LETTER

Genetic evidence for two founding populations of the Americas

Pontus Skoglund[1,2], Swapan Mallick[1,2,3], Maria Cátira Bortolini[4], Niru Chennagiri[1,2], Tábita Hünemeier[5], Maria Luiza Petzl–Erler[6], Francisco Mauro Salzano[4], Nick Patterson[2] & David Reich[1,2,3]

Genetic studies have consistently indicated a single common origin of Native American groups from Central and South America[1–4]. However, some morphological studies have suggested a more complex picture, whereby the northeast Asian affinities of present-day Native Americans contrast with a distinctive morphology seen in some of the earliest American skeletons, which share traits with present-day Australasians (indigenous groups in Australia, Melanesia, and island Southeast Asia)[5–8]. Here we analyse genome-wide data to show that some Amazonian Native Americans descend partly from a Native American founding population that carried ancestry more closely related to indigenous Australians, New Guineans and Andaman Islanders than to any present-day Eurasians or Native Americans. This signature is not present to the same extent, or at all, in present-day Northern and Central Americans or in a ~12,600-year-old Clovis-associated genome, suggesting a more diverse set of founding populations of the Americas than previously accepted.

All Native American groups studied to date can trace all or much of their ancestry to a single ancestral population that probably migrated across the Bering land bridge from Asia more than 15,000 years ago[2], with some Northern American and Arctic groups also tracing other parts of their ancestry to more recent waves of migration[2,9,10]. Ancient genomic evidence has shown that this so-called 'First American' ancestry is present in an individual associated with Clovis technology from North America dating to ~12,600 years ago[3], and mitochondrial DNA has suggested that it was also present by 13,000–14,500 years ago[11,12]. In contrast, some morphological analyses of early skeletons in the Americas have suggested that characteristics of some Pleistocene and early Holocene skeletons fall outside the variation of present-day Native Americans and instead fall within the variation of present-day indigenous Australians, Melanesians and so-called 'Negrito' groups from Southeast Asia (and some sub-Saharan African groups)[7,13]. This morphology has been hypothesized to reflect an initial 'Paleoamerican' pioneer population in the Americas, which according to some interpretations was largely replaced by populations with Northeast Asian affinities in the early Holocene, but may have persisted in some locations[14,15]. However, morphological similarity can arise not only through shared descent but also through convergent evolution or phenotypic plasticity coupled with similar environments[16,17]. Another limitation of morphological data is that it provides very few independent characters that can be analysed. Genome-wide data, with its hundreds of thousands of independent characters that evolve effectively neutrally, should be a statistically powerful and robust way to test whether a distinct lineage contributed to Native Americans.

Analysis of population history in the Americas is complicated by post-Columbian admixture from mainly European and African sources[2]. We identified 63 individuals without discernable evidence of European or African ancestry in 21 Native American populations genotyped at ~600,000 single nucleotide polymorphisms (SNPs) on the Affymetrix Human Origins array[18,19] (Extended Data Fig. 1 and Supplementary Information section 1). We further restricted our studies to individuals from Central and South America that have the strongest evidence of deriving entirely from a homogeneous First American ancestral population[2]. We computed all possible f_4-statistics of the form $f_4(American_1, American_2; outgroup_1, outgroup_2)$, the product of the allele frequency differences between the two American groups and the two outgroups. We represented the Americas by a panel of 7 Central and South American groups, and the outgroups by 24 populations (4 from each of 6 worldwide regions). If the two Native American groups descend from a homogeneous ancestral population whose ancestors separated from the outgroups at earlier times, it follows that the difference in allele frequencies between Native American populations will have developed entirely after their separation from the outgroups, and so the correlation in allele frequency differences is expected to be zero. To evaluate whether all possible f_4-statistics computed in this way are consistent with zero, correcting for multiple hypothesis testing due to the large number of statistics examined, we measured the empirical covariance of the matrix of f_4-statistics using a block jackknife[18], and performed a single Hotelling's T^2 test[2] for consistency with zero. We reject the null hypothesis at high significance ($P = 2 \times 10^{-7}$), suggesting that the analysed Native American populations do not all descend from a homogeneous ancestral population since separation from the outgroups (Extended Data Table 1 and Supplementary Information section 2). The coefficients for which non-American populations contribute the most to the signals separate Native Americans into a cline with two Amazonian groups (Suruí and Karitiana) on one extreme and Mesoamericans on the other (Extended Data Fig. 2). Among the outgroups, the most similar coefficients to Amazonian groups are found in Australasian populations: the Onge from the Andaman Islands in the Bay of Bengal (a so-called 'Negrito' group), New Guineans, Papuans and indigenous Australians (Supplementary Information section 2).

We extended our analysis to 197 non-American populations sampled worldwide[18–20]. We computed D-statistics[21] to test whether a randomly drawn derived allele from each worldwide population has an equal probability of matching a randomly drawn Mesoamerican or Amazonian chromosome at sites where these differ. This test takes as its null hypothesis the tree-like population history (*Test population*, (*Mesoamericans*, *Amazonians*)), and produces a positive D-statistic only in the case of excess affinity between the test population and Amazonians (negative values in the case of an excess affinity with Mesoamericans). Consistent with the signals observed when many populations are analysed together, we find that Andamanese Onge, Papuans, New Guineans, indigenous Australians and Mamanwa Negritos from the Philippines all share significantly more derived alleles with the Amazonians ($4.6 > Z > 3.0$ standard errors (s.e.) from zero) (Extended Data Table 2). No population shares significantly more derived alleles with the Mesoamericans than with the Amazonians. We

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [3]Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, 91501-970 Porto Alegre, RS, Brazil. [5]Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, 05508-090, SP, Brazil. [6]Departamento de Genética, Universidade Federal do Paraná, 81531-980 Curitiba, PR, Brazil.

find consistent results for this test not only for Onge, Papuans, New Guineans and indigenous Australians as representatives of Australasian populations, but also for different outgroups in place of chimpanzee: Africans, Europeans and East Asians ($2.8 < Z < 4.8$) (Supplementary Information section 3). In Fig. 1, we show a quantile–quantile plot of D-statistics contrasting the Mesoamerican Mixe and the Amazonian Suruí, revealing Australasian populations as the only discernible outliers.

We replicated the significant evidence for affinity between Australasians and Amazonians using D-statistics computed on Illumina SNP array data[2] (as an alternative to the Affymetrix Human Origins SNP array data) ($2.6 < Z < 3.0$) and on high-coverage genome sequences from 3 Yoruba, 2 Suruí, 3 Mixe and 16 Papuans (18 of these genomes are reported for the first time here[22,23]; Table 1) ($Z = 4.3$). In addition to the three independent molecular experiments that these data sets represent, we find consistent results for all different mutation classes in the high-coverage genomes ($2.6 < Z < 4.3$), and different ascertainment schemes (for example, in polymorphisms discovered in Africans, New Guineans and East Asians) (Supplementary Information section 3) ($1.1 < Z < 3.3$ for panels with $>20,000$ SNPs). We also find consistent results for two differently genotyped subsets of Suruí individuals from a total of 24 individuals[2] (Table 1 and Extended Data Fig. 3a) ($2.6 < Z < 3.6$). Simulations (Supplementary Information

section 3) show that genotype and sequence errors cannot explain the magnitude of the observed signal (Extended Data Fig. 3b). Finally, we generated new data from 9 populations from present-day Brazil using the Affymetrix Human Origins array, including previously untested individuals from the Amazonian Suruí and Karitiana for which DNA was extracted from blood. These new samples replicate the signal, and furthermore show that the signal is also strong in the Xavante ($1.3 < Z < 3.25$), a population of the Brazilian Central Plateau which speaks a language of the Ge group that is different from the Tupi language group to which the languages of the Karitiana and Suruí both belong. We do not detect any excess affinity to Australasians in the ~12,600-year-old Clovis-associated Anzick individual from western Montana ($Z = -0.6$) (Supplementary Information section 3).

To test if the significant D-statistics have the patterns expected for a genuine admixture event, we stratified the high coverage genomes into deciles of 'B-values'[24], which measures proximity to functionally important regions. Genuinely significant D-statistics are expected to be of larger magnitude closer to genes, as selection increases variability in fitness of haplotypes near functionally important regions, which in turn increases the genetic drift in these regions and the absolute magnitude of D-statistics[25,26], a prediction that we confirmed empirically (Extended Data Fig. 4). We computed D(Yoruba, Papuan; Mixe,
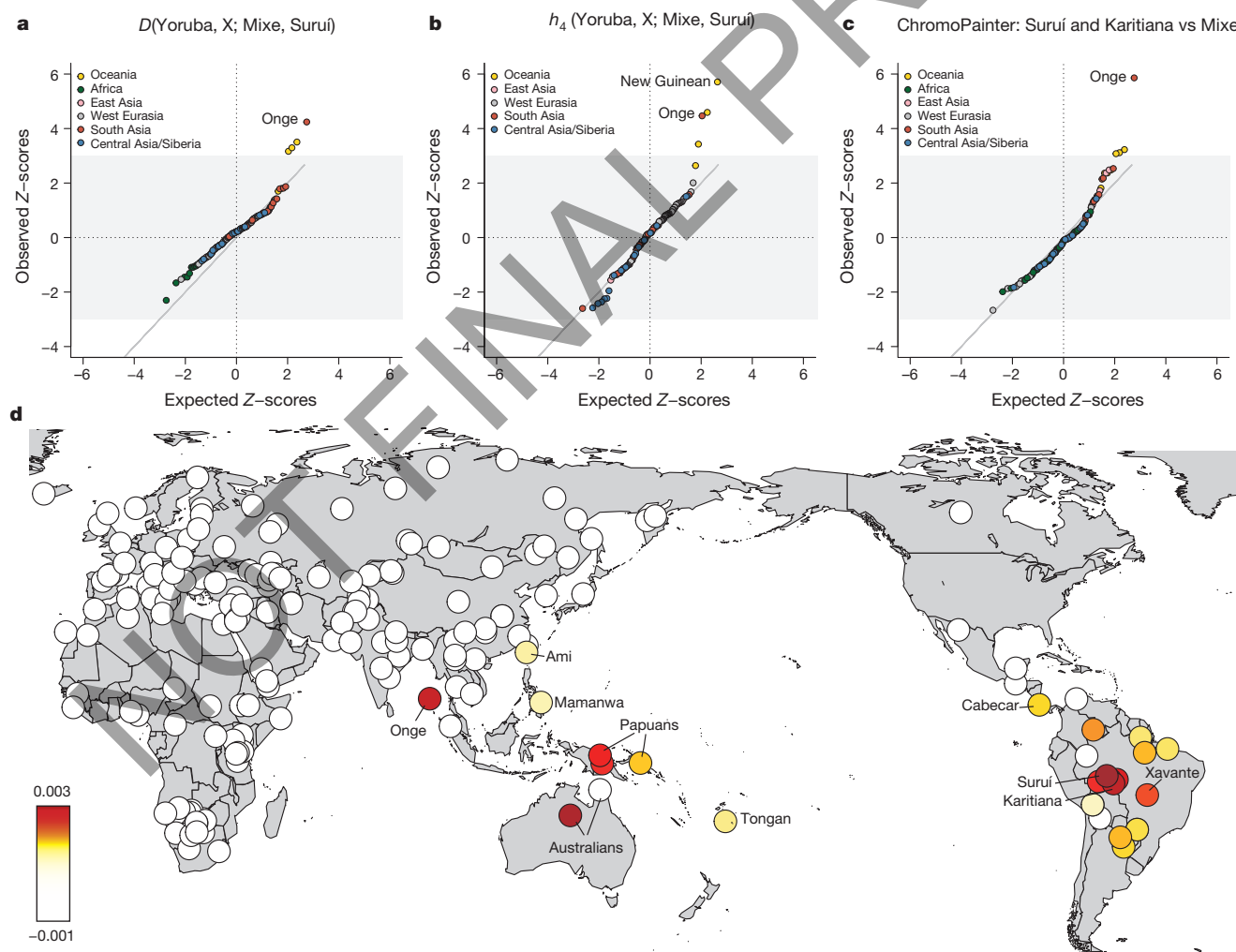
Figure 1 | South Americans share ancestry with Australasian populations that is not seen in Mesoamericans or North Americans. a, Quantile–quantile plot of the Z-scores for the D-statistic symmetry test for whether Mixe and Suruí share an equal rate of derived alleles with a candidate non-American population X, compared to the expected ranked quantiles for the same number of normally distributed values. **b**, Z-scores for the h_4-statistic. **c**, Z-scores for

the ChromoPainter statistic. **d**, Heatmap of ChromoPainter statistics. For non-Americans we display the symmetry statistic S(non-American; Mixe, Suruí and Karitiana) for donating as many haplotypes to Mixe as to Suruí and Karitiana. For the Americas we plot S(Onge; Mixe, American) for receiving as many haplotypes from the Onge as do the Mixe.

Table 1 | Statistics testing the consistency of the tree (Yoruba, (Papuan, (Mixe, Suruí)) with the data

	Test statistic	Z-score	Informative loci
High-coverage genomes	0.0211	4.26	798,873
A/T SNPs	0.0169	2.63	60,538
A/G SNPs	0.0191	3.64	268,962
A/C SNPs	0.0208	3.49	67,210
G/T SNPs	0.0248	4.27	67,623
C/T SNPs	0.0220	4.24	270,133
C/G SNPs	0.0248	4.26	64,951
Illumina array Suruí samples from HGDP	0.0076	2.63	247,814
Illumina array Suruí samples not in HGDP	0.0081	3.02	249,941
Affymetrix Human Origins array (Suruí cell lines)	0.0099	3.63	318,544
Affymetrix Human Origins array (Suruí blood samples)	0.0072	2.57	313,349
h_4-statistic (Affymetrix Yoruba ascertainment)	0.0003	4.60	14,938
Chromosome painting symmetry test	0.0026	5.26	-

Note: except for the new h_4 statistics and chromosome painting symmetry tests which are explicitly noted, all statistics are D-statistics[21]. Z-scores were obtained by computing standard errors using a weighted block jackknife.

Suruí) separately for each bin, and found that it is of larger magnitude close to functionally important regions (Extended Data Fig. 4) ($Z = -2.0$ for the slope of a linear regression model), as expected for a real admixture event. A caveat is that when we formally combine the evidence from the genome-wide D-statistic and the correlation to the B-value, the significance ($Z = 3.6$ s.e. from 0) is not any greater than for the basic $D = 0.021 \pm 0.005$ statistic ($Z = 4.2$ s.e. from 0) because the two statistics co-vary. Nevertheless, the fact that the correlation with B-values is significant by itself and in the expected direction adds to the qualitative evidence for an admixture event.

Alternative approaches for testing for admixture involve detecting admixture linkage disequilibrium in a test population that is correlated to allele frequency differentiation between two populations that are related to the sources[27,28]. We devised a statistic 'h_4' that is analogous to an f_4-statistic, but instead of studying allele frequencies, it tests whether the linkage disequilibrium patterns of two populations are consistent with descending from a common ancestral population since separation from two outgroups. A classic statistic for measuring linkage disequilibrium in a population A is $H^A = p_{12}^A - p_1^A p_2^A$, which measures the extent to which a haplotype of two derived mutations occurring at frequency p_{12}^A is observed more or less frequently than would be expected from the individual frequencies of alleles 1 and 2 (p_1^A and p_2^A). Thus, we define $h_4(A, B; C, D)$ as the average of $(H^A - H^B)(H^C - H^D)$ across the genome, and view a deviation from zero as evidence against the unrooted tree ((A, B), (C, D)). We used loci ascertained as polymorphic in African Yoruba, which is effectively an outgroup to the other populations analysed here, to test h_4(Yoruba, X; Mixe, Suruí) for all SNP pairs within 0.01 centimorgans (cM) and for a large set of worldwide non-African populations, and obtained normalized Z-scores by estimating the number of standard errors this quantity is from zero using a block jackknife. Although Z-scores computed for most of 120 non-American and non-Africans as population X conform to a normal distribution (Fig. 1b), we again found significant evidence of excess affinity of the Suruí to Australasian populations ($Z = 5.7$, $P = 10^{-8}$ for New Guineans; $Z = 4.6$, $P = 10^{-5}$ for Papuans; $Z = 4.4$, $P = 10^5$ for Andamanese). When we exclude the Australasians, we detect no evidence of correlation between Z-transformed h_4- and f_4-statistics for the remaining 114 populations ($R = -0.026$) suggesting that h_4 can provide evidence independent of allele frequency based statistics. Although h_4 can theoretically be biased by loss of polymorphism due to bottlenecks (Supplementary Information section 4), there is no evidence that this is a problem for our analysis as East Asian and Siberian populations with comparable loss of polymorphism do not show an affinity to Amazonians by this statistic (Extended Data Fig. 5). In addition, there is a high degree of correlation between significant h_4- and D-statistics in empirical data (Extended Data Fig. 5). Computing h_4(Yoruba, Onge; Mixe, Suruí) over windows of increasingly large genetic distances reveals that it dissipates at approximately 0.2 cM. This is an order of magnitude smaller than linkage disequilibrium caused by admixture events at

the ~4,000 year upper limit of previous methods[18], but at a larger scale than the signal of admixture between Neanderthals and non-Africans 37,000–86,000 years ago[29] (Extended Data Fig. 5).

As a third population symmetry test, we applied a method for detecting shared haplotypes between individuals ('chromosome painting'[30]) to infer in each Native American individual which non-American chromosome segment each American chromosome segment shares the closest affinity to, using a set of 174 non-American populations as references. We then performed a symmetry test for a candidate population sharing more haplotypes with a given non-American population than the Mesoamerican Mixe do, performing a block jackknife across all chromosomes (weighting to correct for variation in chromosome length) to assess uncertainty. We find that the blood and cell line Suruí are significantly closer to the Onge than the Mixe are ($Z = 5.3$) (Fig. 1c), as are the blood and cell line Karitiana samples ($Z = 4.2$ to 5.0), the Xavante ($Z = 4.3$), and the Piapoco and Guarani ($Z > 3$) (Fig. 1d). In contrast, populations from west of the Andes or north of the Panama isthmus show no significant evidence of an affinity to the Onge ($Z < 2$). An exception to this is the Cabecar, who have previously been shown to be partially admixed from a source south of the Panama isthmus[2].

The geographic distribution of the shared genetic signal between South Americans and Australasians cannot be explained by post-Columbian African, European or Polynesian gene flow into Native American populations. If such gene flow produced signals strong enough to affect our statistics, our statistics would show their strongest deviations from zero for African, European or Polynesian populations, which is not observed. For example, a direct test is significant in showing that the Suruí-specific ancestry component is genetically closer to the Andamanese Onge than to Tongans from Polynesia ($D = 0.0094$, $Z = 3.4$).

To investigate models consistent with the data, we studied admixture graph models relating the ancestry of Native American groups to Han Chinese and Onge Andaman Islanders, incorporating a previously described admixture event into Native American ancestors from a lineage related to a ~24,000-year-old Upper Paleolithic individual from Mal'ta in Siberia[4]. We are unable to fit Amazonians as forming a clade with the Mesoamericans, or as having a different proportion of ancestry related to Mal'ta or present-day East Asians. Thus, our signal cannot be explained by lineages that have previously been documented as having contributed to Native American populations. However, we do find that a model where Amazonians receive ancestry from the lineage leading to the Andamanese fits the data in the sense that the predicted f_4-statistics are all within two standard errors of statistics computed on the empirical data (Extended Data Figs 6 and 7 and Extended Data Table 3). These results do not imply that an unmixed population related anciently to Australasians migrated to the Americas. Although this is a formal possibility, an alternative model that we view as more plausible is that the 'Population Y' (after Ypykuéra, which means 'ancestor' in the Tupi language family spoken by the Suruí and Karitiana) that contributed Australasian-related
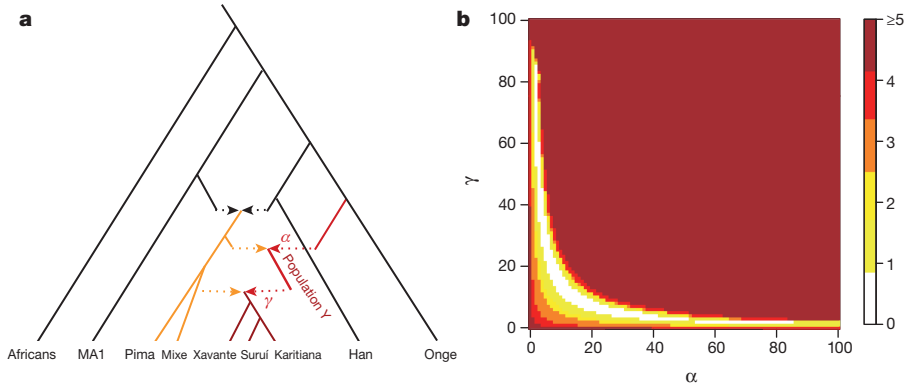
Figure 2 | A model of population history that can explain the excess affinity to Oceanians observed in Amazonian populations. a, We fit an admixture graph model where a population related to the Andamanese Onge contributed a fraction α of the ancestry of 'Population Y', which later contributed a fraction γ to the ancestry of Amazonian groups today

(the remainder of which is related to Mesoamerican Mixe). **b**, Two-dimensional grid of combinations of the admixture proportions α and γ which are compatible with the data in terms of how many predicted f_4-statistics deviate by $Z \geq 3.0$ from empirical values.

ancestry to Amazonians was already mixed with a lineage related to First Americans at the time it reached Amazonia. When we model such a scenario, we obtain a fit for models that specify 2–85% of the ancestry of the Suruí, Karitiana and Xavante as coming from Population Y (Fig. 2). These results show that quite a high fraction of Amazonian ancestry today might be derived from Population Y. At the same time, the results constrain the fraction of Amazonian ancestry that comes from an Australasian related population (via Population Y) to a much tighter range of 1–2% (Fig. 2).

We have shown that a Population Y that had ancestry from a lineage more closely related to present-day Australasians than to present-day East Asians and Siberians, likely contributed to the DNA of Native Americans from Amazonia and the Central Brazilian Plateau. This discovery is striking in light of interpretations of the morphology of some early Native American skeletons, which some authors have suggested have affinities to Australasian groups. The largest number of skeletons that have been described as having this craniofacial morphology and that date to younger than 10,000 years old have been found in Brazil[6], the home of the Suruí, Karitiana and Xavante groups who show the strongest affinity to Australasians in genetic data. However, in the absence of DNA directly extracted from a skeleton with this morphology, our results are not sufficient to conclude that the Population Y we have reconstructed from the genetic data had this morphology.

An open question is when and how Population Y ancestry reached South America. There are several archaeological sites in the Americas that are contemporary to or earlier than Clovis sites. The fact that the one individual from a Clovis context who has yielded ancient DNA had entirely First American ancestry[3] suggests the possibility that Population Y ancestry may be found in non-Clovis sites. Regardless of the archaeological associations, our results suggest that the genetic ancestry of Native Americans from Central and South America cannot be due to a single pulse of migration south of the Late Pleistocene ice sheets from a homogenous source population, and instead must reflect at least two streams of migration or alternatively a long drawn out period of gene flow from a structured Beringian or Northeast Asian source. The arrival of Population Y ancestry in the Americas must in any scenario have been ancient: while Population Y shows a distant genetic affinity to Andamanese, Australian and New Guinean populations, it is not particularly closely related to any of them, suggesting that the source of population Y in Eurasia no longer exists; furthermore, we detect no long-range admixture linkage disequilibrium in Amazonians as would be expected if the Population Y migration had occurred within the last few thousand years. Further insight into the population movements responsible for these findings should be

possible through genome-wide analysis of ancient remains from across the Americas.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Wang, S. *et al.* Genetic variation and population structure in Native Americans. *PLoS Genet.* **3**, e185 (2007).
2. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
3. Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014).
4. Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014).
5. Neves, W. & Pucciarelli, H. The origins of the first Americans—an analysis based on the cranial morphology of early South American remains. *Am. J. Phys. Anthropol.* **81**, 274 (1990).
6. Neves, W. *et al.* Early Holocene human skeletal remains from Cerca Grande, Lagoa Santa, Central Brazil, and the origins of the first Americans. *World Archaeol.* **36**, 479–501 (2004).
7. Neves, W. A., Prous, A., González-José, R., Kipnis, R. & Powell, J. Early Holocene human skeletal remains from Santana do Riacho, Brazil: implications for the settlement of the New World. *J. Hum. Evol.* **45**, 19–42 (2003).
8. González-José, R. *et al.* Late Pleistocene/Holocene craniofacial morphology in Mesoamerican Paleoindians: implications for the peopling of the New World. *Am. J. Phys. Anthropol.* **128**, 772–780 (2005).
9. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010).
10. Raghavan, M. *et al.* The genetic prehistory of the New World Arctic. *Science* 345, (2014).
11. Gilbert, M. T. P. *et al.* DNA from pre-Clovis human coprolites in Oregon, North America. *Science* **320**, 786–789 (2008).
12. Chatters, J. C. *et al.* Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern Native Americans. *Science* **344**, 750–754 (2014).
13. Jantz, R. L. & Owsley, D. W. Variation among early North American crania. *Am. J. Phys. Anthropol.* **114**, 146–155 (2001).
14. Neves, W. A., Hubbe, M. & Correal, G. Human skeletal remains from Sabana de Bogota, Colombia: a case of Paleoamerican morphology late survival in South America? *Am. J. Phys. Anthropol.* **133**, 1080–1098 (2007).
15. González-José, R. *et al.* Craniometric evidence for Palaeoamerican survival in Baja California. *Nature* **425**, 62–65 (2003).
16. Sparks, C. S. & Jantz, R. L. A reassessment of human cranial plasticity: Boas revisited. *Proc. Natl Acad. Sci. USA* **99**, 14636–14639 (2002).
17. Relethford, J. H. Apportionment of global human genetic diversity based on craniometrics and skin color. *Am. J. Phys. Anthropol.* **118**, 393–398 (2002).
18. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
19. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
20. Qin, P. & Stoneking, M. Denisovan ancestry in East Eurasian and Native American populations. *Mol. Biol. Evol.* (2015).
21. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
22. Meyer, M. *et al.* A high-coverage genome sequence from an Archaic Denisovan individual. *Science* **338**, 222–226 (2012).

23. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).
24. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5,** e1000471 (2009).
25. Gillespie, J. H. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* **155,** 909–919 (2000).
26. Coop, G. *et al.* The role of geography in human adaptation. *PLoS Genet.* **5,** e1000500 (2009).
27. Moorjani, P. *et al.* The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet.* **7,** e1001373 (2011).
28. Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343,** 747–751 (2014).
29. Sankararaman, S., Patterson, N., Li, H., Pääbo, S. & Reich, D. The date of interbreeding between Neandertals and modern humans. *PLoS Genet.* **8,** e1002947 (2012).
30. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8,** e1002453 (2012).

Supplementary Information is available in the online version of the paper.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

New Affymetrix Human Origins genotypes. We generated new Affymetrix Human Origins array genotypes for 48 individuals from 9 populations from present-day Brazil (Apalaí, Arara, Guarani_GN, Guarani_KW, Karitiana, Suruí, Urubu Kaapor, Xavante and Zoró). Ethical approval for the sample collection was provided by the Brazilian National Ethics Commission (CONEP Resolution no. 123/98). CONEP also approved the oral consent procedure and the use of these samples in studies of population history and human evolution. Individual and/or tribal informed oral consents were obtained from participants who were not able to read or write. All sampling was coordinated by co-authors of this study (M.L.P.-E. and F.M.S.) and their collaborators, in a manner consistent with the Helsinki Declaration 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 (FUNAI). We curated the data in the same way that was reported in ref. 19 (Supplementary Information section 1). We computationally phased these data together with the previously published Affymetrix Human Origins SNP array data using SHAPEIT2 (ref. 31) with default parameters.

High-coverage genome sequencing and processing. We sent samples from 18 Papuan, Mixe, Suruí and Yoruba individuals to Illumina for deep-coverage sequencing using a non-PCR-based protocol as part of the Simons Genome Diversity Project. The sequence reads were mapped using the 'aln' algorithm of BWA (version 0.5.10)[32] and genotypes were inferred using the unified genotyper from GATK[33] (version 2.5.2-gf57256b) These data are available from (https://www.simonsfoundation.org/life-sciences/simons-genome-diversity-project-dataset/). Briefly, sequence reads were stripped of adapters before alignment to the decoy version of the hg19 reference sequence (hs37d5). Read groups were added for identification and compatibility with GATK tools, before indel realignment and duplicate removal. The genotyping performed thereafter used a reference-free procedure that reduces reference bias. A specially developed filtering engine assigned filtering levels from 0 to 9 for each position in the genome. All population genetic analyses in this paper used the most stringent level of filtering (level 9).

Testing for more than one ancestral population of Central and South Americans. To investigate whether Central and South American populations are consistent with being derived from a single stream of ancestry, we applied the software qpWave[2] to ask the question whether the set of f_4-statistics of the form $f_4(A = American_1, B = American_2; X = outgroup_1, Y = outgroup_1) = (p_A - p_B)(p_X - p_Y)$ forms a matrix that is consistent with being of rank 0 (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). If all these Native American populations descend from the same stream of migration into the Americas, then the f_4-statistic relating each Native American population to each non-Native American population should be the same for all Native American populations, and in particular consistent with 0. Formally, to evaluate whether the f_4-statistic matrix is consistent with being rank 0, we compute a Hotelling's T^2 test that appropriately corrects for the correlation structure of the f_4-statistics. We analysed 7 Native American populations each with at least 3 individuals with no detected post-Columbian admixture, and 4 populations from each of 6 worldwide regions as outgroups (Supplementary Information section 2).

D-statistic tests based on correlation in allele frequencies. To investigate whether a tree-like population history $((A, B),(X, Y))$ is consistent with the data, for example, with A = chimpanzee, B = Onge, X = Mixe and Y = Suruí, we computed D-statistics[18,21]

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

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. We computed standard errors using a block jackknife weighted by the number of SNPs in each 5 cM (5 Mb in the case of high-coverage genome sequences) block in the genome[34,35]. We report Z-scores as normalized Z = D/s.e. and we interpret statistics $|Z| > 3$ as being significantly different from 0. We only considered SNPs that were informative, in the sense that they are polymorphic both within (A, B) and (X, Y).

Correlation of signal to regions of functional importance. We divided the genome into 10 deciles of the 'B-value' described in ref. 24, which integrates multiple genomic annotations into a single estimate of proximity to functional regions for each nucleotide in the genome. We then used linear regression to estimate the coefficient a of the function $y = ax + c$ where $x = B$ (the rank of the decile of B) and $y = D_B$ (D restricted to the particular decile of B). To compute

standard errors, we used a weighted block jackknife procedure where each 5 Mb block of the genome is dropped in turn and a is recomputed. The variability of a across each of these leave-one-out computations, weighting by the number of informative loci in each block, was what we used to estimate a standard error[34,35].

h_4-statistic tests based on correlation in linkage disequilibrium. We devised a linkage disequilibrium statistic that tests for symmetry in linkage disequilibrium between two proposed clades with a pair of populations in each. The statistic, h_4, is:

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

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 $p_{12}^A - p_1^A p_2^A$ thus measures the difference between the observed haplotype frequency and the expected haplotype frequency given the allele frequencies[36]. The motivation for this statistic being informative about population history is that under a tree-like model $((A, B), (C, D))$ with no gene flow, differences in linkage disequilibrium between populations A and B are not expected to correlate to differences in linkage disequilibrium between populations C and D. If there has been gene flow between the two clades, the statistic may be significantly positive or negative like f_4- and D-statistics[18].

In practice, we computed this statistic for each polymorphic locus ('target locus') by identifying all other polymorphic loci 5′ of the target locus at distance interval $d \pm w$ and computing the statistic for each pairing. We then averaged the statistic over all valid pairs of loci in the genome identified in this way. We computed standard errors using a block jackknife over contiguous 5 cM blocks in the genome, where SNP pairs that bridge the boundary of two blocks are assigned to the block in which the target locus is found. For the main analysis we computed h_4-statistics of the form $h_4(Yoruba, X; Mixe, Suruí)$ for all populations X genotyped using the Affymetrix Human Origins SNP array, and all pairs of SNPs within 0.01 cM of each other. We restricted the analysis to populations with at least 10 individuals. We also computed the h_4-statistic for windows of 0.001 cM centred around different genetic distances for selected populations (Extended Data Fig. 5).

Chromosome-painting symmetry tests. We used SHAPEIT to phase 593,142 SNPs with the same set of individuals as described above, using all autosomal SNPs in the Affymetrix Human Origins array. We then 'painted' unadmixed Native American individuals using non-American populations, and 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 of these subsets a different autosome had been dropped. We then used the results of these 22 block jackknife pseudo-replicates 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:

$$S(D; R_2, R_1) = \frac{C_{R1:D} - C_{R2:D}}{C_{R1:D} + C_{R2:D}}$$

If two Native American populations, such as the Suruí and the Mixe, derive all of their ancestry from a single common origin, 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 generated the world map in Fig. 1d by using the R maps package to plot the value of S(X; Mixe, Surui + Karitiana) for each non-American population X, and S(Onge; Mixe, Y) for each American population Y.

Admixture graph models of population relationships. We used ADMIXTURE-GRAPH[18] to fit suggested phylogenies with admixture events to the data. We assessed goodness-of-fit by investigating all possible f-statistics predicted by the fitted model and assessing whether they differed significantly from the empirical data. We chose as a starting point the model relating Mbuti Africans, Andamanese Onge, MA1 and Karitiana fitted by a previous study[19] where lineages related to MA1 and the Onge both contributed ancestry to the Karitiana. We added to this Han Chinese to represent a population that is phylogenetically more closely related to one of the ancestral populations of Native Americans than are the Onge (Extended Data Figs 6 and 7). We find that this model is inconsistent with the data, as the model 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 Fig. 6), and found that this provides an excellent fit to the data, with no $|Z|$-score differences greater than 3. In contrast, alternative models of Han-related or MA1-related gene flow into the Americas are inconsistent with the data (Extended Data Fig. 6 and Extended Data Table 3).

Code availability. A python program for computing h_4 symmetry statistics and other population genetic statistics used in this paper is available at (https://github.com/pontussk/popstats).

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

32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
33. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
34. Busing, F. M., Meijer, E. & Van Der Leeden, R. Delete-m jackknife for unequal m. *Stat. Comput.* **9,** 3–8 (1999).
35. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461,** 489–494 (2009).
36. Robbins, R. B. Some applications of mathematics to breeding problems III. *Genetics* **3,** 375–389 (1918).
37. Becker, R. A. & Wilks, A. R. Maps in S. *AT&T Bell Laboratories Statistics Research Report [93.2]*, (1993).
38. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
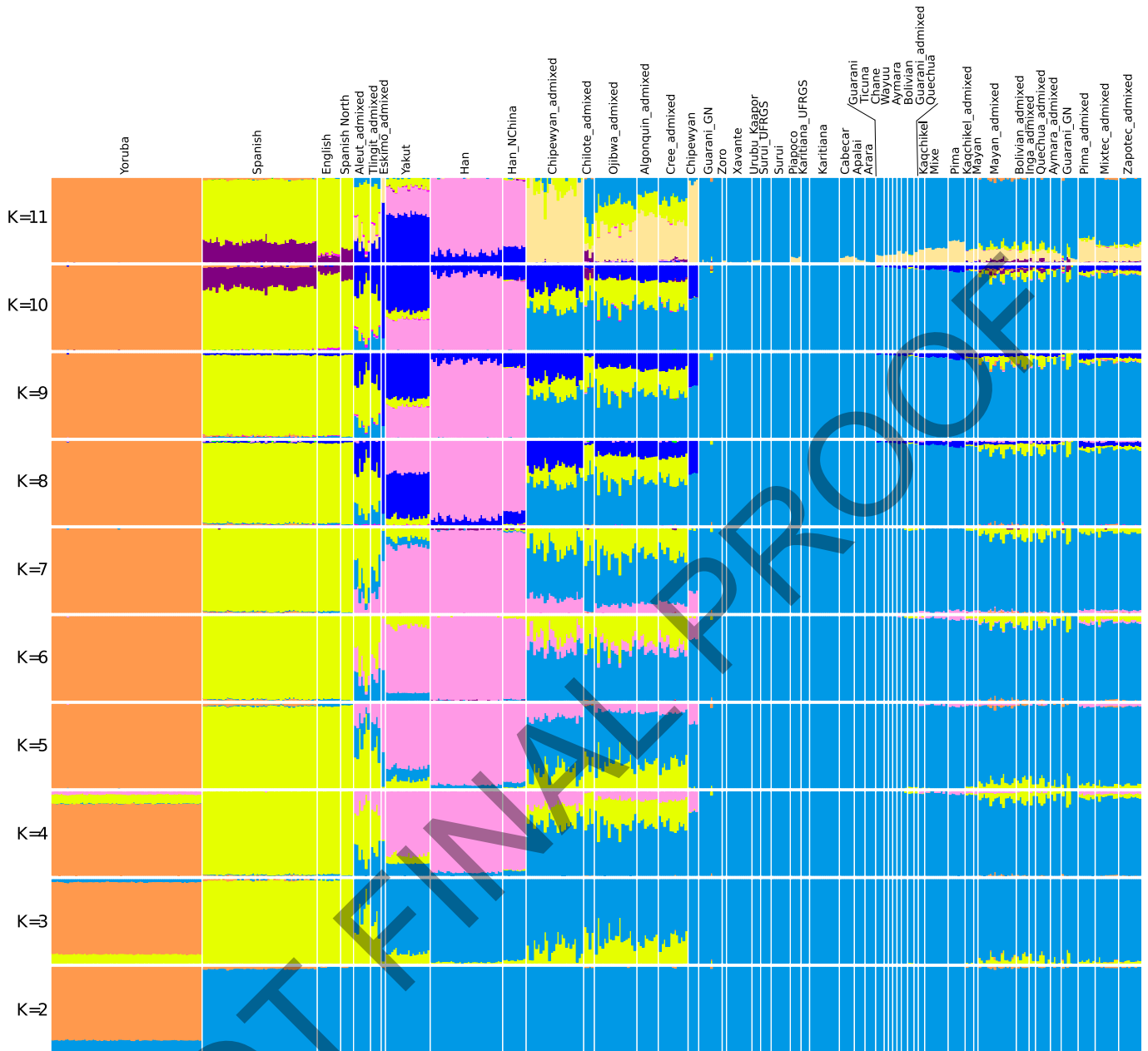
Extended Data Figure 1 | Clustering analysis. ADMIXTURE[38] clustering analysis performed on the Affymetrix Human Origins data used in this study. To aid in visualization, we only show results for Native American samples and for selected samples from Eurasian populations.
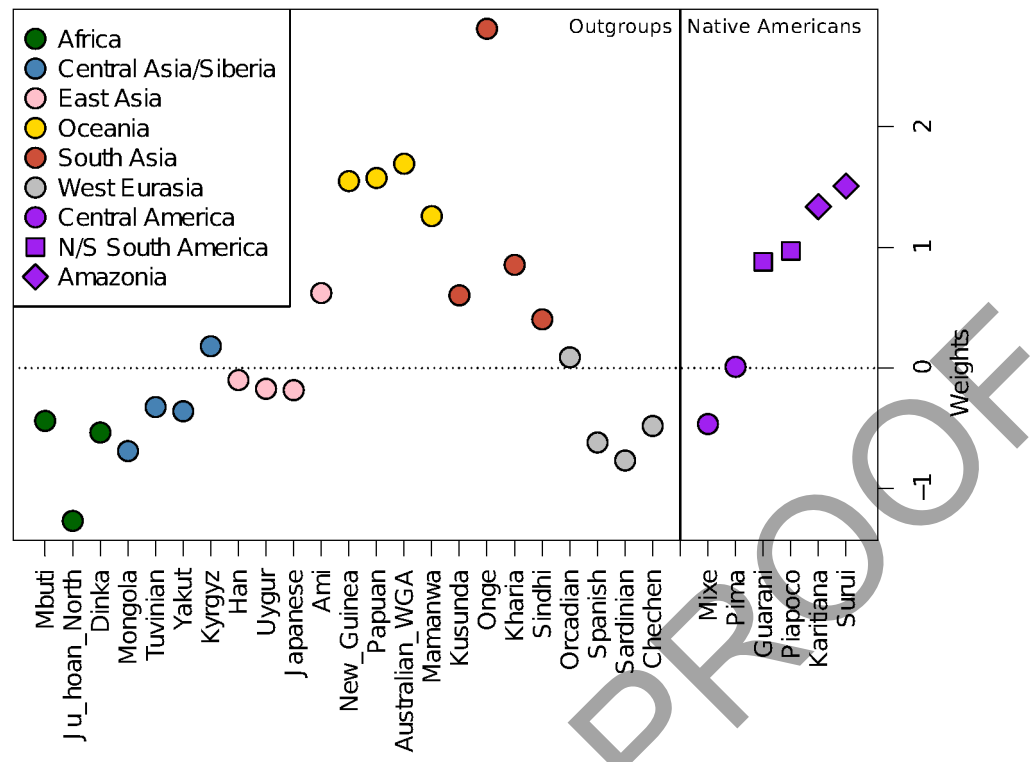
Extended Data Figure 2 | qpWave coefficients. Weights from qpWave for Native American populations and for non-American outgroup populations. No weights are given for Yoruba and Cabecar, as they are used in the computation.

a

b
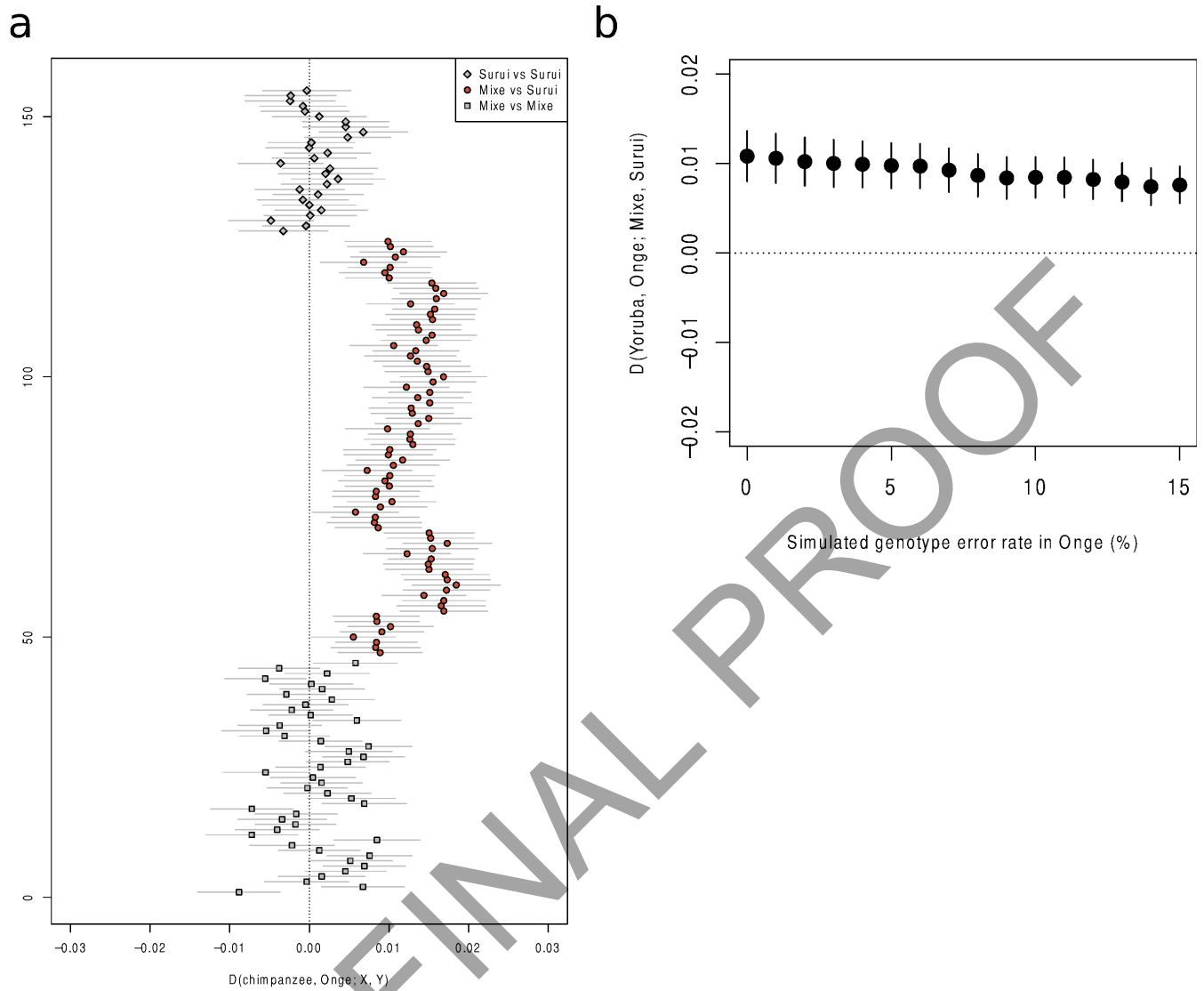
Extended Data Figure 3 | Excess allele sharing between the Surui and the Onge. a, Tests for excess shared derived alleles with the Onge in all possible comparisons of 8 Suruí and 10 Mixe individuals. All Mixe–Suruí comparisons show a positive skew whereas all Mixe–Mixe and Suruí–Suruí comparisons are consistent with 0. Lines correspond to one standard error in either direction. **b**, Random sequence or genotype errors cannot explain the affinity of the Amazonians to Australasians, as simulated increased errors in the Onge do not cause an increased affinity to Suruí.

a

b
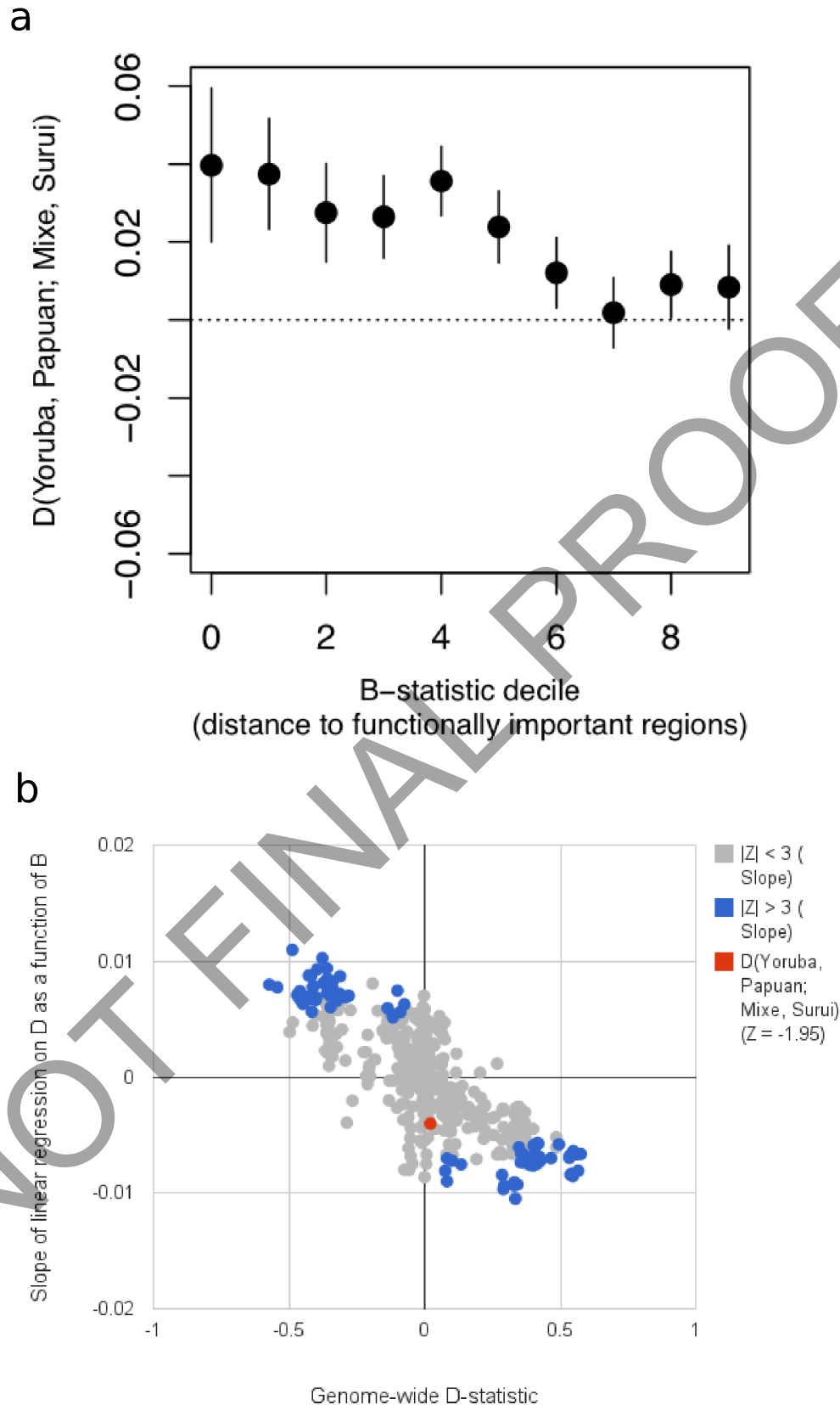
Extended Data Figure 4 | Signals of admixture as a function of proximity to functional regions. a, The affinity of 16 Papuan high-coverage genomes to 2 Amazonian Suruí high-coverage genomes as a function of proximity to regions of functional importance (measured by B-value). **b**, A total of 395 tests of quartets $D(Yoruba, X; Y, Z)$ shows that quartets with significantly positive slopes ($|Z| > 3$) also yield significant genome-wide D-statistics of the opposite sign. This suggests that signals of admixture are systematically stronger close to functionally important regions.
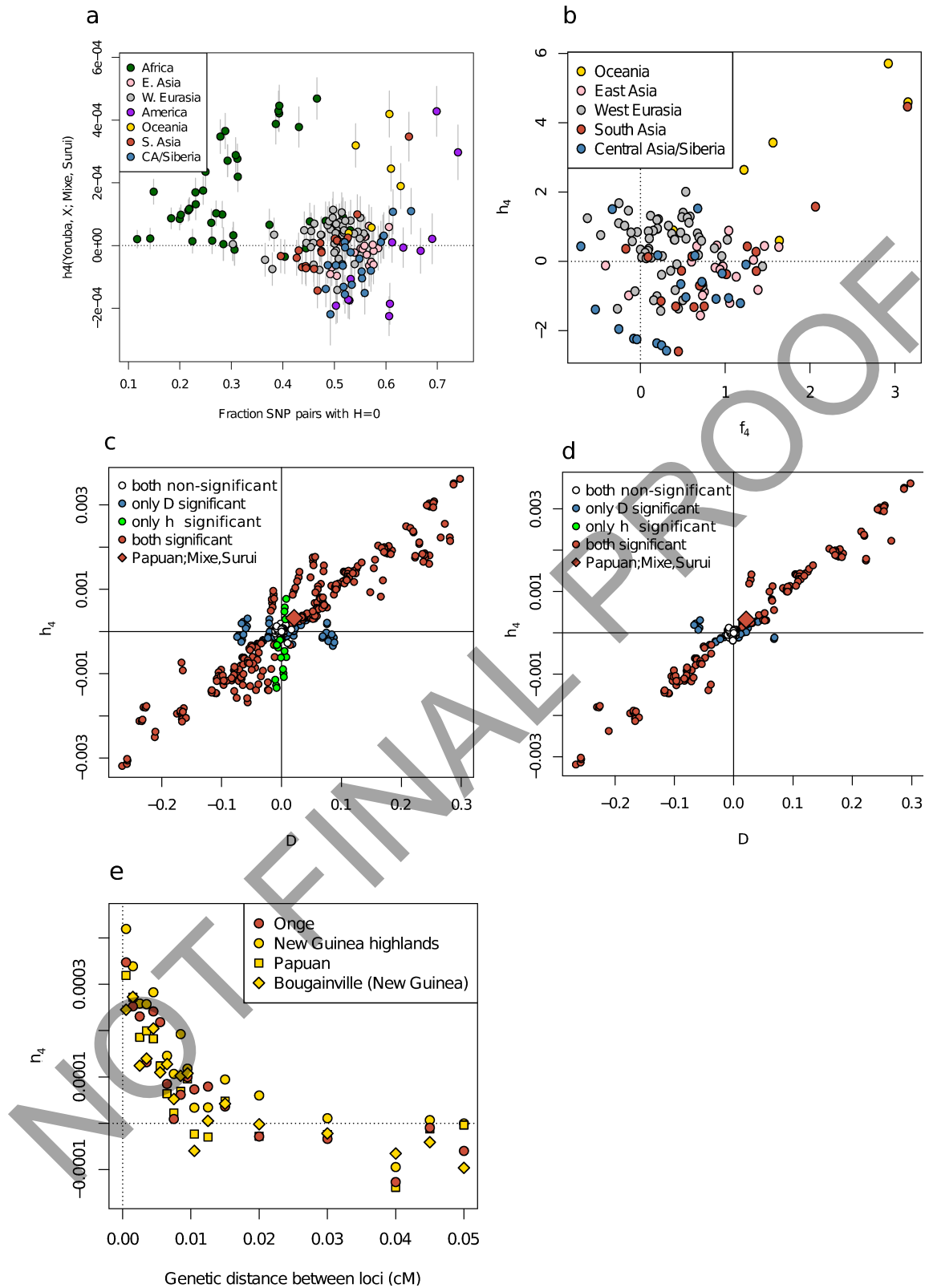
Extended Data Figure 5 | Linkage disequilibrium-based symmetry tests.
a, h_4(Yoruba, X; Mixe, Suruí) for SNP pairs within 0.01 cM of each other contrasted with the fraction of SNP pairs in linkage equilibrium in population X ($H = 0$). Error bars show ± 1 s.e. **b**, Scatterplot of Z-scores for the f_4- and h_4-statistics for the same quartets. For both these panels we only use populations with at least 6 samples. **c, d**, We computed D(Yoruba, X; Y, Z) and h_4(Yoruba, X; Y, Z) for many combinations of populations as X, Y and Z using

phased Affymetrix Human Origins SNP array data ascertained in a Yoruba individual. Except for Africans who have ancestry from lineages that diverged before the Yoruba used for ascertainment and Oceanians (who have archaic Denisovan ancestry) we observe that $|Z| > 3$ h_4-statistics are always associated with a significantly positive D for the same quartet. **e**, Correlation of the h_4-statistic with the genetic distance separation of pairs of SNPs for h_4(Yoruba, X; Mixe, Suruí).

a. Single origin (no fit)
b. Onge-related mixture in Amazonia (fit)
c. Asian mixture in Amazonia (no fit)
d. Asian admixture in Mesoamerica (no fit)
e. MA-1 related mixture in Mesoamerica (no fit)
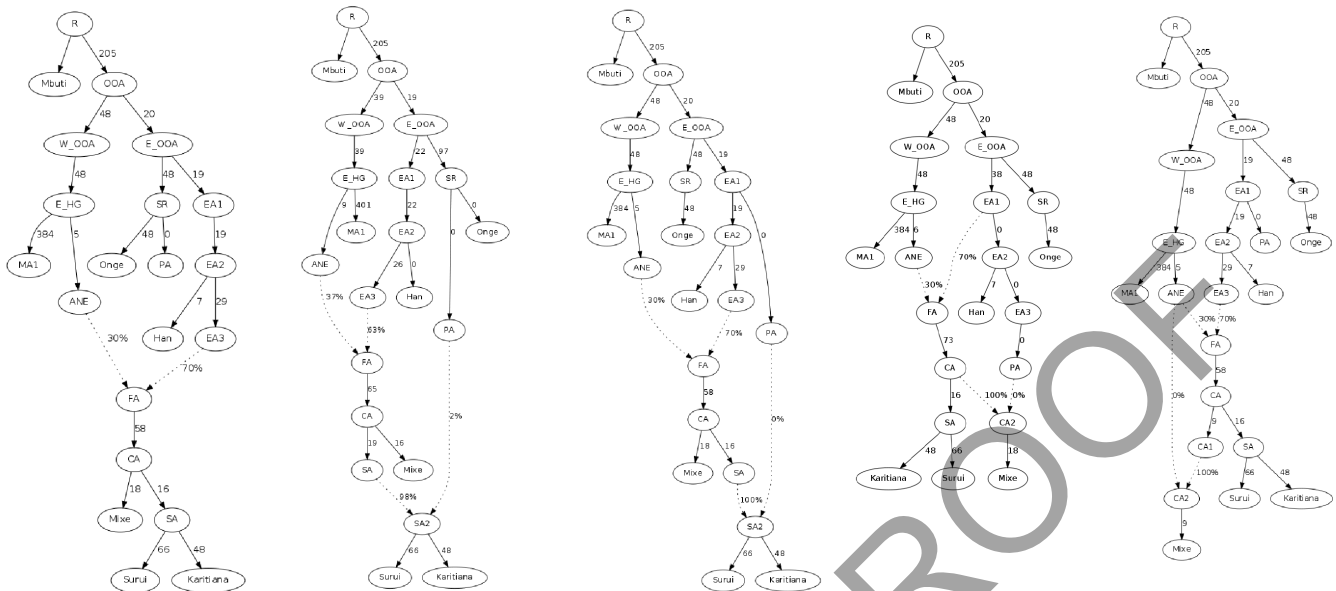
Extended Data Figure 6 | Admixture graphs for fitted population history models. a, An admixture graph where all of Mixe, Suruí and Karitiana are of 100% First American ancestry is rejected with 6 predicted f-statistics at least 3 standard errors from the empirically observed value. **b**, An admixture graph where the ancestors of Suruí and Karitiana receive 2% ancestry from a lineage related to the Onge is consistent with the data with no outliers. **c**, An admixture graph where the distinct ancestry in Amazonians is more closely related to Han than to Onge produces 6 outliers. **d**, An admixture graph with no distinctive ancestry in Karitiana or Suruí but East Asian gene flow into the Mixe produces 7 outliers. **e**, An admixture graph with no distinctive ancestry in Karitiana or Suruí but MA1-related gene flow into the Mixe produces 6 outliers.

a

b
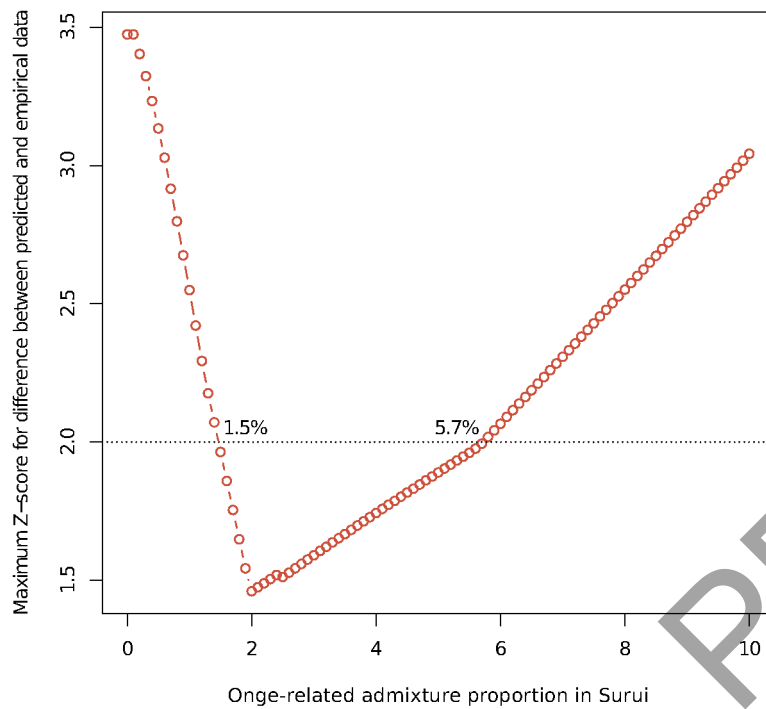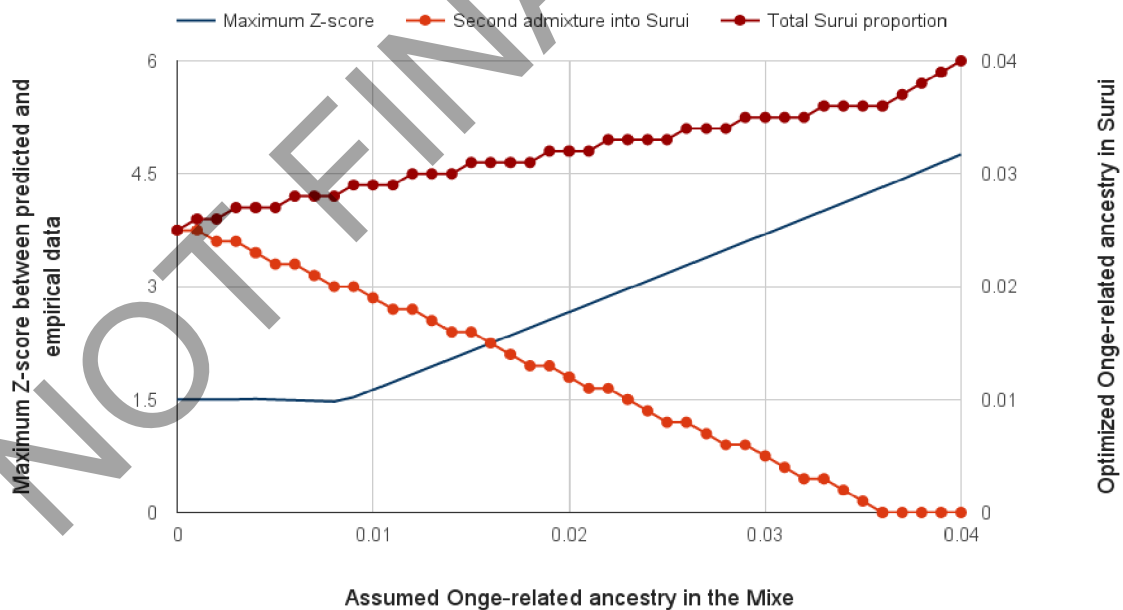
Extended Data Figure 7 | Plausible range for the non-First American admixture proportion in Amazonians. **a**, Range obtained assuming entirely First American ancestry in the Mixe. **b**, The maximum proportion of non-First American ancestry in the Mixe that is consistent with the data.

Extended Data Table 1 | qpWave analysis provides evidence that Central and South American genetic variation is inconsistent with being derived from a single homogeneous population

	P-value for this number of streams			
	1	2	3	4
Full data	2.03E-07	0.09	0.58	0.92
Outgroup region dropped				
Africa	1.67E-04	0.34	0.92	0.95
C. Asia/Siberia	5.91E-07	0.11	0.6	0.89
East Asia	4.46E-09	0.04	0.57	0.92
South Asia	6.95E-05	0.1	0.4	0.82
West Eurasia	1.41E-05	0.06	0.37	0.89
Oceania	4.39E-05	0.43	0.88	0.97
Native American population dropped				
Cabecar	1.13E-08	0.02	0.27	0.73
Guarani	9.50E-07	0.27	0.76	0.99
Karitiana	1.41E-06	0.1	0.61	0.86
Mixe	8.32E-03	0.2	0.91	0.98
Piapoco	1.30E-04	0.55	0.93	0.98
Pima	2.19E-05	0.31	0.78	0.97
Surui	2.35E-06	0.11	0.56	0.84
Africa + 1 other region				
Siberia	0.16	0.56	0.8	0.87
East Asia	0.06	0.29	0.71	0.72
South Asia	0.004	0.57	0.93	0.96
West Eurasia	0.02	0.23	0.62	0.67
Oceania	0.01	0.25	0.82	0.99
Siberia + 1 other region				
Africa	0.16	0.56	0.8	0.87
East Asia	0.41	0.91	0.99	1
South Asia	0.03	0.82	0.91	0.9
W. Eurasia	0.2	0.59	0.77	0.83
Oceania	0.003	0.18	0.75	0.93

Extended Data Table 2 | Top 20 *D*-statistics observed for *D*(chimpanzee, Old World population; Central Americans, Amazonians)

Rank	Population	*D*	SE	*Z*	Region 1	Region 2
1	Onge	0.0101	0.0022	4.60	India	South Asia
2	Papuan	0.0084	0.0022	3.82	Papua New Guinea	Oceania
3	New_Guinea	0.0082	0.0023	3.54	Papua New Guinea	Oceania
4	Australian_WGA	0.0074	0.0024	3.12	Australia (Arnhem Land)	Oceania
5	Mamanwa	0.0068	0.0020	3.40	Philippines (Negrito)	Oceania
6	Bougainville	0.0065	0.0023	2.85	Papua New Guinea	Oceania
7	Kharia	0.0059	0.0020	2.97	India	South Asia
8	Tongan	0.0058	0.0022	2.68	Tonga	Oceania
9	Bengali	0.0058	0.0019	3.00	Bangladesh	South Asia
10	Mala	0.0055	0.0019	2.93	India	South Asia
11	Ami	0.0052	0.0020	2.61	Taiwan	East Asia
12	Lodhi	0.0052	0.0019	2.72	India	South Asia
13	Sindhi	0.0051	0.0019	2.72	Pakistan	South Asia
14	Kusunda	0.0050	0.0020	2.56	Nepal	South Asia
15	Lahu	0.0050	0.0021	2.37	China	East Asia
16	Kinh	0.0049	0.0020	2.46	Vietnam	East Asia
17	Australian	0.0048	0.0025	1.96	Australia	Oceania
18	Balochi	0.0047	0.0019	2.55	Pakistan	South Asia
19	Thai	0.0047	0.0020	2.38	Thailand	East Asia
20	Semende	0.0045	0.0020	2.27	Indonesia (Sumatra)	Oceania

Extended Data Table 3 | f_4-statistics for which the statistic predicted by the fitted admixture graphs deviates by more than $|Z| > 3$ from the statistic computed on the empirical data

A	*B*	*X*	*Y*	**Predicted f_4**	**Empirical f_4**	**Z-score**
Single First American origin (Extended Data Figure 6A)						
Mbuti	Onge	Mixe	Surui	0	0.003506	3.535
Mbuti	Onge	Mixe	Karitiana	0	0.00315	3.431
Onge	Mixe	Mixe	Surui	-0.018466	-0.021724	-3.061
Onge	Mixe	Mixe	Karitiana	-0.018466	-0.021849	-3.226
Onge	Han	Mixe	Surui	0	-0.002902	-3.654
Onge	Han	Mixe	Karitiana	0	-0.00239	-3.279
Onge-related ancestry in the Amazon (Extended Data Figure 6B)						
(No outliers)						
East Asian admixture in South America (Extended Data Figure 6C)						
Mbuti	Onge	Mixe	Surui	0	0.003506	3.535
Mbuti	Onge	Mixe	Karitiana	0	0.00315	3.431
Onge	Mixe	Mixe	Surui	-0.018466	-0.021724	-3.061
Onge	Mixe	Mixe	Karitiana	-0.018466	-0.021849	-3.226
Onge	Han	Mixe	Surui	0	-0.002902	-3.654
Onge	Han	Mixe	Karitiana	0	-0.00239	-3.279
East Asian admixture in Central America (Extended Data Figure 6D)						
Mbuti	Onge	Mixe	Surui	-0.000002	0.003506	3.537
Mbuti	Onge	Mixe	Karitiana	-0.000002	0.00315	3.433
Onge	Mixe	Mixe	Surui	-0.018466	-0.021724	-3.061
Onge	Mixe	Mixe	Karitiana	-0.018466	-0.021849	-3.225
Onge	Han	Mixe	Surui	-0.000004	-0.002902	-3.649
Onge	Han	Mixe	Karitiana	-0.000004	-0.00239	-3.273
Ancient Siberian (MA1) admixture in Central America (Extended Data Figure 6E)						
Mbuti	Onge	Mixe	Surui	0	0.003506	3.535
Mbuti	Onge	Mixe	Karitiana	0	0.00315	3.431
Onge	Mixe	Mixe	Surui	-0.018470	-0.021724	-3.057
Onge	Mixe	Mixe	Karitiana	-0.018470	-0.021849	-3.222
Onge	Han	Mixe	Surui	0	-0.002902	-3.654
Onge	Han	Mixe	Karitiana	0	-0.00239	-3.279