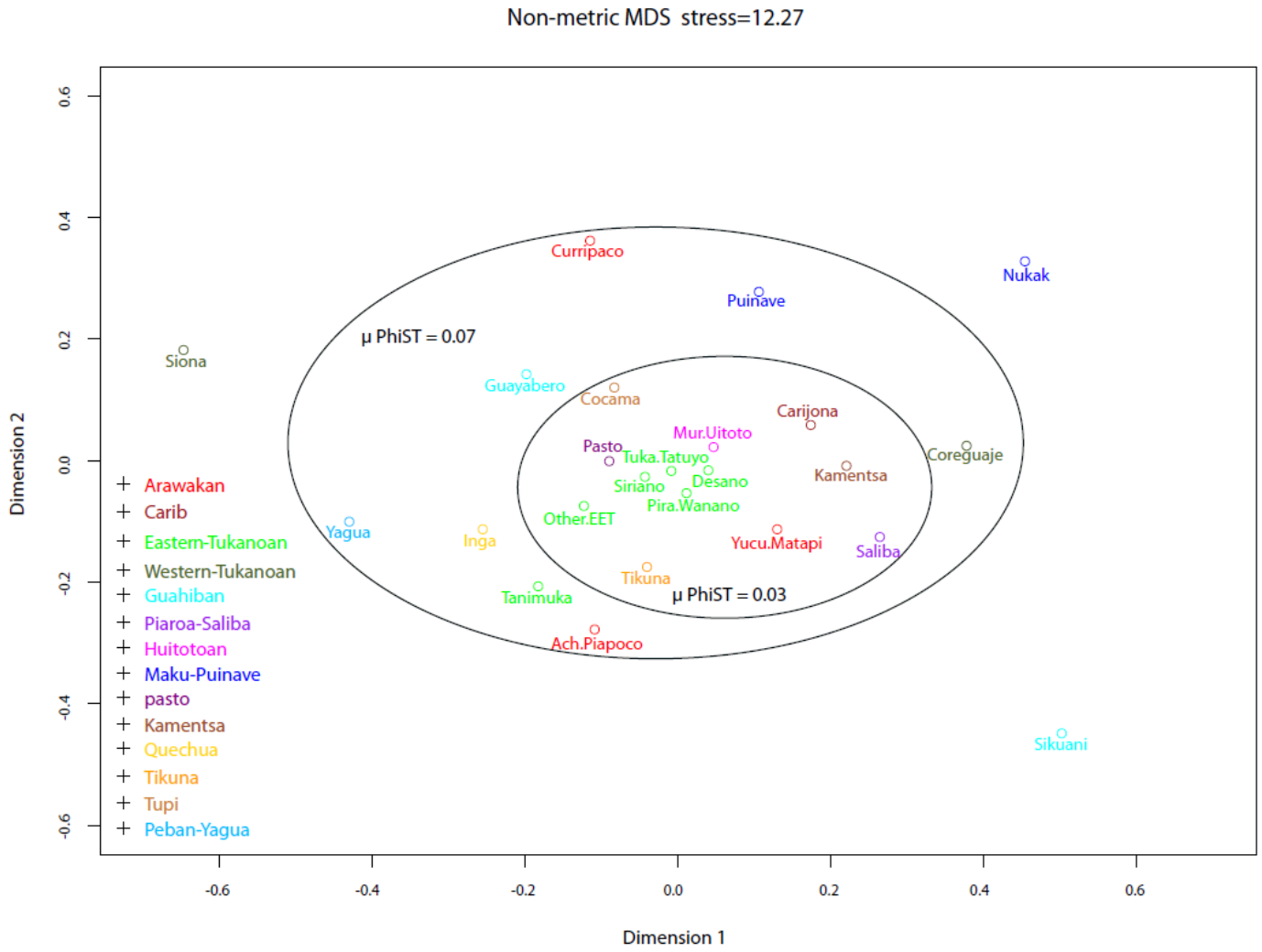

Figure 5. Multidimensional Scaling plot based on Φ_{ST} genetic distances. Stress value is given in percentage. The inner circle indicates populations with low genetic differentiation and the outer circle indicates populations with moderate differentiation. μ PhiST is the average pairwise Φ_{ST} value within each circle.

Table 1. Sampled ethnolinguistic groups with information on merged groups (see Material & Methods: Data Analysis) given below the compound names.

Population	Label in figure 1	N	Census Sizea	Language family	Subsistence Strategyb	River/place of residence
Yucu-Matapi		39		Arawakan	AG	Miritíi-Paraná
Yucuna	4	31	550	Arawakan	AG	Miritíi-Paraná
Matapi	2	8	220	Arawakan	AG	Miritíi-Paraná
Curripaco		17		Arawakan	AG	Atabapo
Curripaco	1	16	7827	Arawakan	AG	Atabapo
Bare	1	1	NAc	Arawakan	AG	Atabapo
Ach-Piapoco		24		Arawakan	AG	Meta
Achagua	3	6	283	Arawakan	AG	Meta
Piapoco	3	18	4926	Arawakan	AG	Meta
Cabiyarid		1	311	Arawakan	AG	Miritíi-Paraná
Carijona	5	8	307	Carib	AG	Upper-Vaupés
Cofan		6	877	Cofan	AG	Guamúez
Barasano		4	2008	Eastern Tukanoan	AG	Upper-Vaupés
Desano	6	17	2457	Eastern Tukanoan	AG	Upper-Vaupés
Kubeo		5	6647	Eastern Tukanoan	AG	Upper-Vaupés
Other-EET		10		Eastern Tukanoan	AG	Upper-Vaupés
Tuyuca	11	7	642	Eastern Tukanoan	AG	Upper-Vaupés
Yuruti	6	1	687	Eastern Tukanoan	AG	Upper-Vaupés
Pisamira	6	1	61	Eastern Tukanoan	AG	Upper-Vaupés
Karapana	6	1	464	Eastern Tukanoan	AG	Upper-Vaupés
Pira-Wanano		13		Eastern Tukanoan	AG	Upper-Vaupés
Piratapuyo	7	8	697	Eastern Tukanoan	AG	Upper-Vaupés
Wanano	7	5	1395	Eastern Tukanoan	AG	Upper-Vaupés
Siriano	8	10	749	Eastern Tukanoan	AG	Upper-Vaupés
Tanimuka	9	10	1247	Eastern Tukanoan	AG	Miritíi-Paraná
Tuka-Tatuyo		10		Eastern Tukanoan	AG	Upper-Vaupés
Tukano	10	8	6996	Eastern Tukanoan	AG	Upper-Vaupés
Tatuyo	11	2	331	Eastern Tukanoan	AG	Upper-Vaupés
Siona	13	17	734	Western Tukanoan	AG	Putumayo
Coreguaje	12	19	2212	Western Tukanoan	AG	Caquetá
Sikuani	15	16	23006	Guahiban	HGP	Guaviare
Guayabero	14	35	1118	Guahiban	HGP	Guaviare
Saliba	23	16	1929	Piaroa-Saliban	AG	Meta
Mur-Uitotoe		26	7343	Huitotoan	AG	Putumayo
Murui	16	18		Huitotoan	AG	Putumayo
Uitoto	17	8		Huitotoan	AG	Putumayo
Puinave	20	19	6604	Maku-Puinave	HGP	Inirida
Nukak	19	16	1483	Maku-Puinave	HGP	Interfluvial
Pasto	21	14	69789	Pasto	AG	Andean

Kamentsa	18	11	4773	Kamentsa	AG	Andean
Inga	24	17	19079	Quechuan	AG	Andean
Tikuna	25	18	7102	Tikuna	AG	Amazonas
Cocama	26	17	792	Tupi	AG	Amazonas
Yagua	22	13	297	Peba-Yaguan	AG	Amazonas
Guambiano		1	23462	Barbacoan	AG	Andean
Nasa		1	138501	Nasa	AG	Andean
Mestizo		9	NA	Mestizo	NA	NA
Total		439				

a. Data from: (Arango and Sánchez 2004). b. AG: agriculturalist; HGP: Hunter-gatherer populations, data from D-PLACE (Kirby et al 2016).and HG (https://huntergatherer.la.utexas.edu/home, accessed on 06.06.2017). d. Populations with label in italics were not considered in the population-based analyses. e. Census data reports the population size including groups that speak five dialectal varieties.

Table 2. Frequency of haplogroups for the 24 NWA ethnolinguistic groups included in the population analyses

Population	N	A2	B2	C1	D1
Yucu-Matapi	39	0.28	0.10	0.56	0.05
Curripaco	17	0.18	0.53	0.24	0.06
Ach-Piapoco	24	0.54	0.04	0.42	0.00
Carijona	8	0.13	0.25	0.63	0.00
Desano	17	0.29	0.12	0.41	0.18
Other-ET	10	0.50	0.00	0.30	0.20
Pira-Wanano	13	0.31	0.08	0.38	0.23
Siriano	10	0.40	0.10	0.40	0.10
Tanimuka	10	0.50	0.10	0.40	0.00
Tuka-Tatuyo	10	0.30	0.10	0.30	0.30
Siona	17	0.59	0.35	0.00	0.06
Coreguaje	19	0.11	0.16	0.74	0.00
Sikuani	16	0.25	0.00	0.75	0.00
Guayabero	35	0.43	0.23	0.34	0.00
Saliba	16	0.19	0.06	0.56	0.19
Mur-Uitoto	26	0.23	0.12	0.31	0.35
Puinave	19	0.11	0.42	0.47	0.00
Nukak	16	0.00	0.31	0.69	0.00
Pasto	14	0.36	0.14	0.21	0.29
Kamentsa	11	0.18	0.09	0.64	0.09
Inga	17	0.59	0.06	0.29	0.06
Tikuna	18	0.44	0.00	0.44	0.11
Cocama	17	0.29	0.35	0.29	0.06
Yagua	13	0.62	0.15	0.23	0.00

Table 3. Shared haplotypes in a worldwide sample of complete mitochondrial sequences sampled at the population level

Geographic region	#Sequences	#Haplotypes	%Unique haplotypes	Shared Within population	Shared Between populations	Source
NW Amazonia	412	216	0.676	0.241	0.144	Present study
Burkina Faso	335	332	0.991	0.006	0.003	(Barbieri et al. 2012)
SW Zambia	169	146	0.897	0.048	0.055	(Barbieri et al. 2013)
Botswana/Namibia	218	128	0.75	0.188	0.133	(Barbieri et al. 2014)
Philippines	365	233	0.734	0.227	0.077	(Delfin et al. 2014)
Sumatra	72	48	0.771	0.229	0.021	(Gunnarsdottir et al. 2011)
Taiwan	549	299	0.669	0.308	0.084	(Ko et al. 2014)
Oceania	1331	650	0.689	0.277	0.106	(Duggan et al. 2014)
Siberia	525	244	0.574	0.336	0.217	(Duggan et al. 2013)
Mexico[a]	113	90	0.867	0.133	0	(Mizuno et al. 2014)

Note: The proportions do not sum up to 1 since some haplotypes are shared both within and between populations. a. The individuals from Mexico are all of Amerindian origin.

Table 4. Analysis of molecular variance (AMOVA)

	# groups	Among groups	Within groups	Within populations	Global FST
One group	1		11.12**	88.88	0.1112
Language[a]	14	-1.33	12.37**	88.96**	0.1104
Geography[b]	6	1.04	10.24**	88.72**	0.1128
Rivers[c]	11	5.42**	5.99**	88.59**	0.1141

** Significant at 0.01 level

[a.] 1.Arawak: Yucu-Matapi, Curripaco, Ach-Piapoco; 2.Carib: Carijona; 3.Eastern-Tukanoan: Desano, Pira-Wanano, Siriano, Tuka-Tatuyo, Other-ET, Tanimuka; 4.Western-Tukanoan: Coreguaje, Siona; 5.Guahiban: Sikuani, Guayabero; 6.Huitoto: Mur-Uitoto; 7.Maku-Puinave: Puinave, Nukak; 8.Kamentsa; 9.Pasto; 10.Piaroa-Saliba: Saliba; 11.Peba-Yaguan: Yagua; 12.Quechua: Inga; 13.Tikuna; 14.Tupi: Cocama.

[b.] 1. Saliba, Ach-Piapoco; 2. Sikuani, Guayabero, Nukak, Desano, Pira-Wanano, Siriano, Tuka-Tatuyo, Other-ET, Carijona; 3. Coreguaje, Siona, Mur-Uitoto, Inga, Kamentsa, Pasto; 4. Curripaco, Puinave; 5. Yucu-Matapi, Tanimuka; 6. Cocama, Tikuna, Yagua.

[c.] 1.Meta: Saliba, Ach-Piapoco; 2.Vaupes: Desano, Pira-Wanano, Siriano, Tuka-Tatuyo, Other-ET, Carijona; 3.Guaviare: Guayabero, Sikuani; 4.Interfluve: Nukak; 5.Atabapo-Inirida: Curripaco, Puinave; 6.High-Putumayo: Inga, Kamentsa, Pasto; 7.Middle-Putumayo: Siona; 8. Lower-Putumayo: Mur-Uitoto; 9. Middle-Caqueta: Coreguaje; 10.Miritíi-Parana: Yucu-Matapi, Tanimuka; 11.Amazon: Cocama, Tikuna, Yagua.

Table 5. Multiple regression analysis on distance matrices

		gen.dist all populations			
		Reg.coefficient	P-val	R.square	P-val
Simple regression	geo.dist	3.28×10^{-8}	0.444	0.009	0.444
Multiple regression	geo.dist	9.64×10^{-10}	0.983	0.036	0.126
	rivers.dist	5.71×10^{-2}	0.011		
		gen.dist without outliers			
Simple regression	geo.dist.no.outliers	6.70×10^{-8}	0.026	0.067	0.026
Multiple regression	geo.dist.no.outliers	5.76×10^{-8}	0.112	0.070	0.040
	rivers.dist.no.outliers	1.38×10^{-2}	0.458		

Supporting information Table S1. Molecular diversity indices by population

Population	N	No.hap	S	GD	GD.SD	ND	ND.SD	MPD	MPD.SD	TajD	Pval
Yucu-Matapi	39	19	150	0.88	0.045	0.0023	0.0011	37.38	16.61	-0.211	0.440
Curripaco	17	11	111	0.88	0.072	0.0023	0.0012	38.54	17.64	-0.141	0.456
Ach-Piapoco	24	8	105	0.77	0.065	0.0021	0.0011	34.60	15.62	0.363	0.719
Carijona	8	6	97	0.89	0.111	0.0026	0.0014	42.71	20.80	-0.257	0.405
Tuka-Tatuyo	10	9	106	0.98	0.054	0.0025	0.0013	40.76	19.37	-0.156	0.467
Desano	17	14	139	0.98	0.027	0.0024	0.0012	40.40	18.47	-0.546	0.322
Pira-Wanano	13	11	119	0.96	0.050	0.0024	0.0012	39.08	18.18	-0.389	0.368
Siriano	10	10	111	1.00	0.045	0.0025	0.0013	41.38	19.66	-0.303	0.388
Other.ET	10	10	93	1.00	0.045	0.0021	0.0011	35.24	16.80	0.006	0.533
Tanimuka	10	4	77	0.73	0.101	0.0023	0.0012	37.91	18.04	1.250	0.928
Siona	17	8	87	0.82	0.071	0.0020	0.0010	32.31	14.84	0.194	0.613
Coreguaje	19	12	104	0.92	0.046	0.0020	0.0010	33.70	15.37	-0.147	0.508
Sikuani	16	7	55	0.87	0.050	0.0015	0.0008	24.08	11.18	1.310	0.941
Guayabero	35	11	80	0.88	0.028	0.0023	0.0011	37.31	16.62	2.318	0.992
Saliba	16	7	108	0.86	0.057	0.0023	0.0012	37.49	17.22	0.046	0.577
Mur-Uitoto	26	23	159	0.99	0.013	0.0023	0.0011	37.46	16.84	-0.841	0.207
Puinave	19	9	93	0.85	0.059	0.0023	0.0012	37.89	17.25	0.630	0.788
Nukak	16	4	56	0.64	0.103	0.0019	0.0010	31.07	14.33	1.925	0.989
Pasto	14	14	131	1.00	0.027	0.0024	0.0013	40.19	18.59	-0.603	0.281
Kamentsa	11	9	99	0.96	0.051	0.0020	0.0010	32.44	15.35	-0.735	0.227
Inga	17	13	112	0.96	0.037	0.0020	0.0010	33.53	15.38	-0.334	0.427
Tikuna	18	13	109	0.94	0.042	0.0022	0.0011	37.02	16.90	0.399	0.722
Cocama	17	14	143	0.98	0.027	0.0027	0.0014	44.64	20.37	-0.580	0.274
Yagua	13	8	98	0.92	0.050	0.0023	0.0012	37.26	17.35	0.205	0.646

No.hap: Number of haplotypes; S: Segregating sites; GD: Gene diversity; GD.SD: GD standard deviation; ND: Nucleotide diversity; ND.SD: ND standard deviation; MPD: Mean number of pairwise differences; MPD.SD: MPD standard deviation; TajD: Tajima's D values; Pval: P values of Tajima's D.