PNAS
www.pnas.org

Supplementary Information for

Deep genetic affinity between Coastal Pacific and Amazonian natives evidenced by Australasian ancestry

Marcos Araújo Castro-Silva[a,1], Tiago Ferraz[a,1], Maria Cátira Bortolini[b], David Comas[c], and Tábita Hünemeier[b,2]

Tábita Hünemeier
Email: hunemeier@usp.br

[a]*Departmento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil;* [b]*Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, 91501-970 Porto Alegre, RS, Brazil;* [c]*Institut de Biologia Evolutiva, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain*

This PDF file includes:

Other supplementary materials for this manuscript include the following:

Summary

Supplementary Information Text

Extended Methods

Data overview

Newly generated data for 37 Brazilian natives from 4 indigenous communities, namely Asurini, Munduruku (both Tupí-speaker groups), Xavánte, and Xikrin (both Jê-speaker groups), were genotyped in the Axiom Human Origins array - Affymetrix/Thermo Fisher (1). These populations are settled in the Amazonian rainforest (Asurini, Munduruku, and Xikrin), or in the Brazilian central plateau tropical savanna (Xavánte).

Ethical approval for sample collection was provided by the Brazilian National Ethics Commission (CONEP Resolution no. 123 and 4599). CONEP also approved oral consent for the use of these samples in population history and human evolution studies. Individual and/or tribal informed oral consent was obtained from participants who were not able to read or write. All sampling was coordinated by Francisco Mauro Salzano (*in memoriam*) and their collaborators, in a manner consistent with the Declaration of Helsinki and Brazilian laws and regulations applicable at the time of sampling. Logistical support for the sample collection was provided by the Fundação Nacional do Índio. The results of this study were presented to the participating communities.

Dataset assembly and quality control

These data were merged with publicly available datasets (Axiom Human Origins array - Affymetrix/Thermo Fisher (1) genotyped or whole-genome sequenced) of populations from Brazil (1–3) and other countries in South America (Colombia, Ecuador, and Peru) and Mexico (4–6). Finally, we also combined the 1240K_HO dataset assembly (v42.4) and the merging procedure was conducted using PLINK 1.9 (7), sharing single nucleotide polymorphisms (SNPs) across merging datasets. The resulting dataset contained 383 individuals from 58 indigenous groups (Dataset S1A), along with the 67 world-wide reference populations (Dataset S2) and a total of 438,443 SNPs. Next, we removed markers with more than a 5% absence rate, and no sample was removed with a 10% absence rate criteria. We also excluded markers with a pairwise correlation above 20% (r^2

> 0.2 inside a sliding window of 50 kb size and step size of 10 kb), obtaining a subset of 127,931 markers and applied an unsupervised ADMIXTURE (8) analysis with K = 3 on the subset of samples from the American continent, Sub-Saharan Africa, and Western Europe. We then estimated the pairwise IBD with PLINK 1.9 (plink --file mydata --genome), which uses the which uses the method-of-moments to calculate the probability of sharing 0 (ZO), 1 (Z1), or 2 (Z2) alleles identical by descent between any given pair of individuals over all the loci, and the total proportion of IBD is estimated between a pair of individuals as PI_HAT = Z2+0.5*Z1. We then used a threshold of PI-HAT < 0.375 to identify the maximum unrelated dataset with PRIMUS (9). Finally, we filtered the data to remove admixed (< 99% inferred non-Native American ancestry; 150 unadmixed samples) and selected the maximum unrelated set of individuals (PI_HAT < 0.375; 312 unrelated samples). The subset of unrelated and unadmixed Native American samples includes 87 individuals. Metadata for every Native American sample (test samples) and every reference population sample are summarized in Dataset S1A and Dataset S2, respectively. The complete set of SNPs and the subset of unrelated and unadmixed Native American samples was used for all analyses, unless otherwise specified.

Exploratory data analysis

Initially, Principal Component Analyses were applied with SNPRelate R/Bioconductor (10) to the LD pruned dataset, obtained as above mentioned, to check data quality, inconsistencies introduced by the merging process and most importantly if any Native American sample was an outlier in relation to the other American indigenous groups. We also applied the ancestry estimates obtained with the unsupervised ADMIXTURE analysis, as previously described, to visualize and evaluate the influence of the proportions genetic ancestry created by the recent post-Contact 3-way admixture between (Native Americans, European conquistadors, and enslaved Sub-Saharan Africans). Next, we performed a PCA on the subset of unadmixed and unrelated Native Americans, in order to examine the broad patterns of ancestry and genetic differentiation, as well as to ensure the absence of outliers in our data set.

Multidimensional scaling of genetic distances

Next, to assess the patterns of allele sharing between individuals, we estimated the *Outgroup* F_3(Y, Z; Mbuti), calculating for every pair of Y and Z indigenous groups. Additionally, a matrix of *Outgroup* F_3(Y, Z; Mbuti) calculated for all Y and Z pairs of individuals, was converted to genetic distances (Genetic distance = 1 - *Outgroup* F_3 estimate). A multidimensional scaling analysis (MDS) was then applied to the matrix of pairwise genetic distances with the "stats" R package.

D-statistics

First, we examined the presence of an excess allele sharing between all Native American groups in the unadmixed and unrelated dataset and present-day indigenous Papuans, Australians, Melanesians, and Andamanese, which was considered to be a signal of the ancestry contribution from the so-called "Population Y" (2). To accomplish this we used Admixtools (1) to estimate D-statistics (Mbuti, Austro-Melanesian; Y, Z) we defined "Australasian" as any Australasian group present in our dataset (i.e., Australian, Melanesian, Onge, or Papuan), and Y and Z as any modern Native American group or individual (e.g., Mixe, Karitiana, or Xavánte). Therefore we estimated the D-statistic for all pairs of Native American individuals and groups. The D-statistic as well as the standard error is estimated by qpDstat program from Admixtools (3), using a weighted Block Jackknife procedure, in which the genome is divided into blocks of 5 cM (default parameter) then multiple runs are executed deleting one block at a time, which allows the estimation of the statistic mean and standard error.

qpWave

Second, we used qpWave (1) to infer how many admixture flows from outside the American continent would be necessary to produce the genetic diversity of present-day Central (represented only by Mixe in this analysis) and South American indigenous groups. The qpWave software infers that if a given set f4(W, X; Y, Z) statistics are consistent with rank 0, 1, or 2 (or more), the test populations (W and X) derive from 1, 2,

and 3 (or more) streams of ancestry from the outgroup populations (Y and Z), respectively. To do so, a set of tests in the form f4($test_1$, $test_2$; $outgroup_1$, $outgroup_2$) was performed, following the original design used by Skoglund et al. (2). As test populations, we analyzed 14 indigenous groups with a minimum of 3 individuals (unadmixed and unrelated) and as outgroups, 4 populations from 6 world regions: Sub-Saharan Africa (Mbuti, Yoruba, Bantu-SouthAfrica, and Bantu-Kenya), Western Europe (Sardinian, French, Orcadian, and Spanish), East Asia (Han, Japanese, Miao, and Uygur), South Asia (Onge [ONG.SG], Sindhi, Cambodian, and Dai), Siberia/Central Asia (Mongola, Yakut, Oroqen, and Hezhen), Oceania (Papuan, Melanesian, and Australian) (Dataset S6A-B). Next, a series of tests were performed by dropping one of the above-mentioned regions at each time, followed by a series of tests performed by dropping one of the test groups (Native Americans) at each time, and finally, two sets of tests were performed by keeping Africa or Siberia/Central Asia plus one of the other regions at each time (Dataset S6A). Furthermore, we also tried to produce a more refined overview of the presence of these deeply divergent ancestries and to evaluate the extent of the variation between contemporary Native Americans; to accomplish this, qpWave was applied to every pair and trio of the test groups, including all worldwide regions mentioned above as outgroups, and the data is summarized in Dataset S6C-D.

qpgraph

Finally, we aimed to assess the population history models to investigate how this deeply diverged shared ancestry between present-day indigenous Australasians and South Native American groups emerged, especially now in the light of the new evidence (D-statistics) pointing to the existence of such affinity, not only amongst Amazonians (Karitiana and Suruí) but also in the Pacific coastal population (Chotuna) and populations from other Brazilian regions (Xavánte from the central Brazilian plateau). In this sense, we applied the models proposed and published by Skoglund et al. (2) as a scaffold admixture graph to model and test these additional groups (i.e., Chotuna and Xavánte). These groups were included by computing all possible positions in the admixture graph scaffold independently. To test the existence of this genetic affinity, we added the Pacific coastal

groups Chotuna, Narihuala, and Sechura to the above-mentioned models; next, we also included Xavánte. We also tried another approach in which we started by first adding Xavánte and the Pacific coastal groups to the scaffold, and only then adding Suruí and Karitiana. Finally, we compared the worst estimated Z-value of all computed models, selecting the candidates with the best fit to the data to represent the population history (i.e., tree topology and admixture events).

Treemix

We also aimed to produce an outline of the population history of the Native American groups here represented, by using an alternative method, distinct from the F-statistics (1, 11) framework. This was done with Treemix (12), which implements an unsupervised method of estimating a Maximum Likelihood tree of the population pairwise allelic covariances and allows the inference of putative gene flow between branches of the tree. First, we inferred the ML tree and then allowed the model to fit a growing number of gene flow events until a plateau of the model likelihood was reached.

Datasets (S1 to S6)

Dataset S1. Metadata and exploratory analysis of test samples.

This dataset includes **(A)** metadata for each Native American individual used in our analyses. The information includes original group name (as used in the data source study), group name (as used in this study), individual ID, major ethnolinguistic group affiliation, country of origin, data source study, data source method (e.g., Axiom Human Origins array or Shotgun sequencing), geographic coordinates, inclusion in the maximum unrelated dataset (True or False), and presence of non-Native American admixture (True or False). This dataset also presents estimates produced by an unsupervised ADMIXTURE (8) analysis on the subset of Native Americans with K = 3. The colors used to represent each individual throughout the study are also included. Additionally, it contains data for the PCAs performed with **(B)** the complete set of Native Americans and **(C)** the subset of unadmixed and unrelated samples. Finally **(D)** a multidimensional scaling analysis was performed on a matrix of the genetic distances (1 - Outgroup F_3) between all pairs of samples in the unrelated and unadmixed subset.

Dataset S2. Metadata for reference samples.

This dataset includes metadata for each individual from a reference population used in our analyses. The information includes group name, individual ID, country of origin, data source study, macro-region of origin, continent of origin, and data source method (e.g., Axiom Human Origins array or Shotgun sequencing).

Dataset S3. Significant Z-values for the D(Mbuti, Australasian; Y, Z) statistic.

The D-statistics (Mbuti, Australasian; Y, Z) (1) for every combination of an Australasian group (i.e., Australian, Australian.DG, Melanesian, Onge, or Papuan) and a pair of Y and Z American indigenous groups, with one set including related samples and another excluding related samples (i.e., maximum unrelated dataset). The complete set of D statistics are accessible in Dataset S4A-B.

Dataset S4. Estimates of D-statistics (Mbuti, Australasian; Y, Z) for every pair of Y and Z indigenous groups and individuals.

This dataset presents the D-statistics D(Mbuti, Australasian; Y, Z) (1) for every combination of an Australasian group (i.e., Australian, Melanesian, Onge, or Papuan) and a pair of Y and Z American indigenous groups, with **(A)** one set including related samples and **(B)** another excluding related samples (i.e., maximum unrelated dataset tab), and also **(C)** for every pair of Y and Z individuals in our dataset. It includes information on D-statistic, Z-Value, number of ABBA and BABA positions, and the total number of shared SNPs across the tested populations. Finally, **(D)** a summary of the number of significant tests per ID when they are at the Y and Z positions of the statistic is provided.

Dataset S5. Number of ancestry streams consistent with the Central and South American genetic diversity.

The consistency of 1 to 4 admixture flows between Central and South American indigenous groups (test populations - Dataset S1A) and other global-wide populations (reference populations - Dataset S2) was tested with qpWave (1). **(A)** A series of tests were performed using different combinations of reference populations as described in the first column. The remaining columns exhibit the *p*-value for each number of tests (i.e., 1 to 4), and significant values are marked with an asterisk. We also present **(B)** the group qpWave weights for the Full dataset analysis, along with a summary of the results of another series of qpWave analysis testing all **(C)** pairs and **(D)** trios of Native American groups.

Dataset S6. Maximum likelihood tree and gene flow events.

A maximum likelihood tree based on the population pairwise allelic covariances matrix was obtained with Treemix (12)) and an increasing number of gene flow events were adjusted by the model (up until 8 events). Here we present (**A**) the likelihood for each of these models, the covariance matrices (**B**) for the ML tree with no gene flow events and (**C**) for the model with 6 gene flow events, which is the first model to reach the likelihood plateau. Finally we (**D**) present the ML tree with no gene flow events and (**E**) the model with 6 gene flow events.

SI References

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

2. P. Skoglund, *et al.*, Genetic evidence for two founding populations of the Americas. *Nature* **525**, 104–108 (2015).

3. M. A. Castro e Silva, *et al.*, Genomic insight into the origins and dispersal of the Brazilian coastal natives. *Proceedings of the National Academy of Sciences* **117**, 2372–2377 (2020).

4. I. Lazaridis, *et al.*, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).

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

6. C. Barbieri, *et al.*, The Current Genomic Landscape of Western South America: Andes, Amazonia, and Pacific Coast. *Mol. Biol. Evol.* **36**, 2698–2713 (2019).

7. C. C. Chang, *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

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

9. J. Staples, D. A. Nickerson, J. E. Below, Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genet. Epidemiol.* **37**, 136–141 (2013).

10. X. Zheng, *et al.*, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

11. D. Reich, K. Thangaraj, N. Patterson, A. L. Price, L. Singh, Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).

12. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).