PNAS
www.pnas.org

Supplementary Information for

The genetic structure and adaptation of Andean highlanders and Amazonian dwellers is influenced by the interplay between geography and culture

Victor Borda[1,2,3,19], Isabela Alvim[1,19], Marla Mendes[1,19], Carolina Silva-Carvalho[1,19], Giordano B Soares-Souza[1], Thiago P Leal[1], Vinicius Furlan[1], Marilia O Scliar[1,4], Roxana Zamudio[1], Camila Zolini[1,5,6], Gilderlanio S Araujo[7], Marcelo R Luizon[1], Carlos Padilla[3], Omar Cáceres[3,8], Kelly Levano[3], César Sánchez[3], Omar Trujillo[9], Pedro O. Flores-Villanueva[3], Michael Dean[10], Silvia Fuselli[11], Moara Machado[1,10], Pedro E. Romero[12], Francesca Tassi[11], Meredith Yeager[10], Timothy D O'Connor [13,14,15], Robert H Gilman [12,16], Eduardo Tarazona-Santos[1,17,20], Heinner Guio[3,8,18,20]

Eduardo Tarazona-Santos, Heinner Guio.

Email: edutars@gmail.com, heinnerguio@gmail.com

This PDF file includes:

Supplementary text

Figures S1 to S30

Legends for Datasets S1 to S10

SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S10

Supplementary Information Text

www.pnas.org/cgi/doi/10.1073/pnas.2013773117

Section 1: Sampling, quality control and Datasets

1.1. Sampling: The present work is part of the Peruvian Genome Diversity Project (PGDP), of the Peruvian Institute of Health (Instituto Nacional de Salud - INS). This is a genomic initiative to explore the genetic composition of native and admixed Peruvians. The protocol for this study was approved by The Research and Ethics Committee (OI003-11 and OI-087-13) of the INS. The PGDP was funded by the Ministry of Health of Peru and involves the collaboration of Tarazona-Santos's Laboratory of Human Genetic Diversity (LDGH) of Universidade Federal de Minas Gerais (UFMG) and INS. Fifteen Native American populations (INS data) were sampled as part of the PGDP including populations from the South Pacific Peruvian Coast (SPPC), Andean and Amazon regions (Dataset S1 and Fig. 1).

Populations from the SPPC region were sampled in northwestern Peru and involved two native communities: Moche and Tallanes. In the Andes, four communities were sampled, two Quechua-speaking communities, Qeros and Chopccas, from South Central Andes, and two Aymara-speaking populations, Jaqarus and Uros. The Amazon populations were sampled in two different ecological areas, Amazon Yunga and lower Amazon. The Amazon Yunga corresponds to a cloudy forest which is a transition between the Andean mountains and the Lower Amazon. Six populations were sampled from Amazon Yunga: Chachapoyas, Lamas, Awajun, Candoshi, Ashaninkas and Matsiguenkas. Chachapoyas comprises several groups that played an important role for the Andean people as an open door to Amazon resources (1). The Lamas population, that lives in the Upper Huallaga river, started as a "reduction" (a forced concentration of Native American groups) during the XVIII century (2). Currently, Both Chachapoyas and Lamas groups speak Quechua. The Awajun population belongs to the Jivaroan linguistic family and its presence in the Amazon Yunga is dated back to around 1,200 years ago (3, 4). This Jivaroan population was involved in several cultural trades with South Pacific Coast populations from Peru and Ecuador but with conflicted interaction with the Inca empire (Andes). The Candoshi group is settled along the tributaries of the Pastaza River and its language has a controversial origin by being considered by some scholars as part of the Jivaro group or even independent (5). For the Arawakan linguistic group, three individuals were collected in the Sepahua district from the Ucayali region that belong to the Matsiguenkas tribe. For the Lower Amazon, three populations of the Panoan linguistic family were collected: Shipibo-Conibo, Matses and Nahuas. Finally, all native participants were required to be over 18 years old, for whom all four grandfathers were born in the selected ancestral native population.

We also included four Native American populations (LDGH data) collected by Universidad Peruana Cayetano Heredia and includes two Andean (one Quechua and one Aymara-speaking) and two Amazonian (Ashaninkas and Matsiguenkas from Shimaa community) (Dataset S1 and Fig. S1). This sampling was conducted under approval of the Institutional Reviews Boards from the Universidad Peruana Cayetano Heredia, Asociación Benéfica PRISMA, Universidade Federal de Minas Gerais and Johns Hopkins University. The Quechua-speaking population was sampled in a large area comprising two continuous regions, Ayacucho and Apurimac, for this reason these individuals were grouped as Quechuas_AA. In the case of the Aymara-speaking population, individuals were collected near the Titicaca lake shore in Puno region. The two Amazonian populations inhabit the Amazon Yunga area and belong to the Arawakan linguistic family.

1.2. DNA sampling: For INS data, after collecting the 10ml blood sample from participants (we only collect blood) we proceed to extract the DNA using commercial kits (Qiagen, USA). For this procedure, we travel to the community with our supplies and equipment to perform the DNA extraction, when we obtain the DNA purified we aliquoted, send to our Laboratory (Laboratorio de Biotecnología y Biología Molecular) in Lima, Perú to be stored in -80 °C until we continued with the next project phase genotyping. Specifically, for LDGH data, we extracted genomic DNA from blood samples using the Gentra Puregene blood kit (Qiagen, USA) in the LDGH.

1.3. Genotyping and Quality Control: A total of 289 individuals were genotyped using the Illumina Human Omni array 2.5M at the INS. The total number of genotyped SNPs was 2,391,739. Quality control was performed using the PLINK 1.7 software (6) and in-house scripts (7). We removed SNPs and individuals with high levels of missing data (>10%), loci with 100% of heterozygous, non chromosomal information and A/T-C/G genotypes. LDGH data was genotyped by the Illumina facility using the HumanOmni2.5–8v1 array for 127 individuals. Quality control for LDGH data was the same as for the INS data. We merged the INS data with LDGH data (Dataset S1 and Fig. S1). populations:.The merged data, INS and LDGH individuals, contain a total of 2,077,858 SNPs for 418 individuals organized in a total of 19 populations (Dataset S1). Both groups, INS and LDGH datasets, include independent samples of the Ashaninka population from the same region, for this reason we merge these individuals in a unique Ashaninka sample and a total of 18 populations.

Before filtering by relatedness, we removed SNPs that were in high linkage disequilibrium (LD) for each population, as it affects the inferences of relatedness, with PLINK 1.7 using the flag --indep-pairwise with the following parameters: 200 25 0.1. The first parameter indicates a window of 200 SNPs, the second indicates that the window steps of 25 SNPs between consecutive windows and the third indicates the LD threshold (r^2).

Family structure affects the analysis of population structure as a familiar cluster can be confounded with a discrete population (8). To overcome this issue, we estimated the kinship coefficients (Φij) for each pair of individuals for each population using autosomal SNPs. For each population, we estimated the kinship coefficients using the option –genome in PLINK 1.7. We considered a thresholds of Φij $\geq$0.25 to define relatedness or not. A pair of individuals with Φij above 0.25 is defined as first-degree relatives (Parent-offspring pair and full sibling). We used a network approach to identify which individuals should be removed preserving a maximum number of unrelated individuals (9). After applying the kinship filter, we kept 358 individuals (Dataset S1) for an unrelated dataset (UDataset).

1.4. Merging datasets: We merged the UDataset with the following datasets:

- 1000 Genomes project (10).
- Human Genome Diversity Project (HGDP) (11).
- Native Americans previously genotyped by Reich *et al.* (unmasked data) (12).
- Native individuals from Guatemala (Kaqchikel population) from Michael Dean-Lab (National Cancer Institute).
- Native American individuals from two public datasets (Simons Genome Diversity Project (13) and Raghavan *et al.,* 2015 (14)).

From the 1000 Genomes Project, we selected individuals of European (IBS, CEU), African (YRI and LWK) and East Asian (CHS, CDX, CHB) ancestries. The unmasked dataset from Reich *et al.* (12) included individuals from HGDP: Yakut, Karitiana, Surui, Pima, Maya, Piapoco, Papuan and Melanesian. From the Simons Genome Diversity Project and individuals generated by Raghavan *et al.* (14), we included all Native Americans. The available dataset of Raghavan *et al.* (14) included the ancient genome of Anzick-1 individual from the Clovis complex (hereafter Clovis). Before merging individuals from Reich *et al.* (12), we applied a relatedness filter. We removed 58 individuals with kinship coefficient above 0.1 using the same procedure employed for our samples. We generated three datasets (Datasets S1-S3, Fig. S1) considering the density of SNPs and sample size:

a) **Natives 1.9M Dataset (**1,927,769 SNPs/673 individuals**):** Dataset with maximum number of genotyped SNPs. This dataset includes just Peruvian Native individuals from INS and LDGH and 107 Iberian (IBS), 108 Yoruba (YRI) and 100 East Asian individuals (CDX) from 1000 Genomes Project (Dataset S1).
b) **Natives 500K Dataset (**567,718 SNPs/849 individuals**)**: This dataset includes individuals from **Natives 1.9M Dataset,** Native American, Siberian, South Asian (Onge) and Oceanian (Bougainville and Papuan), individuals from the Simons Project (13), Raghavan *et al.,* 2015 and 79 individuals from Guatemala of Michael Dean NCI lab (Dataset S2) genotyped for 600K

SNPs. The Guatemalan sample includes individuals from the Kaqchikel native population and non-native individuals with more than 99% of Native American ancestry.

c) **Natives 230K Dataset** (235,352 SNPs/1,286 individuals): Dataset with maximum number of individuals. This data includes individuals from Natives 1.9M dataset and all Native Americans from the unmasked data of Reich *et al.* (2012) (~300K SNPs), which includes HGDP individuals (Dataset S3).

The East and South Asian, Siberian and Oceanian populations were used only for population history analysis of the masked data and genotype based methods and not for the population structure analyses in order to avoid any confounding signal.

Use of datasets masked for Non-Native American local ancestry: For *D* statistics analysis (15) and Admixture Graphs (16), we used a dataset where regions of European and African ancestries were masked. These regions were identified using RFMix software (17) and then masked. Using masked datasets and methods based on allele frequency correlation, we inferred genetic affinity among South American Natives. RFMix identifies regions of a specific ancestry in the genome of admixed individuals using reference panels of individuals of European, African and Native American ancestries. For this purpose, we used the phased **Natives 500K** (Dataset S2) and **Natives 230K** (Dataset S3) **datasets**. We used 100 African (YRI and LWK) and 100 European (CEU and IBS) individuals from 1000 Genomes project as parentals. For the Native American reference panel, we selected individuals with less than 0.002% of Non-Native ancestry (European + African ancestries) using the ADMIXTURE results (see Section 2) for 3 ancestry clusters (K=3). All other Native American individuals that have some level of European or African ancestry were used as targets. We ran RFMix with the option PopPhased to enable the phase correction option. We also used two rounds of the expectation-maximization (EM) algorithm. All other settings were used as default. After running RFMix, we used the forward-backward probability output to set all local ancestry inferences that have less than 0.95 posterior probability of being Native American as missing data. Finally, the genomic regions in each sample that did not contain homozygous high quality Native American ancestry inferences were set as missing data.

In this paper, we will apply several methods on our three datasets to explore the following four

scientific questions:

Question 1 (Section 2): whether the between-population homogenization of Western South America, and the dichotomy Arid Andes/Amazonia extends to the northward Fertile Andes?

Question 2 (Sections 2 and 3): whether gene flow accompanied the cultural and socioeconomic interactions between Andean and Amazon Yunga populations?

Question 3 (Section 4): when this between-population genetic homogenization started in the context of the arid Andean chronology.

Question 4 (Section 5): were there episodes of genetic adaptation to the Arid Andes and the Amazonian tropical forest?

Section 2: Genetic relationships in Western South America

To address our scientific questions:

Question 1: whether the between-population homogenization of Western South America, and the dichotomy Arid Andes/Amazonia extends to the northward Fertile Andes?

and

Question 2: whether gene flow accompanied the cultural and socioeconomic interactions between Andean and Amazon Yunga populations?

We performed population structure analysis using two approaches: genotype based (ADMIXTURE and PCA) and haplotype based (CHROMOPAINTER and fineSTRUCTURE) methods.

2.1. Methods

2.1.1. Population Structure using genotype based methods

We applied genetic clustering analysis and Principal Component Analysis (PCA). For the genetic clustering analysis, we ran ADMIXTURE (18). The ADMIXTURE algorithm assumes that the genetic composition of each individual is made up of up to K parental populations or ancestry clusters, where K is defined by the user. ADMIXTURE estimates the fraction of each K population that contributes to an individual, as well as the allele frequencies of each of the K populations, by fitting the Hardy Weinberg equilibrium in each of the K populations/clusters. We ran ADMIXTURE in unsupervised mode for different values of K and used a cross validation (CV) test to determine the K value with the best model fitting. The ADMIXTURE results are represented as a bar plot where each individual is represented by a vertical bar in which each color corresponds to the ancestry proportion of a specific cluster. The PCA is a non-model based method that reduces a complex data (i.e. genotypes and individuals) to few dimensions (19).

ADMIXTURE analysis and PCA assume independence among SNPs, for this reason we pruned all datasets for linkage disequilibrium (LD). We removed highly linked SNPs using PLINK 1.7 with the option indep-pairwise 200 25 0.4 for each dataset. We generate three datasets pruned by LD:

- **Natives 1.9M dataset_LDpruned** (625,736 SNPs)
- **Natives 500K dataset_LDpruned** (229,895 SNPs)
- **Natives 230K dataset_LDpruned** (136,797 SNPs)

We ran 50 replicates of ADMIXTURE in unsupervised mode with different random seeds for each K value and calculated the cross validation error for each run. We ran ADMIXTURE considering from K=2 ancestral clusters until cross validation error started to increase for each dataset. We plot all ADMIXTURE runs with the higher log likelihood for each K value. We ran the PCA using EIGENSOFT 4.21 (19) for the three LD pruned datasets.

Natives 1.9M dataset

ADMIXTURE results are displayed on Fig. S2. The lower CV error was obtained for the run with five ancestry clusters (K=5). ADMIXTURE run with K=3 infers clusters related to continental ancestry: Native American (green), European (IBS, red) and African (YRI, blue) clusters. This result showed some Native American individuals (Quechuas_AA, Chachapoyas and Moche populations) with European ancestry (~10%). Specifically, for the result with the lowest cross validation error (K=5), we observed the Andean populations as a homogeneous group (brown cluster). On the other hand, we observed an ancestry cluster (light green) predominant in Northern Peruvian populations that is shared between SPPC and Chachapoyas population (Amazon Yunga).

For the PCA (Fig. S3), we excluded Africans (YRI) due to its high level of differentiation that masks the relationships in Native Americans. The first principal component (PC1, Variance explained=2.36%) showed an axis of differentiation between the European and Native American groups. We observed that some Andean, Moches, and Chachapoyas individuals have some degree of European ancestry. The PC2 (Variance explained=1.2%) separated Western (Andean and SPPC populations) and Eastern (Amazon) South American natives. Chachapoyas showed affinity with SPPC populations. Jivaroan populations (Awajun and Candoshi), were intermediate in the axis Western-Eastern. Furthermore, the PC2 showed a cline for the genetic diversity of the Amazon populations, from North (Matses) to South (Matsiguenkas). As in ADMIXTURE, both Matsiguenkas groups, Shimaa (Matsiguenkas 1) and Sepahua (Matsiguenkas 2), showed high genetic affinity.

For this dataset, ADMIXTURE analysis and PCA showed high differentiation between populations within the Amazonia and high genetic affinity among Central Arid Andean groups. Chachapoyas showed a close genetic relationship with SPPC populations. Moreover, North Amazon populations (Awajun and Lamas) share ancestry with SPCC as well as with other Amazonian populations.

Natives 500K dataset

ADMIXTURE results are presented on Fig. S4. For bar plot representation, we grouped Surui and Karitiana as Tupian. Mesoamerican individuals were divided into Guatemalan and Mexican (Mixe, Mixtec, Pima, Zapotec and Mayan), and we grouped the Clovis individual, two Greenland, two Aleutian and two Athabascan individuals as North America. Our ADMIXTURE runs showed the lowest CV error for eight clusters (K=8). Our description was focused on patterns not observed on the 1.9M dataset for the lowest cross validation.

ADMIXTURE run K=8 showed 6 clusters associated with Native American groups, associated with Andes (brown), Mesoamerica (purple), SPPC (pink) and three Amazon related clusters (shades of green). SPPC populations showed a predominant pink ancestry that is also predominant in Chachapoyas population. The Andean populations have a predominant brown cluster. Moreover, central Andean populations (Jaqarus, Quechuas_AA and Chopccas) showed ~10% of SPPC related ancestry. Matsiguenkas individuals were observed as a highly differentiated population since it has a specific ancestry cluster (darkgreen) which is not shared with other populations of the same linguistic group (Ashaninkas). Panoan populations (Shipibo, Matses and Nahua) showed a predominant ancestry associated with the Ashaninkas population. Jivaroan groups showed a specific ancestry which was predominant in the Awajun population.

For the Principal Component Analysis (Fig. S5), the PC1 separated Native Americans from Europeans. Some Chachapoyas, Quechuas_AA, 1 Moche, 1 Mixtec and 1 Shipibo individuals showed affinity to Europeans due to admixture. The PC2 separated Amazon from non-Amazon populations; Jivaroan and Tupian individuals were observed as intermediate between these groups. The PC3 separates a group that includes Andean and Matsiguenkas individuals from other natives. PC4 showed the separation between a group including Mesoamericans and Tupian individuals from other natives. Higher PC values showed population specific differentiation and genetic variation in the IBS population.

For this dataset, both ADMIXTURE and PCA support the similarity between SPPC and Chachapoyas individuals. Awajun and Candoshi were intermediate between Andean-Amazon axis of genetic diversity.

Natives 230K dataset

The following description was focused on clusters related to South American natives. The lower CV error was obtained for 18 ancestral clusters (K = 18, Fig. S6). The ADMIXTURE plots (Fig. S6) from K=3 to K=5 inferred continental ancestry clusters. The ADMIXTURE plot K=5, five ancestry clusters related to each continental region were identified: Africa, Europe, Asia, Oceania (light purple, pops 90-91) and America. ADMIXTURE K=6 showed a cluster (dark green) associated with Arawakan groups (Ashaninkas and Matsiguenkas). ADMIXTURE K=8 identified a cluster (light pink) for Costa Rican natives (Bribri, Cabecar, Chorotega, Guaymi, Huetar, Maleku, Teribe). ADMIXTURE K=9 inferred an ancestral cluster (black) related to Mesoamericans, predominantly in Pima. ADMIXTURE K=10 showed an ancestry cluster (light green) related to the Awajun population which represented most of non Andean natives. ADMIXTURE K=12 identified an ancestry cluster (gray) associated with Tupian populations (Surui and Karitiana) also observed in Mesoamericans and in non Andean natives. Furthermore, a brown cluster is predominant in the Andean populations. ADMIXTURE K=13 showed a light blue cluster associated with Pima natives. ADMIXTURE K=15 showed an ancestry cluster, darkgreen and forest green, for each Peruvian Arawakan individuals (Ashaninkas and Matsiguenkas). ADMIXTURE K=17 detected a cluster (light blue) associated with the SPPC population, and which is also present in Chachapoyas and Lamas. The ADMIXTURE K=18 with lowest cross validation showed differentiation in Asian natives. ADMIXTURE K=21 showed an ancestry cluster related to Lamas sample.

In the Principal Component Analysis (Fig. S7), the PC1 showed the differentiation between the Old world from the Native American ancestry. Old world ancestry included Europeans (IBS), Oceanians, Asians (Russian and Mongolians) and admixed Native Americans. Greenland natives were shown as intermediate between the Old and the New world groups. In the Native American axis, the Matsiguenkas individuals were the most differentiated. Considering PC1 and PC2, we discriminated three blocks of Native ancestry, one Eskimo-Aleut (Greenland), the second Athabaskan and Algic (North American group), and the third including all other Native Americans. This pattern is consistent with Reich *et al. (12)*.

In summary, our genotype frequency approach supports the dichotomic model between the Arid Andes and the adjacent Amazonia. However, this is not valid in Northern Peru in which SPPC populations were closely related to Amazon Yunga Chachapoyas, Moreover Jivaroan populations were observed as more similar to SPPC and Chachapoyas than to other Amazon populations Arawakan and Panoan. Furthermore, The Fertile Andes populations (SPPC and Chachapoyas) and Jivaroan populations, conditioning on K=17 ancestry clusters, shared some level of genetic ancestry with Mesoamericans.

2.1.2. Population Structure using haplotype based methods

We used haplotype-based methods (CHROMOPAINTER and fineSTRUCTURE algorithms) that explore the patterns of haplotype similarity among individuals (20). First, we phased our datasets using shapeit2 software (21). For the phasing process, we used the complete dataset (without LD pruning). To increase the accuracy of the phasing process, we used 200 conditioning states and 30 main iterations of Markov chain Monte Carlo (MCMC).

The haplotype-based methods are based on the identification of the LD patterns along the genome of individuals in order to infer the number and length of DNA chunks (CHROMOPAINTER) shared among them. Then, this information is exploited in the identification of clusters of individuals based on the pattern of genetic similarity at a fine scale (fineSTRUCTURE). The identification of shared DNA chunks between individuals is called chromosome painting and is performed by CHROMOPAINTER (20). In this process each haplotype (recipient) is reconstructed based on chunks shared (or “donated”) with (by) other individual haplotypes (donors) (20). For this inferences CHROMOPAINTER requires phased data and two scalar parameters (inferred in a previous CHROMOPAINTER run): 1) the recombination scaling and 2) mutation parameters. The result of the chromosome painting is summarized in two interindividual matrices called coancestry matrices. These matrices of putative donor-recipient similarity contain as their elements the total number (chunkcounts) and the length (chunklengths) of DNA chunks shared among individuals.

After the chromosome painting, we used the chunkcounts co-ancestry matrix to infer the population structure using the model-based approach fineSTRUCTURE (20). Using a reversible-jump MCMC, fineSTRUCTURE assigns individuals into clusters that may resemble their populations. Like other MCMC algorithms, fineSTRUCTURE is dependent on the number of MCMC iterations so it uses a previous burn-in stage and then several iterations (i.e 2 millions).

Since our main question is about the history of Native Americans, we excluded individuals with >5% of Non-Native ancestry in the Natives 230K dataset except for Chachapoyas individuals. We did this because Chachapoyas population has almost all individuals with more than 5% of Non-Native ancestry, so we maintained all individuals (see ADMIXTURE results). We removed slightly admixed individuals just in the 230K dataset because being the most numerous dataset, the exclusion of 202 individuals did not represent a considerable loss. We determined which individuals had to be removed based on ADMIXTURE results (K=3). For Natives 1.9M and Natives 500K datasets we maintained the complete dataset.

To define two scalar parameters (recombination scaling and mutation) for the entire dataset analysis, we ran the Expectation-Maximization CHROMOPAINTER algorithm for a subset of individuals and chromosomes for each data set, we obtained the following values for the parameters. The recombination scaling and mutation were respectively: 220.324 and 0.00018 for the Natives 1.9M dataset, 144.362 and 0.0002 for the Natives 500K dataset, and, finally, 150 and 0.00045 for the Natives 230K dataset. After obtaining the co-ancestry matrix, we ran fineSTRUCTURE for all datasets considering 1,000,000 of burn-in steps, 2,000,000 of MCMC iterations and 100,000 of sampling. After the MCMC calculations, we construct the tree using 100,000 additional steps of hill-climbing steps. We represented the fineSTRUCTURE results as a tree and the chunklengths coancestry matrix as a heatmap.

Natives 1.9M dataset

The fineSTRUCTURE tree clusterize the native individuals in three main groups: natives with some European admixture (Fig. S8A), Amazon (Fig. S8B), SPPC and Andean individuals (Fig. S8C). Almost all Native Americans (~95%) were grouped in clusters containing individuals of the same population label. Most of the SPPC individuals have a close relationship with Andean populations. In the Amazon, we observed that Jivaroan populations (Candoshi and Awajun) were not closely related among them as is observed in other amazon linguistic groups (Fig. S8B). Chachapoyas population showed a close relationship with the admixed Native Americans.

Natives 500K dataset

The arrangement of the clusters was similar to the resulting tree of Natives 1.9M dataset except (Fig. S9): 1) the Moche cluster and the Chachapoyan cluster were more similar to Andean clusters (Fig. S9A), 2) West Mesoamerican natives (Mixe, Mixtec, Zapotec and Pima) clustered together with some Amazon Tupian (Surui and Karitiana) individuals (Fig. S9B), 3) Peruvian Arawakan natives (Matsiguenkas and Ashaninkas) were shown as the most differentiated clusters (Fig. S9D).

Natives 230K dataset

The fineSTRUCTURE tree (Fig. S10A and S10B) showed a more external cluster that contains IBS, Chipewyan and Greenland natives, reflecting the high level of admixture of these North Native American populations. The second more external clusters include Arawakan speakers as the most differentiated populations (Fig. S10B). Other Native Americans were organized in two macro clusters, Andean (Fig. S10B) and non-Andean clusters (Fig. S10A and S10B). The Andean cluster (Fig. S10B) showed no major differences from the clustering of the other datasets, showing Uros as the most differentiated Andean population. Costa Rican populations (Fig. S10B) showed close affinity to Northern South American populations (Embera, Wayuu, Waunana and Kogi). Moreover, Mayan populations (Mayan and Kaqchikel individuals) showed close affinity to Northern Peruvian (Lamas and SPPC) and Inga individuals (Fig. S10A). Panoan individuals (Matses and Nahua) clusterize with other Eastern populations (Fig. S10A).

Summarizing the fine-scale population structure analyses, we inferred that Mesoamericans (Maya and Kaqchikel) share more ancestry with SPPC natives than with Arid Andes populations. The Arid Andes populations showed high similarity among them. Moreover, Chachapoyas populations were highly similar to SPPC individuals.

2.1.3. Ancestry profiles inferred by GLOBETROTTER and SOURCEFIND

We performed a Chromosome painting inference for each Native American population, setting a population as recipient from all other populations (CHROMOPAINTER "-f" switch). This inference result in a chunklengths matrix that summarizes the contribution (shared DNA) from the donor populations to the recipient. Then, with this matrix, we applied two approaches to infer the ancestry proportions: a regression model (non-negative least squares) implemented on GLOBETROTTER software (MIXTURE MODEL) (22–24) and a Bayesian model implemented in SOURCEFIND (25). To generate the chunklengths output for each Native American groups, we used the same two scaling parameters inferred in the last subsection. Furthermore, since fineSTRUCTURE and ADMIXTURE showed no differences among Matsiguenkas individuals, Matsiguenkas 1 (Shimaa) and Matsiguenkas 2 (Sepahua), for GLOBETROTTER inferences we considered these two groups as one single Matsiguenkas group.

Natives 1.9M dataset

The MIXTURE MODEL (Fig. S11) revealed a particular pattern in Andean populations. Each Andean population shares a high proportion of DNA with other Andean populations showing a homogeneous pattern. Specifically, the ancestry profile of the Uros population showed that 95% of its DNA is shared with the Puno population indicating lower genetic variability of Uros. Moreover, Aimara-speaking Jaqarus showed high genetic affinity with Quechua-speaking groups. The SPPC populations have more affinities with northern Amazon populations. These affinities are related to the similar ancestry proportions and the ancestry related to Chachapoyas. Also, the three North Amazon populations (Candoshi, Awajun, Lamas and Chachapoyas) showed a significant proportion of ancestry related to SPPC populations (>10%). Furthermore, we observed that SPPC and North Amazon populations have a significant proportion of shared DNA with Andean population (>20%). Simulations suggest that SOURCEFIND tends to eliminate contributions that could be due to background noise (25). For this particular dataset, SOURCEFIND identifies the SPPC ancestry in north Amazon populations and that most of the Andean ancestry is explained by Quechua related ancestry (Fig. S11). For this dataset, that has the advantage of being the more dense in terms of number of SNPs, it is important to consider the poor representation of other native populations, Mesoamericans and South Americans. The Amazon populations showed a common pattern of genetic composition among them. Only Quechuas_AA, Jaqarus, Moche and Chachapoyas showed European ancestry.

Natives 500K Dataset

The MIXTURE MODEL (Fig. S12) showed the Andean populations with the same pattern as was observed with the Natives 1.9M dataset, they share a high proportion of DNA with other Andean populations showing a homogeneous pattern. Furthermore, it is possible to observe Mesoamerican related ancestry in almost all Peruvian natives. The ancestry composition of SPPC populations showed more similarity to Chachapoyas population. Moreover, Jivaroan and Lamas populations showed very similar patterns of ancestries. We observed that Moche and Chachapoyas populations have a significant proportion of shared DNA with Andean population (~30%). All SPPC and Amazon populations (except Ashaninkas and Matsiguenkas) showed more ancestry related to Mesoamerican (>18%) than to the Andean populations. Matsiguenkas showed more ancestry related to Ashaninkas, indicating a close genetic affinity among Arawakan populations. SOURCEFIND (Fig. S12) showed that the Mesoamerican related ancestry is particularly high in SPPC and Karitiana populations. The Andean contribution in the Amazon was reduced and restricted to the Chachapoyas population. Moreover, Jivaroan (Awajun and Candoshi) populations showed some contribution from the SPPC group.

Natives 230K Dataset

Using fineSTRUCTURE results (Fig. S10A and S10B), we organize Native American populations in the following clusters:

- Chopccas (n=17)
- Quechuas_Per (n=13) [Quechuas_AA and Quechua_R2]
- Aimaras_PB (n=29)
- Qeros (n=12)
- Uros (n=13)
- Quechuas_Bol (n=10)
- Ashaninkas (n=35)
- Matsiguenkas (n=26)
- Matses (n=11)
- Nahua (n=2)
- Awajun (n=23)
- Lamas (n=21)
- Chachapoyas (n=9)
- Moche (n=25)
- Tallanes (n=34)
- Pima (n=21)
- Tepehuano (n=21)

- Maya (n=29)
- Kaqchikel (n=6)
- Mixe (n=17)
- Zapotec1 (n=6)
- Zapotec2 (n=21)
- Mixtec (n=5)
- Karitiana (n=8)
- Surui (n=10)
- Guahibo (n=6)
- Palikur (n=3)
- Piapoco (n=6)
- Ticuna (n=4)
- Embera (n=4)
- Kogi (n=3)
- Waunana(n=3)
- Wayuu (n=2)
- Toba (n=4)
- East_Amazon_Brazil [Arara (n=1) - Parakana (n=1)]
- Inga (n=2)
- 1 Chaco cluster [Guarani (n=3) - Chane (n=2)]
- Wichi (n=4)
- Maleku (n=2)
- Cabecar (n=16)
- Guaymi (n=5)
- Teribe (n=3)
- Chipewyan (n=3)
- East Greenland (n=3)
- IBS cluster (n=107)
- CDX cluster (n=93)

The number in parenthesis indicates the number of individuals after filtering the ones with more than 5% of European ancestry. We did not include Chono and Huilliche as donors for two reasons 1) Due to the low probability that they were involved in gene flow events with Central Andes or Amazonia (12, 26), and 2) almost all individuals have higher levels of European ancestry (>5%). Furthermore, the Quechuas individuals from our dataset that clustered with Quechuas R2 from Reich *et al.* (2012) (12) were included as Quechuas Per and all Quechuas R1 were considered as Quechuas Bol (from Bolivia). Although Jamamadi samples form a cluster with Arara and Parakana, these individuals are geographically distant from Arara and Parakana, for this reason we excluded them from the GLOBETROTTER analysis. After the contribution inferences, and just to improve the visualization of the results, we made an arbitrary merge of clusters as follow:

- North Amazon 1: Inga and Ticuna.
- North Amazon 2: Guahibo, Palikur and Piapoco.
- Caribbean: Embera, Kogi, Wayuu and Waunana.
- Chaco natives: Wichi, Guarani and Chane.
- Pampas: Toba.
- Central America: Cabecar, Maleku, Guaymi and Teribe.
- Mayan: Maya1 and Maya 2.
- West Mesoamericans: Mixe, Mixtec, Zapotec1 and Zapotec2.
- North American: East Greenland and Chipewyan.

The MIXTURE MODEL (Fig. S13) showed a contrasting pattern between Western (SPPC and Arid Andes) and Amazon groups. The SPPC populations as well as Amazonian groups showed shared haplotypes with Mesoamerican groups. Differently, SOURCEFIND (Fig. S13) only detected sharing with Mesoamerican haplotypes for Chachapoyas, Awajun and Inga groups. Moreover Chachapoyas and Inga groups showed some level of Andean related ancestry.

2.2. Conclusions

Considering our Question 1, we conclude that the genetic dichotomy between populations living on the Arid Andes and adjacent Amazonia **does not extend** to the Fertile Andes. Furthermore, Chachapoyas, an Amazon Yunga population, is more genetically similar to SPPC populations than to Arid Andes or other Amazon populations. Regarding our Question 2, GLOBETROTTER results suggest longitudinal (west/east) gene flow in the northern of Peru.

Section 3: Cultural interactions were accompanied by gene flow events across the Andes

3.1. Introduction

Further evidence of commercial interaction between populations of the Coast, Andes and Amazon of fertile Northern Peru: Archaeological evidence suggests an intensive interaction across the Fertile Andes in contrast to the Arid Andes. Cultural and commercial interactions involve the South Pacific Coast, Andean and Amazonian populations (3, 4, 27, 28). Archaeological data point out that the Fertile Andes (involving Southern Ecuador and Northern Peru) was a main crossroads for the Andes-Amazon interaction (27), which could be facilitated since the Andean mountain chain has its lowest altitude in this region (29, 30). These cultural and commercial interactions involved the trade of Spondylus shells from the Coast and medicinal plants and herbs from the Amazon. Particularly, Chachapoyas, an Amazon Yunga, was part of a significant interregional exchange network during the Early Intermediate period (around 2300 YBPto 1400 YBP) and that extended to the Inca period (around 1500 AD) (1). Also, it was suggested a socioeconomic exchange between Moche (Coast) and Awajun (Amazon Yunga) natives (3). Nowadays, Chachapoyas and Lamas populations both living in the Amazon Yunga speak Quechua, which could be adopted as a lingua franca in the last centuries. Specifically, Lamas population is an intriguing case since it was attributed to have an Andean Origin (related to Chanka population). A recent study demonstrated that Lamas population has a closer genetic affinity to surrounding Amazon populations than with Chanka population, suggesting a possibly Amazonic origin instead of Andean (31), which is consistent with our results in this paper (Figs S2-S13). In the latter section, we showed that Jivaroan populations were more similar to Fertile Andes populations (SPPC and Chachapoyas) than to Arid Andes populations. This genetic proximity could be related to the historical interactions. Here we address:

Question 2: whether gene flow accompanied the cultural and socioeconomic interactions between Andean and Amazon Yunga populations?.

To address this question we performed Patterson's statistics analyses: *D* statistics and Admixture graphs analyses on the masked data. The *D* statistics determine if two populations have an excess of alleles sharing due to gene flow. The admixture graphs explore the best model of relationships among populations taking into account gene flow events in a statistical framework.

3.2. Methods

3.2.1. *D* statistics

The *D* statistics (15, 32), or ABBA-ABAB test, is a method to detect gene flow among closely related populations. It evaluates four populations: P_1, P_2, P_3 and an *Outgroup*:

$$Outgroup\ (P_3 \quad (P_1, \quad P_2))$$

This treeness test considers, as a null hypothesis, that P_1, P_2, P_3 and an outgroup have a tree relationship. In this null hypothesis, P_1 and P_2 diverged earlier from the ancestor of P_3, with no gene flow between P_3 and P_1 or P_2 after the divergence. The alternative hypothesis is that P_3 was involved in gene flow with P_1 or P_2 after the divergence. This analysis is restricted to biallelic sites and considered an allele "A" as the ancestral allele of the outgroup (O) and an alternative allele "B" in P_3. Considering the order O-P_3-P_1-P_2, the *D* test is focused on the ABBA or ABAB pattern. The first pattern (ABBA) corresponds to the total number of sites with the alternative allele (B) shared only by P_1 and P_3. The second configuration (ABAB) is the total number of sites with the alternative allele (B) shared only by P_2 and P_3. The *D* statistic is calculated by the relationship of: the difference of ABBA-ABAB counts (numerator) and the total count ABBA+ABAB (denominator). The numerator of this relationship indicates the signal of the statistic which is interpreted as the direction of gene flow. If the *D* statistic is not significantly different from zero, we accept the null hypothesis of treeness. If the result differs significantly from zero, we reject the null hypothesis and consider the possibility of gene flow between P_3 with P_1 or P_2. For *D* statistics estimation we used ADMIXTOOLS (16). The results are interpreted as follows: negative values of *D* are interpreted as gene flow between P_1 and P_3 and positive values indicate a gene flow between P_2 and P_3. A *D* value is considered statistically significant if the absolute value of the relationship between the *D* value and its standard deviation is equal or above 3 (|Z-score| □ 3). We applied D statistics to the masked Datasets 500K and 230K. Considering the huge number of combinations for *D* inferences and even obtaining highly significant values, we construct Q-Q plots to analyze the Z-score distribution, which is expected to be approximately normal under the hypothesis that our results could be expected by chance (H_0) (33).

Cultural interactions and gene flow across the Andes

Considering the divergence between Western (including Andean, SPPC) and Eastern (Amazon) groups, we tested the following configurations:

- (Outgroup, Eastern (Western$_{2,}$ Western$_1$))
- (Outgroup, Western (Eastern$_1$, Eastern$_{2,}$))

Outgroup: For both masked datasets, we used Africans (YRI).

Natives_500K Dataset (Masked)

Configuration **(Outgroup, Western (Eastern$_{2,}$ Eastern$_1$)):**
We explored the gene flow among populations in South America. This configuration explores if one western population shares more alleles with one population from the Amazon (Amazon Yunga or Lower Amazon) than another. Deviation from the diagonal of the Q-Q plot indicates that some populations in this configuration are involved in gene flow events. We observed that the major signals of gene flow involved Chachapoyas, Lamas and Jivaroan populations (Awajun and Candoshi) with SPPC (Fig. S14A).

Configuration **(Outgroup, Eastern (Western$_{2,}$ Western$_1$)):**
When we explored the possible signal of gene flow from Eastern (Amazon yunga and Lower Amazon) to Western, we found results similar to the first configuration, a significant deviation from the diagonal involving SPPC, with Chachapoyas, Lamas and Jivaroan populations (Fig. S14B).

Natives_230K Dataset (Masked)

For this dataset, both configurations showed congruent results with the Dataset 500K (Fig. S15A - B) and no new signals of gene flow appeared.

In conclusion, regarding **Question 2: whether gene flow accompanied the cultural and socioeconomic interactions between Andean and Amazon Yunga populations**?, *D* statistics show evidence of longitudinal gene flow between the north Coast of Peru (part of the so-called northern Fertile Andes) and Amazonian populations of similar latitudes.

3.2.2. Admixture graphs

Rationale: Population Structure Analysis (Section 2) and *D* statistics showed high genetic affinity among Native American groups of Fertile Andes. Strong signals of *D* statistics involved SPPC and Jivaroan populations, Chachapoyas and Lamas suggesting gene flow among these groups (Fig S14-S15). In order to determine the direction and parameters (contributions) of the gene flow or admixture events we applied admixture graphs using qpGraphs in ADMIXTOOLS (16). qpGraph evaluates the fit of *a priori* suggested tree and the *f*-statistics applied on a set of populations included in the tree.

We test two contrasting hypotheses of gene flow. **The first one** (W→E), Amazon groups (Eastern) receives a contribution from the Coast (Western) and **the second one** (E→W), Coast (Western) receives a contribution from Amazon (Eastern). As a result, qpGraph offers a log-likelihood of the fit, Z-score values for the *f*-statistics and the parameters involved in admixture events (contributions) if these are tested. We accept an hypothetical tree if the absolute value of the highest *f*-statistic is less than 3 (|Z-score| < 3). Moreover, if we obtain trees with similar Z-score values, we select the tree with fewer zeroed branches and lower log-likelihood.

First, we selected a small group of populations from the masked Natives 500K dataset and we merged it with the MA1 sample (34). Our final dataset included 235,402 SNPs. The MA1 sample is important due its relationship with one of the dual ancestries that gave origin to Native Americans (34). To create a draft tree in which admixture events will be fitted, we select the following populations:

African:	YRI
Siberian:	MA1
South Asian:	Onge
East Asian:	CDX
North America:	Clovis
Mesoamerican:	Mixe
Amazon:	Ashaninkas and Matses
Andes:	Uros

The second step was to add the Tallanes population from the Peruvian Coast (Western) and a Northern Amazon population (Awajun, Candoshi, Lamas, Chachapoyas). We test each of these populations as unadmixed and admixed. We selected Tallanes since this population showed the most of the highest values of *D* statistics and therefore, was likely involved in gene flow. We tested our hypotheses for 4 combinations:

-Tallanes-Awajun
-Tallanes-Candoshi
-Tallanes-Lamas
-Tallanes-Chachapoyas

Results: All graphs (data not shown) that fit Tallanes and Northern Amazonas as unadmixed showed poor fit. When we allowed for gene flow considering our two contrasting hypotheses, the best fit was for the hypothesis 1 (W→E) (Fig S16A,S17A,S18A,S19A). Admixture graphs for hypothesis 1 and 2 showed the same value for the worst *f*-statistic (Z-score), but hypothesis 2 (E→W) included zeroed length branches, except for the Tallanes-Chachapoyas combination. In this last combination the hypothesis 1 has the best fit with the lower Z-score (Fig S19A). These admixture graphs support a predominant contribution from Coast related populations to the North Amazon.

3.3. Conclusion

- Populations around the Fertile Andes were involved in gene flow events. Specifically, North Coast populations showed a significant contribution to the genetic ancestry or North Amazon populations.

Section 4: Dating the between-population homogenization of the arid Andes

Question 3 (Section 4): when the Andean between-population genetic homogenization started in the context of the arid Andean chronology.

4.1. Methods

4.1.1 Identical-by-descent segment analysis

We analyzed the pattern of segments identical-by-descent (IBD) to infer the relationship among populations across the time. If two DNA segments are identical and have the same ancestral origin they are considered Identical-by-descent (35, 36). From one generation to another, large segments of DNA are inherited, but in successive generations recombination events break these regions (37). The relationship between the size of an IBD segment found between two individuals and the time in generations until coalescence have the following approximation (38, 39):

$$E \simeq 3/2L$$

Where:
E: time in generations to the most recent common ancestor.
L: length of IBD segments (in units of Morgan).

To infer the pattern of gene flow along the time, we ran RefinedIBD software (40) with the **Natives 1.9M Dataset** and **Natives 500K Dataset**. To analyze the demographic evolution in Central Andes, we used IBDne software (41) with the **Natives 1.9M Dataset**, both approaches are described below.

RefinedIBD

To infer IBD segments, we used RefinedIBD (40). This software performs two steps: first, it uses the GERMLINE algorithm (42) for IBD detection, and second, a refinement step, that calculates the probability of each segment to be IBD (40). We removed all missing data in the specific dataset selected for this analysis using PLINK (--geno parameter). We used the genetic map GRCh37 from HapMap and we restricted our analyses to segments larger than 3.2cM. We organized the IBD segments in four intervals that could be related to historical periods, considering one generation as a period of 28 years:

1)	3.2 to 4.2cM	(50 to 36 generations before present)
2)	4.2 to 7.8cM	(36 to 19 generations before present)
3)	7.8 to 9.3cM	(19 to 16 generations before present)
4)	all segments greater than 9.3cM	(16 generations before present to present day)

The first interval is related to pre-Inca times, more specifically to the Middle Horizon and Late Intermediate, that correspond to the Wari-Tiwanaku Empire. The second interval involves the rise and fall of the Inca Empire. Finally, the last interval is related to colonial times until the present day (Fig. 2, Fig. S20).

We calculated the average amount of shared DNA between two individuals from the same (aaIBD) or different populations (abIBD), for each interval(40). Considering a specific pair of populations (a and b), we calculated the total amount of shared DNA between one sample from “a” and another from “b”. After that, we sum all pairwise values and divided by the number of pairs between a and b:

$$\text{abIBD} = \frac{\Sigma' ij Lij}{Npairs}$$

Where:
abIBD: average of the total shared IBD length between two individuals from different populations (or the same population if it is aaIBD).
i and *j*: the two individuals.
L: total IBD length shared between each pair of individuals

Npairs: Na * Nb (for different populations), and Na(Na-1)/2 (For same population). Where N is the number of individuals in the respective population (a or b).

The representation of IBD relationships was presented as a similarity heatmap constructed with the log of the abIBD values (Fig. 2, Fig. S21).

Natives 1.9M dataset

In the first interval (3.2 to 4.2 cM, Fig. 2B) it is possible to observe homogeneous patterns among Andean populations. We did not observe differences between the intra and interpopulation sharing in the arid Andes. In a temporal view, this interval coincides with the Middle Horizon (43) that included the expansion of Tiwanaku-Wari and its falling. This society, which dominated the political landscape of the central highlands of the Andes, was probably an ancestor of Quechua-speaking populations (44), which may be related to the fact that 3 of the 6 Andean studied populations speak this language today. Posteriorly, the difference between intra and interpopulational sharing ratio for Andean populations gradually increased until the most recent interval, but remains smaller than other groups. However, the hypothesis that the Andean homogenization already existed before the Incas was evidenced by the visualization of the high degree of sharing of IBD segments between these groups during the Tiwanaku-Wari expansion.

Natives 500K dataset

In the earliest interval (Fig. S21A), the Andean region already appears homogeneous, corroborating the **Natives 1.9M dataset** results. The SPPC populations showed high internal affinity degrees. The Amazon group in general has some relations with other groups, but the diagonal is very intense, evidencing its high degree of intrapopulation IBD. In the second interval (Fig. S21B), corresponding to the period between the falling of Tiwanaku-Wari Empire and the beginning of the Inca Empire, the arid Andes stay homogeneous. The next period (Fig. S21C) comprises the entire duration of the Inca Empire, which remains, in general, homogeneous. In the last interval (Fig. S21D), after the Europe conquest, Andes is apparently more structured. SPPC populations remain connected since the first interval, as do Matses and Lamas. Like the first dataset, the genetic affinity between Chachapoyas and Andean and SPPC is constant along the intervals.

4.1.2. IBDne

To understand the demographic dynamics of populations in the arid Andean, we calculated the pattern of effective population size (Ne) with software **IBDne** (41). This algorithm infers the pattern of Ne along the generations, allowing us to study how demographic changes make the genetic diversity of a population vulnerable to genetic drift. This method has some particularities that need to be taken into account: 1) it tends to smooth over sudden changes in Ne, 2) it assumes a closed population, 3) it assumes a homogenous population. For this reason we performed the analysis just for the arid Andean group. To avoid the underestimate of effective population size, we restricted the analysis to segments larger than 4cM, as suggested by the authors for array data (41). We inferred this parameter (Ne) only between 4 and 50 generations before the present, because segments related to the last 3 generations are not informative for the dynamic of the population. As our arid Andean populations are genetically homogeneous, we grouped as a unique population for the IBDne inference, which would not be acceptable for the other groups. Moreover, as the density of SNPs is also an important factor for these inferences, we only applied this method for **Natives 1.9M dataset**.

Natives 1.9M dataset

In the earliest heatmap interval, approximately between 50 to 36 generations before present, we can see an expansion period in the population effective population size. After approximately 27 generations, the Ne decreased (Fig. S22), which can mean a bottleneck or continuous population structuring, this reduction stopped in the last 10 generations.

4.2. Conclusions

- The Andean homogenization already existed before the Late Intermediate. Probably related to the Tiwanaku-Wari expansion.

.

- Inferences on the dynamics of the effective population size based on IBD suggest that the decline in population size that followed the European conquest (~1500 AD) affected the genetic diversity of Andean populations, making it more vulnerable to be affected and lost by genetic drift.

- The effective population size (Ne), estimated from IBD segments, shows the dynamics (Fig. S22) characterized by a Post-Contact decline to around one-third the level observed around 1250 years ago (Middle Horizon), when it was rising likely due to an increase in population size in the arid Andean regions.

Section 5: Genetic Differentiation and Natural Selection in the Andes and Amazon

5.1. Introduction

The evolutionary mapping of genetic variants is an efficient approach to identify functional genomic regions that have played an essential role in survival, and possibly have consequences for human health (45, 46). The evolutionary history of modern humans is marked by major migration events for environments with different climates, diets, and diseases (47). These factors compose the selective pressures that act on variants that affect biological mechanisms that influence the adaptation process (48, 49). The process of natural selection leaves genetic signatures that can be detected, making it possible to identify regions of the human genome related to these mechanisms.

In the following section, we applied statistical methods based on population differentiation (Population Branch Statistic - PBS) and linkage disequilibrium (cross population extended haplotype homozygosity - xpEHH) to identify genomic regions under natural selection in Andean and Amazon populations. For this purpose, we used the **Natives 1.9M dataset** considering only the following populations organized in **two groups**:

1) **Arid Andean group:** Chopccas, Quechuas_AA, Qeros, Puno, Jaqarus, and Uros.We excluded 2 Quechua individuals who had more than 10% non-native ancestry according to the ADMIXTURE analysis.

2) **Amazon group :** Ashaninkas, Matsiguenkas (including Matsiguenkas 1 and 2), Matses and Nahua. We did not include Awajun, Candoshi, Lamas and Chachapoyas in this analysis because our previous results (Section 2 and 3) demonstrated that these populations were involved in gene flow and this may mask differentiation signals.

5.2. Methods

Natural Selection Candidate SNPs

5.2.1. Population Branch Statistic

PBS is a statistical test to identify changes in the allele frequencies of a target population since its divergence from an ancestral population. PBS is based on the comparison of differentiation (F_{ST}) values among three groups: 1) the target population; 2) a population closely related to the target, and 3) an outgroup (50).

Before the PBS analysis we applied a MAF (Minimum Allele Frequency > 0.05) filter with PLINK. Since we are searching for evidence of differentiation between Andes and Amazon, we considered only SNPs with low differentiation inside these groups (F_{SC}<0.15) (51). The F_{SC} for each SNP for each group was estimated with varcomp function from the hierfstat R package (52). 4P software (53) was used to calculate F_{ST} for each SNP. The F-statistics estimated through varcomp function and 4P rely on the Weir and Cockerham (1984) algorithm (54). Subsequently, the F_{ST} values were transformed as following (55):

$$F_{ST}T = -\log(1-F_{ST})$$

To the transformed F_{ST} values, we applied the PBS formula (50):

$$PBS = (F_{ST}T1+F_{ST}T2-F_{ST}T3)/2$$

Where:
$F_{ST}T1$: transformed F_{ST} between the target population and the closely related population.
$F_{ST}T2$: transformed F_{ST} between the target population and the distant population.
$F_{ST}T3$: transformed F_{ST} between the close population and the distant population.

To avoid spurious outliers when the branches were long or short in all groups, we applied a normalized version from PBS (56):

$$\text{PBSn} = \text{PBS1} / (1 + \text{PBS1} + \text{PBS2} + \text{PBS3})$$

Where:
PBSn: normalized PBS.
PBS1: estimated PBS when the PBS is calculated for the target population.
PBS2: estimated PBS when PBS is focused on the closely related population.
PBS3: estimated PBS when PBS is focused on the distant population.
Our final result is based on the PBSn.

We performed PBS with the following configurations: 1) Andes as a target group, Amazon as a closely related group; and 2) Amazon as a target group, and Andes as a closely related group; in both approaches the CDX (Chinese Dai in Xishuangbanna, China), a population from 1000 Genomes (57) was used as an outgroup. We analyzed the results in windows of 20 SNPs with 5 SNPs of overlap. To determine the probability that a PBS value occurs under the null hypothesis of genetic drift, we simulated 10,000 chromosomal regions of 1Mb under the neutral model for the three populations involved (Andes, Amazon and CDX) (Fig. S23) using the Recosim program to simulate the recombination maps and Cosi2 (58) to simulate the genetic data under a neutral model as described:

```
####################### NEUTRAL MODEL #######################

#DETAILS: In this model the split in Native Americans is Andean (source) and Amazon (new population)
#Andean and Coast events are based on the inference of Ne performed on IBDNe based on IBD segments
#split <label> <source pop id> <new pop id> <T>  516 generations ~ 12900 Andes-Amazon 12700 AndesCosta years
estimated by Harris et al. 2018 (1 generation = 25 years)

gene_conversion_relative_rate 0.0000000045

# mu,
mutation_rate         1.5e-8
length 1000000
# population info
# for each population, include a line:
# pop_define pop-index pop-label

pop_define 1 amazon
pop_define 2 andean
pop_define 3 asian
pop_define 4 coast

#init sample pops
# for each sample set, include
# pop_size pop-label pop-size
# sample_size pop-label sample-size

#amazon
pop_size 1 2749
## Ne is the mean of three values obtained for Matses population in Harris et al. 2018 (N1=2848,N2=2881,N3=2518)
sample_size 1 206
## 206 Considering 103 samples

#andean
pop_size 2 8064
## Ne is the mean of six values obtained for Chopccas  (N1=7774,N2=7070,N3=9348), populations in Harris et al. 2018
sample_size 2 166
## 166 considering 83 diploid samples

#asian
pop_size 3 7700
sample_size 3 240
# 240 considering 120 diploid samples

#coast
```

```
pop_size 4 6975
sample_size 4 62
# 62 considering 31 diploid samples

pop_event exp_change_size "Andean second expansion" 2 4 9 8064 2500
pop_event bottleneck "Andean bottleneck due to European conquest" 2 29 0.067
pop_event bottleneck "Coast bottleneck due to European conquest" 4 29 0.067
pop_event bottleneck "Amazon bottleneck" 1 479 0.067
pop_event exp_change_size "Andean expansion" 2 30 450 7000 2426
pop_event exp_change_size "Coast expansion" 4 4 9 6975 1500
pop_event split "andean and amazon split" 2 1 516
pop_event split "andean and coast split" 2 4 508
pop_event bottleneck "native bottleneck" 2 959 0.067
pop_event split "asian and native split" 3 2 960
pop_event bottleneck "asian bottleneck" 3 1998 0.067

random_seed 2022747205
###################### END OF FILE ######################
```

After this, we estimated the PBSn values for the simulated data with the same methodology used for empirical data. For each observed PBSn result, we calculated the p value as a proportion of simulated PBSn values that are equal or greater than the observed value (50). We considered as candidates for natural selection those SNPs in the 0.05% higher values of PBSn (PBSn > 0.150 for the Andes and PBSn > 0.191 for the Amazon) that were encompassed in the windows in the 0.05% higher PBSn mean values (PBSn mean > 0.095 for the Andes and PBSn mean > 0.116 for the Amazon). We found 142 signals comprising 16 genes in the Andes and 137 signals comprising 15 genes in the Amazon (Tables S1, S2; Fig. 3).

5.2.2. xpEHH: cross population extended haplotype homozygosity

Positive selection events increase the frequency of a genetic variant and, consequently, the frequency of the variants around it (59). This process occurs faster than the haplotypes are broken down by recombination, leading to the emergence of an unusual high frequency long-range haplotype. Considering this, we decided to perform an extended haplotype homozygosity (EHH) test to select the most likely candidates for natural selection.
Sabeti *et al.* (60, 61) developed methods to detect natural selection signatures calculating the EHH, defined as the probability of finding homozygosity of all SNPs around the haplotype of interest choosing two random chromosomes containing this haplotype in a population:

$$EHH(x_i) = \sum_{h \in C(x_i)} \frac{\binom{n_h}{2}}{\binom{n}{2}}$$

Where $C(x_i)$ is the number of all possible distinct haplotypes considering the extension from de core SNP to the i-th SNP, and n_h is the number of observed haplotypes of a specific type h (62).

The method xpEHH defines a core SNP and calculates the EHH for all SNPs in 1MB of distance forwards and backwards considering the chromosomes of two target populations, A and B. When the EHH decays to 0.03-0.05 before reaching 1MB of distance this point is defined as SNP X, if this score is not reached in this range the core SNP is discarded from the next analyzes. Next, the populations are separated and the EHH parting from the core SNPs selected in the first step is calculated again until it reaches the value of 0.03-0.05 (SNP X) in each population. Then the integral of the EHH in respect to the distance from the core SNP to the SNP X is calculated giving the results called I_A (for population A) and I_B (for population B). The xpEHH log ratio is defined as $\ln(I_A/I_B)$. The results are genome-wide normalized. Extreme positive values are indicative of selection in population A, and negative values in population B. The xpEHH analysis was performed with the software Selscan[62].

We considered as positive signals for natural selection the SNPs representing the 99.5 percentile of the xpEHH results of an Andean vs Amazon comparison (xpEHH > 2.97 for the Andes and xpEHH <

-3.34 for the Amazon). For each observed xpEHH result, we calculated the empirical p value as the proportion of values that are equal or greater than the observed value. Only concordant results between PBS and xpEHH were considered as strong candidates for natural selection, with this approach we found 22 candidate SNPs comprising 3 genes in the Andes and 21 SNPs comprising 1 gene in the Amazon (Tables S3, S4; Fig. S26, S27).

To assess whether the results obtained from xpEHH analysis between Andean and Amazon populations are corroborated when comparing each group with an outgroup, we performed xpEHH analysis between each group and East Asian populations from 1000 genomes ($xpEHH_{ANDvsEAS}$ and $xpEHH_{AMZvsEAS}$). The results for Andean populations show a high score for gene HAND2-AS1 (higher $xpEHH_{ANDvsEAS}$=1,59, p-value=0.001), but not for gene RARS (higher $xpEHH_{ANDvsEAS}$=0,79 p-value=0,105). For Amazon populations, the highest signal was from an intergenic region near the gene PTPRC (rs1326288 higher $xpEHH_{AMZvsEAS}$=-1,23 p-value=0.026)(Tables S3, S4).

The candidate loci for natural selection were annotated with MASSA (Multi-agent Annotation System) (63), that mines the following datasets for SNPs (based on rs code): dbSNP (64), OMIM (65) [Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 2020. World Wide Web URL: https://omim.org/], Reactome (66) , HGNC (HGNC Database (67)), GWAS Catalog (68), PolyPhen2 (69), Provean (70), SIFT (71); and the following datasets for genes: UCSC (72), Gene Ontology (73, 74) , PharmGKB (75).

It is interesting to note that the strongest signal for PBS analysis for the Andean populations was from the *DUOX* genes (Fig. S24), previously suggested as a candidate gene for natural selection by Jacovas *et al*. (76). However, this signal does not appear in the xpEHH results.

5.2.3 Linkage Disequilibrium Patterns in natural selection signals

To find other possible candidate SNPs under natural selection, we look for linkage disequilibrium (LD) patterns associated with our strongest natural selection signals. We had access to sequencing data from 60 Andean native individuals (30 Chopccas and 30 Uros) for the genes *DUOX2* and *HAND2-AS1* from Harris et al. (77) . The Amazon sample (12 individuals) was not large enough to allow LD inferences. We calculated LD using the software Haploview 4.2 version (7821). We only consider in LD those variants with $r^2 \geq 0.80$. We found 37 no genotyped SNPs in LD with our two missense signals of natural selection (rs269868 and rs57659670) in the *DUOX2* region in chromosome 15 including three missense mutations: one in the *DUOX2* gene (rs2001616: G>A,T: Pro138Leu), one in the *DUOXA2* gene (rs2252371: C>T: Pro126Leu), and one in the *DUOXA1* gene (rs61751061: C>G,T: Arg478Pro). The SNP rs2001616 is located in the Peroxidase Homologue Domain (aa 26-601) of DUOX2 protein, and rs2252371 is located in an extracellular strep of the DUOXA2 protein (aa 78-183). In these domains occur disulfide bridges between specific cysteines that are essential to the stability and function of the DUOX complex (7989).

In the gene *HAND2-AS1* we found 23 no genotyped SNPs in LD with our 4 natural selection signals, all of them in intronic regions, including rs3775587, mapped within a putative enhancer (Fig. S28). These SNPs were used in the analysis of regulatory elements described below.

5.2.4 Identification of regulatory elements located around HAND2-AS1 locus

GeneHancer track available in the UCSC Genome Browser (90 80) was used to identify active regulatory elements (enhancers and promoters) that may target *HAND2-AS1* (Fig. S29). GeneHancer database was created by integrating >1 million regulatory elements from seven genome-wide databases: ENCODE project Z-Lab Enhancer-like regions (version v3); Ensembl regulatory build (version 92); functional annotation of the mammalian genome (FANTOM5) atlas of active enhancers; VISTA Enhancer Browser; dbSUPER super-enhancers; Eukaryotic Promoter Database (EPDnew) promoters; and UCNEbase ultra-conserved noncoding elements. Genes were linked to enhancers by GeneHancer using five methods: eQTLs from GTEx (v6p); Capture Hi-C promoter-enhancer long range interactions; FANTOM5 co-expression of enhancers in the form of noncoding enhancer RNA;

transcription factor co-expression; and gene target distance. For this analysis a “double elite” dataset was considered, which is composed of regulatory elements derived from more than one database (elite enhancers) that are associated to genes from more than one method (elite association).

Highly Differentiated Variants Between Andean and Amazon Populations and its Medical Relevance

5.2.5 F-statistics

We searched for functionally relevant SNPs differentiated between the Arid Andes and Amazon populations of similar latitudes (i.e. the same groups of populations tested for natural selection) using the classical F statistics for each SNP as defined by Weir and Cockerham (54). These SNPs are differentiated between the two groups not necessarily due to the action of natural selection, but their sharp differences in frequencies in the Andean vs. the Amazon environments may be biomedically relevant. We found 1,985 highly differentiated SNPs between the two groups of populations (0.1% highest values of FCT distribution: > 0.318, estimated with the 4P software19), but relatively homogeneous within the groups (FSC<0.15, estimated with hierfstat software20). We annotated the 1,985 SNPs usIng our bioinformatics MASSA platform (63), that mines the following datasets based on SNPs rs code): dbSNP (64), Ensembl (81), GWAS Catalogue (68), PharmGKB (75), SIFT (71), PolyPhen (69). SNPs that are GWAs hits, related to drug response and missense mutations, are listed in Tables S5-7.

5.3. Conclusions

- We have confirmed a natural selection signal from a gene previously reported in Andean populations, *DUOX2* (76) (PBSn=0.22 p-value=0.002, xpEHH=-2.647 p-value=0.991).

- We identified Natural selection signals Andeans in genes related to (Tab. S4):
    - High altitude adaptation: ***SULT1A1***: PBSn=0.167 p-value=0.007; ***RARS***: PBSn=0.15 p-value=0.010, xpEHH=2.980 p-value=0.0025 (82, 83),
    - Heart development: ***HAND2-AS1***: PBSn=0.21 p-value=0.003, xpEHH=4.481 p-value<2e-5 (84),
    - Immune response: ***UBQLN4***: PBSn=0.17 p-value=0.007, xpEHH=-0.217 p-value=0.607; ***SSR2***: PBSn=0.17 p-value=0.007, xpEHH=-0.215 p-value=0.606; ***DUOX2***: PBSn=0.22 p-value=0.002, xpEHH=-2.647 p-value=0.991 (85–87).

- We identified Natural selection signals in Amazon populations related to (Tab. S5):
    - Immune response: ***PTPRC***: PBSn=0.265 p-value=0.004, xpEHH=-4.222 p-value=0.0003 (88),
    - Food intake regulation: ***MCHR1***: PBSn=0.26 p-value=0.004 (89),
    - Lipid transport: ***ABCA9***: PBSn=0.21 p-value=0.008, xpEHH=-1.570 p-value=0.060, ***ABCA6***: PBSn=0.19 p-value=0.011, xpEHH=-1.362 p-value=0.084 (90, 91).

Supplementary Figures

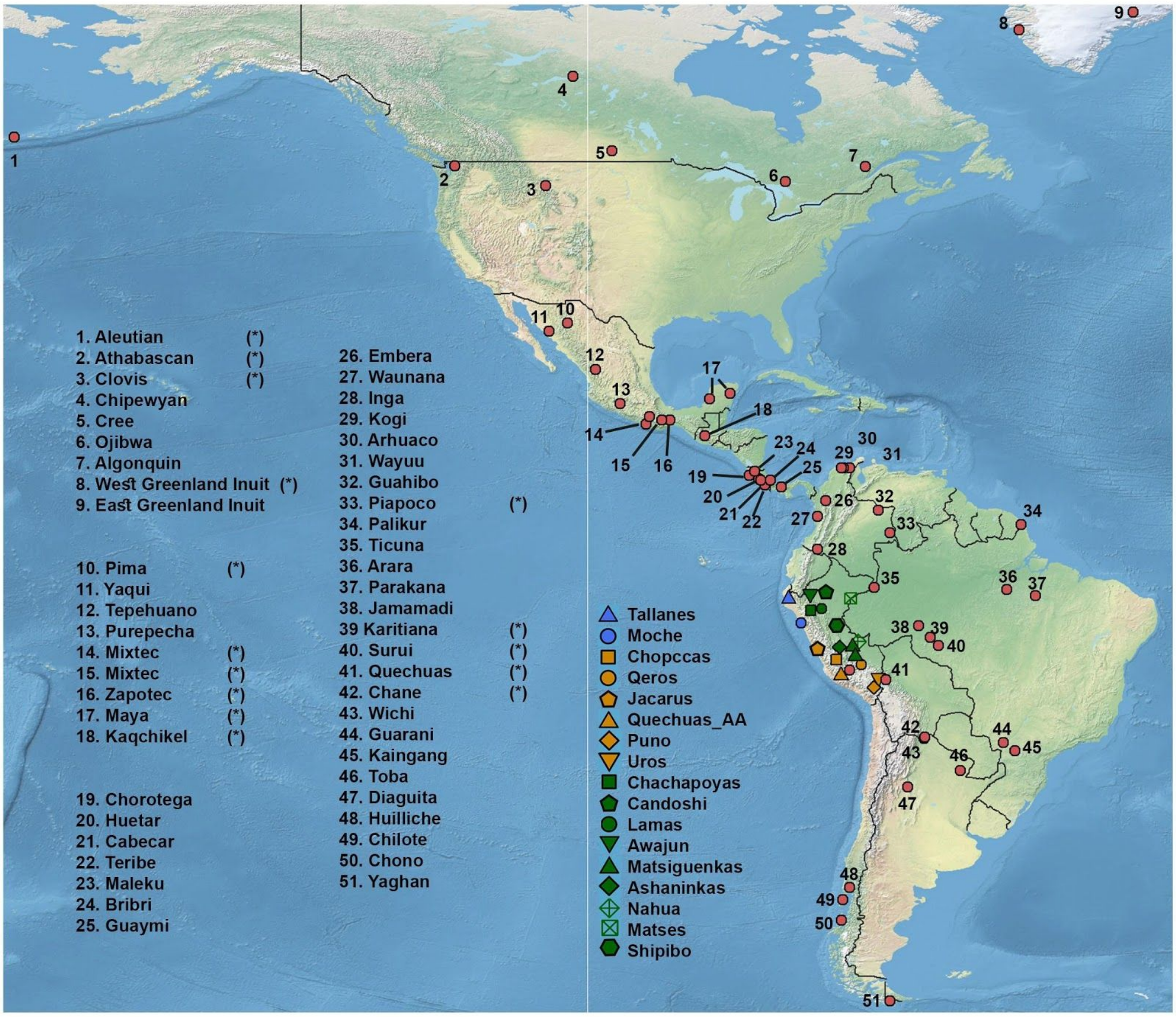


Figure S1. Geographical distribution for the 18 Peruvian Native populations sampled, plus the 65 sampled Native American populations and public data sets (Mallick *et al.* 2016, Raghavan *et al.* 2015, Reich *et al.* 2012). All samples except Clovis and Athabascan were included in a data set of ~ 230,000 SNPs. Peruvian samples and (*) were included in a data set of ~ 500,000 SNPs.

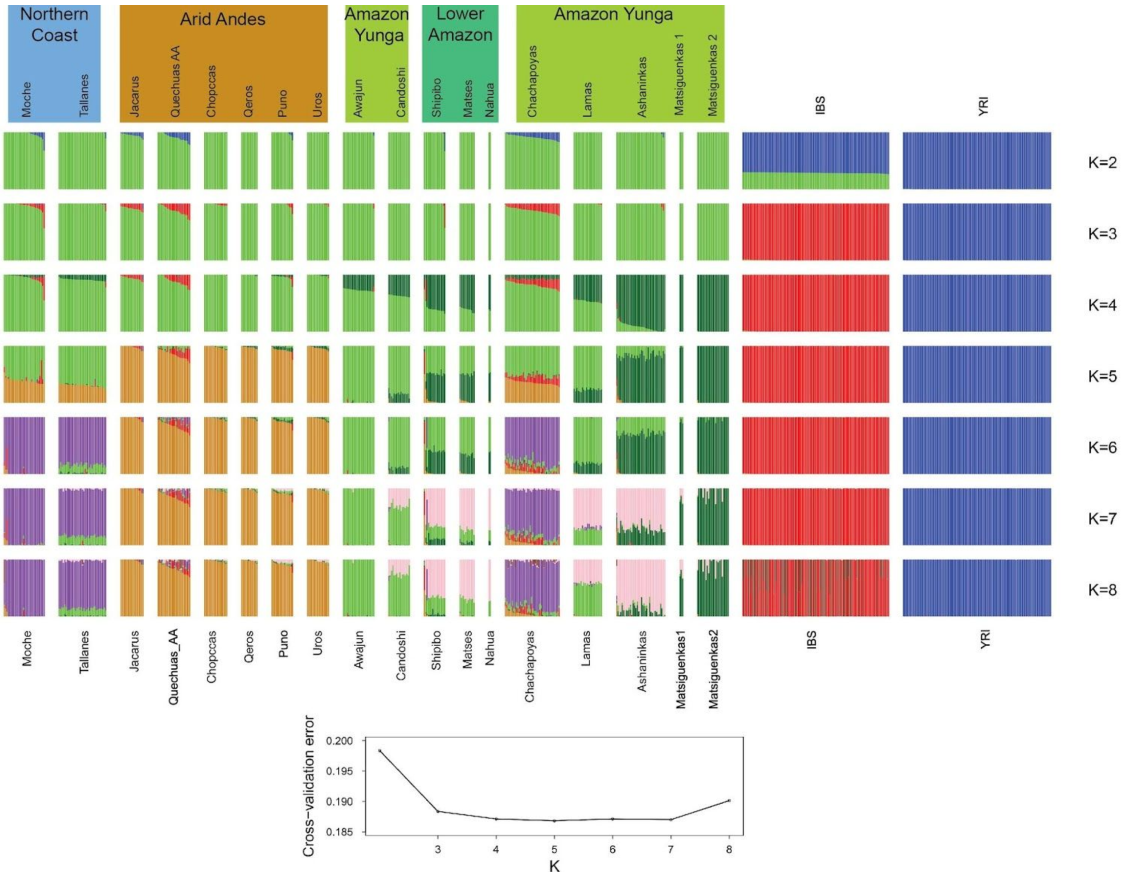


Figure S2. ADMIXTURE analysis for 18 Native American populations, as well as Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 1.9M Dataset). Figure shows results for 2 to 8 ancestral clusters (K) and a plot (Bottom) with the ADMIXTURE cross-validation errors as a function of K. The lowest cross validations error corresponds to K=5 in which we observed four Native American, one European and one African cluster.

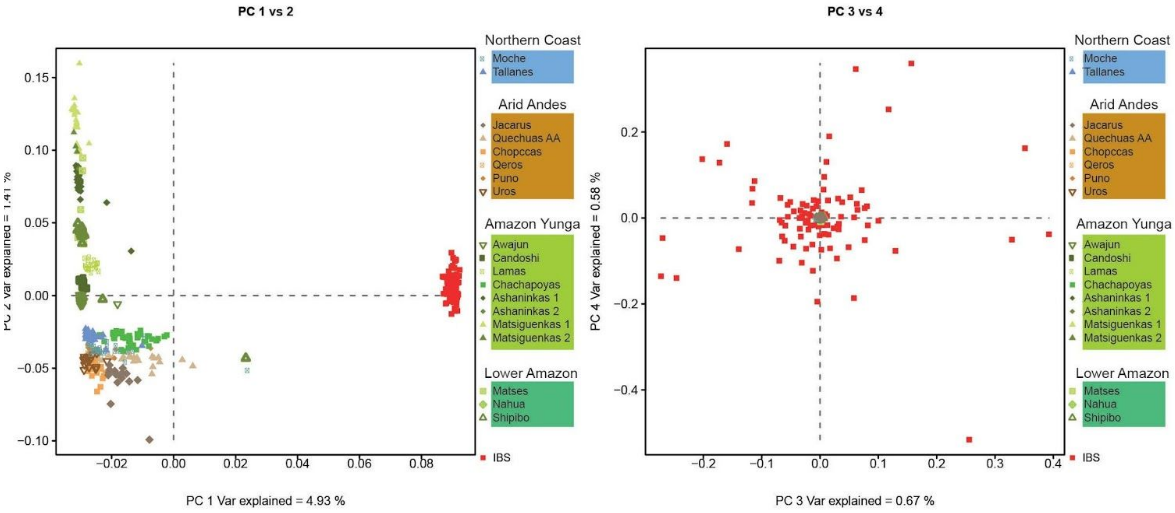


Figure S3. Principal Component Analysis for 18 Native American Peruvian populations and Iberian individuals (IBS) from 1000 Genomes Project (Natives 1.9M Dataset). Shades of blue are related to Coast populations. Orange-brown colors are related to Andean populations and green colors are related to Amazon.

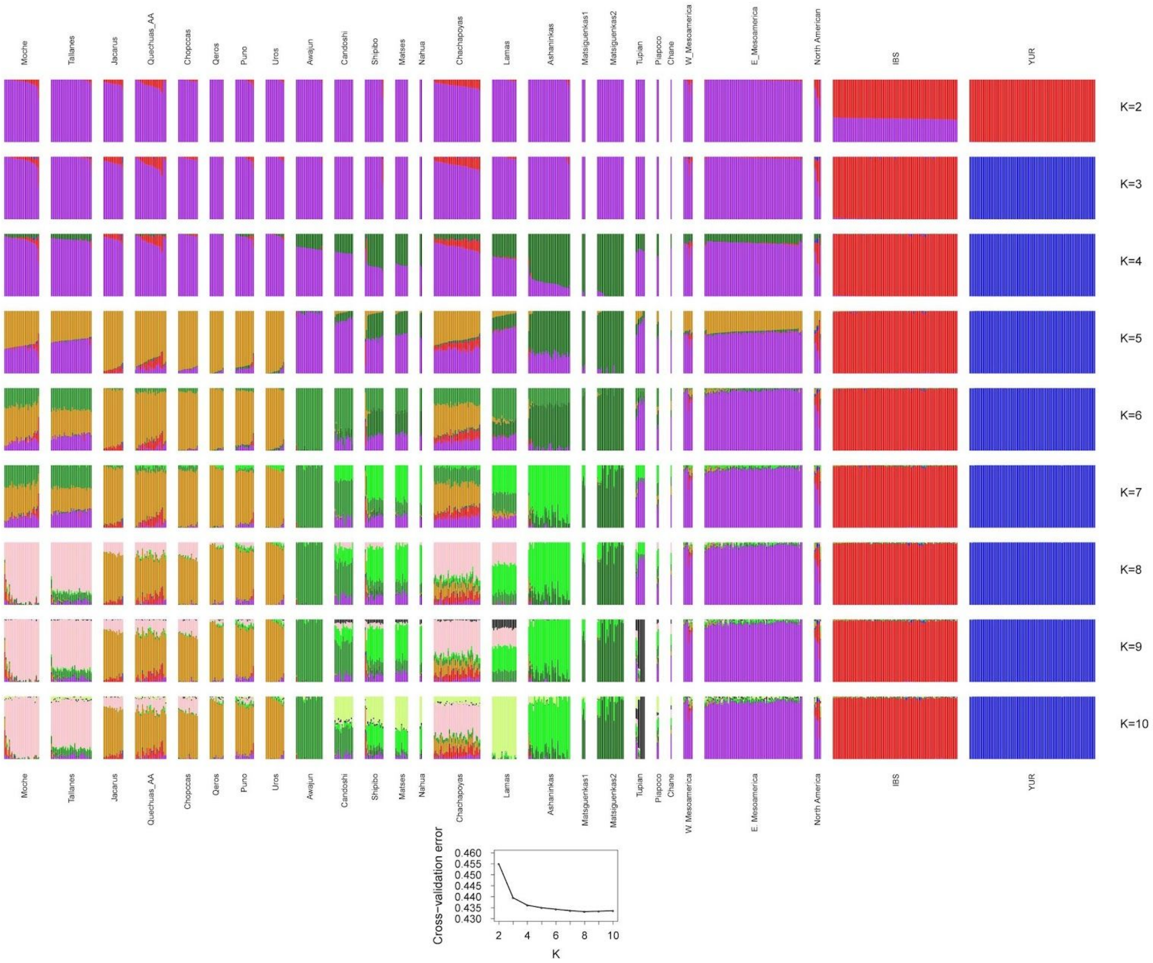


Figure S4. ADMIXTURE analysis for 18 Natives American Peruvian populations, Guatemala samples, Native Americans from Raghavan *et al.* 2015 and the Simons Project (Mallick *et al.* 2016) Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 500K Dataset). Figure shows results for 2 to 10 ancestral (K) clusters and a plot (Bottom) with the ADMIXTURE cross-validation errors as a function of K. The lowest cross validation error corresponds to K=8.

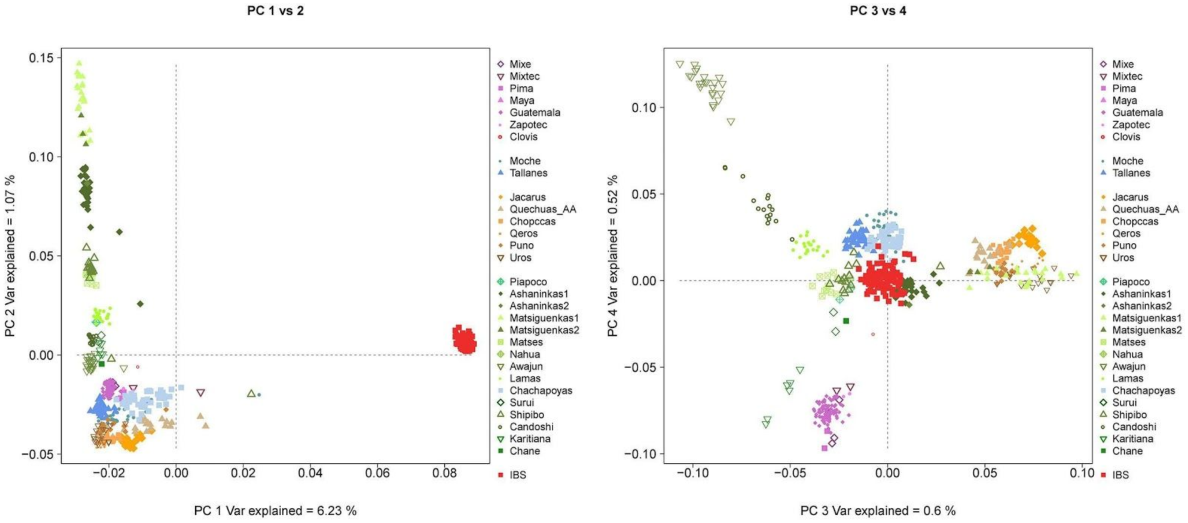
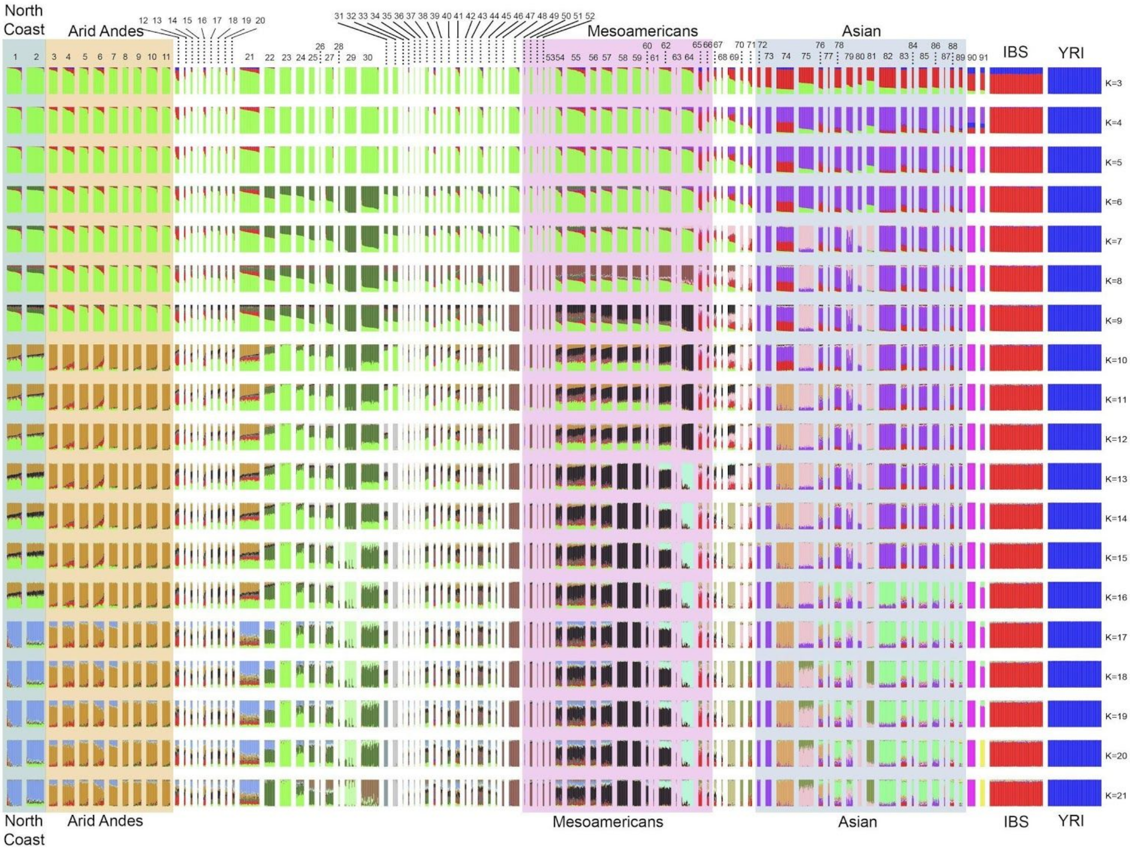
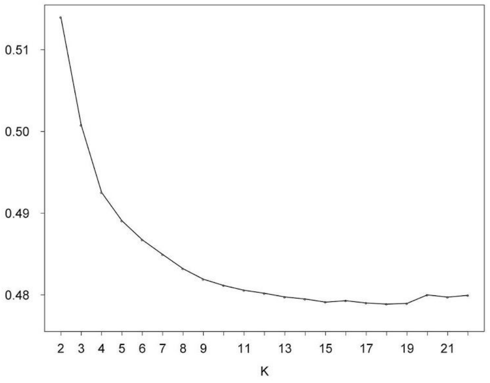


Figure S5. Principal Component Analysis for 18 Native American Peruvian populations, Guatemala samples, Native Americans from Raghavan *et al.* 2015 and the Simons Project (Mallick *et al.* 2016) and Iberian (IBS) populations from 1000 Genomes Project (Natives 500K Dataset). Shades of blue are related to Peruvian Coast populations. Orange-brown colors are related to Andean populations and green colors are related to Amazon. Shades of purple are related to Mesoamericans. Shades of beige are related to North American natives.



1. Moche
2. Tallanes
3. Jacarus
4. Quechuas
5. Quechua_R1
6. Quechua_R2
7. Chopccas
8. Qeros
9. Aimaras
10. Aimaras_R
11. Uros
12. Chilote
13. Chono
14. Hulliche
15. Chane
16. Diaguita
17. Toba
18. Wichi
19. Yaghan
20. Guarani
21. Chachapoyas
22. Lamas
23. Awajun
24. Candoshi
25. Matses
26. Nahua
27. Shipibo
28. Matsiguenkas
29. Shimaa
30. Ashaninkas
31. Karitiana
32. Surui
33. Arara
34. Kaingang
35. Parakana
36. Jamamadi
37. Piapoco
38. Arhuaco
39. Embera
40. Guahibo
41. Inga
42. Kogi
43. Waunana
44. Wayuu
45. Ticuna
46. Palikur
47. Bribri
48. Cabecar
49. Chorotega
50. Guaymi
51. Huetar
52. Maleku
53. Teribe
54. Kaqchikel
55. Maya1
56. Maya2
57. Zapotec1
58. Zapotec2
59. Mixe
60. Mixtec
61. Purepecha
62. Tepehuano
63. Yaqui
64. Pima
65. Aleutian
66. Cree
67. Ojibwa
68. Algonquin
69. Chipewyan
70. EastGreenland
71. WestGreenland
72. Nganasan1
73. Nganasan2
74. Khanty
75. Chukchi



76. Selkup
77. Tundra_Nentsi
78. Tuvinians
79. Yukaghir
80. Koryak
81. Naukan
82. Yakut
83. Altaian
84. Dolgan
85. Buryat
86. Evenki
87. Ket
88. Mongolian
89. Mongola
90. Papuan
91. Melanesian
92. IBS
93. YRI

Figure S6. (Top) ADMIXTURE analysis for 90 worldwide populations including 71 Native American populations, 18 Asian, 2 Oceanian, Iberian (IBS) and Yoruba (YRI) populations from 1000 Genomes Project (Natives 230K Dataset). Figure shows results for 3 to 21 ancestral clusters (K). (Bottom) ADMIXTURE cross-validation errors as a function of K and list of populations included. The lowest cross validation corresponds to ADMIXTURE K=18.

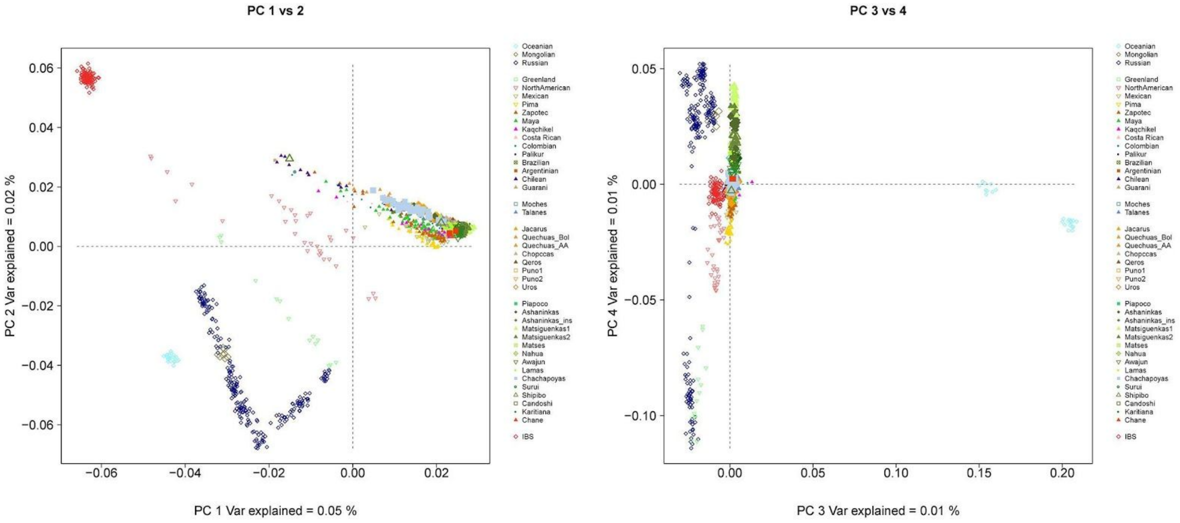


Figure S7. Principal Component Analysis for 89 worldwide populations including 68 Native American populations, 18 Asian populations, 2 Oceanian populations, Iberian (IBS) populations (Native 230K Dataset).

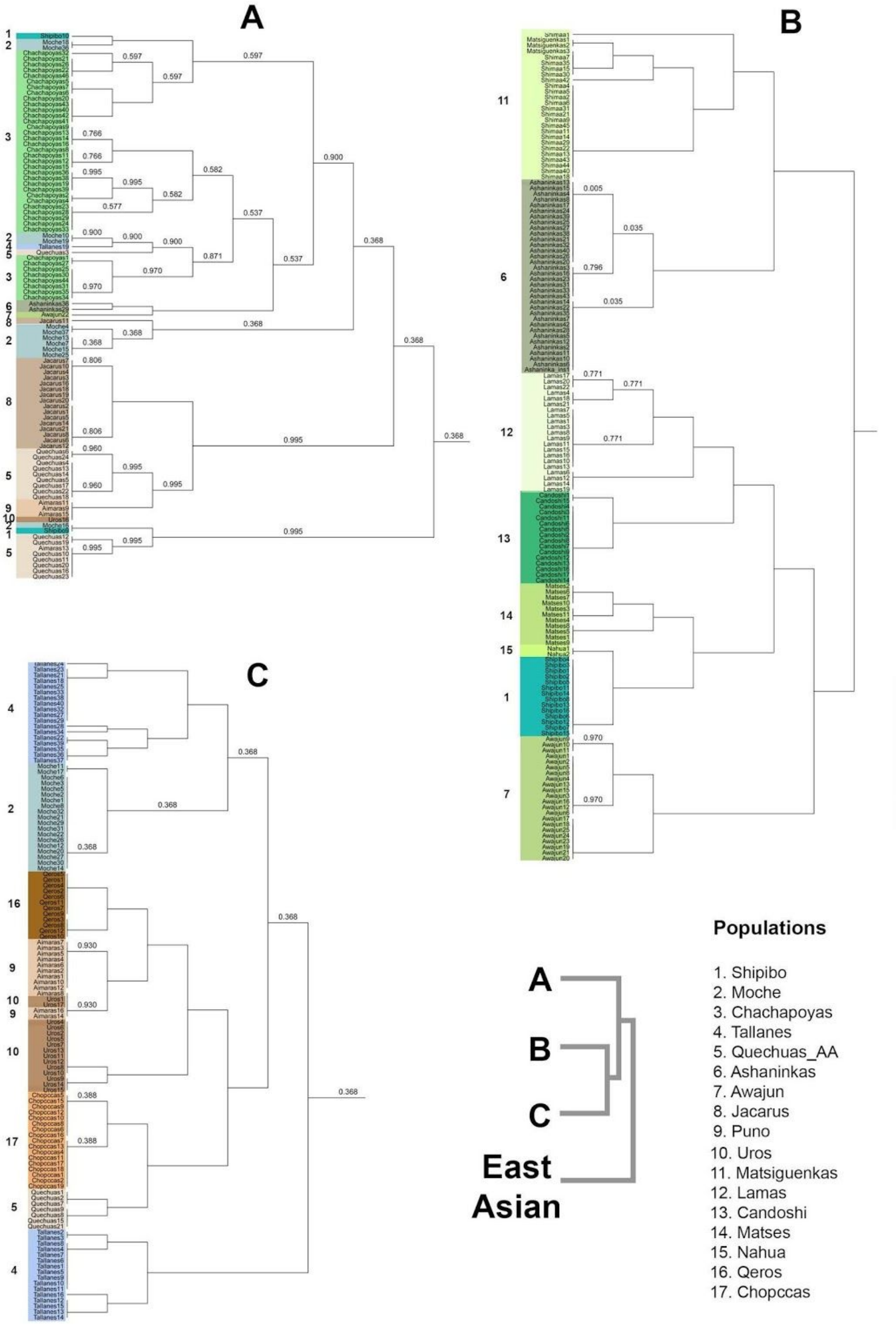


Figure S8. fineSTRUCTURE clustering analysis for the Dataset 1.9M dataset. The tree shows the haplotype sharing between Native Americans and East Asian samples. Figures A, B, and C represent the clusters A, B, and C, respectively, in the tree on the right. East Asian clusters grouped all Asian samples. Shades of blue are related to Peruvian Coast populations. Orange-brown colors are related to Andean populations and green colors are related to Amazon. Shades of purple are related to Mesoamericans. Shades of beige are related to North American natives.

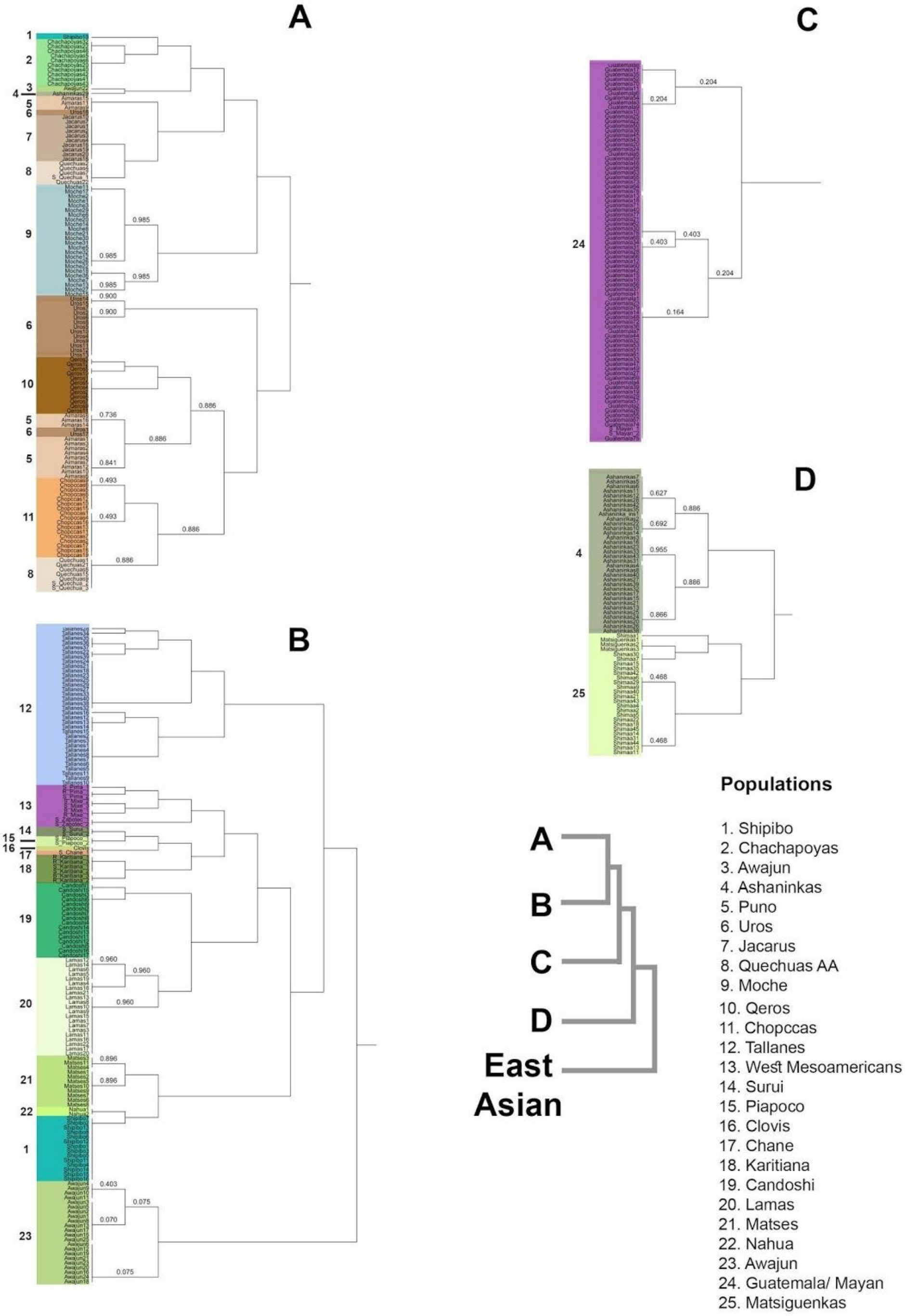


Figure S9. fineSTRUCTURE clustering analysis for the Dataset 500K dataset. The tree shows the haplotype sharing between Native Americans and East Asian samples. Figures A, B, C and D represent the clusters A, B, C and D, respectively, in the tree on the right. East Asian clusters grouped all Asian samples. Shades of blue are related to Peruvian Coast populations. Orange-brown colors are related to Andean populations and green colors are related to Amazon. Shades of purple are related to Mesoamericans. Shades of beige are related to North American natives.

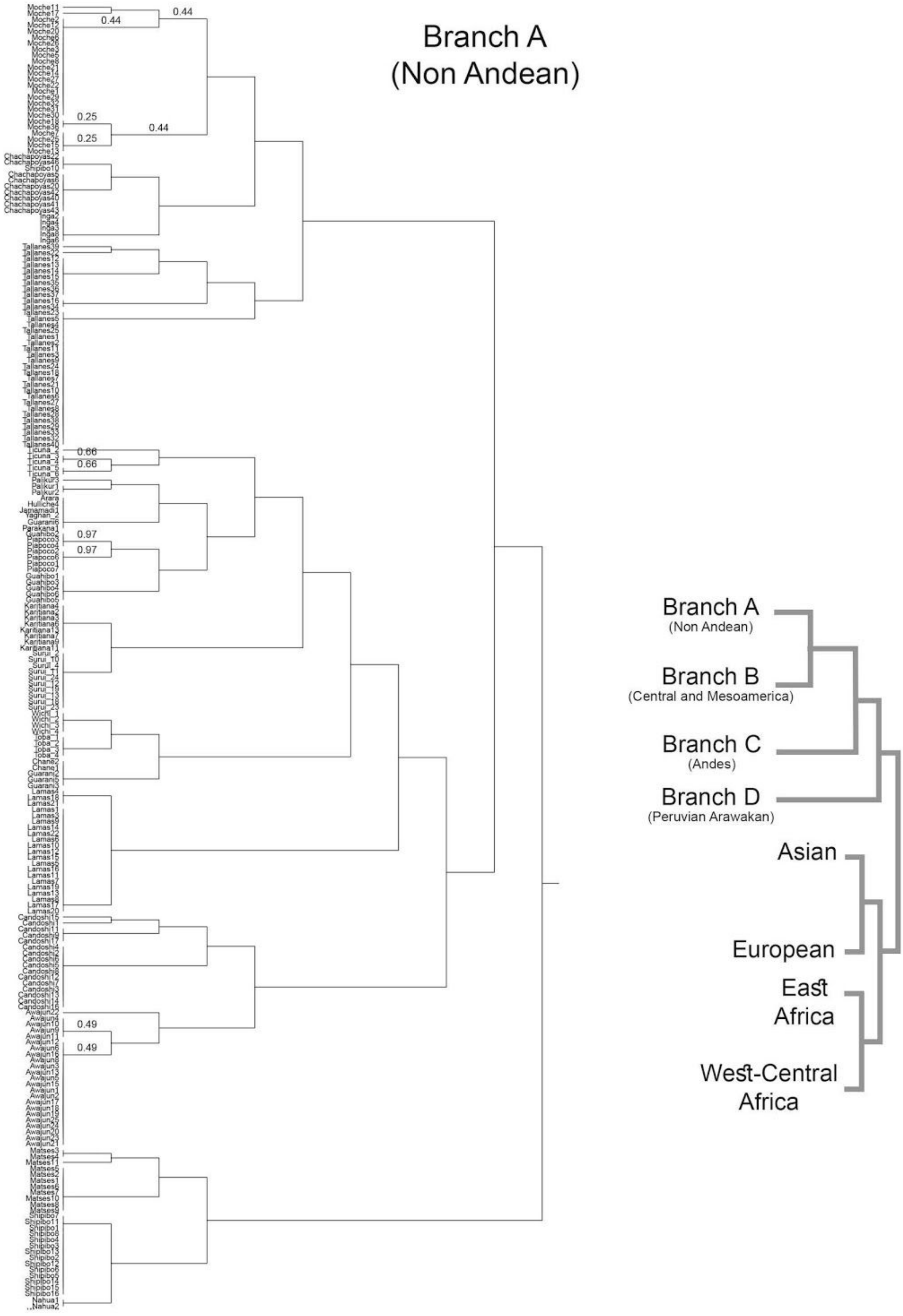


Figure S10. fineSTRUCTURE clustering analysis for the Dataset 230K dataset. A) Branch A of the tree showing the clustering of the Non Andean populations of South America. B) Branches B (Central Americans and Mesoamericas), C (Andean populations) and D (Peruvian Arawakan Ashaninkas and Matsiguenkas).

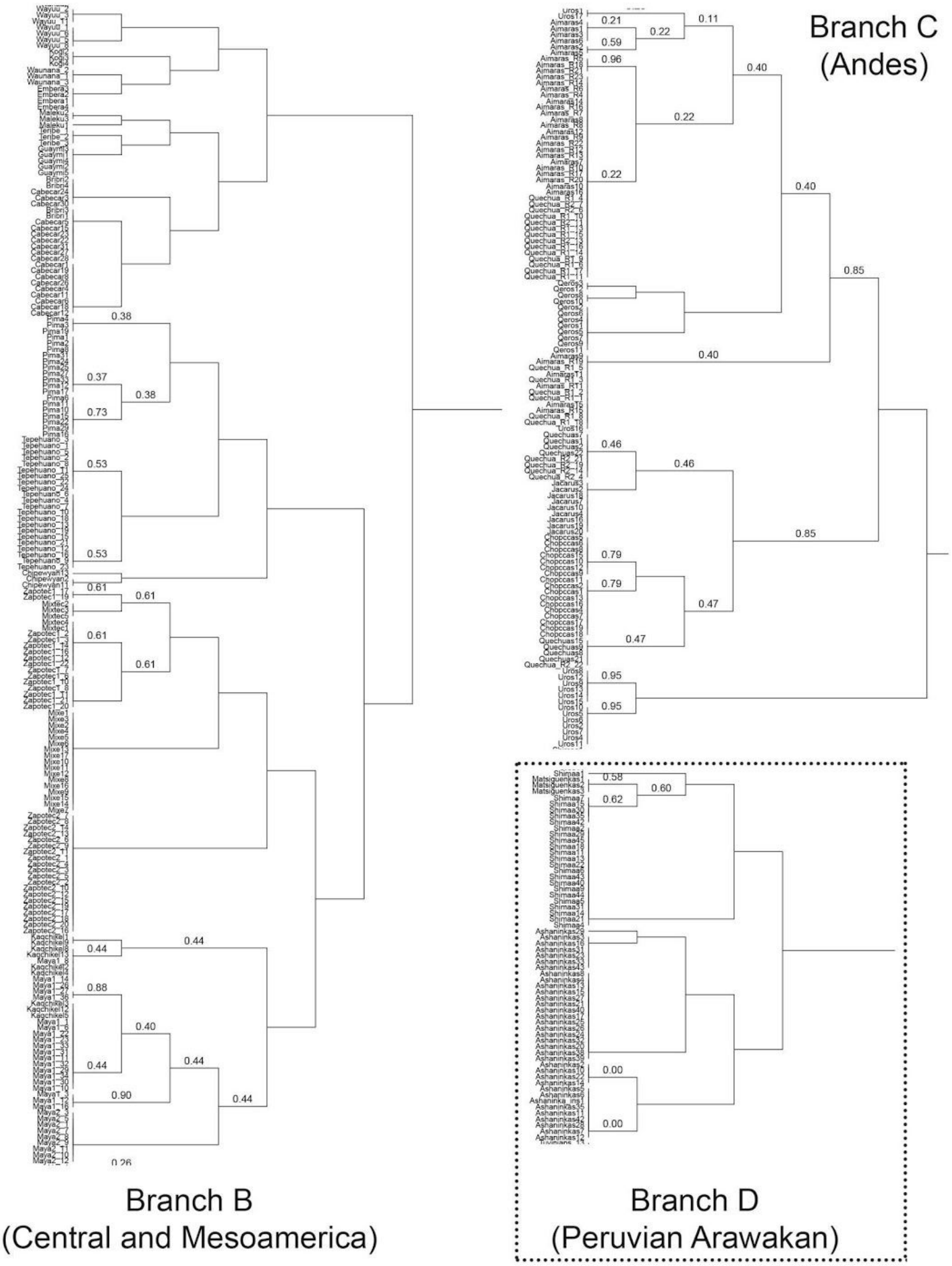


Figure S10 (Continued). fineSTRUCTURE clustering analysis for the Dataset 230K dataset. A) Branch A of the tree showing the clustering of the Non Andean populations of South America. B) Branches B (Central Americans and Mesoamericas), C (Andean populations) and D (Peruvian Arawakan Ashaninkas and Matsiguenkas).

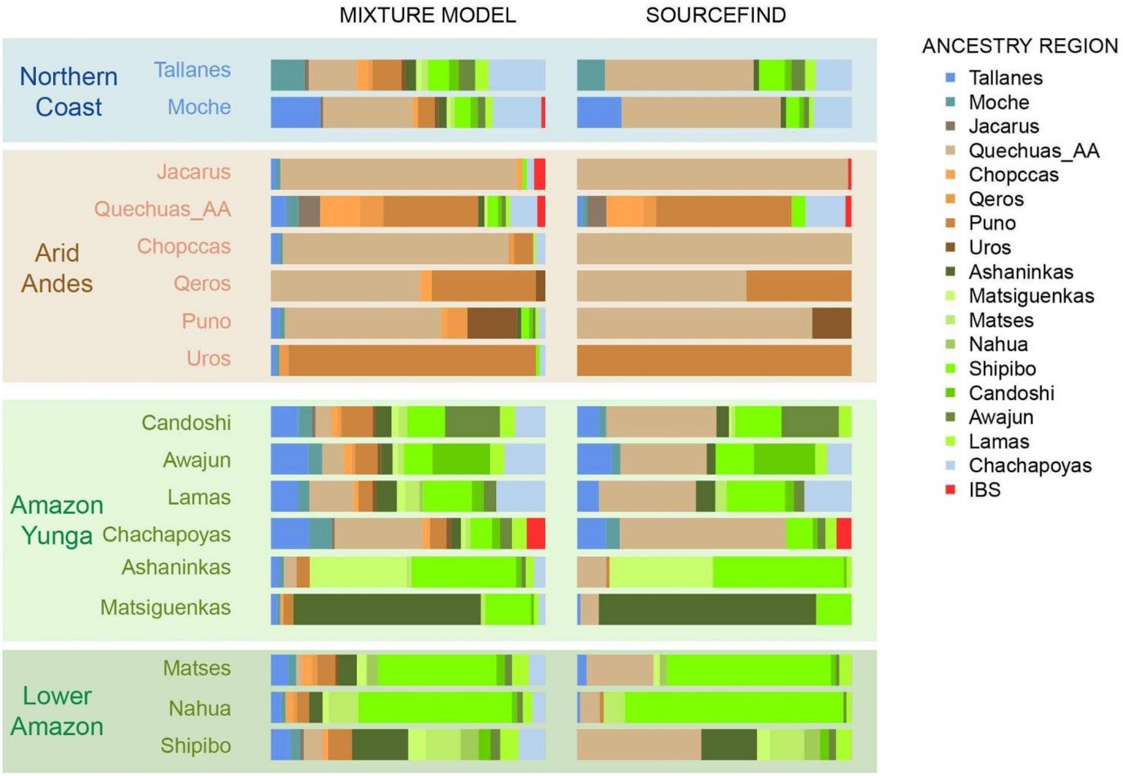


Figure S11. Proportions of haplotype sharing for each target population respect to Native Americans, Europeans and Africans donors populations, for the Dataset Natives 1.9M, inferred by two approaches: A non negative regression (MIXTURE MODEL) and a Bayesian approach (SOURCEFIND). Colored bars indicate a proportion of shared haplotypes shared DNA between the target population and a specific donor.

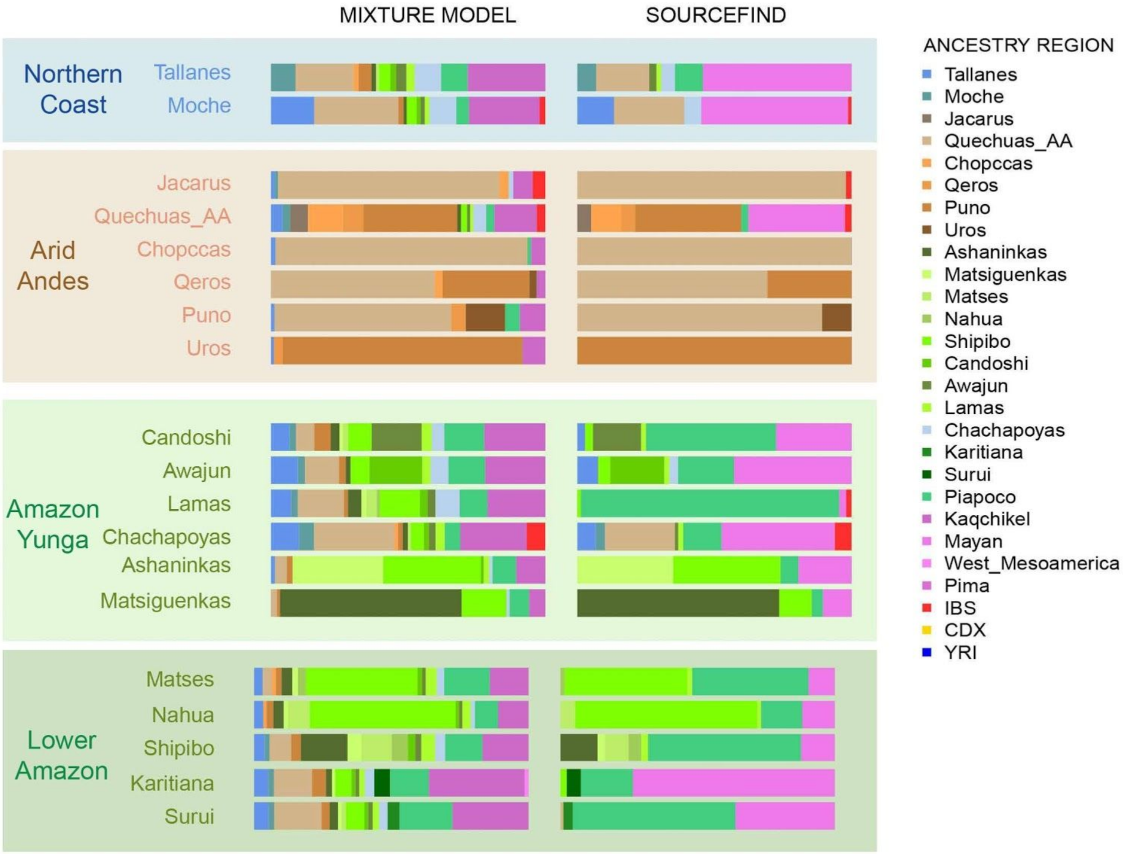


Figure S12. Proportions of haplotype sharing for each target population respect to Native Americans, Europeans and Africans donors populations, for the Dataset Natives 500K, inferred by two approaches: A non negative regression (MIXTURE MODEL) and a Bayesian approach (SOURCEFIND). Colored bars indicate a proportion of shared haplotypes shared DNA between the target population and a specific donor.

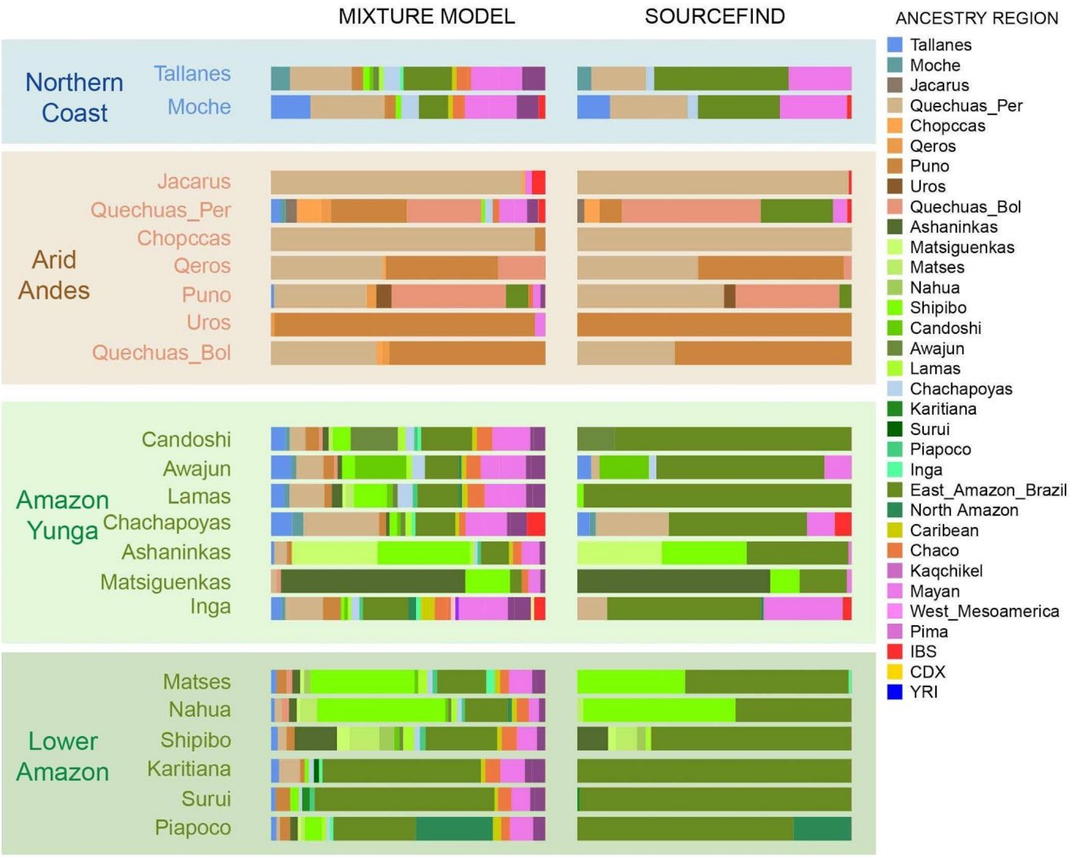


Figure S13. Proportions of haplotype sharing for each target population respect to Native Americans, Europeans and Africans donors populations, for the Dataset Natives 230K, inferred by two approaches: A non negative regression (MIXTURE MODEL) and a Bayesian approach (SOURCEFIND). Colored bars indicate a proportion of shared haplotypes shared DNA between the target population and a specific donor.

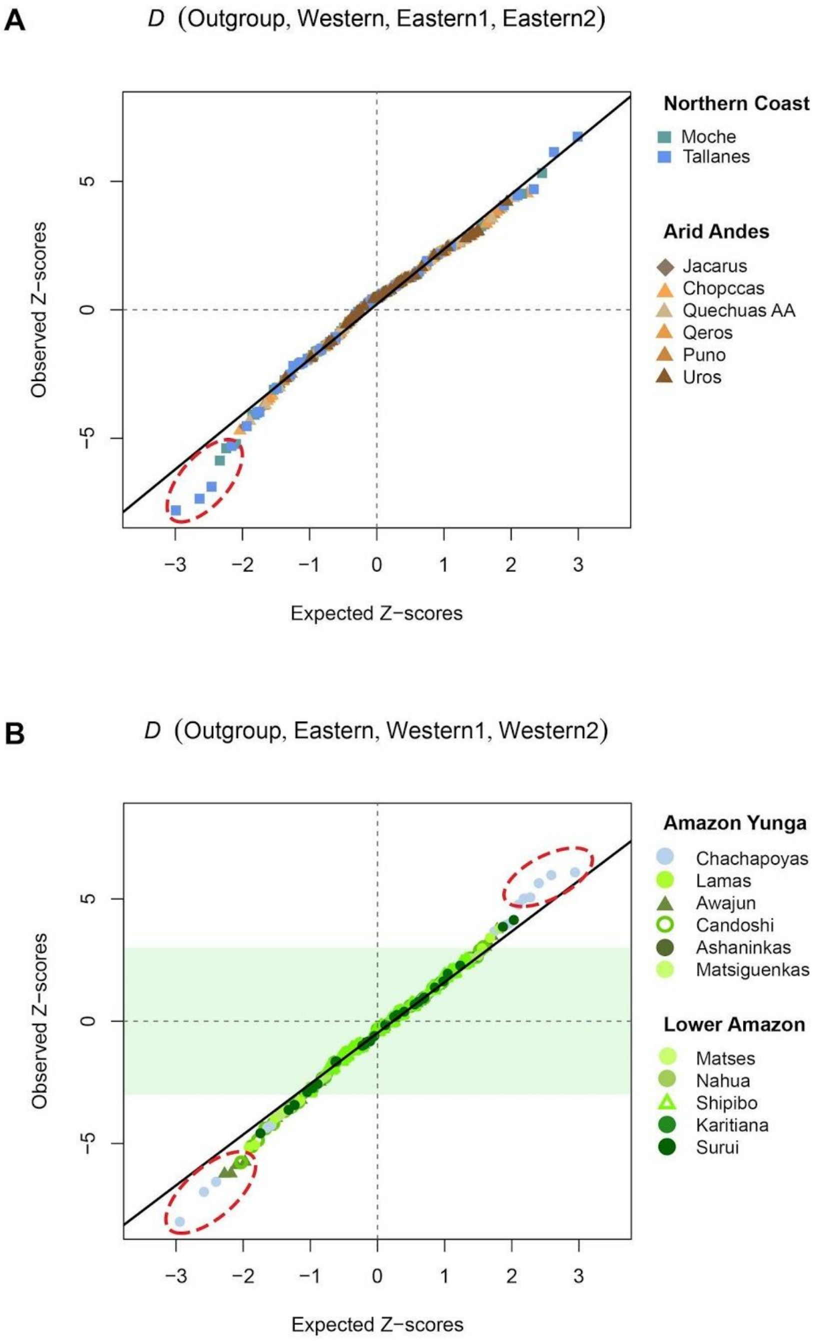


Figure S14. Quantile-quantile plot comparing Z-scores from *D*-statistics relating Western (Northern Coast and Arid Andes) and Eastern Andean slope (Amazon Yunga and Amazon Yunga) populations to those expected under a normal distribution (green diagonal) for the Dataset 500K. Red dashed circles show the Eastern populations with significant values of *D* statistics . **A)** We tested the configuration (outgroups (Western (Eastern1, Eastern2))). We detected evidence of gene flow between the Northern Coast and Amazon Yunga populations. **B)** We test the configuration (outgroups (Eastern (Western1, Western2))). We detected strong genetic affinity between Awajun, Candoshi, Lamas and Chachapoyas with Western populations.

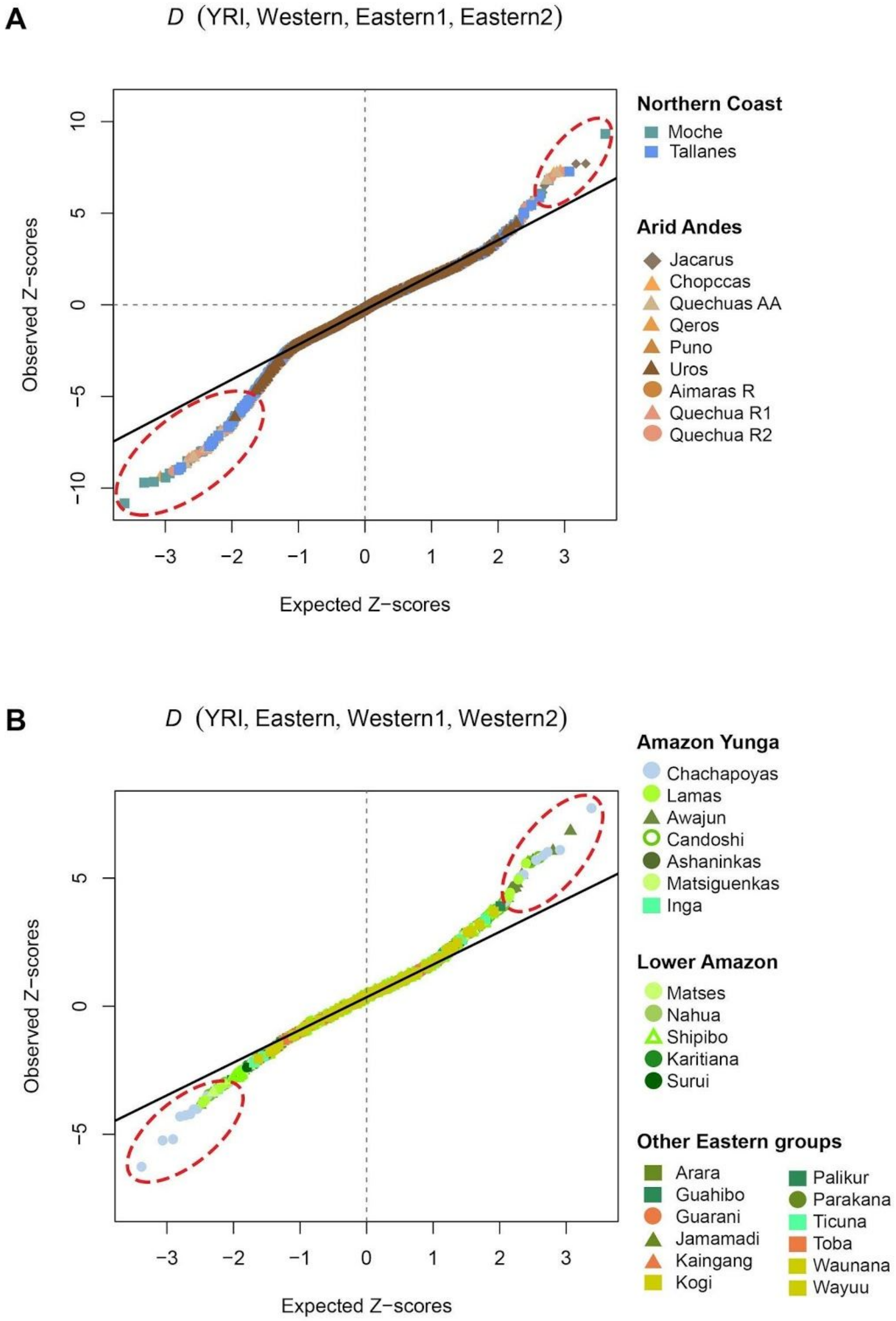
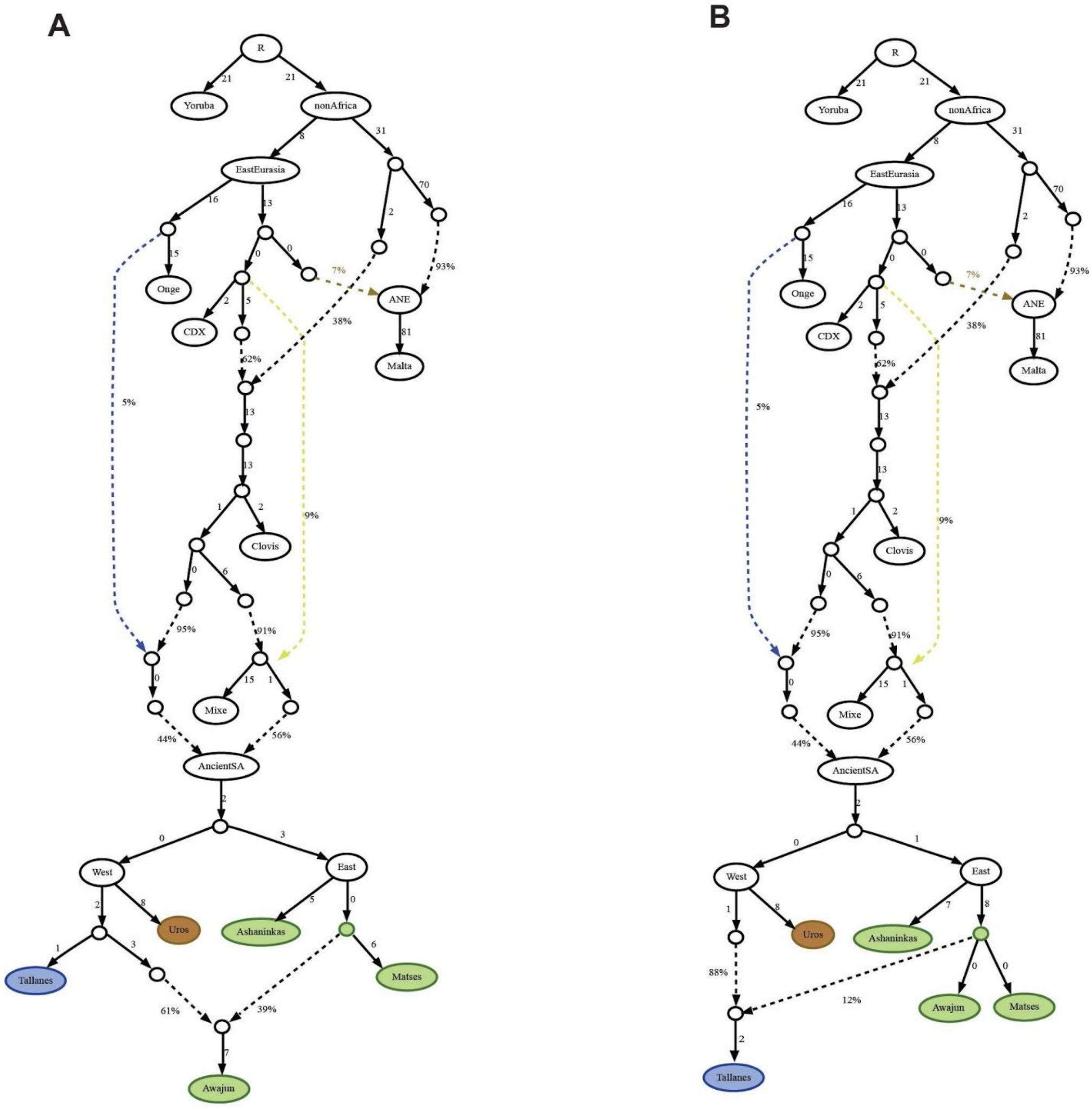


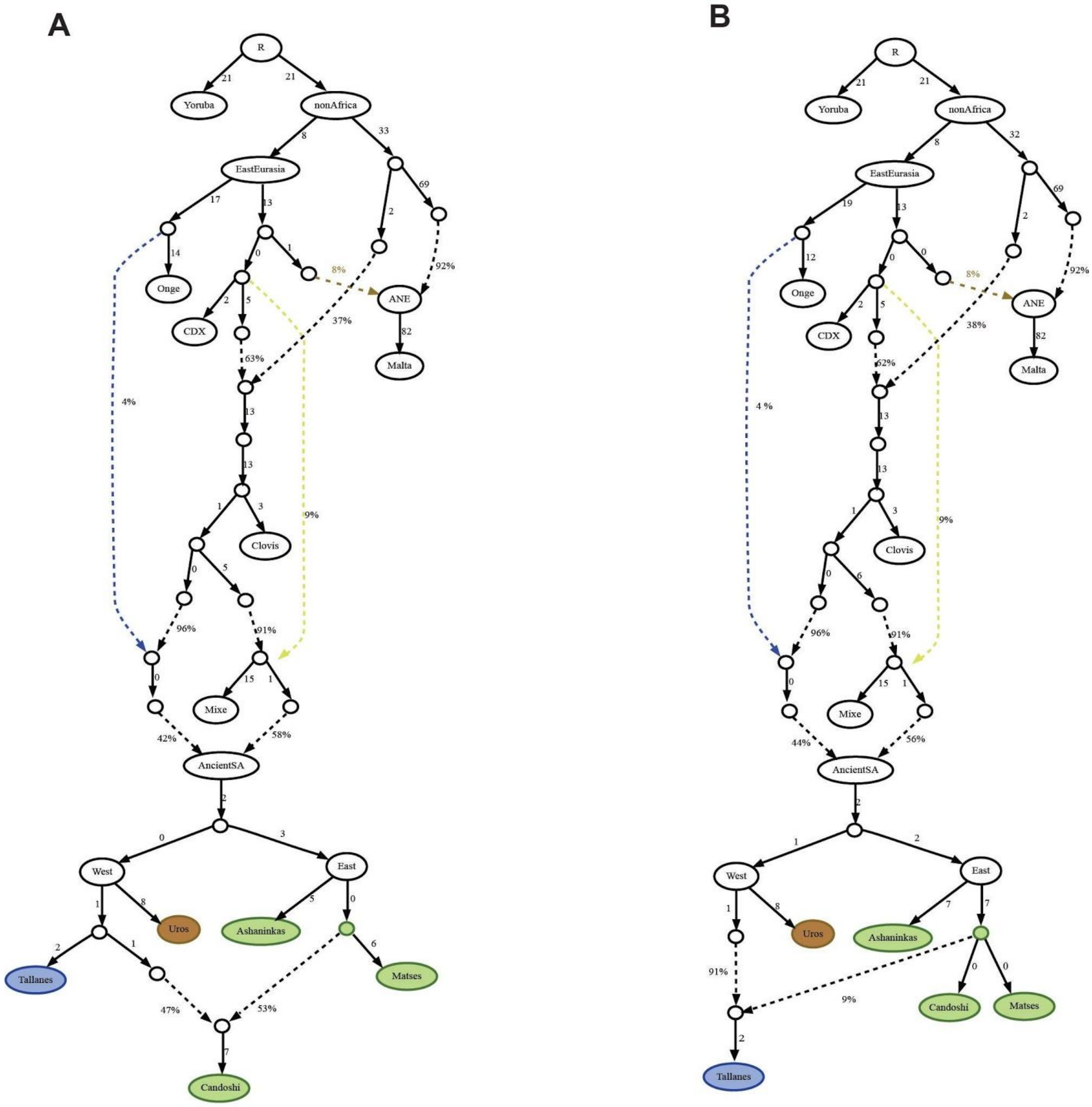
Figure S15. Quantile-quantile plot comparing Z-scores from D-statistics relating Western (Northern Coast and Arid Andes) and Eastern Andean slope populations (Amazon Yunga, Lower Amazon and other eastern groups) to those expected under a normal (green diagonal) distribution for the Dataset 230K. Red dashed circles show the Eastern populations with significant values of *D* statistics. **A)** We test the configuration (outgroups (Western (Eastern1, Eastern2))). We detected evidence of gene flow between Peruvian Coast and Eastern populations in the North Fertile Andes. **B)** We test the configuration (outgroups (Eastern (Western1, Western2))). We detected strong genetic affinity between Awajun, Candoshi, Lamas and Chachapoyas with Western populations.



Hypothesis 1 (Western into Eastern)
Worst f-stat : 2.822
Final score : 25.23
Zero branch lengths: 1

Hypothesis 2 (Eastern into Western)
Worst f-stat : 2.822
Final score : 31.99
Zero branch lengths: 3

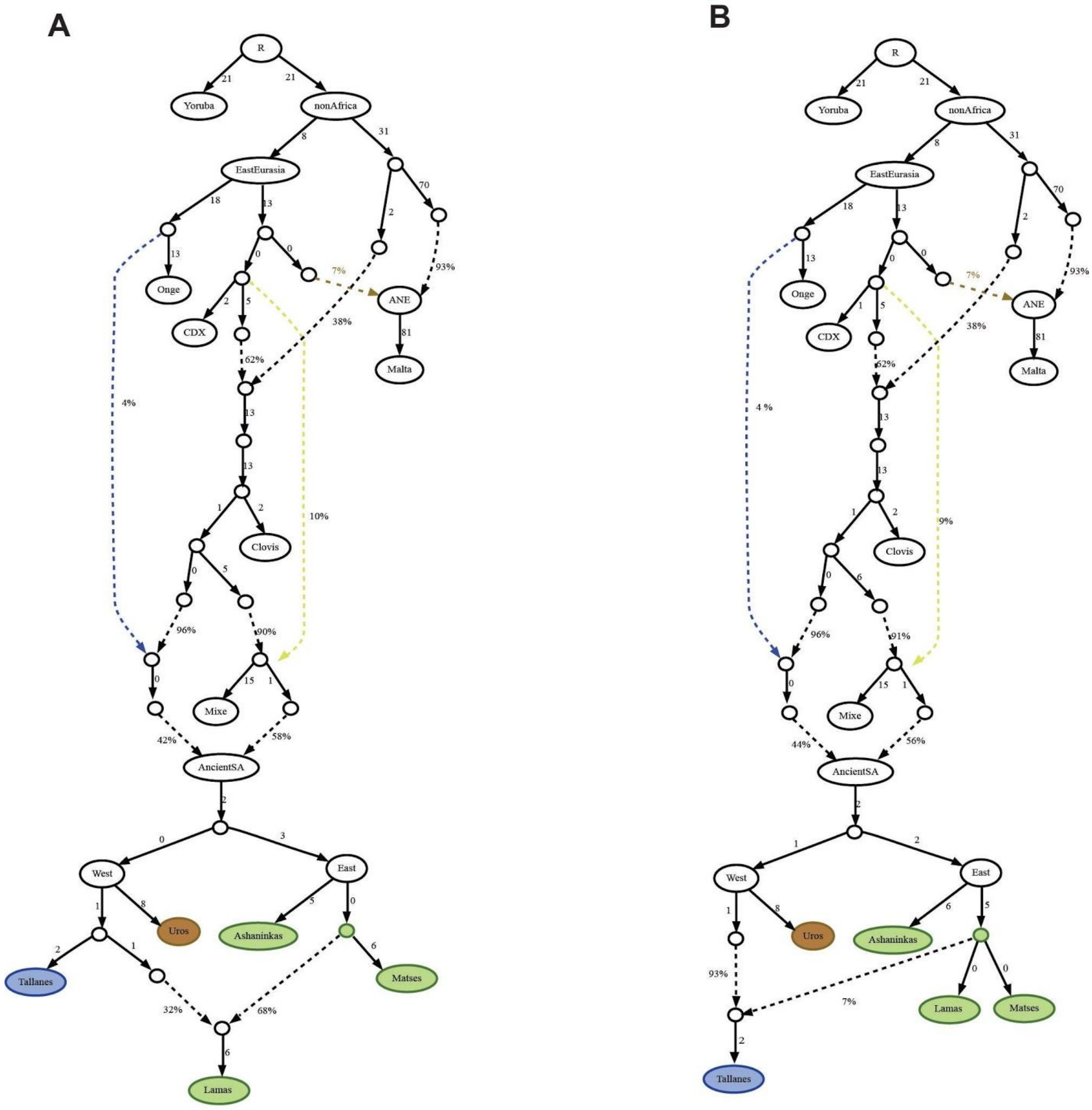
Figure S16. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes. We explore the relationship of the Tallanes-Awajun and the direction of the gene flow. White balls in the intermediate nodes represent hypothetical ancestors for each divergence event. **A)** Admixture graph for testing the Hypothesis 1 (from Western to Eastern): the gene flow from the Northern Coast to Awajun. **B)** Admixture graph for testing the Hypothesis 2 (From Eastern to Western) the gene flow event into Tallanes. Hypothesis 1 is better supported considering its lower final score and number of zeroed branches in contrast to Hypothesis 2.



Hypothesis 1 (Western into Eastern)
Worst f-stat : 2.822
Final score : 29.46
Zero branch lengths: 1

Hypothesis 2 (Eastern into Western)
Worst f-stat : 2.822
Final score : 25.003
Zero branch lengths: 3

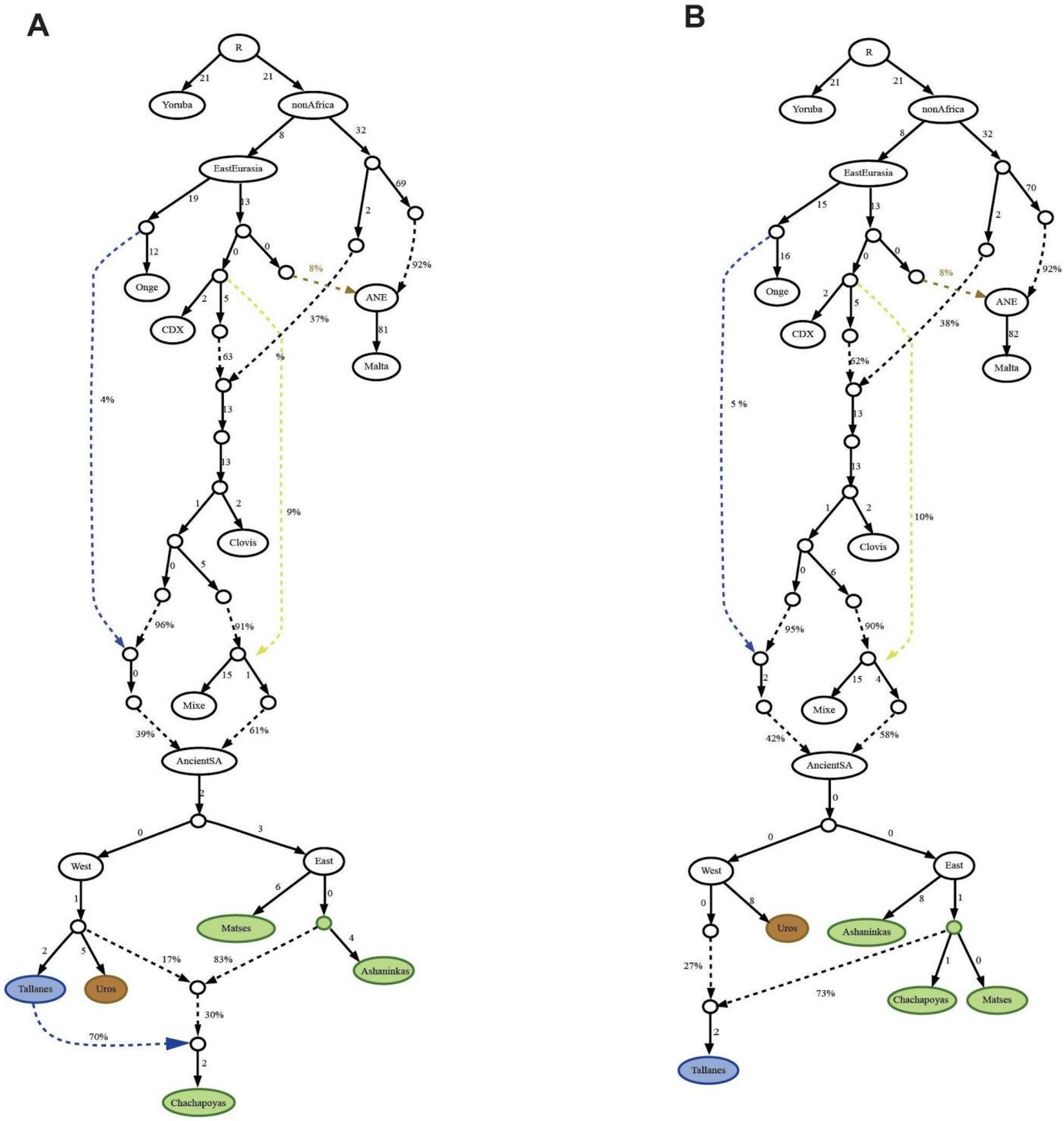
Figure S17. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes. We explore the relationship of the Tallanes-Candoshi and the direction of the gene flow. White balls in the intermediate nodes represent hypothetical ancestors for each divergence event. **A)** Admixture graph for testing the Hypothesis 1 (from Western to Eastern): the gene flow from the Northern Coast into Candoshi. **B)** Admixture graph for testing the Hypothesis 2 (From Eastern to Western) the gene flow event from Candoshi into Tallanes. Hypothesis 1 is better supported considering its number of zeroed branches in contrast to Hypothesis 2.



Hypothesis 1 (Western into Eastern)
Worst f-stat : 2.822
Final score : 28.684
Zero branch lengths: 0

Hypothesis 2 (Eastern into Western)
Worst f-stat : 2.823
Final score : 25.863
Zero branch lengths: 2

Figure S18. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes. We explore the relationship of the Tallanes-Lamas and the direction of the gene flow. White balls in the intermediate nodes represent hypothetical ancestors for each divergence event. **A)** Admixture graph for testing the Hypothesis 1 (from Western to Eastern): the gene flow from the Northern Coast into Lamas. **B)** Admixture graph for testing the Hypothesis 2 (From Eastern to Western) the gene flow event from Lamas into Tallanes. Hypothesis 1 is better supported considering that it does not include any zeroed branches in contrast to Hypothesis 2.



Hypothesis 1 (Western into Eastern)
Worst f-stat : -1.888
Final score : 16.115
Zero branch lengths: 1

Hypothesis 2 (Eastern into Western)
Worst f-stat : 5.689
Final score : 75.605
Zero branch lengths: 4

Figure S19. Admixture graphs and their parameters to test two hypotheses for gene flow across the Fertile Andes. We explore the relationship of the Tallanes-Chachapoyas and the direction of the gene flow. White balls in the intermediate nodes represent hypothetical ancestors for each divergence event. **A)** Admixture graph for testing the Hypothesis 1 (from Western to Eastern): the gene flow from the Northern Coast into Chachapoyas. **B)** Admixture graph for testing the Hypothesis 2 (From Eastern to Western) the gene flow event from Chachapoyas into Tallanes. Hypothesis 1 is better supported considering its significant f statistic, lower final score and just one zeroed branch lengths.

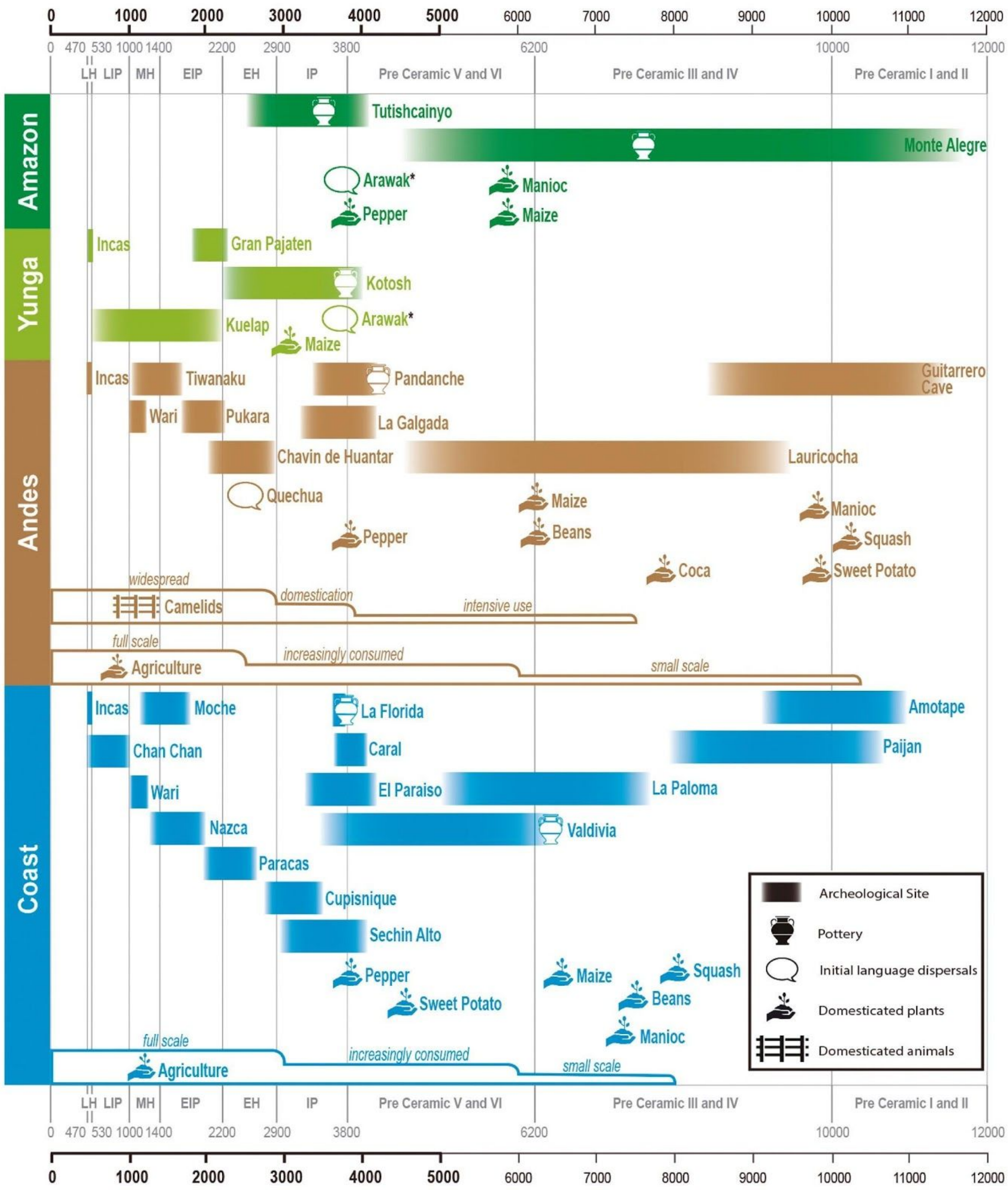


Figure S20. Key historical events of Peruvian prehistory in four longitudinal regions: Peruvian Coast, Andes, Amazon Yunga and Amazonia. Pottery and cultivars symbols represent the earliest archaeological record for the region. To account for time uncertainties, This figure showed the events in the chronology plot without clearly defined chronological borders. Timeline on the top and bottom is represented in Years before present. LH: Late Horizon, LIP: Late Intermediate Period, MH: Middle Horizon, EIP: Early Intermediate Period, EH: Early Horizon, IP: Initial Period. *Controversial geographic region of Arawak origin. Each step in Agriculture and Camelids representations shows an increase in their relative importance. Adapted from ref. 92, which is licensed under CC BY 4.0.

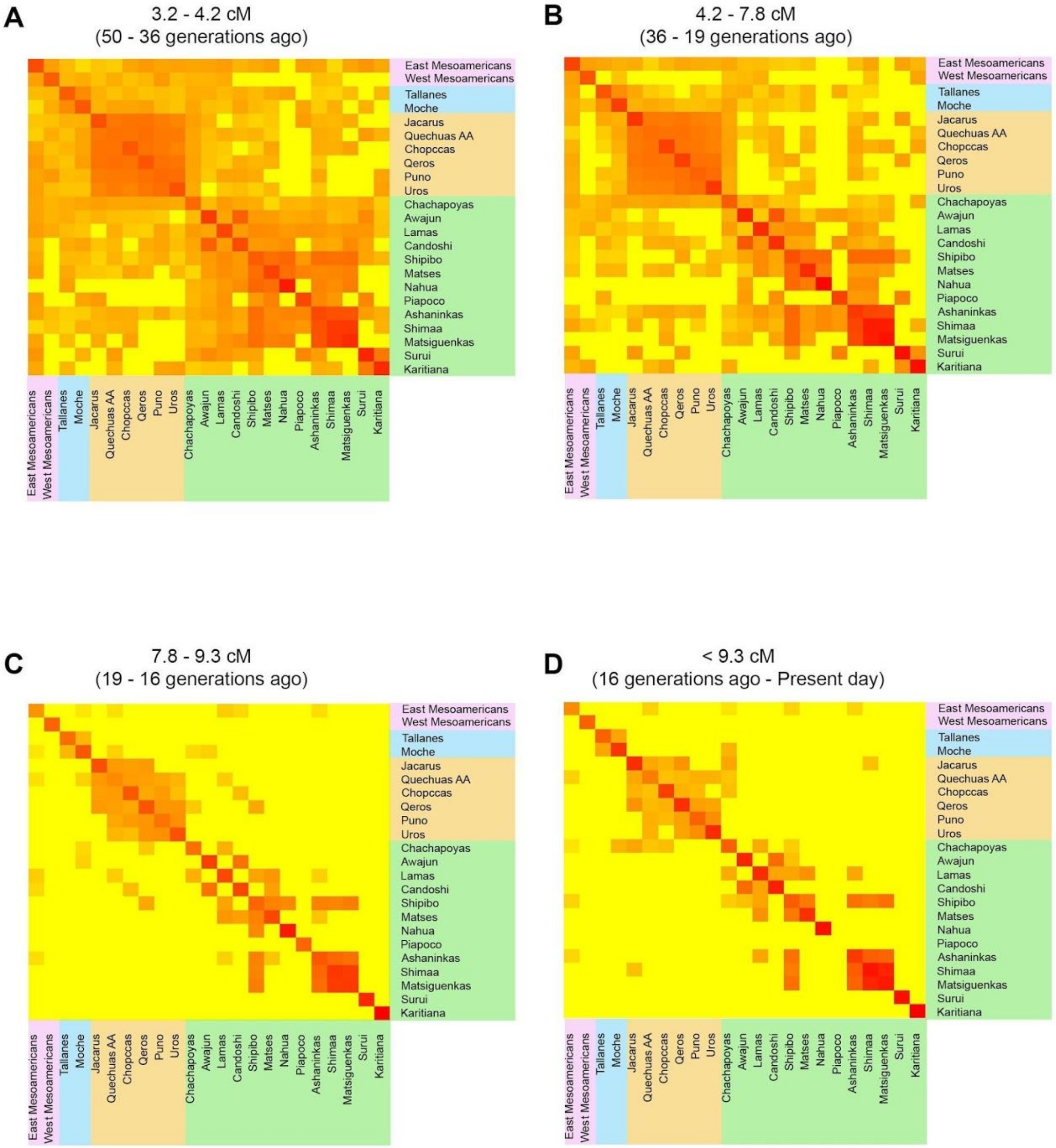


Figure S21. Heatmap representation of the shared Identical by descent (IBD) segments among Native Americans of the Natives 1.9M dataset. Each heatmap represents an interval of segments size and is correlated with time generation for the most common recent ancestor. A) An interval from 3.2 to 4.2 cM correlated with 50 to 36 generations ago. B) The second interval from 4.2 to 7.8 cM correlated with 36 to 19 generations ago. C) The third interval from 7.8 to 9.3 cM correlated with 19 to 16 generations ago. D) And the last interval for all segments longer than 9.3 cM correlated with 16 generations ago to the present day.

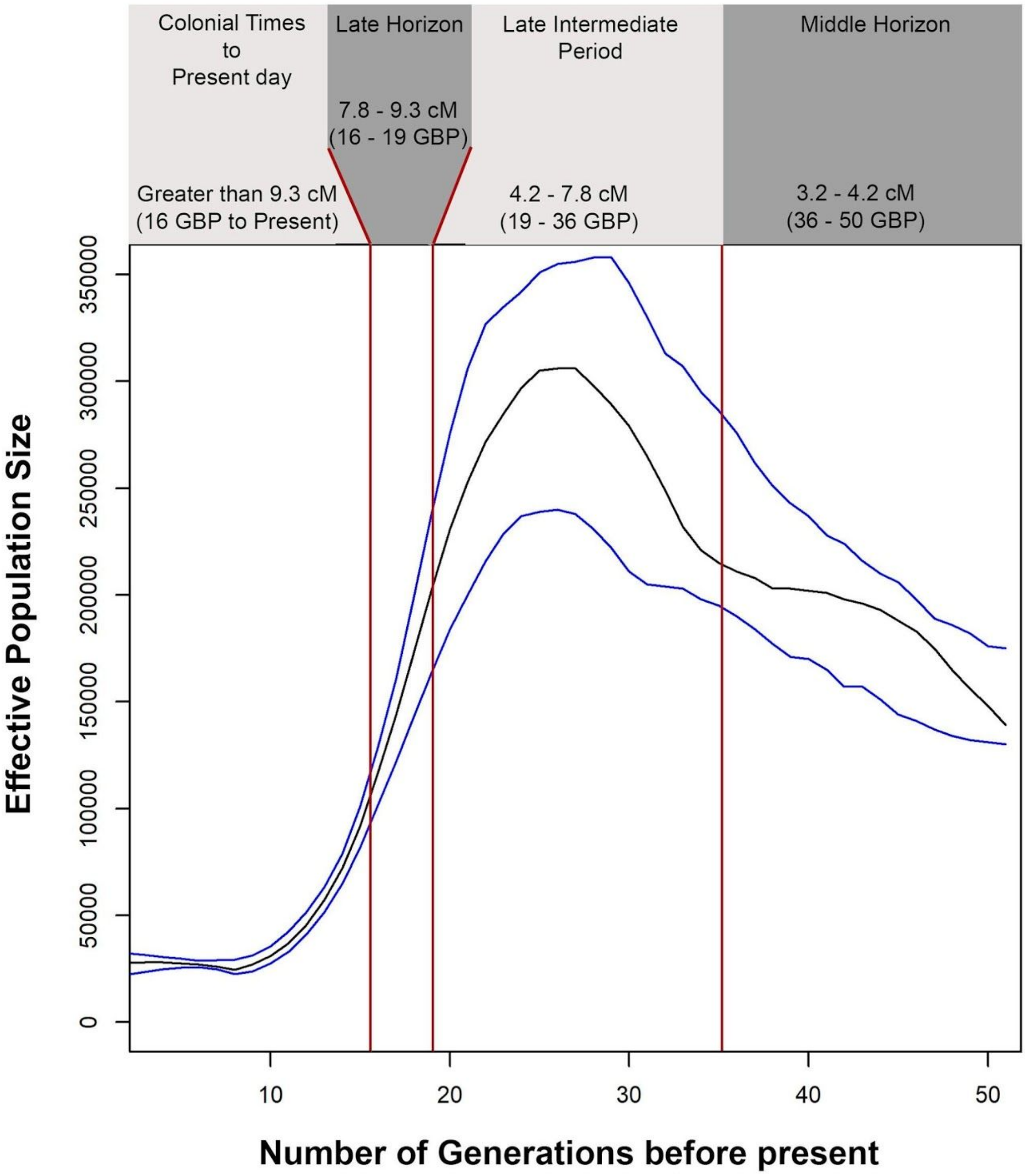


Figure S22. IBDNe analysis to infer the dynamic of the effective population size (Ne) from 4 generations ago to the last 50 generations for the Andean populations (Quechuas_AA, Aimaras_P, Chopccas, Qeros and Uros) as a whole. We used the Natives 1.9M dataset. The x axis represents the number of generations from the present to the past. The y axis represents the estimated value of the Ne. Blocks separated by red lines in the graph correspond to the intervals of the IBD heatmaps (Fig 2). GBP: Generations before present.

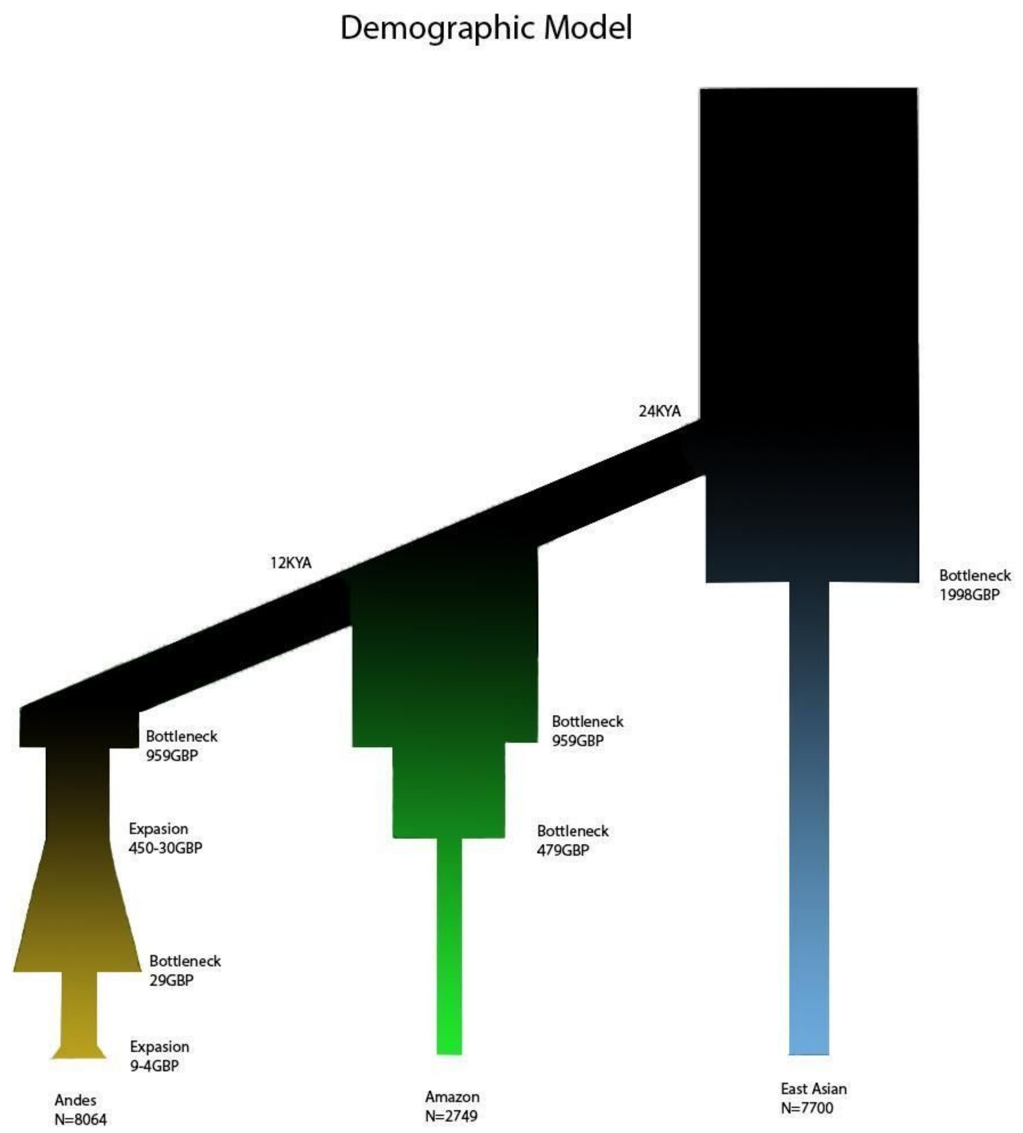


Figure S23. Demographic model of the Andean, Amazonian and East Asian populations. This model was used for the simulations made to calculate the p-value of the obtained PBSn values.

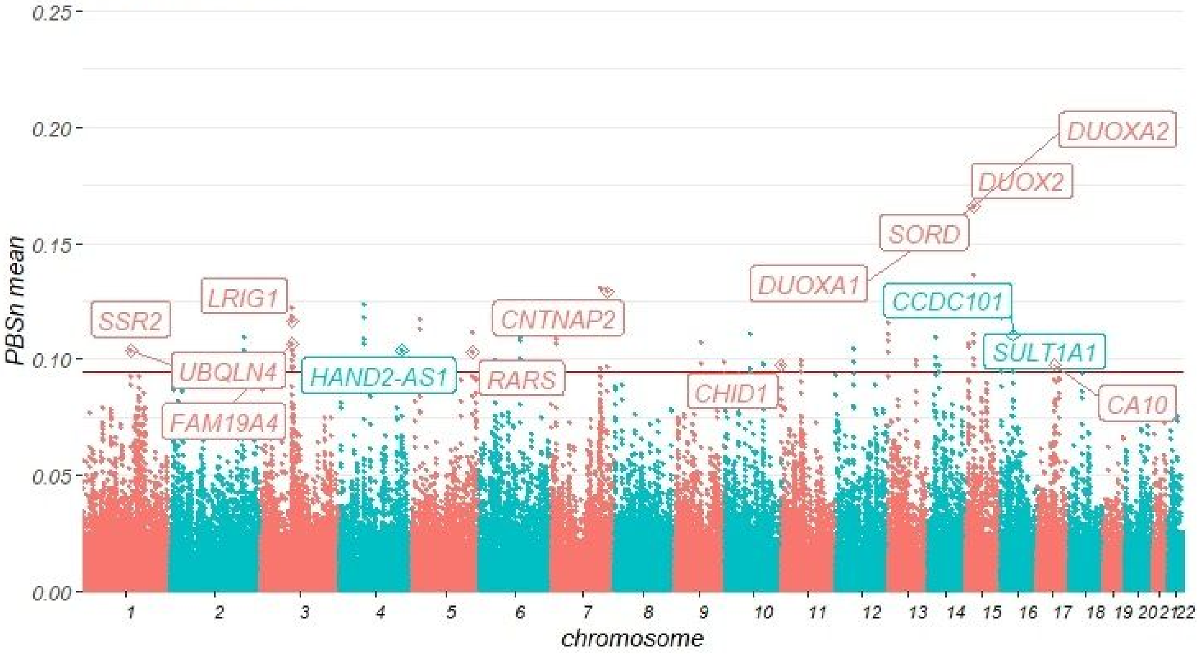


Figure S24. PBSn mean values Andean populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of PBSn mean (red line).

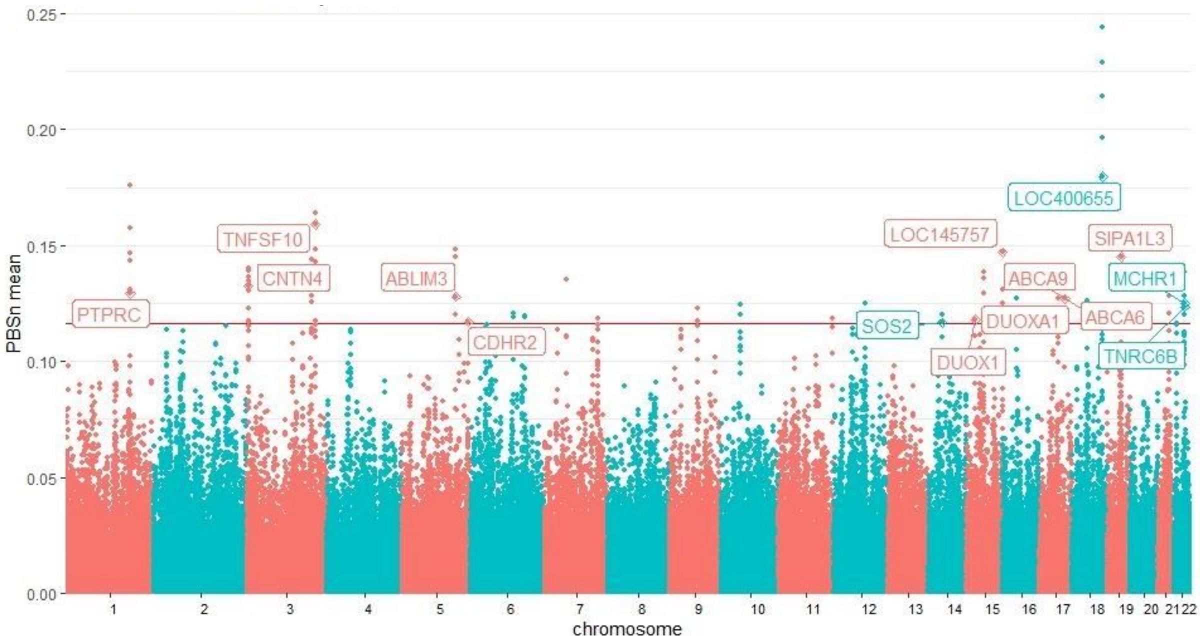


Figure S25. PBSn mean values Amazon populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of PBSn mean (red line).

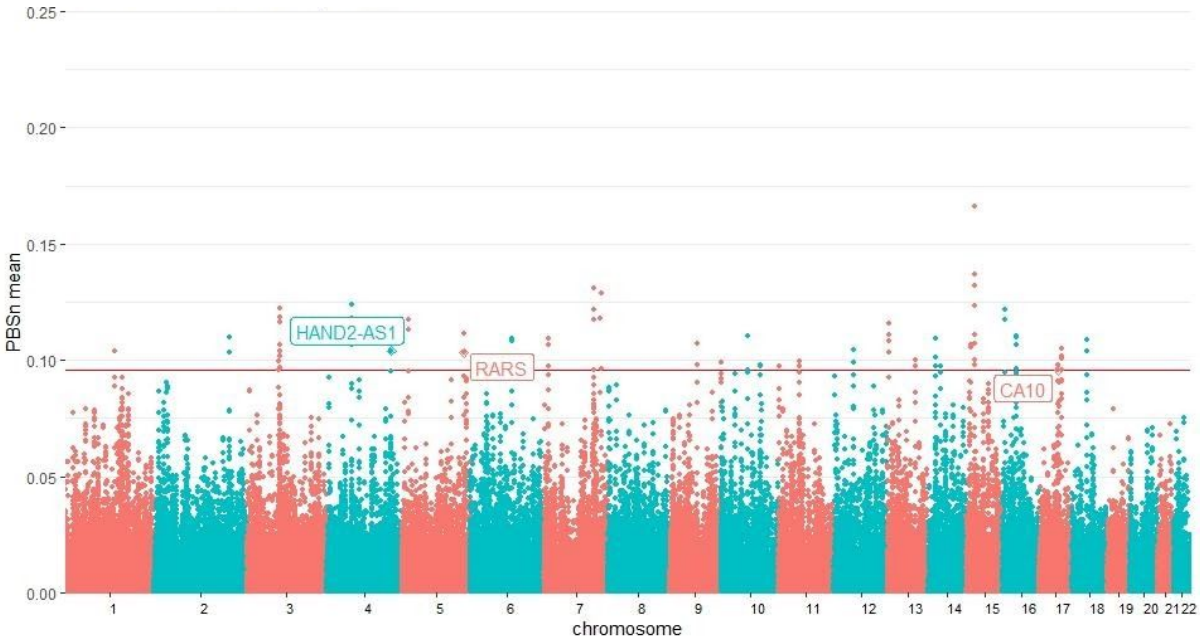


Figure S26. PBSn mean values for windows of 20 SNPs with five SNPs of overleap in Andean populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of windows PBSn mean (red line) that also present high values for xpEHH are labeled.

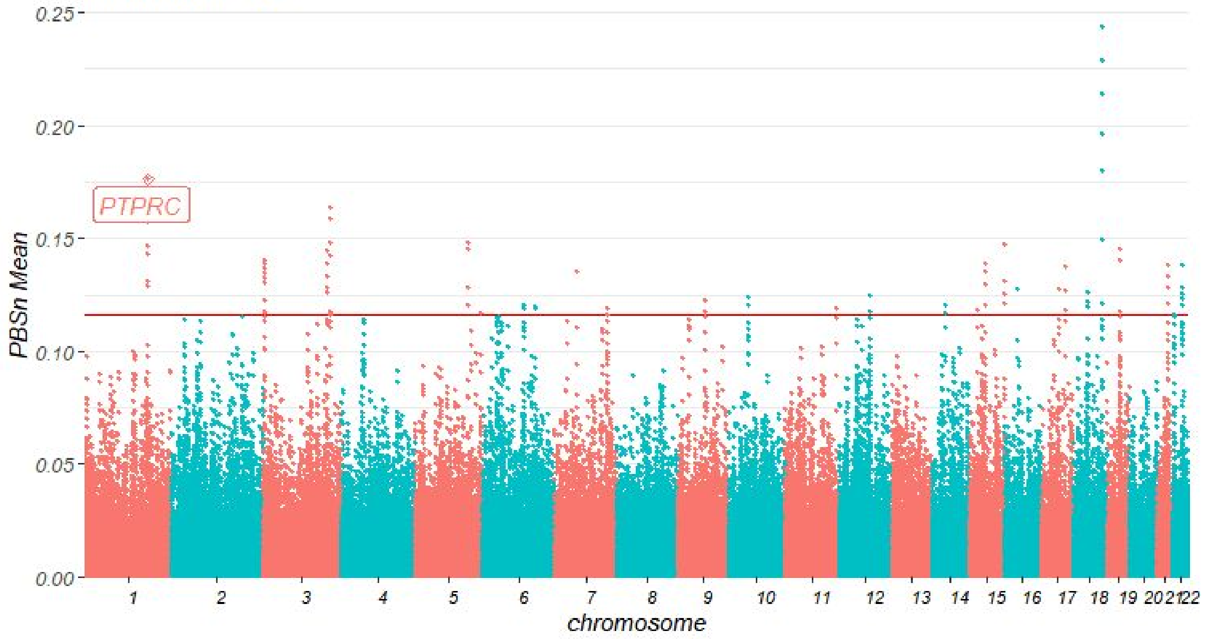


Figure S27. PBSn mean values for windows of 20 SNPs with 5 SNPs of overleap in Amazon populations. Genes related to SNPs inside the 99.95th percentile of PBSn values and the 99.95th percentile of windows PBSn mean (red line) that also present high values for xpEHH are labeled.

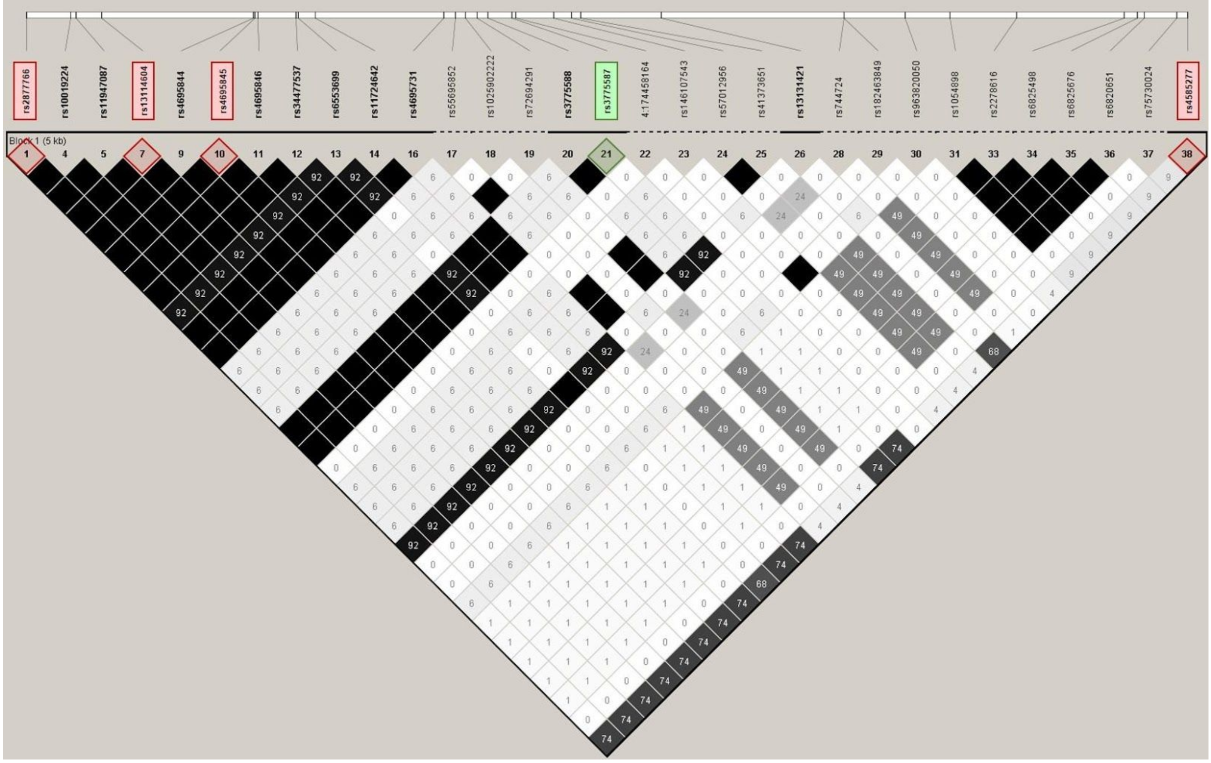


Figure S28. Linkage Disequilibrium between rs3775587 and the SNPs found to be under selection in the gene HAND2-AS1. SNPs with signals in PBS and xpEHH analysis are in red and SNP rs3775587, mapped within the putative enhancer GH04J173536 is in green.

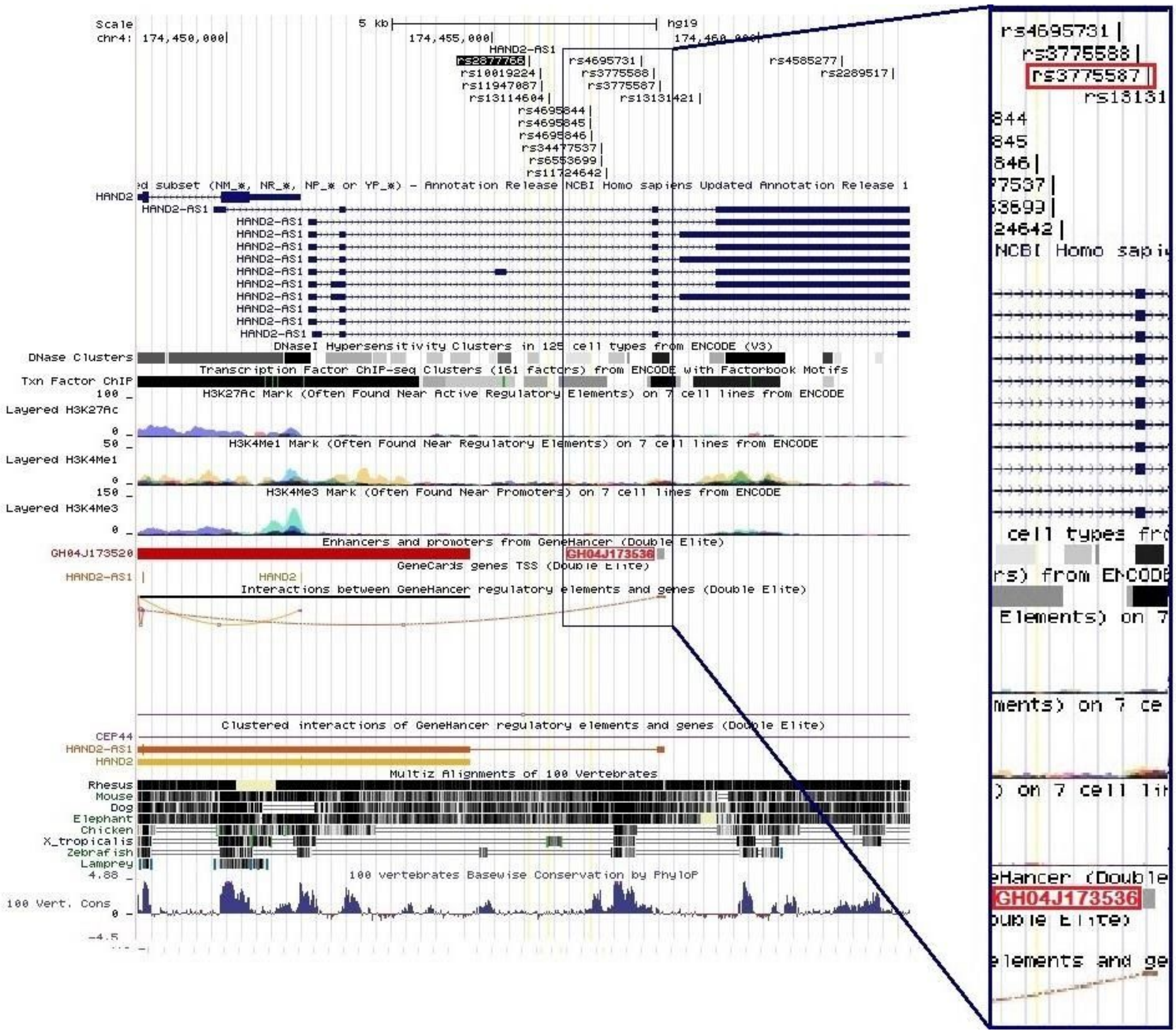

Scale
chr4:
5 kb
hg19
174,450,000
174,455,000
174,460,000
HAND2-AS1
rs2877766
rs10019224
rs11947087
rs13114604
rs4695844
rs4695845
rs4695846
rs34477537
rs6553699
rs11724642
rs4695731
rs3775588
rs3775587
rs13131421
rs4585277
rs2289517
NCBI Homo sapiens Updated Annotation Release 1
HAND2
HAND2-AS1
DNase Clusters
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Txn Factor ChIP
Transcription Factor ChIP-seq Clusters (161 factors) from ENCODE with Factorbook Motifs
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me1
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me3
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
100
50
150
0
Enhancers and promoters from GeneHancer (Double Elite)
GH04J173520
GH04J173536
GeneCards genes TSS (Double Elite)
Interactions between GeneHancer regulatory elements and genes (Double Elite)
Clustered interactions of GeneHancer regulatory elements and genes (Double Elite)
CEP44
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
4.88
-4.5

Figure S29. UCSC Genome Browser view of HAND2-AS1 locus with the SNPs located in regions found to be under selection, DNase I hypersensitivity clusters, Transcription Factor ChIP-seq binding sites, and the histone modifications H3K27ac (Often Found Near Active Regulatory Elements), H3K4me1 (Often Found Near Regulatory Elements) and H3K4me3 (Often Found Near Promoters) on cell lines from the ENCODE Project, and GeneHancer (see Supplementary Methods) and vertebrate conservation data. According to GeneHancer, rs2877766 and other SNPs lie within an ~2.5Kb intronic region of HAND2-AS1, which is located between a promoter/enhancer (GH04J173520) and an enhancer (GH04J173536).

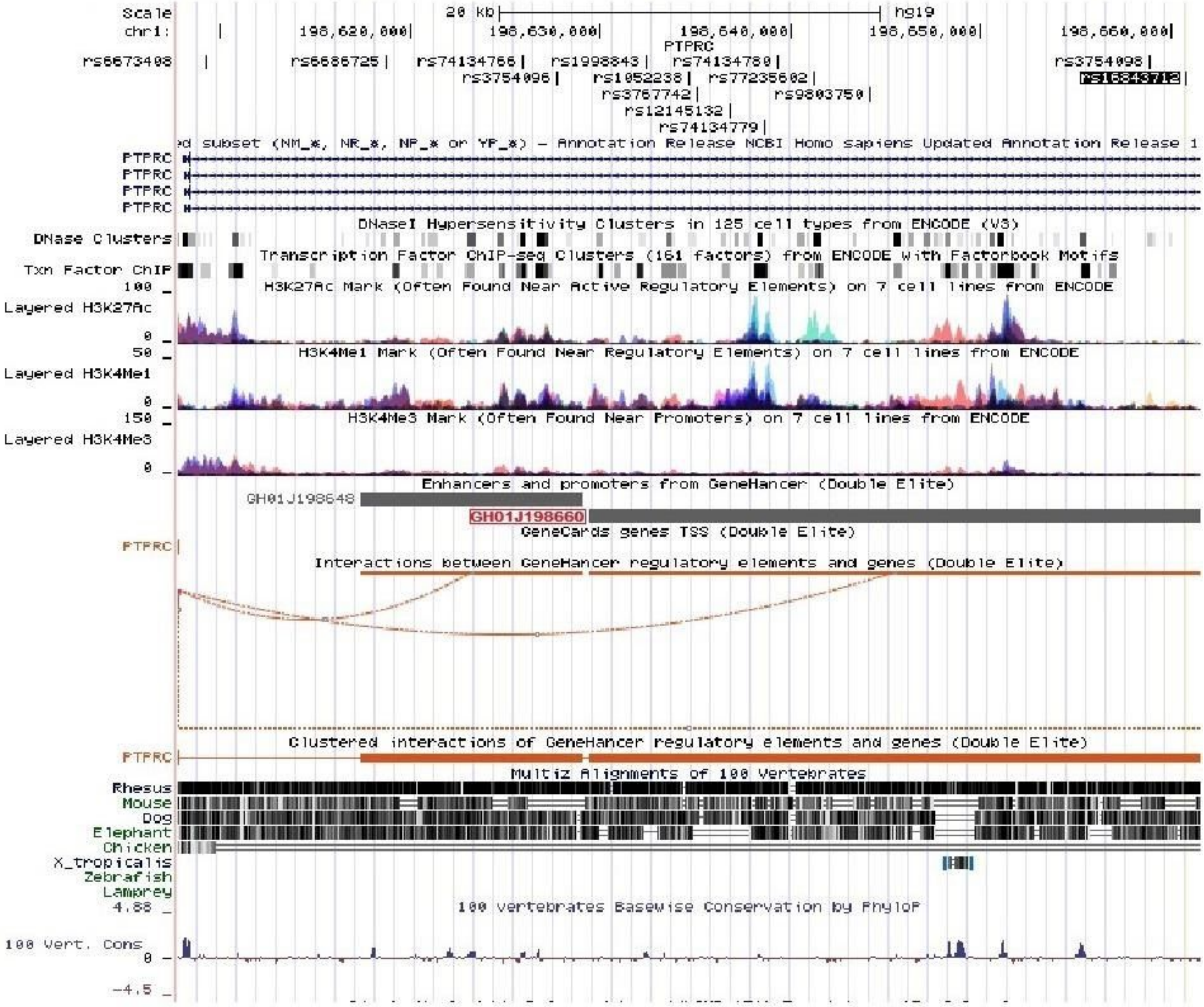

Scale
20 kb
hg19
chr1:
198,620,000
198,630,000
198,640,000
198,650,000
198,660,000
PTPRC
rs6673408
rs6686725
rs74134766
rs1998843
rs74134780
rs3754098
rs3754096
rs1052238
rs77235602
rs16843712
rs3767742
rs9803750
rs12145132
rs74134779
ed subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release NCBI Homo sapiens Updated Annotation Release 1
PTPRC
PTPRC
PTPRC
PTPRC
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq Clusters (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
50
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me1
0
150
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Layered H3K4Me3
0
Enhancers and promoters from GeneHancer (Double Elite)
GH01J198648
GH01J198660
GeneCards genes TSS (Double Elite)
PTPRC
Interactions between GeneHancer regulatory elements and genes (Double Elite)
Clustered interactions of GeneHancer regulatory elements and genes (Double Elite)
PTPRC
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
4.88
100 Vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5

Figure S30. UCSC Genome Browser view of PTPRC locus with the SNPs located in regions found to be under selection, DNase I hypersensitivity clusters, Transcription Factor ChIP-seq binding sites, and the histone modifications H3K27ac (Often Found Near Active Regulatory Elements), H3K4me1 (Often Found Near Regulatory Elements) and H3K4me3 (Often Found Near Promoters) on cell lines from the ENCODE Project, and GeneHancer (see Supplementary Methods) and vertebrate conservation data. According to GeneHancer, rs16843712 and other SNPs lie within an intronic enhancer (GH01J198660) of PTPRC.

Legends for Datasets S1 to S3

Dataset S1: Description of 19 studied Native American populations from Peruvian National Institute of Health and from Laboratory of Human Genetic Diversity. Ashaninka population was sampled twice independently, for this reason, we merge these samples in a unique Ashaninka group and a total of 18 studied populations.

Dataset S2: List of all samples included in the Native 500K dataset.

Dataset S3: List of all samples included in the Native 230K dataset.

Dataset S4: SNPs under selection in Andean populations according to Population Branch Statistic (PBS) test.

Dataset S5: SNPs under selection in Amazon populations according to Population Branch Statistic (PBS) test.

Dataset S6: SNPs under selection in Andean populations according to Population Branch Statistic (PBS) and Cross-Population Extended Haplotype Homozygosity (XP-EHH) tests

Dataset S7: SNPs under selection in Amazon populations according to Population Branch Statistic (PBS) and Cross-Population Extended Haplotype Homozygosity (XP-EHH) tests

Dataset S8: Highly Differentiated Variants Between Andean and Amazon Populations: Annotation from GWAs Catalog. CHR: chromosome, FST: Level of genetic differentiation between groups, A1: alternative allele, AMZ: Amazon populations, AND: Andean populations, PEL: Peruvians from Lima, EAS: East asian populations, EUR: European populations, WAFR: West African populations.

Dataset S9: Highly Differentiated Variants Between Andean and Amazon Populations: Annotation from PharmGKB. CHR: chromosome, FST: Level of genetic differentiation between groups, A1: alternative allele, AMZ: Amazon populations, AND: Andean populations, PEL: Peruvians from Lima, EAS: East asian populations, EUR: European populations, WAFR: West African populations.

Dataset S10: Highly Differentiated Variants Between Andean and Amazon Populations: Annotation from Sift and Polyphen. CHR: chromosome, Wild.AA: Wild Aminoacid, Mutant.AA: Mutant Aminoacid, FST: Level of genetic differentiation between groups, A1: alternative allele, AMZ: Amazon populations, AND: Andean populations, PEL: Peruvians from Lima, EAS: East asian populations, EUR: European populations, WAFR: West African populations.

SI References

1. W. B. Church, A. von Hagen, "Chachapoyas: Cultural Development at an Andean Cloud Forest Crossroads" in *The Handbook of South American Archaeology*, H. Silverman, W.

H. Isbell, Eds. (Springer New York, 2008), pp. 903–926.

2. I. Schellerup, Wayko-Lamas: a Quechua community in the Selva Alta of North Peru under change. *Geografisk Tidsskrift*, 199–208 (1999).

3. G. Seitz, *Cultural Discontinuity: The New Social Face of the Awajun* (Amakella Publishing, 2017).

4. J. M. Guallart, *La tierra de los cinco ríos* (Pontificia Universidad Católica del Perú, Instituto Riva Agüero, 1997).

5. L. Campbell, “Language isolates and their history” in *Language Isolates*, (Routledge, 2017), pp. 1–18.

6. S. Purcell, *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

7. W. C. S. Magalhães, *et al.*, EPIGEN-Brazil Initiative resources: a Latin American imputation panel and the Scientific Workflow. *Genome Res.* **28**, 1090–1095 (2018).

8. A. L. Price, N. A. Zaitlen, D. Reich, N. Patterson, New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).

9. F. S. G. Kehdy, *et al.*, Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8696–8701 (2015).

10. 1000 Genomes Project Consortium, *et al.*, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

11. J. Z. Li, *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).

12. D. Reich, *et al.*, Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).

13. S. Mallick, *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

14. M. Raghavan, *et al.*, Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).

15. R. E. Green, *et al.*, A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).

16. N. Patterson, *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).

17. B. K. Maples, S. Gravel, E. E. Kenny, C. D. Bustamante, RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).

18. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

19. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

20. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).

21. O. Delaneau, J. Marchini, J.-F. Zagury, A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).

22. G. Hellenthal, *et al.*, A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).

23. S. Leslie, *et al.*, The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).

24. L. van Dorp, *et al.*, Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. *PLoS Genet.* **11**, e1005397 (2015).

25. J.-C. Chacón-Duque, *et al.*, Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat. Commun.* **9**, 5388 (2018).

26. G. A. Gnecchi-Ruscone, *et al.*, Dissecting the Pre-Columbian Genomic Ancestry of Native Americans along the Andes–Amazonia Divide. *Mol. Biol. Evol.* **36**, 1254–1269 (2019).

27. D. W. Lathrap, The antiquity and importance of long-distance trade relationships in the moist tropics of pre-Columbian South America. *World Archaeol.* **5**, 170–186 (1973).

28. H. Silverman, W. Isbell, *Handbook of South American Archaeology* (Springer Science & Business Media, 2008).

29. C. Quintana, R. T. Pennington, C. U. Ulloa, H. Balslev, Biogeographic Barriers in the Andes: Is the Amotape—Huancabamba Zone a Dispersal Barrier for Dry Forest Plants? *Ann. Mo. Bot. Gard.* **102**, 542–550 (2017).

30. J. Guffroy, "Cultural Boundaries and Crossings: Ecuador and Peru" in *The Handbook of South American Archaeology*, H. Silverman, W. H. Isbell, Eds. (Springer New York, 2008), pp. 889–902.

31. J. R. Sandoval, *et al.*, The Genetic History of Peruvian Quechua-Lamistas and Chankas: Uniparental DNA Patterns among Autochthonous Amazonian and Andean Populations. *Ann. Hum. Genet.* **80**, 88–101 (2016).

32. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).

33. A. Bergström, *et al.*, A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science* **357**, 1160–1163 (2017).

34. M. Raghavan, *et al.*, Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014).

35. S. R. Browning, B. L. Browning, High-resolution detection of identity by descent in

unrelated individuals. *Am. J. Hum. Genet.* **86**, 526–539 (2010).

36. D. Speed, D. J. Balding, Relatedness in the post-genomic era: is it still useful? *Nat. Rev. Genet.* **16**, 33–44 (2015).

37. E. A. Thompson, Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* **194**, 301–326 (2013).

38. S. Baharian, *et al.*, The Great Migration and African-American Genomic Diversity. *PLoS Genet.* **12**, e1006059 (2016).

39. V. Pankratov, *et al.*, East Eurasian ancestry in the middle of Europe: genetic footprints of Steppe nomads in the genomes of Belarusian Lipka Tatars. *Sci. Rep.* **6**, 30197 (2016).

40. B. L. Browning, S. R. Browning, Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).

41. S. R. Browning, B. L. Browning, Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *Am. J. Hum. Genet.* **97**, 404–418 (2015).

42. A. Gusev, *et al.*, Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009).

43. J. Haas, S. Pozorski, T. Pozorski, *The Origins and Development of the Andean State* (Cambridge University Press, 1987).

44. C. Stanish, The Origin of State Societies in South America. *Annu. Rev. Anthropol.* **30**, 41–64 (2001).

45. S. C. Stearns, R. M. Nesse, D. R. Govindaraju, P. T. Ellison, Evolutionary perspectives on health and medicine. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1691–1695 (2010).

46. E. Vasseur, L. Quintana-Murci, The impact of natural selection on health and disease: uses of the population genetics approach in humans. *Evol. Appl.* **6**, 596–607 (2013).

47. R. Lewin, *Human Evolution: An Illustrated Introduction* (John Wiley & Sons, 2009).

48. S. Fan, M. E. B. Hansen, Y. Lo, S. A. Tishkoff, Going global by adapting local: A review of recent human adaptation. *Science* **354**, 54–59 (2016).

49. F. M. Salzano, The role of natural selection in human evolution - insights from Latin America. *Genet. Mol. Biol.* **39**, 302–311 (2016).

50. X. Yi, *et al.*, Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).

51. D. L. Hartl, A. G. Clark, A. G. Clark, *Principles of population genetics* (Sinauer associates Sunderland, MA, 1997).

52. J. Goudet, Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Resour.* **5**, 184–186 (2005).

53. A. Benazzo, A. Panziera, G. Bertorelle, 4P: fast computing of population genetics

statistics from large DNA polymorphism panels. *Ecol. Evol.* **5**, 172–175 (2015).

54. C. C. Cockerham, B. S. Weir, Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics* **40**, 157–164 (1984).

55. L. L. Cavalli-Sforza, Human diversity in *Proc. 12th Int. Congr. Genet*, (1969), pp. 405–416.

56. J. E. Crawford, *et al.*, Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans. *Am. J. Hum. Genet.* **101**, 752–767 (2017).

57. 1000 Genomes Project Consortium, *et al.*, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

58. I. Shlyakhter, P. C. Sabeti, S. F. Schaffner, Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics* **30**, 3427–3429 (2014).

59. S. W. Buskirk, R. E. Peace, G. I. Lang, Hitchhiking and epistasis give rise to cohort dynamics in adapting populations. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8330–8335 (2017).

60. P. C. Sabeti, *et al.*, Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).

61. P. C. Sabeti, *et al.*, Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

62. Z. A. Szpiech, R. D. Hernandez, selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Molecular Biology and Evolution* **31**, 2824–2827 (2014).

63. G. Soares-Souza, “Novas Abordagens para Integração de Bancos de Dados e Desenvolvimento de Ferramentas Bioinformáticas para Estudos de Genética de Populações,” Universidade Federal de Minas Gerais. (2014).

64. S. T. Sherry, *et al.*, dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

65. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–7 (2005).

66. B. Jassal, *et al.*, The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).

67. B. Braschi, *et al.*, Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res.* **47**, D786–D792 (2019).

68. A. Buniello, *et al.*, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

69. I. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7**, Unit7.20

(2013).

70. Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, A. P. Chan, Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).

71. R. Vaser, S. Adusumalli, S. N. Leng, M. Sikic, P. C. Ng, SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).

72. F. Hsu, *et al.*, The UCSC Known Genes. *Bioinformatics* **22**, 1036–1046 (2006).

73. M. Ashburner, *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

74. The Gene Ontology Consortium, The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

75. M. Whirl-Carrillo, *et al.*, Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92**, 414–417 (2012).

76. V. C. Jacovas, *et al.*, Selection scan reveals three new loci related to high altitude adaptation in Native Andeans. *Sci. Rep.* **8**, 12733 (2018).

77. D. N. Harris, *et al.*, Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proc. Natl. Acad. Sci. U. S. A.*, 201720798 (2018).

78. J. C. Barrett, B. Fry, J. Maller, M. J. Daly, Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).

79. A. Carré, *et al.*, When an Intramolecular Disulfide Bridge Governs the Interaction of DUOX2 with Its Partner DUOXA2. *Antioxid. Redox Signal.* **23**, 724–733 (2015).

80. W. J. Kent, *et al.*, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

81. T. J. P. Hubbard, *et al.*, Ensembl 2007. *Nucleic Acids Res.* **35**, D610–7 (2007).

82. Y. Ahmad, *et al.*, The proteome of Hypobaric Induced Hypoxic Lung: Insights from Temporal Proteomic Profiling for Biomarker Discovery. *Sci. Rep.* **5**, 10681 (2015).

83. Y. Shen, *et al.*, Ischemic preconditioning inhibits over-expression of arginyl-tRNA synthetase gene Rars in ischemia-injured neurons. *J. Huazhong Univ. Sci. Technolog. Med. Sci.* **36**, 554–557 (2016).

84. X. Cheng, H. Jiang, Long non-coding RNA HAND2-AS1 downregulation predicts poor survival of patients with end-stage dilated cardiomyopathy. *J. Int. Med. Res.* **47**, 3690–3698 (2019).

85. Y.-Z. Fu, *et al.*, Human Cytomegalovirus Tegument Protein UL82 Inhibits STING-Mediated Signaling to Evade Antiviral Immunity. *Cell Host Microbe* **21**, 231–243 (2017).

86. D. Xie, *et al.*, Exploring the associations of host genes for viral infection revealed by genome-wide RNAi and virus-host protein interactions. *Mol. Biosyst.* **11**, 2511–2519 (2015).

87. A. van der Vliet, K. Danyal, D. E. Heppner, Dual oxidase: a novel therapeutic target in allergic disease. *Br. J. Pharmacol.* **175**, 1401–1418 (2018).

88. S. Meer, Y. Perner, E. D. McAlpine, P. Willem, Extraoral plasmablastic lymphomas in a high human immunodeficiency virus endemic area. *Histopathology* (2019) https:/doi.org/10.1111/his.13964.

89. A. S. Motani, *et al.*, Evaluation of AMG 076, a potent and selective MCHR1 antagonist, in rodent and primate obesity models. *Pharmacol Res Perspect* **1**, e00003 (2013).

90. E. M. van Leeuwen, *et al.*, Genome of The Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. *Nat. Commun.* **6**, 6065 (2015).

91. A. Piehler, W. E. Kaminski, J. J. Wenzel, T. Langmann, G. Schmitz, Molecular structure of a novel cholesterol-responsive A subclass ABC transporter, ABCA9. *Biochem. Biophys. Res. Commun.* **295**, 408–416 (2002).

92. Scliar, M.O., Gouveia, M.H., Benazzo, A. *et al.* Bayesian inferences suggest that Amazon Yunga Natives diverged from Andeans less than 5000 ybp: implications for South American prehistory. *BMC Evol Biol* **14,** 174 (2014).