

1. Data preparation

New Genotyping Data from 9 Brazilian Populations

We genotyped 48 individuals from 9 Brazilian Native American populations on the Affymetrix Human Origins array. All DNA was extracted from blood. We curated the data as in Lazaridis et al¹. Table S1.1 lists information on these new samples.

Table S1.1. New Human Origins Array genotypes for 48 Brazilian samples.

Sample ID	Sex	Population ID	Region	Language	Lat.	Long.
Apalai147	U	Apalai	Pará	Karib	-54.67	-1.33
Apalai185	M	Apalai	Pará	Karib	-54.67	-1.33
Apalai222	M	Apalai	Pará	Karib	-54.67	-1.33
Apalai228	M	Apalai	Pará	Karib	-54.67	-1.33
Arara1	M	Arara	Pará	Karib	-53.6	-3.9
Arara11	F	Arara	Pará	Karib	-53.6	-3.9
Arara23	M	Arara	Pará	Karib	-53.6	-3.9
Arara49	M	Arara	Pará	Karib	-53.6	-3.9
GuaraniGN5	F	Guarani_GN (Adm)	Mato Grosso do Sul	Tupi	-54.5	-23.33
GuaraniGN28	F	Guarani_GN (Adm)	Mato Grosso do Sul	Tupi	-54.5	-23.33
GuaraniGN405	F	Guarani_GN	Mato Grosso do Sul	Tupi	-54.5	-23.33
GuaraniGN841	F	Guarani_GN (Adm)	Mato Grosso do Sul	Tupi	-54.5	-23.33
GuaraniGN837	F	Guarani_GN (Adm)	Mato Grosso do Sul	Tupi	-54.5	-23.33
GuaraniGN845	F	Guarani_GN	Mato Grosso do Sul	Tupi	-54.5	-23.33
GuaraniGN852	F	Guarani_GN	Mato Grosso do Sul	Tupi	-54.5	-23.33
GuaraniKW203	F	Guarani_KW	Mato Grosso do Sul	Tupi	-55.2	-23.33
GuaraniKW220	F	Guarani_KW	Mato Grosso do Sul	Tupi	-55.2	-23.33
GuaraniKW223	F	Guarani_KW	Mato Grosso do Sul	Tupi	-55.2	-23.33
GuaraniKW224	F	Guarani_KW	Mato Grosso do Sul	Tupi	-55.2	-23.33
GuaraniKW230	F	Guarani_KW	Mato Grosso do Sul	Tupi	-55.2	-23.33
GuaraniKW644	F	Guarani_KW	Mato Grosso do Sul	Tupi	-55.2	-23.33
GuaraniKW645	F	Guarani_KW	Mato Grosso do Sul	Tupi	-55.2	-23.33
GuaraniKW646	M	Guarani_KW	Mato Grosso do Sul	Tupi	-55.2	-23.33
GuaraniKW650	F	Guarani_KW	Mato Grosso do Sul	Tupi	-55.2	-23.33
GuaraniKW626	F	Guarani_KW (Adm)	Mato Grosso do Sul	Tupi	-55.2	-23.33
Karitiana12	M	Karitiana	Rondônia	Tupi	-64.25	-9.33
Karitiana19	M	Karitiana	Rondônia	Tupi	-64.25	-9.33
Karitiana27	M	Karitiana	Rondônia	Tupi	-64.25	-9.33
Karitiana37	M	Karitiana	Rondônia	Tupi	-64.25	-9.33
Surui14	F	Surui	Rondônia	Tupi	-61.17	-10.33
Surui20	F	Surui	Rondônia	Tupi	-61.17	-10.33
Surui72	M	Surui	Rondônia	Tupi	-61.17	-10.33
Surui307	F	Surui	Rondônia	Tupi	-61.17	-10.33
UKaapor95	F	UrubuKaapor	Maranhão	Tupi	-45.22	-2.33
UKaapor150	M	UrubuKaapor	Maranhão	Tupi	-45.22	-2.33
UKaapor164	F	UrubuKaapor	Maranhão	Tupi	-45.22	-2.33
Xavante107	F	Xavante	Mato Grosso	Ge	-52.5	-14.33
Xavante214	U	Xavante	Mato Grosso	Ge	-52.5	-14.33
Xavante1004	M	Xavante	Mato Grosso	Ge	-52.5	-14.33
Xavante1104	F	Xavante	Mato Grosso	Ge	-52.5	-14.33
Xavante1105	F	Xavante	Mato Grosso	Ge	-52.5	-14.33
Xavante1513	F	Xavante	Mato Grosso	Ge	-52.5	-14.33
Xavante2302	M	Xavante	Mato Grosso	Ge	-52.5	-14.33
Xavante2304	M	Xavante	Mato Grosso	Ge	-52.5	-14.33
Xavante401	F	Xavante	Mato Grosso	Ge	-52.5	-14.33
Xavante506	F	Xavante	Mato Grosso	Ge	-52.5	-14.33
Xavante606	F	Xavante	Mato Grosso	Ge	-52.5	-14.33
Zoro51	M	Zoro	Rondônia	Tupi	-60.33	-10.33

Note: Samples with "(Adm)" at the end of their population ID have population genetic evidence of European or African admixture.

The informed consent associated with these samples is not consistent with public posting of data. The data are available to researchers who send a PDF of a signed letter containing the text below to David Reich (reich@genetics.med.harvard.edu).

Box S1.1. Text that needs to be included in a letter to access the new data.

I affirm that

- (a) I will not distribute the data outside my collaboration
- (b) I will not post it publicly
- (c) I will make no attempt to connect the genetic data to personal identifiers
- (d) I will use the data only for studies of population history
- (e) I will not use the data for any commercial purposes

Identification of individuals without post-Columbian admixture

Previous genomic studies have identified substantial amounts of European and African ancestry due to post-Columbian admixture into Native American populations^{2,3}. However, since this admixture process is recent and ongoing there exists substantial intra-population variation in European and African ancestry in most admixed groups. To avoid the confounding factor of this admixture, we restricted analyses of Native Americans previously genotyped on the Affymetrix Human Origins array to individuals with <0.1% cluster membership in both the European- and African-modal components^{1,4} based on the $K=3$ ADMIXTURE analysis of Lazaridis et al¹. We excluded these individuals from subsequent analyses (Table S1.2).

Table S1.2. Identification of admixed Native Americans.

Population	Individuals	Admixed	Not admixed
<i>Published data</i>			
Chipewyan	30	26	4
Cree	13	13	0
Algonquin	9	9	0
Ojibwa	19	19	0
Pima	14	7	7
Mayan	18	17	1
Mixtec	10	10	0
Mixe	10	0	10
Kaqchikel	5	3	2
Wayuu	1	0	1
Cabecar	6	0	6
Piapoco	4	0	4
Inga	2	2	0
Ticuna	1	0	1
Karitiana	12	0	12
Surui	8	0	8
Quechua	7	6	1
Aymara	5	4	1
Chane	1	0	1
Guarani	5	2	3
Chilote	4	4	0
Bolivian	7	5	2
<i>New data</i>			
Apalai	4	0	4
Arara	4	0	4
Guarani_GN	6	4	2
Guarani_KW	11	1	10
Karitiana	4	0	4
Surui	4	0	4
Urubu_Kaapor	3	0	3
Xavante	11	0	11
Zoro	1	0	1
Total	240	133	107

For the newly genotyped Brazilians (Table S1.1) we used ADMIXTURE⁵ to infer cluster memberships. We co-analyzed the samples with other samples previously genotyped on the Human Origins array, after removing SNPs in high linkage disequilibrium using PLINK v1.07 (--indep-pairwise 200 25 0.4). We performed

clustering with ADMIXTURE for 2 to 12 clusters (K), using default parameters. We identified a Native American cluster at $K=3$ and excluded 4 Guarani_GN (Guarani-GN_5, Guarani-GN_28, Guarani-GN_837 and Guarani-GN_841), and 1 Guarani_KW (Guarani-KW_626) (Extended Data Figure 1).

Identification of unadmixed Dakelh

We analyzed data from 20 Athabascan-speaking Dakelh individuals from British Columbia⁶. These samples were genotyped on an Illumina SNP array which has only modest overlap with the Affymetrix Human Origins array, and so these samples had to be curated separately. To exclude individuals with evidence of recent European or East Asian admixture, we merged the samples with a diverse panel of individuals who had been previously genotyped using Illumina arrays⁷, and computed three statistics $D(\text{Yoruba, French; } X, \text{ Karitiana})$, $D(\text{French, Han; Karitiana, } X)$ and $D(\text{Yoruba, Han; } X, \text{ Karitiana})$ for each Dakelh sample X in turn. We interpreted all statistics with $|Z|>3$ as providing significant evidence of post-Columbian admixture. Eleven samples had no evidence of mixture we restricted analyses to these (Table S1.3).

Table S1.3. Identification of unadmixed Dakelh Athabascan-speakers.

	D(Yoruba, French; X, Karitiana)		D(French, Han; Karitiana, X)		D(Yoruba, Han; X, Karitiana)		Status
	D	Z	D	Z	D	Z	
athabaskSV6	0.003	1.10	-0.005	-1.66	-0.002	-0.81	
athabaskCN27	0.005	2.02	-0.014	-4.15	-0.009	-2.81	admixed
athabaskTL1	-0.004	-1.31	0.000	-0.12	-0.004	-1.26	
athabaskHD2	0.029	10.35	-0.072	-18.51	-0.044	-13.04	admixed
athabaskCN42	0.041	14.52	-0.105	-28.31	-0.065	-19.78	admixed
athabaskCA6	-0.005	-2.02	0.010	2.91	0.004	1.39	
athabaskCA12	-0.003	-1.15	0.007	2.36	0.004	1.23	
athabaskCA16	-0.001	-0.20	0.002	0.78	0.002	0.54	
athabaskCA24	0.003	1.03	-0.006	-2.06	-0.004	-1.24	
athabaskCA26	0.025	9.72	-0.064	-18.38	-0.039	-12.33	admixed
athabaskCA85	-0.003	-0.94	0.004	1.18	0.001	0.36	
athabaskCN9	0.013	4.24	-0.029	-8.06	-0.017	-5.36	admixed
athabaskCN40	0.000	0.03	-0.005	-1.61	-0.005	-1.56	
athabaskCA93	0.034	13.64	-0.099	-35.32	-0.065	-22.82	admixed
athabaskCA13	-0.006	-2.12	0.005	1.69	-0.001	-0.43	
athabaskSV3	-0.003	-1.11	0.008	2.64	0.005	1.57	
athabaskCN15	-0.002	-0.62	0.007	2.49	0.006	1.80	
athabaskCN36	0.011	3.57	-0.026	-7.71	-0.016	-4.98	admixed
athabaskHD4	0.048	18.63	-0.127	-37.98	-0.080	-25.46	admixed
athabaskHD3	0.034	12.66	-0.094	-25.09	-0.060	-18.28	admixed

Siberian groups with no evidence of recent back-from-America gene flow

A potential confounding factor for studies of migrations into the Americas is the use of far eastern Siberian populations who derive some ancestry from back-from-America gene flow². If the Native American ancestry in these Siberian populations is not symmetrically related to the Native American groups being studied, it has the potential to generate artifactual signals of distinct migrations into the Americas². In order to avoid using Siberian groups with such gene flow, we used ALDER⁸ to investigate whether there is evidence of mixture related to Native Americans in each Central Asian Siberian group in turn for which there is published Human Origins genotyping data. Specifically, we used the test for mixture implemented in ALDER, which tests for linkage disequilibrium (LD) that is correlated to the allele frequency

differences between two populations that are proposed to be related to the admixing populations (here we use Han Chinese and Mixe Native Americans)⁸. We only consider populations with a 2-reference population Z-score of <3 for inclusion in the *qpWave* analysis in SI 2.

Table S1.4. ALDER test for admixture LD in Central Asians / Siberians.

Mixe and Han Chinese are used as source populations.

Test population	2-ref z-score	1-ref Z for Mixe	1-ref Z for Han	2-ref decay
Altaiian	2.65	0	1.1	49 ± 14
Chukchi	2.1	3.49	7.83	28 ± 13
Eskimo	2.97	1.17	3.11	83 ± 18
Even	1.3	1.37	2.69	8 ± 6
Itelmen	1.38	0.39	0.17	84 ± 55
Kalmyk	2.5	2.17	3.99	42 ± 17
Koryak	2.53	1.7	1.63	51 ± 18
Kyrgyz	3.54	0	2.85	26 ± 7
Mansi	0.11	0.19	2.69	29 ± 262
Mongola	1.17	1.02	0	23 ± 20
Nganasan	0	1.16	3.29	n/a
Selkup	2.91	3.21	1.39	8 ± 3
Tajik_Pomiri	1.75	2.38	4.97	186 ± 66
Tubalar	3.55	2.17	9.7	33 ± 9
Turkmen	2.44	1.71	1.66	101 ± 41
Tuvinian	2.09	0	1.14	29 ± 13
Ulchi	1.35	2.09	4.03	167 ± 88
Uzbek	0.57	2.51	5.78	6 ± 11
Yakut	1.13	4.52	5.52	17 ± 15
Yukagir	3.01	2.66	4.44	9 ± 3

2. Test of the number of Native American founding populations

To investigate whether all Native American groups from Central and Southern America^{1,4} are consistent with being derived from a single stream of ancestry, we applied *qpWave*² to ask the question whether the set of f_4 -statistics of the form $f_4(\text{American}_1, \text{American}_2; \text{Outgroup}_1, \text{Outgroup}_2)$ forms a matrix that is consistent with being of rank 0. Intuitively, if all these Native American populations descend from the same stream of migration into the Americas, then all these statistics should be consistent with 0. We test for deviations from this null hypothesis on all the f_4 -statistics jointly. We compute a single P -value that appropriately corrects for the correlation structure of the statistics using a Hotelling t -test.

The Outgroups in our analysis were 4 populations from each of 6 worldwide regions.

Africa: Yoruba, Ju_hoan_North, Dinka, Mbuti

Siberia/Central Asia: Tuvinian, Mongola, Yakut, Kyrgyz

East Asia: Han, Uygur, Japanese, Ami

Oceania: New_Guinea, Papuan, Australian_WGA, Mamanwa

South Asia: Kusunda, Onge, Kharia, Sindhi

West Eurasia: Orcadian, Spanish, Sardinian, Chechen

For Native American, we restrict to 7 groups with at least 3 unadmixed individuals.

Native Americans: Cabecar, Guarani, Karitiana, Mixe, Piapoco, Pima, Surui

We used *qpWave* to perform likelihood ratio tests for whether the matrix of statistics $f_4(\text{American}_1, \text{American}_2; \text{Outgroup}_1, \text{Outgroup}_2)$ is consistent with rank 0, 1, 2, or 3, which corresponds to 1, 2, 3 or 4 ancestral populations being required to explain the data. We find that 1 ancestral population was rejected ($P = 2 \times 10^{-7}$), but 2 or more were consistent with the data ($P \geq 0.09$) (Extended Data Table 1). The rejection of a single ancestral population is robust to dropping each of the 6 large geographic regions from which the Outgroups were derived ($P < 10^{-3}$), as well as dropping each of the 7 Native American populations ($P < 0.01$) (Extended Data Table 1).

The evidence for two ancestral populations is driven by Amazonian and Australasian populations

To determine which outgroup and Native American populations contributed the most to the rejection of rank 0, we examined the weight coefficients assigned to each population by *qpWave*. We find that the greatest coefficient among the outgroups is assigned to the Onge from the Andaman Islands. The next strongest coefficients are assigned to South Asians and Oceanians. The greatest coefficients among Native Americans are assigned to the Amazonian Suruí and Karitiana, whereas the lowest are assigned to the Central American Mixe and Pima (Extended Data Figure 2).

An alternative approach to assessing which outgroup populations contribute the most to the rejection of rank 0 is to test different subsets of the 24 outgroup populations separately. We tested different sets of 8 populations where we either paired the 4 African populations or 4 Central Asian/Siberian populations with 4 populations from another region. We find that the lowest P -values are found for pairings with South Asia and Oceania (Extended Data Table 1). This is contrary to what would be expected if the detected signal is due to recent gene flow between the Americas and Siberia, in which case we would expect to see the lowest P -values for Siberian populations. It is also contrary to what would be expected if the signal is due to cryptic post-Columbian European or African ancestry, in which case we might expect to find the lowest P -values for those populations.

3. Allele frequency symmetry tests

In SI 2 we found that genome-wide data from Central and Southern Native American populations were inconsistent with a single ancestral population, and that the major inconsistencies were driven by asymmetrical affinities to South Asians and Oceanians. Within the Americas, our analyses suggested that major differences could

be found between the Amazonian populations Suruí and Karitiana on the one hand, and Central American populations such as Pima and Mixe on the other.

To investigate whether there is an asymmetrical relationship to Non-American populations between Central Americans and Amazonians, we computed f_4 -statistics of the form $f_4(\text{Chimpanzee, Non-American; Central Americans, Amazonians})$ where Central Americans consisted of Pima, Mixe and Kaqchikel (see SI 1) and Amazonians consisted of the Karitiana and Suruí. The f_4 -statistic⁹ is computed as the product

$$f_4(A, B; X, Y) = (p_A - p_B)(p_X - p_Y)$$

averaged over all SNPs, where p_A , p_B , p_X , and p_Y are the frequencies of an arbitrarily chosen allele in populations A , B , X and Y at each locus. Another way to think of the f_4 -statistic is as the numerator of the D -statistic¹⁰. In practice, f_4 -statistics and D -statistics give qualitatively indistinguishable results for tests for consistency with zero¹⁰. Intuitively, the f_4 -statistic can be thought of as testing whether allele frequency differences between A and B are correlated with those between X and Y , an observation that is not consistent with a tree-like model where X and Y are from a single ancestral population. In some instances we also compute the D -statistic

$$D(A, B; X, Y) = \frac{(p_A - p_B)(p_X - p_Y)}{(p_A + p_B - 2p_A p_B)(p_X + p_Y - 2p_X p_Y)}$$

with the corresponding notation. We compute standard errors (SEs) for all statistics using a Block Jackknife weighted by the number of SNPs in each 5 cM block in the genome^{9,11}. We report Z -scores based on the ratio of D/SE . We interpret statistics $|Z| > 3$ as being significantly different from 0.

A genetic affinity between native Amazonians and native Oceanians

We find evidence for a significant excess in shared derived allele frequencies between Amazonians and five populations: Onge, Papuans, New Guinean highlanders, Australians from Arnhem Land, and Mamanwa Negritos from the Philippines ($Z > 3$). In addition, we find positive statistics for other Oceanian populations such as Bougainville Papuans, Tongans, and Ami from Taiwan, and some Indian populations such as Kharia and Bengali (Extended Data Table 2).

Robustness to different Non-American outgroups

We tested the consistency of this signal for other outgroups than Chimpanzee, replacing A in the test $f_4(A, B; \text{Mixe, Surui})$ with one of Chimpanzee, Mbuti and Biaka pygmies from Central Africa, Yoruba from West Africa, Dinka from East Africa, Ju_hoan_North (San) from southern Africa, Yakut and Yukagir from Siberia, and Han Chinese. We also varied B among Onge, Papuans, New Guineans, and Australians. We find that the signal is highly consistent for these different combinations. In many cases it is even stronger for the non-Chimpanzee outgroups (Table S3.2).

Table S3.2. Significant statistics of the form $f_4(A, B; \text{Mixe, Surui})$.

<i>A</i>	<i>B</i>	$f_4(A, B; \text{Mixe, Surui})$	<i>Z</i>
Chimpanzee	Onge	0.00100490	4.01
Chimpanzee	Papuan	0.00083341	3.29
Chimpanzee	New_Guinea	0.00083884	3.17
Chimpanzee	Australian_WGA	0.00085497	3.23
Mbuti	Onge	0.00088618	4.06
Mbuti	Papuan	0.00071469	3.30
Mbuti	New_Guinea	0.00072013	3.11
Mbuti	Australian_WGA	0.00073710	3.19
Biaka	Onge	0.00091768	4.36
Biaka	Papuan	0.00074620	3.56
Biaka	New_Guinea	0.00075163	3.33
Biaka	Australian_WGA	0.00076856	3.41
Yoruba	Onge	0.00085923	4.21
Yoruba	Papuan	0.00068774	3.37
Yoruba	New_Guinea	0.00069317	3.13
Yoruba	Australian_WGA	0.00071020	3.28
Dinka	Onge	0.00104014	4.90
Dinka	Papuan	0.00086865	4.14
Dinka	New_Guinea	0.00087409	3.88
Dinka	Australian_WGA	0.00089104	4.01
Ju_hoan_North	Onge	0.00102585	4.79
Ju_hoan_North	Papuan	0.00085436	3.94
Ju_hoan_North	New_Guinea	0.00085979	3.69
Ju_hoan_North	Australian_WGA	0.00087661	3.81
Yakut	Onge	0.00078838	3.95
Yakut	Papuan	0.00061689	3.21
Yakut	New_Guinea	0.00062233	2.92
Yakut	Australian_WGA	0.00063896	2.91
Han	Onge	0.00079784	4.10
Han	Papuan	0.00062635	3.34
Han	New_Guinea	0.00063179	2.95
Han	Australian_WGA	0.00064859	2.99
Yukagir	Onge	0.00090198	4.64
Yukagir	Papuan	0.00073050	3.87
Yukagir	New_Guinea	0.00073593	3.50
Yukagir	Australian_WGA	0.00075238	3.51

Robustness to different Native American contrasts

We tested different combinations of Central American and Amazonian populations, and found consistent results for all Central Americans (Kaqchikel, Mixe and Pima), as well as for the Amazonian Karitiana and Suruí. Karitiana shows a non-significantly attenuated signal compared to the Suruí (approximately 1 SE difference) (Table S3.3).

We tested each individual of the 8 Suruí and 10 Mixe against each other in the test $D(\text{Chimpanzee, Onge; individual 1, individual 2})$. We find that comparisons between

two Suruí individuals are generally consistent with 0, whereas all $8 \times 2 = 16$ comparisons between Mixe and Suruí individuals show a positive skew. Comparisons between Mixe individuals are more dispersed, but do not show a systematic pattern. This suggests that the affinity to Onge in the Suruí is not due to sequence errors or unusual ancestry in some individuals (Extended Data Figure 3).

Table S3.3. Different contrasts between Central Americans and Amazonians

A	B	X	Y	$f_4(A, B; X, Y)$	Z
Chimpanzee	Onge	Mixe	Surui	0.001005	4.01
Chimpanzee	Onge	Kaqchikel	Surui	0.001156	3.90
Chimpanzee	Onge	Pima	Surui	0.001089	4.01
Chimpanzee	Onge	Mixe	Karitiana	0.000631	2.79
Chimpanzee	Onge	Kaqchikel	Karitiana	0.000784	2.76
Chimpanzee	Onge	Pima	Karitiana	0.000716	2.78

We analyzed Native American individuals from other regions that have been genotyped on the Human Origins array, as well as individuals genotyped on Illumina arrays². The study that reported the Illumina array data masked genomic segments of post-Columbian European or African ancestry, allowing us to carry out analyses including individuals with post-Colombian mixture (and analyses restricting to individuals without any evidence of post-Columbian mixture in their genome).

We computed $f_4(\text{Yoruba, Papuan; Mixe, Surui})$ in these 3 data sets, choosing Papuans as the reference Australasian population since Illumina data for the Onge have not previously been published. We find no clear evidence ($|Z| > 3$) for any population beyond the Karitiana and Suruí as having excess affinity to Papuans in the unadmixed individuals on the Illumina array. In the ancestry masked version of the Illumina data, the Maya2 show a significant signal. While these results are intriguing, we choose not to draw strong conclusions based on the masked data, since it is possible that the local ancestry masking could affect a fine-scale signature such as the affinity between some Native Americans and Australasians that we are investigating.

Table S3.4. Generalization of findings to more populations and different datasets

Affinity of Papuans to Native Americans identified as entirely of First American ancestry by Reich et al. (2012) and genotyped on Illumina and Affymetrix arrays. We present results for the statistic $f_4(\text{Yoruba, Papuan, Mixe, Test})$.

	Illumina masked		Illumina unadmixed		Human Origins unadmixed	
	f_4	Z	f_4	Z	f_4	Z
Cree	-0.0032	-1.07
Algonquin	-0.0050	-1.72
Ojibwa	-0.0050	-1.89
Pima	-0.0007	-0.41	-0.0008	-0.38	0.0001	0.31
Yaqui	0.0111	2.54
Tepehuano	0.0011	0.75	0.0000	-0.02
Maya1	0.0035	2.61	0.0051	1.62
Maya2	0.0066	4.06

Purepecha	-0.0070	-1.84
Zapotec2	0.0020	1.38	0.0000	0.01
Mixtec	0.0030	1.58
Zapotec1	0.0039	2.91	0.0059	2.3
Kaqchikel	0.0027	1.72	0.0061	1.91
Wayuu	0.0028	1.57	0.0049	2.06	0.0031	0.92
Kogi	0.0010	0.37	0.0014	0.53
Arhuaco	-0.0043	-1.53
Maleku	0.0008	0.27	0.0011	0.35
Chorotega	-0.0011	-0.24
Huetar	0.0024	0.53
Cabecar	0.0044	2.22	0.0046	2.13	0.0040	1.6
Bribri	0.0052	2.15	0.0064	2.29
Teribe	0.0063	2.39	0.0072	2.5
Guaymi	0.0047	1.94	0.0043	1.69
Embera	-0.0013	-0.59	-0.0009	-0.4
Guahibo	-0.0004	-0.2	-0.0001	-0.03
Waunana	0.0015	0.63	0.0019	0.79
Palikur	0.0019	0.75	0.0023	0.81
Piapoco	-0.0011	-0.56	-0.0005	-0.26	0.0002	0.1
Inga	0.0023	1.18
Ticuna	0.0023	0.99	0.0026	1.07	0.0076	2.14
Arara	0.0027	0.74	0.0027	0.72
Parakana	0.0055	1.67	0.0055	1.67
Jamamadi	0.0005	0.13	0.0008	0.23
Karitiana	0.0036	1.63	0.0041	1.77	0.0054	2.38
Surui	0.0075	3.15	0.0080	3.19	0.0087	3.36
Quechua	0.0019	1.29	0.0004	0.12	-0.0060	-1.83
Aymara	0.0018	1.18	0.0021	0.98	-0.0005	-0.15
Chane	0.0028	1.11	0.0018	0.68	0.0046	1.46
Wichi	0.0002	0.07	0.0018	0.74
Guarani	-0.0002	-0.12	-0.0021	-0.93	0.0001	0.03
Kaingang	0.0008	0.27
Toba	0.0028	1.41	0.0021	0.87
Diaguita	0.0004	0.16
Hulliche	0.0030	1.33
Chilote	0.0025	1.01
Chono	-0.0022	-0.74
Yaghan	0.0010	0.39	0.0013	0.37

Comparison with the Athabaskan Dakelh

The three northern North American groups in the ancestry masked data showed somewhat negative statistics in Table S3.4, which could be consistent with an excess of Papuan affinity not only in South Americans but also in Central Americans, and the possibility that northern North Americans might form a better baseline for detecting excess affinities to Oceanians. However, no Northern Amerind individuals in this data set were identified as unadmixed. We therefore analyzed data from Athabaskan-speaking Dakelh from Raghavan et al⁶, and used the unadmixed individuals identified in SI 1. We find no significant evidence of greater affinity to Papuans in the Pima compared to the Dakelh (Table S3.5). However we do observe $Z \sim 2.1$ for an affinity between the Dakelh and Yakut, which hints at an affinity to Siberians such as the Yakut in the Dakelh, who like the Chipewyan are also Athabaskan-speakers. An affinity to Siberians would also cause an attraction between the Dakelh and Papuans, since Papuans share more genetic drift with Siberians than with Yoruba. This would

tend to diminish any signals of difference in Papuan-relatedness between the Pima and Dakelh. In summary, while we do not find any clear evidence for a difference in First American ancestry proportions in the Dakelh compared with more southern Native Americans, we also cannot exclude the possibility that northern North Americans (not Dakelh) have a lower baseline affinity to Australasians than do Central Americans.

Table S3.5. No evidence for differences in Papuan-related affinity between unadmixed British Columbian Dakelh and Pima from Central America.

<i>A</i>	<i>B</i>	<i>X</i>	<i>Y</i>	<i>D</i>	<i>Z</i>
Yoruba	Papuans	Dakelh	Surui	0.0101	3.82
Yoruba	Papuans	Dakelh	Pima	0.0034	1.50
Yoruba	Papuans	Pima	Surui	0.0071	2.35
Yoruba	Yakut	Dakelh	Surui	-0.0018	-0.88
Yoruba	Yakut	Dakelh	Pima	-0.0040	-2.06
Yoruba	Yakut	Pima	Surui	0.0023	1.02

Comparison with the ancient Clovis-associated Anzick individual

We investigated the evidence for differential relatedness to Australasians using ancient Native American samples overlapped with San and Yoruba ascertained SNPs on the Human Origins array. We use Dinka as outgroup, as they are less differentiated from non-Africans than Yoruba are. This minimizes the effects of errors in the ancient DNA causing an attraction to the outgroup. We computed $f_4(\text{Dinka}, \text{Onge/New_Guinea/Papuan}; \text{Ancient}, \text{Mixe/Surui})$. We find no evidence for a greater affinity to Onge/New_Guinea/Papuan in any ancient sample than is found in the Mixe. In contrast, we find that the Suruí show an affinity to Onge/New_Guinea/Papuan compared to the ancient Anzick individual (Table S3.6).

Table S3.6. Comparison with the Anzick Clovis sample.

Z-scores greater than 2 are highlighted and only observed for comparisons with Suruí.

<i>A</i>	<i>B</i>	<i>X</i>	<i>Y</i>	$f_4(\text{A}, \text{B}; \text{X}, \text{Y})$	<i>Z</i>	SNPs
Dinka	Onge	Clovis	Surui	0.000933	2.64	349,435
Dinka	New Guinea	Clovis	Surui	0.000787	2.22	349,435
Dinka	Papuan	Clovis	Surui	0.000826	2.51	349,435
Dinka	Onge	Clovis	Mixe	-0.000032	-0.10	349,435
Dinka	New Guinea	Clovis	Mixe	-0.000108	-0.33	349,435
Dinka	Papuan	Clovis	Mixe	-0.000036	-0.12	349,435

Comparisons between different Australasian groups

We tested whether there is evidence for either Oceanians or Andamanese being significantly more strongly related to Amazonians using direct contrasts of the form $D(\text{A}, \text{B}; \text{Mixe}, \text{Surui})$. We found that the Philippine Mamanwa show less affinity than the other populations (many $|Z| > 2$), consistent with their known Austronesian admixture¹². There are no consistent patterns for the other populations (Table S3.7).

Table S3.7. Direct comparisons between Andamanese and other Oceanians.

<i>A</i>	<i>B</i>	<i>X</i>	<i>Y</i>	<i>D(A, B; X, Y)</i>	<i>Z</i>
Onge	Papuan	Mixe	Surui	-0.0023	-0.85
Australian	Onge	Mixe	Surui	0.0069	2.20
Mamanwa	Onge	Mixe	Surui	0.0069	2.76
Australian_WGA	Onge	Mixe	Surui	0.0020	0.67
New_Guinea	Onge	Mixe	Surui	0.0022	0.77
Australian	Papuan	Mixe	Surui	0.0052	2.06
Mamanwa	Papuan	Mixe	Surui	0.0048	2.00
Australian_WGA	Papuan	Mixe	Surui	-0.0003	-0.15
New_Guinea	Papuan	Mixe	Surui	-0.0001	-0.08
Australian	Mamanwa	Mixe	Surui	-0.0002	-0.07
Australian	Australian_WGA	Mixe	Surui	0.0059	2.33
Australian	New_Guinea	Mixe	Surui	0.0055	1.95
Australian_WGA	Mamanwa	Mixe	Surui	-0.0050	-1.83
Mamanwa	New_Guinea	Mixe	Surui	0.0048	1.78
Australian_WGA	New_Guinea	Mixe	Surui	-0.0003	-0.11

Consistency of the link between Australasians and Amazonians in different data

We assessed whether the affinity of Amazonians to Australasians was robust in Human Origins data as well entirely genetic data sets generated using different technological platforms. We use $f_4(\text{Yoruba, Papuan; Mixe, Surui})$ throughout since these four populations are present in each of the datasets. In Table S3.8 we list this statistic for Illumina SNP arrays, Affymetrix Human Origins (HO) SNP arrays, and Illumina high-coverage sequencing. For all these data we find evidence of significant ($Z > 3$) affinity between the Suruí and Papuans.

Table S3.8. $f_4(\text{Yoruba, Papuan; Mixe, Surui})$ in different data sets. n_Y , n_P , n_M and n_S refer to the sample size of Yoruba, Papuans, Mixe, and Suruí in each data set.

	f_4	<i>Z</i>	SNPs	n_Y	n_P	n_M	n_S
Illumina masked	0.000776	3.15	364,428	21	16	17	24
Illumina unadmixed	0.000823	3.21	364,470	21	16	9	24
SGDP genomes	0.001265	4.00	9,873,045	3	16	3	2
Affymetrix HO	0.000688	3.37	593,142	70	26	10	8

Consistency of the signal for different mutation classes

We stratified the *D*-statistics for the complete genome data into each separate nucleotide substitution class, and found highly consistent results for all 6 classes (within 2 SEs) (Table S3.9). This suggests that the results are not due to convergent evolutionary processes, for example, correlations in the effectiveness of gene conversion on the lineages leading to Suruí and Papuans.

Table S3.9. Stratification of D (Yoruba, Papuan; Mixe, Surui) by mutation class.

These analyses are performed on the full genome sequencing data.

Mutation class	D	Z	Informative SNPs
A / T	0.0169	2.63	60,538
A / G	0.0191	3.64	268,962
A / C	0.0208	3.49	67,210
G / T	0.0248	4.27	67,623
C / T	0.0220	4.24	270,133
C / G	0.0248	4.26	64,951

Consistency of the signal of relatedness between Australasians and Amazonians for different ascertainment schemes

We studied all 13 different ascertainment panels that underlie the full Human Origins array SNP set. Some of these panels contain relatively few SNPs. However, we find strong statistics for ascertainments as different as Han Chinese and Yoruba, and no evidence for any one panel contributing disproportionately to the signal (Table S3.10).

Table S3.10. An affinity between Suruí and Papuans is seen across multiple ascertainment schemes of the Human Origins array. Rows are ordered by the number of SNPs analyzed (more SNPs gives more precision).

Ascertainment individual	f_i (Yoruba, Papuan; Mixe, Surui)	Z	SNPs (autosomal)
Union of 13 panels	0.000688	3.37	593,142
San	0.000525	2.24	156,365
Denisova-San differences	0.000611	2.81	140,044
French	0.000478	1.13	108,311
Yoruba	0.000894	3.16	119,765
Han	0.001587	3.27	75,390
Papuan1	0.001009	2.14	46,676
Cambodian	0.00026	0.40	16,442
Melanesian	0.001247	1.82	14,449
Sardinian	0.000843	1.16	12,470
Mbuti	0.001186	2.34	11,745
Papuan2	0.001753	2.30	11,720
Mongolian	-0.000546	-0.65	10,389
Karitiana	0.000003	0.00	2,555

Admixture signals as a function of proximity to functionally important regions

To further characterize the excess genomic affinity between Amazonians and southeast Asian populations, we divided the genome into 10 deciles of the ‘ B -value’ proposed by McVicker et al.¹³ which integrates multiple genomic annotations into a single score for functional importance for each base pair in the genome. We computed D (Yoruba, Papuan; Mixe, Surui) separately for each bin. We observe that the bin that is closest to functionally important elements shows the most asymmetry ($D = 0.0397 \pm 0.0197$) and than the bin that is most distant from functional elements shows the least ($D = 0.00837 \pm 0.0106$) (Extended Data Figure 4A). We fit a linear regression to D as a function of B , and estimate the slope to be -0.0004.

To compute standard errors, we used a weighted Block Jackknife procedure where one 5 Mb block of the genome is dropped in turn and the linear regression is recomputed. The variability of this statistic can be used to obtain an estimate of the standard error^{9,11}, which we weighted using the number of informative loci in each block. We estimate the Z -score for the linear coefficient being different from zero to be -1.95 (one-tailed P -value = 0.026). To test if this observation is independent from the observation of a genome-wide significant D -statistic, we implemented a two-dimensional jackknife for both the linear coefficient and the D -statistic.

To understand the expected behavior of the correlation between B and D for known signals of asymmetry, we examined the relationship between B and D for a larger number of population comparisons. For this analysis, we used previously published genomes from Yoruba, San, Mbuti, Han, Dai, Sardinians, French, Australians, Papuans, Karitiana and Mixe^{14,15}, together with the 18 newly reported genomes from Yoruba, Papuans, Mixe and Suruí reported in this study. We computed all possible linear coefficients for D as a function of B for all 495 quartets (Yoruba, X; Y, Z). We found that all quartets of populations for which the Z -score for the slope was significant ($|Z| > 3$) also showed a significant genome-wide D -statistic with the opposite sign (Extended Data Figure 4B).

We hypothesize that the explanation for this phenomenon is that allele frequencies in isolated populations become more differentiated in the vicinity of functionally important regions due to linked selection imposing increased evolutionary stochasticity. This emulates genetic drift (or a low effective population size), which makes admixture between more differentiated populations easier to detect in the sense that the magnitude of allele frequency skews that are used by statistics such as D is increased. This suggests that genome-wide signals of admixture in modern humans are systematically stronger near functional regions, but does not imply that ancestry itself is systematically depleted in these regions.

Having validated this test, we applied it to our signal of interest. This yielded a P -value of 0.00014 , which is in fact less significant (larger P -value) than the genome-wide D -statistic alone. This suggests that the monotonic relationship between D and B does not provide any statistical evidence for admixture above and beyond that of the genome-wide D -statistic alone. Thus, the value of these results is not to increase the strength of the signal, but rather to show that the direction of the signal is what is expected for a real biological effect.

Analysis of 9 newly genotyped populations from Brazil

We also performed allele frequency based symmetry specifically using the 9 newly genotyped Brazilian populations. We find that the new Suruí and Karitiana data both show strong signals with African Yoruba as outgroup. When we use Han Chinese as outgroup, the Xavante also show a strong signal ($Z = 3.25$) and all Amazonians except the Urubu_Kaapor show moderate signals ($Z > 2$) (Table S3.11).

Table S3.11. Affinity to Onge in 9 newly genotyped Brazilian populations

<i>X</i>	D(Yoruba, Onge; Mixe, X)		D(Han, Onge; Mixe, X)	
	<i>D</i>	<i>Z</i>	<i>D</i>	<i>Z</i>
Apalai	0.0037	1.55	0.0066	2.92
Arara	0.0043	1.49	0.0072	2.64
Guarani_GN	0.0021	0.74	0.0078	2.96
Guarani_KW	0.0032	1.35	0.0060	2.66
Karitiana	0.0082	3.10	0.0107	4.14
Suruí	0.0092	3.27	0.0084	3.01
Urubu_Kaapor	0.0042	1.49	0.0037	1.34
Xavante	0.0044	1.98	0.0071	3.25
Zoro	0.0052	1.37	0.0099	3.24

We find no significant difference in affinity to the Onge comparing the new Karitiana and Suruí genotypes obtained from blood samples to those obtained from Human Genome Diversity Project (HGDP) cell lines. This is an important observation, as it shows that our signal cannot be a cell line artifact (Table S3.12).

Table S3.12. The signal is not a cell line artifact

<i>A</i>	<i>B</i>	<i>X</i>	<i>Y</i>	<i>D(A, B; X, Y)</i>	<i>Z</i>
Yoruba	Onge	Suruí _{blood}	Suruí _{cell_line}	-0.00209	-0.92
Yoruba	Onge	Karitiana _{blood}	Karitiana _{cell_line}	0.00039	0.21
Han	Onge	Suruí _{blood}	Suruí _{cell_line}	-0.00242	-1.11
Han	Onge	Karitiana _{blood}	Karitiana _{cell_line}	0.00293	1.66

4. Linkage disequilibrium symmetry tests

We devised a novel linkage disequilibrium statistic that measures symmetry in linkage disequilibrium between two proposed clades with a pair of populations in each. The statistic, which we refer to as h_4 , is:

$$h_4 = ((p_{12}^A - p_1^A p_2^A) - (p_{12}^B - p_1^B p_2^B)) \times ((p_{12}^C - p_1^C p_2^C) - (p_{12}^D - p_1^D p_2^D))$$

where 1 and 2 are arbitrarily chosen reference alleles at two different loci, respectively, and A , B , C , and D denote four different populations. Thus, p_{12}^A is the frequency of the 12 haplotype in population A , and p_1^A is the frequency of the 1 allele in population A . The quantity $H_{12}^A = p_{12}^A - p_1^A p_2^A$ measures the difference between the observed haplotype frequency and the expected haplotype frequency given the allele frequencies¹⁶, and corresponds to the classical population genetic quantity “ D ” (this should not be confused with the D -statistic used to test for consistency with a tree elsewhere in this study).

The motivation for this statistic being informative about population history is that under a tree-like model $((A, B), (C, D))$, differences in linkage disequilibrium between populations C and D are not expected to be correlated to differences in LD between populations A and B . If there has been gene flow between the two clades however, the statistic may be significantly positive or negative, much like f_4 and D -statistics⁴.

In practice, we compute this statistic for each polymorphic locus ('target locus') by identifying all other polymorphic loci 5' of the target locus in a window extending w cM from a focal point at distance interval d . We average the statistic over all valid pairs of loci in the genome identified in this way. We compute standard errors by a Block Jackknife over contiguous 0.5 cM blocks, where SNP pairs that bridge the boundary of two blocks are assigned to the block in which the target locus is found.

Test simulation

To test the h_4 statistic, we simulated a history modified from Lipson et al¹⁷. We simulated 500 chromosomes of 50 kb each. At $0.02 \times 4N_e$ generations ago, 60% of the ancestry of population 6 was contributed by population 5 and the remainder by population 3. We ascertained polymorphisms in 'pop1', made up of 2 chromosomes from pop2 that were not used in h_4 computations. The *ms* command line was

'Base model':

```
ms 254 500 -t 500 -r 500 50000 -I 7 2 50 50 50 50 2 -ej 0.0 2 1 -es 0.02 6 0.4 -ej
0.06 6 3 -ej 0.04 8 5 -ej 0.08 5 4 -ej 0.12 4 3 -ej 0.2 3 1 -ej 0.3 1 7 -en 0.3 7 1
```

We also simulated a bottleneck scenario where the effective population size of population 5 was reduced by a quarter. The following command was added to the above command line: '-en 0.0 5 0.25 -en 0.04 5 1'

Table S4.1. Simulation results from applying the h_4 (pop2, pop3; pop4, pop5) statistic for SNP pairs within a distance of 10,000 bp. The numbers are Z-scores.

Model	NO admixture		Admixture	
	phased	unphased	phased	unphased
Basic model	0.185	0.354	9.475	5.689
Bottleneck in pop5	-1.459	-1.739	7.124	3.368

We computed the h_4 -statistic for all SNP pairs within 10,000 base pairs of each other both for the perfectly phased simulated data as well as for versions where the phase was randomized in pairs of 2 chromosomes (to mimic unphased human SNP data). The h_4 -statistic finds significant evidence of gene flow when testing populations that were involved in the admixture event, with the signal still present (albeit weakened) when phase information is scrambled (Table S4.1). As we would hope for a useful test, sets of populations that are related according to a simple tree give non-significant statistics.

Application to data from the Human Origins array

We phased individuals typed on the Human Origins array panel 5 (119,765 autosomal SNPs ascertained in a Yoruba individual) using SHAPEIT with default parameters (SI 1). We then computed $h_4(\text{Yoruba}, B; X, Y)$ with B , X and Y drawn from the populations: Australian, Dai, French, Han, Ju_hoan_North, Karitiana, Mbuti, Mixe, Onge, Papuan, Pima, Sardinian, Suruí, and Yakut. We find that D and h_4 are largely highly correlated, but that there are also a few cases where they give conflicting significant results (Extended Data Figure 5C). When all Africans and Oceanians (Mbuti, San, Papuan, Australian) are removed, there are no such conflicts (Extended Data Figure 5D). One possible explanation is that these 4 populations are the only ones that harbor ancestry that is basal to Yoruba (which we have used to ascertain SNPs). The h_4 -test thus does not seem to be appropriate when populations used in the test branched off prior to the population(s) used in SNP ascertainment. However the significant h_4 -statistic we observe in the Onge as described below cannot be explained by this bias.

We computed h_4 -statistics of the form $h_4(\text{Yoruba}, X; \text{Mixe}, \text{Surui})$ for all populations X in the Human Origins array, and all pairs of SNPs within 0.01 cM of each other. We restrict the analysis to populations with at least 6 individuals. We find support for an affinity between Oceanians and Onge to Suruí, with larger Z-scores than is the case for the standard D -statistic based on allele frequencies. Four non-African populations, including the Onge and native New Guineans and Papuans, show $Z > 3$ (Table S4.2). We also computed the h_4 -statistic for windows of 0.001 cM centered around different genetic distances for selected populations (Extended Data Figure 5E). We find that the signal dissipates at approximately 0.02 cM.

Table S4.2. Significant statistics for $h_4(\text{Yoruba}, \text{non-African}; \text{Mixe}, \text{Surui})$.

	Population	h_4	SE	Z	Loci	N	Region
1	New_Guinea	0.0004195	7.3E-05	5.71	14938	38	Oceania
2	Onge	0.0003474	7.8E-05	4.43	14938	22	S. Asia
3	Papuan	0.0003193	6.9E-05	4.60	14938	52	Oceania
4	Bougainville	0.0002452	7.2E-05	3.43	14938	20	Oceania

The most negative statistics in non-Americans are found for Northeast Asians and Siberians (*e.g.* $Z = -2.6$ for Evens), which would be expected if the other founding population of the Americas (the population without the strong affinity to Austroriparians) was related to present-day Siberians (Figure 1B). This is as expected for conventional models for the ancestry of First Americans.

Caveats

One qualitative feature that differs between f_4 -statistics and h_4 -statistics is that h_4 -statistics do not have the Martingale property in relation to genetic drift. Specifically, the h_4 -statistic can be biased by different degrees of genetic drift since divergence.

For example, for $h_4(A, B; X, Y)$, if Y has more SNP pairs without polymorphism than X , this will in turn result in more SNP pairs with $H_{12}^Y = 0$ than $H_{12}^X = 0$. Thus we might expect that if B also has many SNP pairs with $H_{12}^B = 0$, we could see false-positive genome-wide h_4 -statistics indicating affinity between B and Y . To assess the impact of this in our empirical data, we estimated the fraction of SNP pairs with $H_{12} = 0$ in all populations. Extended Data Figure 5A shows the fraction of SNP pairs with $H_{12} = 0$ on the x-axis and $h_4(\text{Yoruba, Test; Mixe, Surui})$ on the y-axis. Papuans, New Guineans and Onge are clearly outliers in that they show a strong affinity to Suruí, like the Amazonian Karitiana. We also see significant statistics for Africans, but we note that we have ascertained in Yoruba for this analysis, and so it is not valid to use populations that may have ancestry basal to Yoruba for the purpose of this test since it breaks the assumption of polymorphism in the ancestral population. We see that there are several East Asian and Siberian populations that have similar fractions of SNP pairs with $H_{12} = 0$ as the populations with significant h_4 -statistics. However, these East Asians and Siberians do not show significant evidence of attraction to the Suruí. This suggests that the fraction of SNP pairs with $D = 0$ is not severely impacting the significant h_4 -statistics that we are detecting.

5. Chromosome painting symmetry tests

We used SHAPEIT to phase 593,142 SNPs with the same set of individuals as described above along with the 48 newly genotyped Brazilian individuals, using all panels of the Human Origins array. We ‘painted’ the chromosomes of unadmixed Native American individuals using non-American populations as donors, but excluded the Yukagir and the Chukchi since they have evidence of back-migration from the Americas. We ran CHROMOPAINTER v2 using default parameters, painting each recipient individual separately, but using all donor populations as candidates to paint each recipient haplotype. To assess statistical uncertainty, we repeated this procedure for each recipient individual using 22 subsets of the data where for each separate subset a different chromosome had been dropped. We then used the results of these 22 block jackknife pseudoreplicates to obtain a weighted Block Jackknife estimate of the standard error for our test statistic (see below).

To test if the recipient populations copied equally from the donor populations, we computed the average chunk count $C_{R:D}$ copied from a given donor population D in each recipient population R (averaged over individuals). We then computed a $S(R_1, R_2; D)$ statistic that quantifies the symmetry between two Native American populations in their copying from each donor population

$$S(D; R_2, R_1) = \frac{C_{R_1:D} - C_{R_2:D}}{C_{R_1:D} + C_{R_2:D}}$$

If two Native American populations, such as the Suruí and the Mixe, derive all of their ancestry from a single ancestral population, we expect that they would copy from the donor populations at an equal rate. We computed the standard error of this statistic using the 22 subsets of the data where each autosome had been dropped, weighted using the number of SNPs on each chromosome.

We fixed R_2 =Mixe and computed all combinations of non-Americans other than Yukagir and Chukchi as D and Native Americans other than Mixe as R_1 . Table

S5.1 lists all top Z -scores for the S -statistics obtained for each population R_1 , and we find that the great majority of Native American populations with a Z -score of 3 or greater show an Australasian population as maximizing this symmetry statistic (Onge, Australian, Bougainville, New_Guinea or Tongan). The exceptions are Chane, Zoro and Pima. The Chane shows a highly significant rate of copying from Turkish_Jew and other European populations such as Norwegians, suggesting that this single Chane individual might have cryptic European ancestry. Similarly, the single Zoro individual show a top Z -score for African Wambo. We caution against strong interpretations of this since single individuals represent these two groups. The Pima show an affinity to African Tshwa ($Z = 3.71$), possibly suggesting cryptic African ancestry. In Table S5.2 we show statistics for all Native American populations when Onge is used as donor, which is the statistic underlying the heatmap in Figure 1D.

Table S5.1. Donor populations that maximize the excess copying in a Native American population when Mixe is used as a baseline.

R_1	D	$S(D;R_2,R_1)$	Z
Karitiana_UFRGS	Australian_WGA	0.0059	5.58
Ticuna	Bougainville	0.0028	3.44
Surui	Onge	0.0027	4.86
Chane	Turkish_Jew	0.0027	6.21
Karitiana	Onge	0.0027	5.04
Wayuu	New_Guinea	0.0027	3.19
Surui_UFRGS	Onge	0.0026	5.26
Zoro	Wambo	0.0023	3.73
Urubu_Kaapor	Tongan	0.0023	4.31
Xavante	Onge	0.0022	4.27
Piapoco	Onge	0.0019	3.09
Guarani_KW	New_Guinea	0.0018	2.95
Arara	Onge	0.0017	2.76
Guarani_GN	New_Guinea	0.0016	2.26
Cabecar	Onge	0.0016	3.61
Guarani	Onge	0.0016	3.19
Bolivian	Kinh	0.0016	2.76
Pima	Tshwa	0.0014	3.71
Aymara	Tshwa	0.0012	2.24
Mayan	Turkmen	0.0010	2.73
Apalai	Gui	0.0009	2.76

Kaqchikel	Ukrainian	0.0007	2.65
Quechua	Egyptian	0.0007	2.14

Table S5.2. Symmetry statistics for haplotype copying from the Onge using Mixe as baseline for diverse Native American populations

X	S(Onge; Mixe, X)	Z
Karitiana_UFRGS	0.00295	4.25
Surui	0.00274	4.86
Karitiana	0.00270	5.04
Surui_UFRGS	0.00260	5.26
Zoro	0.00232	2.06
Xavante	0.00222	4.27
Piapoco	0.00189	3.09
Chane	0.00176	2.31
Arara	0.00175	2.76
Cabecar	0.00162	3.61
Guarani	0.00162	3.19
Guarani_KW	0.00151	2.76
Urubu_Kaapor	0.00147	1.97
Apalai	0.00142	2.04
Aymara	0.00119	1.16
Guarani_GN	0.00111	1.60
Bolivian	0.00105	1.59
Ticuna	0.00098	0.85
Pima	0.00084	1.74
Quechua	0.00083	0.72
Chipewyan	0.00073	1.42
Mayan	0.00062	0.66
Wayuu	0.00010	0.11
Kaqchikel	-0.00008	-0.13

6. Models of population history

Our analyses suggest that Amazonian groups such as Karitiana and Suruí share more derived alleles with Oceanian aboriginal groups and Negritos than with other Native Americans from Central and South American. This suggests that the history of Amazonians and other Americans cannot be accurately described as a simple tree. To investigate which possible alternative models of population history could fit the data, we used an admixture graph framework to test formally different hypotheses.

We used ADMIXTUREGRAPH^{4,9} to fit suggested phylogenies with admixture events to the data, and assessed goodness-of-fit by investigating all possible f_4 -statistics predicted by the fitted model and assessing whether they differed significantly from

the empirically observed statistics. We chose as a starting point the model relating Mbuti Africans, Onge, MA1 and Karitiana found by Lazaridis et al¹ where MA1 is basal to the other non-Africans but contributed ancestry to the ancestral population of Karitiana. We added to this Han Chinese as being more closely related to the ancestral population of Native Americans than the Onge are, as well as Suruí as a sister group to Karitiana and Mixe (Extended Data Figure 6A). We find that this model is inconsistent with the data, in line with the previously reported results in SI 2 and SI 3, since it predicts that Mixe and Suruí/Karitiana are equally related to Onge, and indeed we observe several statistics for which the Z-score for the difference between the predicted and empirical statistics is $|Z| > 3$ (Extended Data Table 3). To account for this, we fitted a model in which the ancestors of Amazonians received admixture from a population related to the Onge (Extended Data Figure 6B), and found that this provided an excellent fit to the data, with no $|Z|$ -scores greater than 3.

To investigate if alternative admixture models could explain the data, we tested a model in which the gene flow is instead from a population closer to Han Chinese than the Onge, into Amazonians or Central Americans (Extended Data Figure 6C). This model predicts several statistics that are inconsistent with the data (Extended Data Table 3), such as the observation that Amazonians are closer to the Onge than Han Chinese (*e.g.* $f_4(\text{Han}, \text{Onge}; \text{Mixe}, \text{Suruí}) \gg 0$). Finally, we tested a model in which the Mixe received admixture from an East Asian source more closely related to the Han (Extended Data Figure 6D), or an ancient Siberian source most closely related to MA1 (Extended Data Figure 6E), but found that these models were also unable to reproduce the empirical evidence of the Amazonians being closer to Onge.

Plausible range of the admixture fraction related to Australasians

The topology of the admixture graph that we infer makes it impossible to use a measure such as an f_4 -ratio to infer the proportion of ancestry related to Australasians and obtain confidence intervals. Instead, we tested different proportions of this ancestry in the Suruí, assuming that the proportion in Mixe is 0, and determined the proportions for which the maximum Z-score between predicted and observed f_4 -statistics is less than 3. We find that the plausible range of Australasian-related ancestry in the Suruí is 0.7-9.8% and that the lowest maximum Z-score is obtained for an admixture proportion of 2.3% (Extended Data Figure 7). If we instead find the proportion for which no Z-score is greater than 2, we obtain a range of 1.5-5.7%.

We also fitted models where the Mixe received Australasian-related gene flow. We fixed the proportion of this ancestry in the Mixe, while allowing ADMIXTUREGRAPH to fit a secondary and larger proportion in the Suruí. We find that a mixture proportion of 2.3% or more the Mixe results always produces $|Z|$ -scores larger than 3, and so we can rule out proportions of Australasian-related ancestry in the Mixe larger than this. In the model where Mixe have 2.3% Australasian-related ancestry, the Suruí are fitted as having an extra 1.0% such ancestry, above and beyond

that in the Mixe (Extended Data Figure 7). Thus, even if the Mixe, too, have Australasian related ancestry, we can put a stringent upper bound on it.

Fitting models in which Population Y is itself admixed

We added the Africa Dinka as an additional outgroup to the Admixture Graph. We also added the Mesoamerican Pima, the northernmost population in our data with no evidence of being other than 100% First American in previous studies. We finally added in the Amazonian Xavante, which were newly genotyped for this study.

We used this extended Admixture Graph model to explore scenarios in which the Australasian related ancestry in Amazonians derived from a Native American founding “Population Y” that was admixed with a First American lineage at the time that it contributed to the ancestors of Amazonians. While we have no statistical support for such a model, the motivation for exploring it is that from a human migration point of view, it seems plausible. It seems unlikely that population sharing 100% of its ancestry with the Onge reached South America without admixing with other populations who were inhabiting the same region.

To explore this family of models, we modeled Population Y as deriving a proportion α of their ancestry from the Onge-lineage and $(1-\alpha)$ from a basal First American lineage. We then allowed for the common ancestral population of the three Amazonian populations to carry a proportion γ from the Population Y lineage and a proportion $(1-\gamma)$ from a Mesoamerican population more closely related to the Mixe than the Pima. Figure 2A illustrates this model. We fitted admixture graphs for $\gamma = 0-100\%$ and $\alpha = 0-100\%$ (with a grid size of 1%) and used the number of f_4 -statistics predicted by the model that deviated by more than 3 sigma from the empirical statistics to evaluate model fit. We also fitted γ and α automatically and obtained point estimates of $\alpha=2\%$ and $\gamma=63\%$, but find that a much broader range of parameter combinations are consistent with the data. These results show that the proportion of Population Y ancestry in Amazonians can plausibly be quite high: indeed, as high as 63% or higher.

Importantly, while the proportion of Population Y ancestry in Amazonians is poorly determined by our analyses, the proportion of ancestry related to the Onge is tightly constrained. In our model, the proportion of Onge-related ancestry in Amazonians is the product of γ and α , which our model fitting shows to be constrained between 1% and 2% (no parameter combinations with a proportion outside this range fitted the data). This is similar to the estimates we obtained with modeling that used fewer populations and assumed Population Y to be an unadmixed sister group of the Onge.

Supplementary References

- 1 Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409-413, doi:10.1038/nature13673 (2014).
- 2 Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488**, 370 - 374 (2012).
- 3 Wang, S. *et al.* Genetic Variation and Population Structure in Native Americans. *PLoS Genet* **3**, e185, doi:10.1371/journal.pgen.0030185 (2007).
- 4 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).
- 5 Alexander, D., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655 - 1664 (2009).
- 6 Raghavan, M. *et al.* The genetic prehistory of the New World Arctic. *Science* **345**, doi:10.1126/science.1255832 (2014).
- 7 Li, J. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100 - 1104 (2008).
- 8 Loh, P.-R. *et al.* Inference of admixture parameters in human populations using weighted linkage disequilibrium. (2012).
- 9 Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489-494 (2009).
- 10 Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328**, 710-722 (2010).
- 11 Busing, F. M., Meijer, E. & Van Der Leeden, R. Delete-m jackknife for unequal m. *Statistics and Computing* **9**, 3-8 (1999).
- 12 Reich, D. *et al.* Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. *The American Journal of Human Genetics* **89**, 516-528, (2011).
- 13 McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics* **5**, e1000471 (2009).
- 14 Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).
- 15 Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49, doi:10.1038/nature12886 (2014).
- 16 Robbins, R. B. Some applications of mathematics to breeding problems III. *Genetics* **3**, 375 (1918).
- 17 Lipson, M. *et al.* Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. *Molecular Biology and Evolution* **30**, 1788-1802, doi:10.1093/molbev/mst099 (2013).