Article

Native American gene flow into Polynesia predating Easter Island settlement

Alexander G. Ioannidis[1,2,18 ✉], Javier Blanco-Portillo[2,18], Karla Sandoval[2], Erika Hagelberg[3], Juan Francisco Miquel-Poblete[4], J. Víctor Moreno-Mayar[5], Juan Esteban Rodríguez-Rodríguez[2], Consuelo D. Quinto-Cortés[2], Kathryn Auckland[6], Tom Parks[6], Kathryn Robson[7], Adrian V. S. Hill[6,8], María C. Avila-Arcos[9], Alexandra Sockell[10], Julian R. Homburger[10], Genevieve L. Wojcik[10], Kathleen C. Barnes[11], Luisa Herrera[12], Soledad Berríos[12], Mónica Acuña[12], Elena Llop[12], Celeste Eng[13], Scott Huntsman[13], Esteban G. Burchard[13], Christopher R. Gignoux[11], Lucía Cifuentes[12], Ricardo A. Verdugo[12,14], Mauricio Moraga[12,15], Alexander J. Mentzer[6,16], Carlos D. Bustamante[10,17] & Andrés Moreno-Estrada[2 ✉]

The possibility of voyaging contact between prehistoric Polynesian and Native American populations has long intrigued researchers. Proponents have pointed to the existence of New World crops, such as the sweet potato and bottle gourd, in the Polynesian archaeological record, but nowhere else outside the pre-Columbian Americas[1–6], while critics have argued that these botanical dispersals need not have been human mediated[7]. The Norwegian explorer Thor Heyerdahl controversially suggested that prehistoric South American populations had an important role in the settlement of east Polynesia and particularly of Easter Island (Rapa Nui)[2]. Several limited molecular genetic studies have reached opposing conclusions, and the possibility continues to be as hotly contested today as it was when first suggested[8–12]. Here we analyse genome-wide variation in individuals from islands across Polynesia for signs of Native American admixture, analysing 807 individuals from 17 island populations and 15 Pacific coast Native American groups. We find conclusive evidence for prehistoric contact of Polynesian individuals with Native American individuals (around AD 1200) contemporaneous with the settlement of remote Oceania[13–15]. Our analyses suggest strongly that a single contact event occurred in eastern Polynesia, before the settlement of Rapa Nui, between Polynesian individuals and a Native American group most closely related to the indigenous inhabitants of present-day Colombia.

A perennial question in Oceanian history concerns the possibility of prehistoric contacts between Polynesian and Native American populations. Previous researchers investigating this question through genetics have focused on Easter Island (Rapa Nui). As the closest inhabited Polynesian island to the Americas, and the Polynesian island with the most elaborate megalithic culture[16], Rapa Nui has been considered a likely locus for contact. High-resolution analyses of human leukocyte antigen (HLA) alleles have revealed a Native American component in modern individuals with self-identified Rapanui ancestry[8,9]. However, in the only two genome-wide studies of Rapanui variation—one of eight modern individuals[10], and one of five skeletal remains (three from pre-European contact era and two from post-European contact)[11]—a Native American component was found in all samples of the former, but none of the

latter. As a consequence, these studies reached opposing conclusions about pre-European contact between Polynesian individuals on Rapa Nui and Native American individuals[10,11]. To date no genome-wide DNA studies have considered the possibility of pre-European Native American contact on other Polynesian islands. Here we investigate both of these questions through high-density genome-wide analyses of a large data set of 166 Rapanui and 188 additional individuals from islands spanning the Pacific (Fig. 1a and Supplementary Tables 1, 2).

Multiple admixture events in Polynesia

We first performed a global ancestry analysis of our Polynesian and coastal Native American samples together with continental reference

[1]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. [2]National Laboratory of Genomics for Biodiversity (LANGEBIO), Unit of Advanced Genomics, CINVESTAV, Irapuato, Mexico. [3]Department of Biosciences, University of Oslo, Blindern, Oslo, Norway. [4]Departamento de Gastroenterología, Facultad de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile. [5]National Institute of Genomic Medicine (INMEGEN), Mexico City, Mexico. [6]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [7]MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [8]The Jenner Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [9]International Laboratory for Human Genome Research (LIIGH), UNAM Juriquilla, Queretaro, Mexico. [10]Center for Computational, Evolutionary and Human Genomics (CEHG), Stanford University, Stanford, CA, USA. [11]Division of Biomedical Informatics and Personalized Medicine, University of Colorado, Denver, CO, USA. [12]Human Genetics Program, Institute of Biomedical Sciences, Faculty of Medicine, University of Chile, Santiago, Chile. [13]Program in Pharmaceutical Sciences and Pharmacogenomics, Department of Medicine, University of California San Francisco, San Francisco, CA, USA. [14]Basic-Applied Oncology Department, Faculty of Medicine, University of Chile, Santiago, Chile. [15]Department of Anthropology, Faculty of Social Sciences, University of Chile, Santiago, Chile. [16]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. [17]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [18]These authors contributed equally: Alexander G. Ioannidis, Javier Blanco-Portillo. ✉e-mail: ioannidis@stanford.edu; andres.moreno@cinvestav.mx
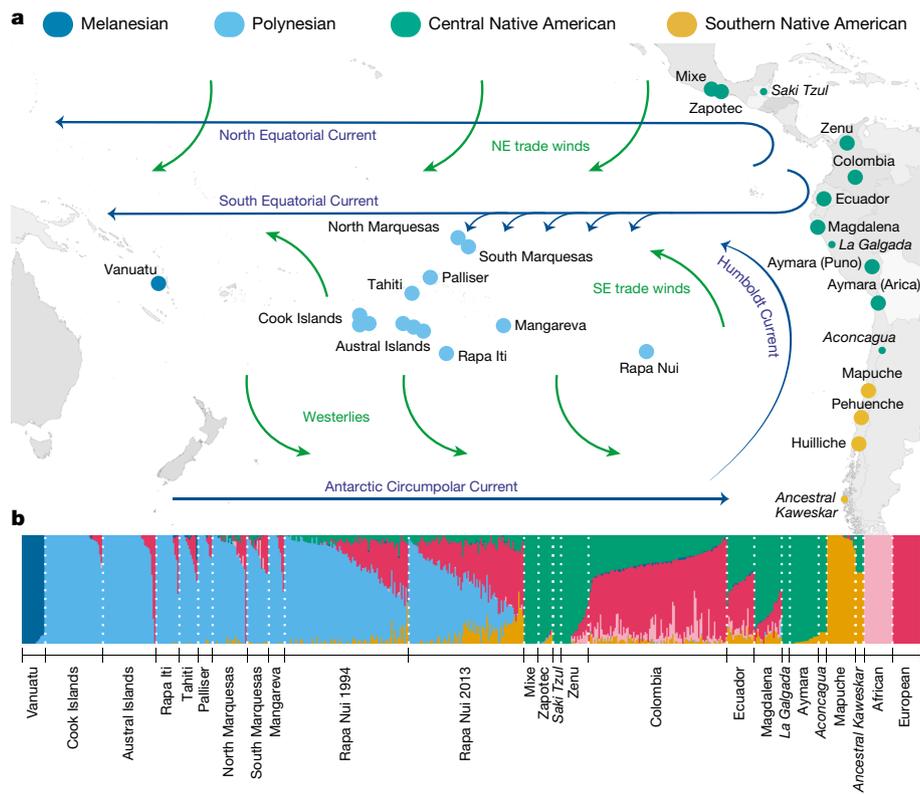
Fig. 1 | Sampled populations with unsupervised (iterative) ADMIXTURE analysis. a, Map, including prevailing currents and winds, showing the location of each sampled population with dots coloured by region (see key at top). **b**, $K = 6$ clustering analysis of Pacific islanders and reference populations using the ADMIXTURE method[17]. The reference panels include populations from: Europe (UK and Spain) and Africa (Yoruba)[43], the Americas (Mapuche, including Pehuenche and Huilliche, from central and south Chile[44]; Aymara from southern Peru and northern Chile; northern Peruvian individuals from Magdalena de Cao; Zenu from Colombia; and Zapotec and Mixe from southern Mexico[45]), and at the far-left Melanesian individuals from Vanuatu (see Supplementary Fig. 5). Each individual is represented as a narrow column, coloured to show the proportion of each ancestry cluster in that individual. We included modern Colombian and Ecuadorian individuals[46] as well as four ancient (pre-European contact) individuals (italics, wide columns), spaced along the coast (small dots), to further illustrate the Native American component[44,47,48], but did not use these populations/individuals as reference panels owing to their lower marker density. See Supplementary Table 5 for the distinction between the nomenclature for early modern and ancient Oceanian clusters. NE, northeast; SE, southeast.

populations using the ADMIXTURE algorithm[17] (Fig. 1b, Supplementary Figs. 1–8, Supplementary Tables 3–5 and Supplementary Discussion) and principal component analysis (PCA; Supplementary Fig. 9). We followed these variant-frequency-based analyses with an independent, sequence-matching-based analysis (local ancestry inference[18]) in order to identify precise genomic regions of Polynesian, Native American, European and African origin in each individual for use in later ancestry-specific analyses (Fig. 2a and Supplementary Table 6).

In all of these ancestry analyses, the Pacific island populations are characterized by a large Polynesian component, but with many islanders also having a European component from colonial admixture. Remarkably, in both of these independent analyses, as well as in f_4 and D-statistic analyses ($P < 0.001$), we also detect admixture in eastern Polynesia from Native American individuals, even when using pre-European-contact Native American reference panels (Supplementary Figs. 10, 11). Looking at the ADMIXTURE plot, in the easternmost Polynesian islands (Palliser, Marquesas, Mangareva and Rapa Nui), but on no other Polynesian islands, two Native American ancestry components can be seen. These components are characteristic of central (green in the figures) and southern (yellow) Native American populations (both modern and ancient). The southern Native American component—highest in the Mapuche and Pehuenche native peoples of Chile—increases in present-day Rapanui individuals in proportion to their European (red) ancestry component (Fig. 1b, Supplementary Fig. 12 and Supplementary Tables 7, 8). This is consistent with the idea that the Native American

component arrived on Rapa Nui together with a Spanish European component via immigration of admixed Chilean individuals following Chile's annexation of the island. By contrast, the central Native American component, characteristic of indigenous Mexican individuals (Mixe and Zapotec) and indigenous Colombian individuals (Zenu), is associated on Rapa Nui only with the Polynesian component, not with the European or southern Native American components, according to the log-ratio variances of those components (Fig. 2b and Supplementary Tables 7–10). This suggests that the central Native American component arrived onto Rapa Nui independently from the European component. Furthermore, in contrast to the southern Native American component (Chilean), the central Native American component varies little between Rapanui individuals, indicating that it stems from an older admixture event[19,20]. Indeed, the Native American DNA segments in Rapanui individuals have an aggregate length distribution that indicates initial contact several centuries before European individuals entered the Pacific (Fig. 2c). Intriguingly, the central Native American component (green) is found in the other remote eastern Polynesian islands (Palliser, Marquesas and Mangareva) and has a similarly early date (Fig. 2d).

When we investigate the European ancestry in Polynesian individuals, we see correspondences with the European nations that colonized each island. For example, the European component in French Polynesian individuals clusters with French reference panels in our new ancestry-specific multidimensional scaling (MDS) analysis (Fig. 3a).
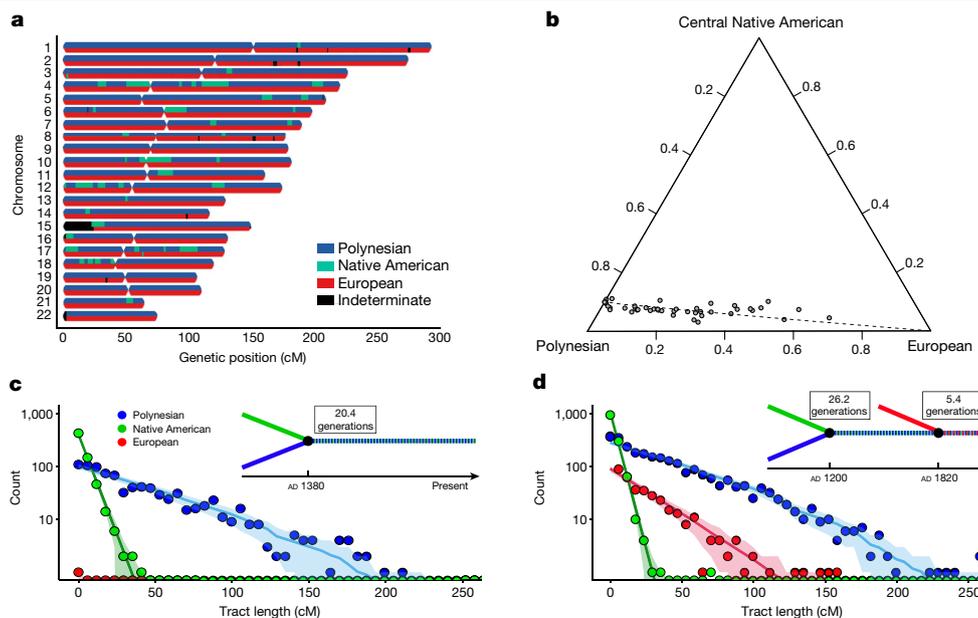
Fig. 2 | Relationship between Polynesian, Native American and European ancestries. a, Random-forest-based local ancestry inference of a Rapanui individual, showing small (old) Native American ancestry tracts embedded in Polynesian ancestry tracts. The ancestry of each haploid genome is coloured (top and bottom for each homologous chromosome of a pair); the autosome pairs are numbered along the vertical axis and genetic position in centimorgans (cM) is on the horizontal axis. **b**, Ternary plot of ADMIXTURE ancestry fractions in Rapanui individuals having Polynesian, European and central Native American, but no other, ancestries (each point corresponds to an individual). The first principal component in the centred log-ratio transform space[49] is projected onto the figure as a dashed curve. The ancestries' log-ratio variances are discussed in Supplementary Tables 7–10. **c**, **d**, Length-distribution analyses for ancestry tracts in the six Rapanui individuals with no European ancestry (**c**) and in North Marquesan individuals (**d**). Plotted points show the aggregate tract length counts; lines show the maximum-likelihood best-fit tract length distributions; and shading shows one standard deviation confidence intervals, assuming Gaussian noise. The best-fit admixture chronology is plotted above the timeline as a line history, with each colour representing an ancestry, as indicated in the key, and with generation times given from the admixture points (black) to the present.

Those Rapanui individuals with southern Native American ancestry in the ADMIXTURE analysis are shifted towards the Spanish reference panels in this European-specific MDS analysis, consistent with those two ancestries arriving together via immigration of admixed Chilean individuals. The remaining Rapanui individuals that have European ancestry, but no southern Native American ancestry, cluster largely with French reference panels, which is consistent with the French origin of the first European residents on Rapa Nui[21]. We also analyse long, shared DNA segments (larger than 7 centimorgans) that are inherited by related individuals and are termed 'identical by descent' (IBD). Analysing genomic regions of only European ancestry, we find that the IBD relationship network mirrors European settlement patterns, with many French Polynesian islands forming one connected component, separate from Rapa Nui (Fig. 3b and Supplementary Table 11). Rapa Nui's single European connection to Mangareva may reflect the transfer[21] of French Catholic missionaries from Rapa Nui to Mangareva in AD 1871.

Native American ancestry in Polynesia

The Native American ancestry in eastern Polynesian individuals shows a very different pattern of interisland IBD sharing, indicating a different history of Native American contact (Fig. 4a, Supplementary Fig. 13 and Supplementary Table 12). To characterize the origin of this ancestry in Polynesia more precisely, we applied a new ancestry-specific PCA to the Native American component (see Methods and Supplementary Figs. 14, 15). The first principal component is found to order the Native American reference individuals along a north–south axis, coinciding roughly with the Pacific coast of the Americas (Supplementary Fig. 14). We plot the density of the Native American reference individuals along this first principal component axis together with the location of the aggregate Native American components for each of the eastern

Polynesian islands possessing such ancestry (Fig. 4b). Consistent with our ADMIXTURE analysis, which showed a central Native American component in Pacific islanders, in this analysis the Native American ancestries of the Pacific islanders all fall within, or beside, the Zenu people—an indigenous Colombian population. The localization of the Native American component to Colombia–Ecuador is shown clearly by our new, lower-noise, ancestry-specific MDS analysis, as well as by PCA, and is consistent with the less-sensitive traditional Procrustes analysis and outgroup-f_3 statistic (Supplementary Figs. 16–22). The only exceptions are the Rapanui individuals with high European ancestry. As expected, their Native American component, which probably came together with their European component through immigration of admixed Chilean individuals to Rapa Nui, is located squarely within the Pehuenche and Mapuche native populations of central Chile (Fig. 4b and Supplementary Figs. 14, 16, 18–22). The Native American ancestry component in Rapanui individuals with no European ancestry, by contrast, clusters with the Colombian Zenu people, just as with the other eastern islands.

Apart from the Chilean annexation of Rapa Nui in AD 1888 and sporadic interactions with ships' crews, the only recorded events potentially connecting Pacific islanders with Native American ancestry are the Peruvian slave raids of AD 1862–1863. During this year, thousands of Pacific islanders were kidnapped and taken to Peru as forced labourers, including 1,407 Rapanui individuals[22]. Following an international outcry, a few repatriation voyages were organized, but smallpox outbreaks onboard meant that only a handful of passengers made it back to Polynesia alive. Only two of the islands in our data set received any recorded returnees: Rapa Nui (15 repatriated) and Rapa Iti (9 captive individuals from other islands resettled). With very few individuals, all self-identifying as islanders, returning to Polynesia, and with their captivity in Peru lasting only a few months, it is unlikely that this episode
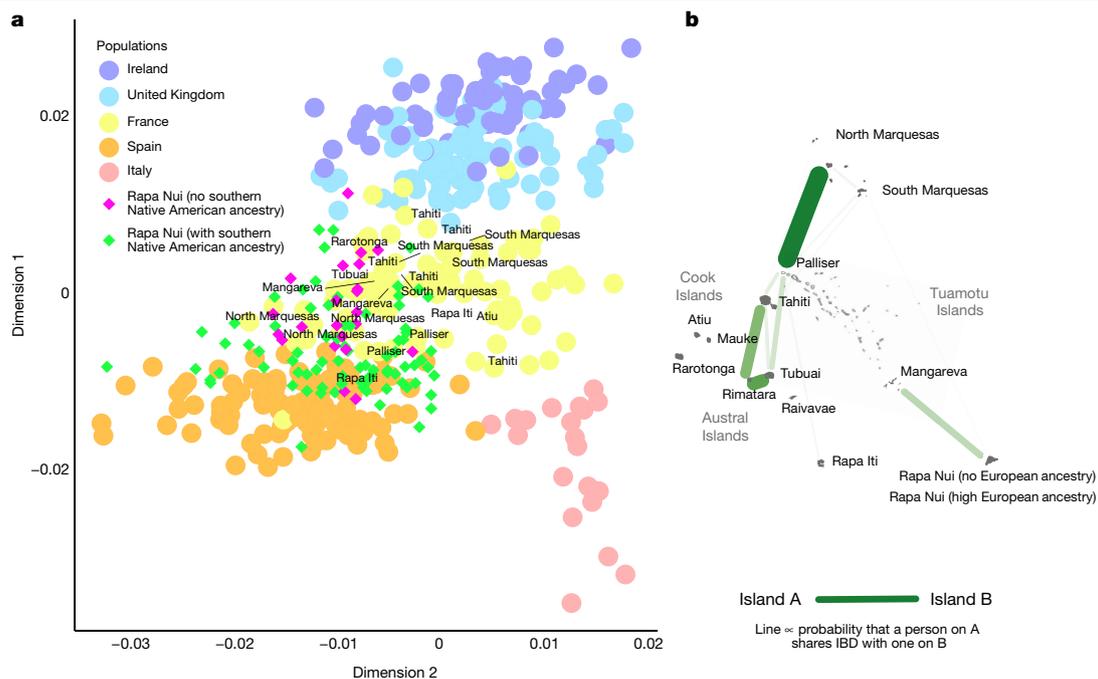
Fig. 3 | Analysis of European ancestry in Pacific islanders. a, Our new ancestry-specific MDS applied to the European ancestry of each admixed sample from the Pacific islands, together with European reference individuals from the Population Reference Sample (POPRES) data set[50], shows French Polynesian islanders (labelled within the plot) clustering with French individuals. In addition, Rapanui individuals (diamonds) cluster with Spain or France, depending on whether (green) or not (violet) they also have southern Native American (Chilean) ancestry. The numbers of samples from each country are given in Supplementary Table 3. **b**, IBD sharing of European ancestry segments in Polynesia is strongest (darker and thicker lines indicate a higher probability of sharing; see key at bottom) between island clusters with the same European colonial backgrounds. The islands' sample sizes are given in Supplementary Table 1.

resulted in any introgression of Native American ancestry into Polynesia. However, such explanations have been advanced[11,23]. In any case, the Native American component that we observe in the easternmost islanders, including on distant islands untouched by returnee voyages, derives from an indigenous American population lying to the north of both of our Peruvian Native American reference individuals, namely the southern Peruvian Aymara and the northern Peruvian Magdalena (Fig. 4b and Supplementary Figs. 16, 18).

Our localization of the Native American ancestry found in Polynesia is consistent with several linguistic, historical and geographical observations that support an origin in northern South America. Although superficial similarities between the monolithic statues of the Pacific islands (found only in the remote eastern Polynesian islands) and those of the pre-Columbian site of San Augustín, Colombia, have long been noted[2], stronger evidence has come from the Polynesian word for the sweet potato, 'kumala'. This word has been linked to names for the food in northern South America, where it originated[2,3,24]. The coastal languages that use these related names lie to the north of Peru—for example, 'cumal' is used by the Cañari people of Ecuador[25]—whereas the Peruvian languages that use such names are Andean and located far from the coast. It is to the north of Peru that the Pacific coast changes from desert to forests suitable for boat construction, and it is from Pacific Ecuador and Colombia that Native American voyagers are believed[26–29] to have embarked for trade with Mesoamerica in large ocean-going sailing rafts made of balsa wood during the period AD 600 to AD 1200. Wind and current simulations from the Pacific coast of the Americas have shown that drift voyages departing from Ecuador and Colombia are the most likely to reach Polynesia, and that they arrive with the highest probability in the South Marquesas islands, followed by the Tuamotu Archipelago[4]. Both of these archipelagos lie at the heart of the region of islands where we have found a Colombian Native American component. The trade winds and the south equatorial current move

east to west at these latitudes, funnelling boaters from northern South America to the archipelagos[26] (Fig. 1a). (In Thor Heryerdahl's famous drift voyage from Peru to Polynesia, his Kon-Tiki raft had to be towed 80 miles offshore from Peru, because the southern current along the Peruvian coast was so unfavourable; once in the trans-Pacific currents the Kon-Tiki raft landed in the Tuamotu Archipelago.) For the same reason, these archipelagos would be the most likely origin for Polynesian individuals discovering the Americas using their characteristic upwind exploration[30,31].

Dating Native American–Polynesian contact

To determine when the Native American component was introduced into each of the affected Polynesian populations in our data set (Nuku Hiva in the North Marquesas, Fatu Hiva in the South Marquesas, Palliser in the Tuamotu Archipelago, Mangareva, and Rapa Nui), we modelled the length distribution of the Native American, European and Polynesian ancestry segments of the islanders using the Tracts method[32] (Supplementary Figs. 23, 24). For all island populations—with one expected exception discussed below—we find that the model with the highest likelihood involves an initial Native American–Polynesian admixture event, followed centuries later by European introgression (Supplementary Table 13). Those later estimated European admixture dates (North Marquesas AD 1820, South Marquesas AD 1830, Mangareva AD 1750 and Palliser AD 1790) fall within the period of European colonization of Polynesia. By contrast, the dates estimated for Native American–Polynesian admixture on the islands are much earlier, and they are similar across the different islands (Mangareva AD 1230, Palliser AD 1230, North Marquesas AD 1200 and South Marquesas AD 1150).

The only exception to these consistently early dates is on Rapa Nui itself, where Rapanui individuals with no European (colonial) ancestry have a slightly later estimated Native American introgression date
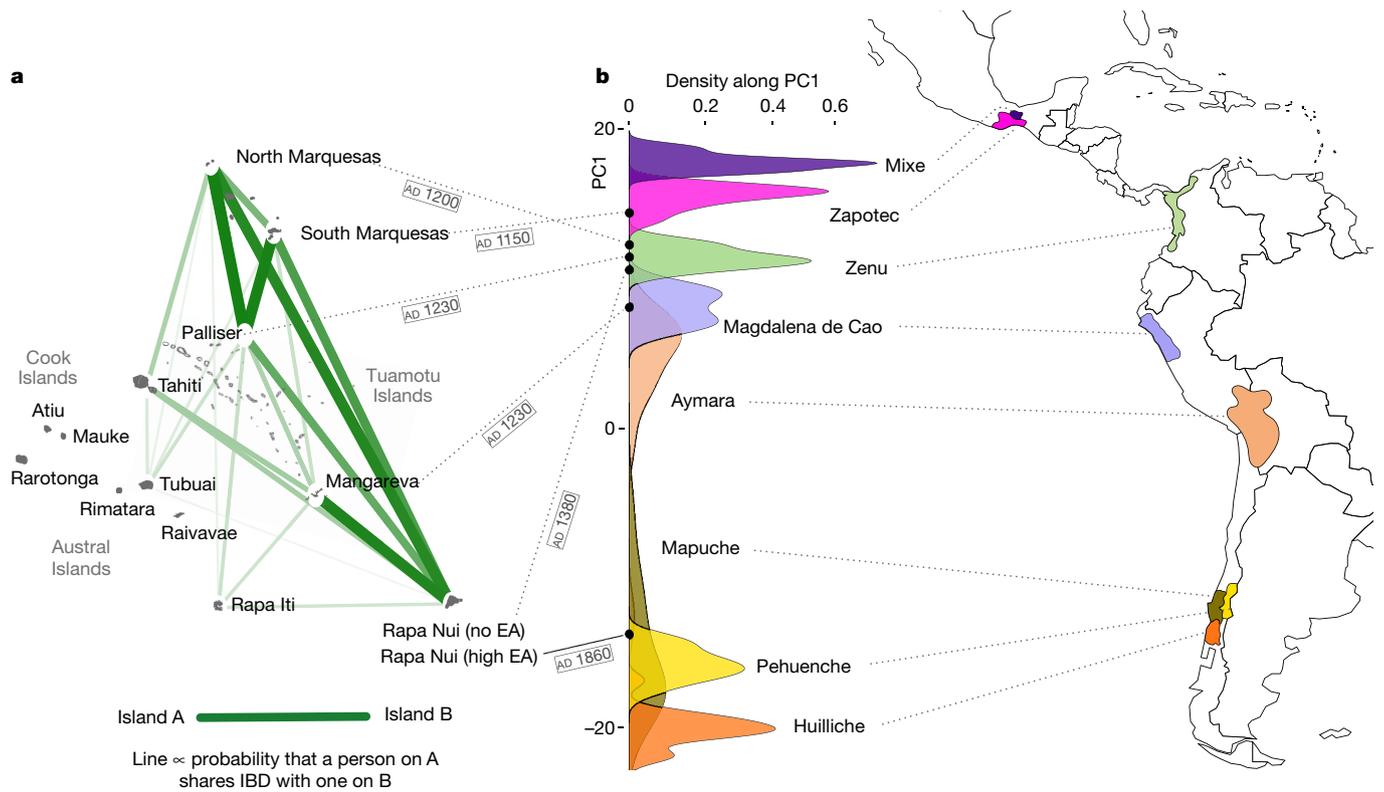
Fig. 4 | Origin and spread of early Native American ancestry in Polynesia. **a**, Results of a Native American specific IBD analysis reflect the common ancestry and origin of the Native American component in easternmost Polynesia. EA, European ancestry. **b**, Our new ancestry-specific PCA (left) separates Pacific Rim Native American reference individuals along a north–south axis, as shown in this kernel density plot of the numbers of individuals from select reference populations along the first principal component (PC1) axis. See Supplementary Figs. 14, 18 for the full two-dimensional plot. Colours indicate the reference populations' locations in the Americas (right). The locations of the aggregate Native American specific components for each Pacific island are also plotted (black dots connected by dashed lines to their source island in **a**). The maximum likelihood date for the Native American introgression event in each island population, as determined by a Tracts analysis (Supplementary Fig. 23), is displayed under the corresponding dashed line. The numbers of samples used from each island and each American population are given in Supplementary Table 1.

(AD 1380; Supplementary Fig. 23a). However, this inferred date may be shifted later owing to more recent Native American introgression from Chile, as already discussed. Indeed, Rapanui individuals who have high European and Native American ancestry (Supplementary Fig. 23b) show Native American introgression predominantly during the colonial period (best fit, AD 1720). According to the best-fitting model, this represents Native American introgression into European ancestry first, probably occurring in Chile, followed later by addition of Polynesian ancestry (best fit, AD 1860), probably when admixed Chilean individuals began immigrating to Rapa Nui (Fig. 4b). That latter date (AD 1860) is slightly before the annexation of Rapa Nui by Chile (AD 1888); however, by this time 12 Chilean individuals (out of approximately 100 total inhabitants) were already recorded to be living on Rapa Nui[33]. Indeed, because of its relation to modern Chile, we find that Rapa Nui is one of the most complicated places to study and date the prehistoric Native American contact in eastern Polynesia (see Supplementary Fig. 24).

To confirm our Tracts dating of the early Native American introgression in Polynesia, we used an alternative, linkage-disequilibrium-based, dating method[34] (Supplementary Fig. 23g and Supplementary Table 14). Unlike the Tracts approach, this method (ALDER) does not rely on phasing or local ancestry inference, instead fitting the exponential decay of linkage disequilibrium within an admixed target population directly, using two reference populations as proxies for the ancestral sources[35]. We used data from unadmixed Native American individuals from Peru and indigenous Austronesian individuals from Taiwan as reference populations, and islanders with Native American and Polynesian admixture only (no European ancestry) as targets (six Rapanui, four Mangareva, two Palliser and one North Marquesas). For these pooled individuals, we obtained an estimated admixture date of AD 1234 ± 90 years (Supplementary Fig. 23g). We note that all of our Tracts date estimates are contained within the confidence interval of this ALDER estimate, except for the aforementioned special case of the Rapanui individuals with recent Chilean ancestry.

Discussion

The Native American component within each of these widely separated remote eastern Polynesian islands has a similar introgression date, a common source in the indigenous peoples of Colombia, and a dense shared IBD network indicating shared ancestors. Each of these results is most parsimoniously explained by a single prehistoric contact event between eastern Polynesian and Native American individuals. Although the island contacted is not yet clear, and perhaps is not found in our present data set, it is likely that the contact occurred during the original period of discovery and settlement of remote eastern Polynesia by Polynesian individuals. Descendants of the initial contact probably transmitted their dual ancestry to new islands upon settling them; interisland trade contact may also have played a part. Thus, the prehistoric Native American component on Rapa Nui, upon which so much research has focused[9–11], is likely to have originated from a contact event not on Rapa Nui, but somewhere upstream in the Polynesian settlement process. This would explain a human-mediated spread of the sweet potato throughout Polynesia, if, as some have speculated,

Article

the Polynesian settlement of Rapa Nui involved no return voyaging or trade links[31,36–38].

Our earliest estimated date of contact is AD 1150 for Fatu Hiva, South Marquesas. This is close to the date estimated by radiocarbon dating for settlement of that island group[13], raising the intriguing possibility that, upon their arrival, Polynesian settlers encountered a small, already established, Native American population. It was on the island of Fatu Hiva—the easternmost island in equatorial Polynesia—that Thor Heyerdahl hypothesized that Native American and Polynesian individuals might have contacted one another, based on islanders' legends stating that their forefathers had come from the east[39]. The Marquesas lie at the latitude of Ecuador, and wind- and current-based simulations indicate that they are the islands most likely to be reached from South America via the strong east-to-west currents and winds at these equatorial latitudes[4,40,41].

We cannot discount an alternative explanation: a group of Polynesian people voyaged to northern South America and returned[42] together with some Native American individuals, or with Native American admixture, as speculated in ref.[10]. We have dated the contact event to the time when Polynesian explorers were, according to some studies, making their longest-range voyages (the century surrounding AD 1200)—a time when these studies suggest that the Polynesian settlers discovered all remaining island groups in the Pacific, from Hawaii to New Zealand to Rapa Nui[13,38,42]. The Tuamotu Archipelago, which lies at the centre of the Polynesian islands in which we found a Native American component, is known to have been a Polynesian voyaging hub, and according to simulations it is the second most likely location to be reached when voyaging from South America[4]. Further population genetics collaborations with these genetically understudied island populations are needed to resolve these alternative hypotheses.

In conclusion, we find strong genetic evidence for pre-Columbian human trans-Pacific voyaging contact (at the turn of the twelfth century), contemporaneous with the Polynesian voyages of discovery in the remote eastern Pacific[13,14]. Previous studies of putative Polynesian–Native American contact have focused on Rapa Nui, whose modern genetic history has been influenced by a recent Chilean admixture event, and have missed the possibility, which we show to be more likely, that prehistoric contact occurred before the settlement of Rapa Nui. We show that evidence for early Native American contact is found on widely separated islands across easternmost Polynesia, including islands not influenced by more recent Native American contact events. Our results show the usefulness of genetic studies of modern populations, which allow for large sample sizes to unravel complex prehistoric questions, and demonstrate the importance of combining anthropological, mathematical and biological approaches to answer these questions.

Online content

1. Brown, J. M. *The Riddle of the Pacific* (T. F. Unwin, 1924).
2. Heyerdahl, T. *American Indians in the Pacific* (Allen & Unwin, 1952).
3. Yen, D. E. *The Sweet Potato and Oceania* (Bishop Museum Press, 1974).
4. Montenegro, A., Avis, C. & Weaver, A. Modeling the prehistoric arrival of the sweet potato in Polynesia. *J. Archaeol. Sci.* **35**, 355–367 (2008).
5. Roullier, C., Benoit, L., McKey, D. B. & Lebot, V. Historical collections reveal patterns of diffusion of sweet potato in Oceania obscured by modern plant movements and recombination. *Proc. Natl Acad. Sci. USA* **110**, 2205–2210 (2013).
6. Clarke, A. C., Burtenshaw, M. K., McLenachan, P. A., Erickson, D. L. & Penny, D. Reconstructing the origins and dispersal of the Polynesian bottle gourd (*Lagenaria siceraria*). *Mol. Biol. Evol.* **23**, 893–900 (2006).
7. Muñoz-Rodríguez, P. et al. Reconciling conflicting phylogenies in the origin of sweet potato and dispersal to Polynesia. *Curr. Biol.* **28**, 1246–1256 (2018).
8. Lie, B. A. et al. Molecular genetic studies of natives on Easter Island: evidence of an early European and Amerindian contribution to the Polynesian gene pool. *Tissue Antigens* **69**, 10–18 (2007).
9. Thorsby, E. The Polynesian gene pool: an early contribution by Amerindians to Easter Island. *Phil. Trans. R. Soc. Lond. B* **367**, 812–819 (2012).
10. Moreno-Mayar, J. V. et al. Genome-wide ancestry patterns in Rapanui suggest pre-European admixture with Native Americans. *Curr. Biol.* **24**, 2518–2525 (2014).
11. Fehren-Schmitz, L. et al. Genetic ancestry of Rapanui before and after European contact. *Curr. Biol.* **27**, 3209–3215 (2017).
12. Hagelberg, E., Quevedo, S., Turbon, D. & Clegg, J. B. DNA from ancient Easter Islanders. *Nature* **369**, 25–26 (1994).
13. Wilmshurst, J. M., Hunt, T. L., Lipo, C. P. & Anderson, A. J. High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. *Proc. Natl Acad. Sci. USA* **108**, 1815–1820 (2011).
14. Hunt, T. L. & Lipo, C. P. Late colonization of Easter Island. *Science* **311**, 1603–1606 (2006).
15. Mulrooney, M. A. An island-wide assessment of the chronology of settlement and land use on Rapa Nui (Easter Island) based on radiocarbon data. *J. Archaeol. Sci.* **40**, 4377–4399 (2013).
16. Martinsson-Wallin, H., Wallin, P. & Anderson, A. Chronogeographic variation in initial East Polynesian construction of monumental ceremonial sites. *J. Island Coastal Archaeol.* **8**, 405–421 (2013).
17. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
18. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
19. Liang, M. & Nielsen, R. The lengths of admixture tracts. *Genetics* **197**, 953–967 (2014).
20. Gravel, S. et al. Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci. USA* **108**, 11983–11988 (2011).
21. Fischer, S. R. *Island at the End of the World* (Reaktion Books, 2006).
22. Maude, H. E. *Slavers in Paradise: The Peruvian Slave trade in Polynesia, 1862–1864* (Stanford Univ. Press, 1981).
23. Hurles, M. E. et al. Native American Y chromosomes in Polynesia: the genetic impact of the Polynesian slave trade. *Am. J. Hum. Genet.* **72**, 1282–1287 (2003).
24. Seemann, B. *Flora Vitiensis* (L. Reeve, 1865).
25. Scaglion, R. in *The Sweet Potato in Oceania: A Reappraisal* (eds Ballard, C. et al.) 35–41 (Oceania Publications, 2005).
26. Hornell, J. Was there pre-Columbian Contact between the Peoples of Oceania and South America? *J. Polynesian Soc.* **54**, 167–191 (1945).
27. Dewan, L. & Hosler, D. Ancient maritime trade on balsa rafts: an engineering analysis. *J. Anthropol. Res.* **64**, 19–40 (2008).
28. Hosler, D. West Mexican metallurgy: revisited and revised. *J. World Prehist.* **22**, 185–212 (2009).
29. Callaghan, R. T. Prehistoric trade between Ecuador and West Mexico: a computer simulation of coastal voyages. *Antiquity* **77**, 796–804 (2003).
30. Irwin, G. J. Against, across and down the wind: a case for the systematic exploration of the remote Pacific Islands. *J. Polynesian Soc.* **98**, 167–206 (1989).
31. Lewis, D. *We, the Navigators* (Univ. Hawaii Press, 1994).
32. Gravel, S. Population genetics models of local ancestry. *Genetics* **191**, 607–619 (2012).
33. Porteous, J. D. *The Modernization of Easter Island* (Univ. Victoria, 1981).
34. Loh, P. R. et al. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233–1254 (2013).
35. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
36. Walworth, M. Eastern Polynesian: the linguistic evidence revisited. *Oceanic Linguistics* **53**, 256–272 (2014).
37. Kirch, P. V. & Green, R. C. *Hawaiki, Ancestral Polynesia* (Cambridge Univ. Press, 2001).
38. Hunt, T. & Lipo, C. *The Statues That Walked* (Free Press, 2011).
39. Heyerdahl, T. *Fatu-Hiva: Back to Nature* (Allen & Unwin, 1974).
40. Montenegro, Á., Callaghan, R. T. & Fitzpatrick, S. M. Using seafaring simulations and shortest-hop trajectories to model the prehistoric colonization of Remote Oceania. *Proc. Natl Acad. Sci. USA* **113**, 12685–12690 (2016).
41. Di Piazza, A., Di Piazza, P. & Pearthree, E. Sailing virtual canoes across Oceania: revisiting island accessibility. *J. Archaeol. Sci.* **34**, 1219–1225 (2007).
42. Kirch, P. V. *On the Road of the Winds* (Univ. California Press, 2017).
43. Durbin, R. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010); corrigendum **473**, 544 (2011).
44. de la Fuente, C. et al. Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia. *Proc. Natl Acad. Sci. USA* **115**, E4006–E4012 (2018).
45. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
46. Bryc, K. et al. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc. Natl Acad. Sci. USA* **107** (Suppl 2), 8954–8961 (2010).
47. Posth, C. et al. Reconstructing the deep population history of Central and South America. *Cell* **175**, 1185–1197 (2018).
48. Moreno-Mayar, J. V. et al. Early human dispersals within the Americas. *Science* **362**, eaav2621 (2018).
49. Aitchison, J. *The Statistical Analysis of Compositional Data* (Chapman and Hall, 1986).
50. Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).

Methods

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

Ethical approval

Written informed consent was obtained from all participants, and research/ethics approval and permits were obtained from the following institutions: the Stanford University Institutional Review Board (IRB approval no. 20839), the Oxford University Tropical Research Ethics Committee (reference no. 537-14), and the Ethical Scientific Committee at the Pontificia Universidad Católica de Chile (reference no. 1971092), conducted in accordance with the guidelines of the National Commission on Science and Technology (CONICYT-Chile).

Sample collection and genotyping

This work combines publicly available genotype data and newly generated single-nucleotide polymorphism (SNP) array data from samples collected over different time periods by the participating institutions (Supplementary Tables 1–3). Sampled populations and genotyping platforms are detailed in Supplementary Table 1. A total of 25 populations were genotyped at the University of California, San Francisco (UCSF) using Affymetrix Axiom LAT-1 arrays. Another seven populations were genotyped using an Illumina multi-ethnic genotyping array (MEGA) or Illumina 610-Quad arrays (see Supplementary Table 1). Genotype calling was performed following default parameters using Affymetrix's Genotyping Console software or Illumina's GenomeStudio application, as appropriate. The average call rate was 98.5% for all newly genotyped samples. Before filtering and merging, the total numbers of SNPs called on the Axiom LAT-1 and Illumina MEGA platforms were 813,036 and 1,738,289 respectively. To remove genotyping errors, all samples genotyped on the same array were filtered together using Plink 1.9, eliminating the following: individuals missing more than 1% of genotype sites (mind .01), SNPs missing in more than 1% of individuals (geno .01), and SNPs out of Hardy–Weinberg equilibrium with a P value of less than 1×10^{-110}.

Data preparation

We used the UCSC tool liftover to bring all data onto the same genome build, GRCh37 (hg19)[51]. When merging data from different genotyping arrays, strand flips were detected and corrected, with ambiguous SNPs (SNPs whose strand definition could not be definitively matched between arrays) removed. This typically resulted in a loss of fewer than 10% of SNPs. Hence, the resulting SNP density after merging with different reference panels varied across working data sets for downstream analyses, as detailed throughout the Methods. Genetic positions were assigned using the interpolated recombination map generated by the 1000 Genomes project[43]. Given the depth and quality heterogeneity of the ancient samples, we called pseudohaploid genotypes for all ancient individuals to minimize potential bias derived from calling diploid genotypes[48]. For each ancient genome, we discarded reads with mapping quality below 30 and bases with quality below 20.

Global ancestry analysis

Principal component analysis. PCA was performed using EIGEN-SOFT 7.2.1 (ref. [52]) by merging the genotyped Polynesian individuals together with reference panels from Africa, Europe, Taiwan, Melanesia and the Americas (Supplementary Fig. 9). For African and European reference panels, we used genotypes from 1000 Genomes individuals: 60 Yoruban (YRI), 30 British (GBR) and 30 Spanish (IBS) individuals[43]. For Melanesian reference individuals, we used 16 individuals from Vanuatu; for Taiwan, we used 20 individuals from the Atayal and Paiwan indigenous groups; and for the Americas, we used 60 individuals with only Native American ancestry (as indicated by our ADMIXTURE

analysis, Fig. 1b; see below) originating from Puno, Peru (Supplementary Table 1). Merging of the sequence and filtered genotype data (689,899 SNPs) was done with PLINK 1.9 (ref. [53]), as was linkage disequilibrium pruning (LD pruning) (–indep-pairwise 50 10 .5), which was used to greedily remove successive variants with a squared correlation greater than 0.5 in 50 SNP sliding windows with 10 SNP steps. Plotting was performed in R version 3.5.2 using the ggplot2 3.1.0 package[54].

Unsupervised ADMIXTURE. To explore Native American substructure in our Polynesian individuals, we merged the Pacific island individuals above together with European (10 UK, 10 Spain), African (20 Yoruba) and Pacific coastal Native American reference populations, which included Mapuche (6 Pehuenche, 14 Huilliche), Aymara (10 Puno, 10 Arica), Magdalena de Cao (19), Zenu (19) and Mexico (10 Mixe, 10 Zapotec). Samples from the two latter locations were genotyped on a second array (see Supplementary Tables 1, 3), so the merged data set of 489 individuals had an overlap of 134,281 SNPs. ADMIXTURE 1.3.0 was run on this data set using unsupervised mode[17] (Supplementary Figs. 1, 2). According to the elbow[55] in the cross-validation plot (Supplementary Fig. 2), a good clustering is found around $K = 7$. Because our data set is heavily imbalanced, with the bottlenecked Rapa Nui samples ($n = 166$) comprising nearly as much of our Pacific island data set (47%) as all other islands combined, a cluster corresponding to Rapa Nui related Polynesian ancestry emerges (see Supplementary Fig. 4). This issue was addressed using the iterative ADMIXTURE approach described below.

Iterative unsupervised ADMIXTURE. To avoid the spurious clustering[56,57] that can be introduced by imbalanced sampling—such as in our Pacific island data set, which comprises 47% Rapanui individuals (described above; see Supplementary Figs. 1, 2 and Supplementary Discussion)—without having to downsample our Rapanui population, we used a new iterative unsupervised ADMIXTURE approach. Previous studies have addressed such spurious clustering, if properly recognized, by using supervised or semisupervised (projection) approaches[58] or by simple downsampling of the overrepresented populations. We found none of those approaches to be fully satisfactory. Supervised learning requires a researcher to subjectively define clusters a priori, which does not allow ancestry patterns to emerge naturally from the data. A semisupervised approach—for example, running unsupervised ADMIXTURE on an evenly sampled data set, followed by projecting the remaining samples onto the clusters found—avoids these subjective biases, but generates noise in the projected samples. This noise manifests itself as small spurious proportions of all ancestries found in the projected samples, and stems from the fact that variants in the projected individuals were not able to inform the original clustering.

We solve both of these problems at once, albeit in a computationally intensive fashion, by using an iterative approach that allows every sample to participate in a fully unsupervised ADMIXTURE run, while ensuring that no one run suffers from a highly imbalanced data set. In particular, we chose evenly sampled reference numbers as is standard for a projection-type analysis (with additional representation from Native American populations, owing to their admixture with other ancestries). These were selected according to the original unsupervised ADMIXTURE (Supplementary Fig. 1) components: African, 20 Yoruban individuals; European, 10 Spanish and 10 British individuals; central Native American, 10 Mixe, 10 Zapotec, 19 Zenu, 20 Aymara, 19 Magdalena individuals; southern Native American, 20 Mapuche individuals (6 Huilliche, 14 Pehuenche); Polynesian, 2 individuals from each island; and Melanesian, 16 Vanuatuan individuals (Supplementary Fig. 5). Within the reference populations, those samples without recent admixture (typically less than 10%) according to our autosomal haplotype-based local ancestry analyses (see below) were chosen. This further eliminated imbalances in the ancestry cluster sizes represented by the reference panels. We then iteratively ran the reference panels

Article

together with each of the remaining Polynesian samples in a series of separate fully unsupervised ADMIXTURE analyses until all samples had been analysed. Each individual run was a standard downsampled unsupervised ADMIXTURE analysis. By repeating many such runs, all of the overrepresented Rapanui samples could be analysed, providing sufficient samples for our later compositional ancestry analyses (see below). The results for all individuals were then plotted[59] using ggplot 2 3.1.0 and Pophelper 2.2.9.

Because our admixed Colombian and Ecuadorian reference panels were genotyped on a third array (Illumina 610-Quad), different from both of the two merged above (Affymetrix Axiom LAT-1 and Illumina MEGA), combining them with our panel would have resulted in further loss of common SNP markers giving an even lower resolution for rare ancestry components. Thus, when these samples were run iteratively (separately), they were run in their own lower-SNP-density (32,872 SNPs) three-way array merge with the reference panels. Owing to the lower density of SNPs, slightly more noise is evident in the ancestry assignments of these samples as compared with the higher-density, neighbouring American samples (see Fig. 1b). The same strategy was used for each of our ancient Native American samples. Each ancient sample was merged separately with the reference panels to maximize the SNP overlap (48,666 SNPs for the La Galgada sample; 114,927 SNPs for the Aconcagua sample; 25,429 SNPs for the best Saki Tzul sample; and 129,612 SNPs for the best Ancestral Kaweskar sample), and then unsupervised ADMIXTURE was run on each merge. Because the ancient sample genotypes were called pseudohaploid, the reference panel individuals were also treated as pseudohaploid for consistency in each of these ancient sample iterative runs.

Marker-frequency-based statistics. To further confirm the existence of Native American ancestry in Polynesia, we conducted genome-wide admixture f_4 and D-statistic tests across 689,899 SNPs. For the target Polynesian individuals, we pooled individuals from those islands with a greater than 1% average Native American component in our ADMIXTURE analysis (Supplementary Table 4), selecting all individuals without later European or African admixture—that is, with a proportion of European and/or African ancestry of no more than 0.005: Mangareva (3), North Marquesas (1) and Rapanui (6) individuals. The f_4 statistics were computed using comparison populations from Europe (UK and Spain), Africa (Yoruba), China, and Vietnam from 1000 Genomes[18], along with Native American individuals from Peru (Aymara) and Polynesian individuals from Mauke, all of which had shown no admixture in our previous analyses. With the program fourpop[60], we computed f_4 statistics of the following form: f_4(target Polynesians, Mauke; X, Y), where X and Y represent all possible combinations of the other populations (Supplementary Fig. 10). Standard errors were estimated by block-jackknife with a block size of 500. As an additional verification of significance, we ran 100 coalescent simulations via fastsimcoal using the method of ref. [61] to estimate the proportion of simulated jackknife blocks larger than those observed for f_4(target Polynesians, Mauke; Aymara, Yoruba) (Supplementary Fig. 10). None were observed. To further test whether the target Polynesian individuals carry Native American ancestry, we computed D-statistics[35] of the following form: D(Mauke, target Polynesians; H_3, Yoruba). In this case, H_3 is a set of reference populations and individuals that include precontact ancient Native American genomes (Supplementary Tables 1, 3, 4). We again estimated standard errors through a block-jackknife procedure (Supplementary Fig. 11).

Compositional analyses of ancestry proportions

Thanks to the large sample size of Rapanui individuals ($n = 166$), we were able to conduct statistical analyses of the ancestry proportions of this population, and thus to characterize the associations between the different ancestries. We consider first all four of the ancestry proportions identified by our iterative ADMIXTURE analysis in

the Rapanui population: central Native American, southern Native American, European and Polynesian. We neglect the African component in the Rapanui individuals, as it is present in only 12 individuals with a proportion above 0.005, and so these dozen individuals are simply excluded.

Because these ancestry components (p_i for each ancestry i) are constrained to live on a simplex (that is, $\Sigma_i p_i = 1$; termed compositional data), computing raw covariances and correlations between ancestry components is not informative. (As one ancestry proportion rises, the others must fall, leading to intrinsic negative covariances and correlations.) Thus, we rely on statistical methods developed for compositional data[49] to characterize associations between ancestry components. In particular, we compute the log-ratio variance for each pair of ancestries i and j, $\tau_{ij} = \mathrm{Var}[\ln(p_i/p_j)]$, which together completely characterize the covariance structure of the composition[49] (see Supplementary Table 7). Smaller values of τ_{ij} indicate that one component does not vary much relative to the other, and larger values indicate that the ancestry components do vary freely relative to one another. We also compute the compositional analogue to correlation[62], $\rho_{ij} = \exp(-\tau_{ij}^2/2)$ (Supplementary Table 8). To visualize these associations, we plot the ancestry composition of each individual inside the four-component simplex, namely a tetrahedron (Supplementary Fig. 12).

We next analyse the subset of the Rapanui individuals without a southern Native American (Chilean) component—that is, the 64 Rapanui individuals with less than 1% southern Native American in the $K = 6$ iterative ADMIXTURE analysis (Fig. 1b). These individuals lie on the triangular simplex (Fig. 2b) that forms the base of the tetrahedral simplex above. Within this subset of individuals, we also compute τ_{ij} and ρ_{ij} for each of their ancestry pairings i, j (Supplementary Tables 9, 10). To confirm the observed association between the central Native American component and the Polynesian component, we also perform a compositional (log-contrast) PCA on these individuals. As the points lie on a two-dimensional compositional simplex within R^3, we map them to a two-dimensional linear subspace of R^3 (the subspace orthogonal to the vector [1,1,1]) using the centred log-ratio transform (clr)[62]. This isometry transforms each individual's vector of ancestry components $[p_i, p_j, p_k]$ by replacing each ancestry proportion with the log of that proportion divided by the geometric mean of all ancestry proportions, that is, $\mathrm{clr}(p_i) = \ln(p_i/(p_i p_j p_k)^{1/3})$.

We then perform a standard singular value decomposition on the centred compositional vectors in this space to determine the principal components. Because we are now in a two-dimensional Euclidean subspace, we find exactly two principal components: the first component ($\mathbf{v}_1$) and the vector orthogonal to it ($\mathbf{v}_2$) in this subspace, $\mathbf{v}_1 = $ [Polynesian, European, Native American] = [0.411, 0.816, 0.405], and $\mathbf{v}_2 = $ [Polynesian, European, Native American] = [−0.705, −0.0037, 0.709], having corresponding singular values of $\sigma_1 = 2.26$ and $\sigma_2 = 0.219$ respectively. Thus, less than 1% of the variance in the clr-transformed space occurs along the second principal component $\sigma_1^2/(\sigma_1^2 + \sigma_2^2) = 0.009 < 1\%$. As there is almost no variation along the second principal component, the projection of ancestries along this direction in clr-space is approximately constant, so $[−0.705, −0.0037, 0.709] \cdot \ln(1/(PEN)^{1/3} [P, E, N])$ is approximately constant, or equivalently $P^{−0.705} E^{−0.0037} N^{0.709}$ is approximately constant, where P, E and N are respectively the Polynesian, European and central Native American ancestry proportions in an individual. Exponentiating on both sides of the latter equation by $(1/0.705) = 1.42$, we have constant $\approx N^{1.006} E^{−0.005}/P \approx N/P$. In other words, the central Native American component (N) varies directly with the Polynesian component (P) in these Rapanui individuals, and both vary freely relative to the European component (E). Compositional analyses and plots were made in R using the Compositions 1.4.0 package[62].

Local ancestry inference

Reference panels for local ancestry. For our Axiom LAT-1 array analyses (689,899 SNPs), we used a balanced set of reference individuals

consisting of: African (60 Yoruban individuals), European (30 Spanish and 30 British individuals), Native American (60 unadmixed Native American Aymara individuals genotyped on the Axiom LAT-1 array) and Polynesian (60 individuals identified by ADMIXTURE (Fig. 1b) as having less than 1% non-Polynesian ancestry) (Supplementary Tables 1, 3). Note that our local ancestry inference method, RFMix, can identify admixture in its reference panels, if such admixture exists, through its expectation maximization iterations[18]. For our Illumina MEGA array analysis (896,557 SNPs), we used as reference those same European and African reference panels together with 60 unadmixed Native American Aymara individuals genotyped on the MEGA Illumina array (Supplementary Table 3). For our Illumina 610-Quad array (620,901 SNPs) analyses, we used the local ancestry results of ref. [63].

Phasing. Phasing was performed together on all samples using SHAPEITv2.837 with default parameter settings[64]. Population phasing has been shown to be particularly effective in such highly related, small, founder populations[64], as on these remote Polynesian islands.

RFMix. The program RFMix v1.5.4 uses a conditional random field smoother to stitch together the results of random forest classifiers applied to successive windows of SNP markers to recognize local autosomal haplotype variant patterns (linked sequences of SNPs) characteristic of different ancestries[18] (Fig. 2a). Methods that ignore SNPs' relative positions and linkage (for example, f_4 and D-statistics, ADMIXTURE and PCA) are blind to such characteristic sequence patterns. This is a semisupervised learning approach that requires reference panels from each ancestry of interest, as described in detail above. We ran RFMix with the recommended two expectation maximization iterations and a 2-millimorgan window size to identify genomic regions of Polynesian, European, African and Native American ancestry in our Pacific island samples and to identify genomic regions of European, African and Native American ancestry in our populations from the Pacific coast of the Americas. We chose these reference ancestries on the basis of our unsupervised admixture analyses (Supplementary Figs. 1, 5), which had indicated the presence of these continental ancestries in our samples.

Ancestry-specific analyses

For the ancestry-specific analyses below, all ancestries except the ancestry of interest are 'masked' within each sample by thresholding the posterior probabilities returned by RFMix at a 0.99 probability level for the ancestry of interest. In other words, all haploid markers along each individual's genome that are inferred to come from a different ancestry than the one of interest are treated as missing. In addition, on Rapa Nui a high Native American ancestry population group is defined to be those Rapanui individuals with greater than 40% of their genome in inferred Native American ancestry segments according to RFMix. This group also has much higher European ancestry than average for the island of Rapa Nui, as southern Native American ancestry and European ancestry are associated on the island (see above and Supplementary Tables 7, 8), and so we also refer to this group as 'high European Rapa Nui' in later analyses.

Ancestry-specific PCA. The two masked haploid genomes (haplotypes) for each individual are combined to generate a genotype frequency vector, with 0 representing no alternate allele seen at a marker, 0.5 representing one alternate allele and one reference allele seen, and 1 representing no reference allele seen. For markers at which both haploid genomes had missing data in a given individual, a missing value is recorded for that marker site. These genotype frequency vectors for each individual are assembled to create a masked ancestry-specific genotype frequency matrix X, with n samples (rows) and p SNPs (columns). This masked matrix is then completed using the singular value decomposition (SVD), with cross-validation used to determine the optimal reconstruction dimensionality (Supplementary Figs. 15, 17),

to produce a Y matrix[65]. Some individuals (rows) have large numbers of masked markers, so to reduce noise in the PCA we perform a weighted, rather than typical unweighted, PCA. The weights allow the principal components to be defined more heavily by the less masked, more precisely known, samples. As the variance of an estimated sample increases with the number of missing (masked) sites in that sample, we compensate by weighting each row (sample) proportionally to the fraction of non-missing sites present in that row (sample). Thus, we compute the SVD of $W^{1/2}Y_c$, where $Y_c = Y - 1/n \times \mathbf{1}_n (\mathbf{1}_n)^T \times W \times Y$, with Y_c being the weight-centred, completed sample matrix, W the diagonal matrix of weights, and $\mathbf{1}_n$ the n-element vector of ones. (Computing the SVD of $W^{1/2}Y_c$ is equivalent to diagonalizing the $Y_c^T W Y_c$ matrix.) The diagonal elements of W are given by $\{w_i = f_i / \Sigma_j f_j\}_{i=1...n}$, with f_i being the fraction of SNP sites present (not masked) in row i.

We apply this algorithm to the samples genotyped on the Affymetrix Axiom LAT-1 array (689,899 SNPs) and to the merge of samples genotyped on this array together with additional American Pacific coast reference individuals genotyped on the Illumina MEGA array (two-array intersection of 91,835 SNPs) (Supplementary Figs. 14, 16, respectively). In the first PCA, only individuals with at least 90,000 SNP markers in Native American tracts (unmasked SNPs) were plotted, as individuals with fewer SNPs suffer from greater noise (scatter) in their projections. For higher resolution (less noise), Native American genomic regions from all individuals on an island are also used to plot island-specific genotype frequency vectors. These genotype frequency vectors are formed by aggregating the Native American ancestry fragments from all individuals on the same island and calculating, for each marker, the ratio of the number of alternate alleles seen at that marker to the total number of unmasked alleles at that marker on that island. In the second PCA, which has far fewer SNPs from the outset, only island-specific genotype frequency vectors were plotted, except for the high Native American Rapa Nui, who each individually have sufficient numbers of Native American SNPs to be plotted separately without excessive noise.

Ancestry-specific MDS. Ancestry-specific MDS makes use of the fact that distances can be computed between pairs of genotypes, even if some markers are missing (masked) in each individual, simply by normalizing by the number of markers present for comparison in each individual. This approach was first pursued by Browning et al.[66], who noticed that the resulting distance matrix may still contain missing elements—namely, when two samples have no non-missing ancestry segments in common. In this case, Browning et al.[66] suggest completing each missing distance matrix entry by using the average distance of that individual against all others (mean imputation). However, one can construct a better estimate by noting that distance matrices have a high degree of structure—in particular, their elements must obey the triangle inequality. This allows missing values to be estimated by finding all possible triangles formed by the two samples that have no overlap and a third sample with which both do overlap. The common missing leg is then taken to be the minimum, over all these triangles, of the sum of their two known legs[67]. This triangulation allows the missing distance to be estimated from the known distances, rather than simply replacing it with a population-wide mean, giving much more accurate estimates for individuals with large amounts of masked ancestry (as found in Pacific islanders in Native American ancestry-specific analyses). As an additional advantage, none of the inferred distances will violate the triangle inequality; this is not true for the method of Browning et al.[66].

We implement this triangle-based algorithm to create an ancestry-specific approach to MDS that is accurate even for highly admixed samples. We use the average number of pairwise differences as a distance metric, as it is proportional to genetic drift[68].

We apply our ancestry-specific MDS method to the Native American ancestry-specific genotypes of Polynesian and American samples genotyped on Affymetrix Axiom LAT-1, Illumina MEGA and Illumina 610-Quad (Supplementary Fig. 14) and also to the European ancestry-specific

genotypes of Polynesian indivdiuals genotyped on Affymetrix Axiom LAT-1 together with European samples from POPRES genotyped on the Affymetrix GeneChip 500K and European full genomes from the 1000 Genomes Project (Supplementary Table 3 and Fig. 3a).

Procrustes. In order to confirm the findings of our new high-resolution ancestry-specific MDS and PCA methods described above, we carried out a traditional Procrustes analysis[69] to combine two separate Native American ancestry-specific PCAs (ASPCAs) that were constructed by the older ASPCA method[70]. The first ASPCA (Supplementary Fig. 19) was constructed using American reference populations genotyped on Illumina MEGA combined with American reference populations and Polynesian populations genotyped on Axiom LAT-1 (a 91,835 SNP two-array intersection). The second ASPCA (Supplementary Fig. 20) was constructed using American reference populations genotyped on Axiom LAT-1, Illumina MEGA and Illumina 610-Quad (a 28,653 SNP three-array intersection). The local ancestry inference used for the masking of non-Native American ancestries for these ASPCAs was performed using the full-density SNP set of each of the three arrays, that is, before intersection (see 'Local ancestry inference' section above). The initial coordinates of the Pacific island individuals' Native American ancestry were determined using the first ASPCA with the higher-density (91,835 SNP) two-array intersection. These positions were then mapped onto the lower-density (28,653 SNPs) three-array intersection of the second ASPCA, containing the full panel of American reference populations, using a Procrustes transform. The linear Procrustes mapping was identified by comparing the positions of the American reference individuals shared between the first ASPCA and the second ASPCA. Because these reference individuals have high Native American ancestry, they have few masked sites and suffer less from reduced SNPs in array intersections than the Pacific island samples (see Supplementary Fig. 20), resulting in less noisy positions.

Identity-by-descent segment analysis

Germline. IBD segments were identified using GERMLINE 1.5.3 on the SHAPEIT phased haploid genomes using the haploid flag, allowing a maximum of four homozygous marker mismatches per IBD slice (-err_hom), a maximum of one heterozygous marker mismatch per IBD slice (-err_het), and a minimum length for IBD detection of 3 cM (-min_m)[71].

Ancestry-specific filtering. IBD segments were then filtered on the basis of their overlap with local ancestry segments. For example, IBD segments located entirely within European ancestry segments, as previously determined by the local ancestry methods described above, are binned separately from those in Polynesian and Native American ancestries.

Ancestry-specific IBD networks. Previous publications on IBD networks have inferred the edge connections on the basis of the average total sum of IBDs shared between individuals within each pair of populations[72,73], or on the total sum shared within a specific IBD segment length range[74]. Here we consider instead the number of individuals from two populations who are connected by IBD segments above a threshold length. Specifically, we consider the probability that an individual selected at random from population A shares substantial IBD (greater than 7 cM, to ensure no spurious matches) with an individual selected at random from population B. This can be easily computed by dividing the total number of such interisland individual pairs connected by more than 7 cM of IBD by the total number of possible interisland individual pairs. We construct two networks with edges reflecting these probabilities, one for IBD segments located entirely in European ancestry segments of the genome and another for Native American ancestry segments (Supplementary Fig. 13 and Supplementary Tables 11, 12). We do not plot Polynesian segment IBD

probabilities, as we found all islands to share Polynesian ancestors with a probability of near one. Networks were plotted in R using the package qgraph[75].

Dating analyses

Tract-length distribution analysis. The timing of admixture events between different ancestral populations can be inferred by analysing the length distributions of genomic segments inherited from each ancestry, aggregated over all individuals in the studied population[32]. Here (Supplementary Fig. 23) we conduct our analysis separately on each island that possesses at least 1% average Native American component in both our ADMIXTURE (Supplementary Table 5) and RFMix (Supplementary Table 6) analyses. We considered Polynesian, Native American and European ancestries with genomics segments assigned by local ancestry inference using RFMix as described above. The small number of individuals with African ancestry were excluded, as above, because such ancestry is rare, likely to be post-colonial, and in any case not the focus of our present dating analysis. We excluded 12 individuals with above 0.005% African ancestry from Rapa Nui, 3 from South Marquesas and 6 from North Marquesas.

We used the Tracts method[32] to fit three models with different sequences of historical admixture for each island: first, Polynesian–Native American admixture followed by later European admixture; second, European–Native American admixture followed by later Polynesian admixture; and third, Polynesian–European admixture followed by later Native American admixture. To optimize model parameters over the nonlinear likelihood surfaces, we ran Python's COBYLA optimizer one hundred times each with different random starts for every population and model. The best-likelihood runs were chosen (Supplementary Table 13) and, although some random starts failed to converge, of those that did, most converged to similar maximum likelihoods and similar model parameters. The admixture model with the highest likelihood (Supplementary Table 13) was then selected. This method gives estimates of the time since each admixture event, measured in number of generations. To convert these to admixture dates, we used a generation time of 30 years (see Supplementary Discussion) and the sample collection dates in Supplementary Table 2. Tracts was also run separately, as described above, on the 64 Rapanui individuals identified as having less than 1% of a southern Native American component in our iterative ADMIXTURE analysis (see Supplementary Fig. 24).

Linkage disequilibrium decay analysis. We also performed a complementary analysis of linkage disequilibrium decay using ALDER 1.0, which requires neither phasing nor local ancestry inference[34]. Observing that the Native American ancestry-specific IBD clustering network indicated common Native American ancestry in eastern Polynesia, we pooled individuals across islands for this analysis. In particular, from the islands that had a greater than 1% average Native American component in our ADMIXTURE analysis (Supplementary Table 5), we pooled all individuals without later European or African admixture, that is, with no more than 1% European and/or African ancestry: Mangareva (4), Palliser (2), North Marquesas (1) and Rapa Nui (6). As ancestral reference proxies, we used 30 unadmixed Native American Aymara individuals and the 22 Austronesian individuals from the Atayal and Paiwan of Taiwan (Supplementary Tables 1, 3). We used a total of 690,692 SNPs in this analysis (see Supplementary Fig. 23g).

Because we had a large number of samples from Rapa Nui, we hoped to have increased resolution in our dating analysis there. However, all but six of the Rapanui individuals have European admixture, so we could not use the two-population model of ALDER on the full set of Rapanui individuals, and instead used MALDER 1.0 (ref. [76]) (Supplementary Table 14). For the Rapanui individuals, we could not pool those sampled in 1994 with those sampled in 2013 for this dating

analysis, because almost one generation separates these two collections. We focused on the 1994 Rapa Nui samples, as this collection had lower amounts of the modern southern Native American and European ancestries (Supplementary Table 5). In addition, to reduce the complexity of the admixture model, we excluded the 13 Rapanui individuals from the 1994 samples having African ancestry (greater than 1% in our ADMIXTURE analysis; Fig. 1b), leaving 73 individuals. For our ancestral reference proxies, we used 30 unadmixed Native American Aymara individuals, 30 European individuals from Spain, and the 22 Austronesian individuals (Atayal and Paiwan) from Taiwan (Supplementary Tables 1, 3).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Individual-level genotypes for new data presented here are available through a data-access agreement to respect the privacy of the participants for the transfer of genetic data from the European Genome Archive (EGA; https://www.ebi.ac.uk/ega/home), under accession number EGAS00001004209.

Code availability

Details regarding the packages and versions used are included in the Methods. Codes for the matrix completions described in the paper are available at https://github.com/AlexIoannidis/completion.

51. Tyner, C. et al. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res*. **45** (D1), D626–D634 (2017).
52. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet*. **2**, e190 (2006).
53. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
54. Wickham, H. *ggplot2* (Springer, 2016).
55. Holmes, S. & Huber, W. *Modern Statistics for Modern Biology* (Cambridge Univ. Press, 2019).
56. Lawson, D. J., van Dorp, L. & Falush, D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat. Commun*. **9**, 3258 (2018).
57. van Dorp, L. et al. Evidence for a common origin of blacksmiths and cultivators in the Ethiopian Ari within the last 4500 years: lessons for clustering-based inference. *PLoS Genet*. **11**, e1005397 (2015).
58. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).
59. Francis, R. M. pophelper: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour*. **17**, 27–32 (2017).
60. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*. **8**, e1002967 (2012).
61. Meyer, B. S., Matschiner, M. & Salzburger, W. Disentangling incomplete lineage sorting and introgression to refine species-tree estimates for Lake Tanganyika cichlid fishes. *Syst. Zool*. **66**, 531–550 (2017).
62. van den Boogaart, K. G. & Tolosana-Delgado, R. *Analyzing Compositional Data with*. (Springer Science & Business Media, 2013).
63. Homburger, J. R. et al. Genomic insights into the ancestry and demographic history of South America. *PLoS Genet*. **11**, e1005602 (2015).
64. O'Connell, J. et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet*. **10**, e1004234 (2014).
65. Troyanskaya, O. et al. Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
66. Browning, S. R. et al. Local ancestry inference in a large US-based Hispanic/Latino study: Hispanic community health study/study of Latinos (HCHS/SOL). *G3 Genes Genom. Genet*. **6**, 1525–1534 (2016).
67. Makarenkov, V. & Lapointe, F.-J. A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics* **20**, 2113–2121 (2004).
68. Peter, B. M. Admixture, population structure, and F-statistics. *Genetics* **202**, 1485–1501 (2016).
69. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer Science & Business Media, 2009).
70. Moreno-Estrada, A. et al. Reconstructing the population genetic history of the Caribbean. *PLoS Genet*. **9**, e1003925 (2013).
71. Gusev, A. et al. Whole population, genome-wide mapping of hidden relatedness. *Genome Res*. **19**, 318–326 (2009).
72. Moreno-Estrada, A. et al. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* **344**, 1280–1285 (2014).
73. Longobardi, G. et al. Across language families: genome diversity mirrors linguistic variation within Europe. *Am. J. Phys. Anthropol*. **157**, 630–640 (2015).
74. Han, E. et al. Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nat. Commun*. **8**, 14238 (2017).
75. Epskamp, S., Cramer, A., Waldorp, L. J., Schmittmann, V. D. & Borsboom, D. qgraph: network visualizations of relationships in psychometric data. *J. Stat. Softw*. **48**, 1–18 (2012).
76. Pickrell, J. K. et al. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl Acad. Sci. USA* **111**, 2632–2637 (2014).

Author contributions A.M.-E. and K.S. conceived the study. A.M.-E., A.G.I., A.J.M. and C.D.B. provided overall project supervision and management. A.M.-E., E.H., R.A.V., M.M., A.J.M., C.D.B. and C.R.G. contributed to study design. A.S., G.L.W., K.C.B., C.E., S.H., E.G.B., C.D.B. and A.M.-E. carried out genotyping experiments and quality control. A.G.I., J.B.-P., J.V.M.-M., J.E.R.-R., C.D.Q.-C., J.R.H. and A.M.-E. analysed the data. A.G.I. developed the analytical methods. A.M.-E., A.G.I. and J.B.-P. interpreted the results. A.M.-E., K.S., E.H., J.F.M.-P., K.A., T.P., K.R., A.V.S.H., M.C.A.-A., A.S., G.L.W., K.C.B., L.H., S.B., M.A., E.L., C.E., S.H., E.G.B., L.C., R.A.V., M.M., A.J.M. and C.D.B. contributed to acquisition of the data. A.G.I. wrote the manuscript, and A.G.I., A.M.-E., E.H., J.B.-P., J.V.M.-M. and M.M. edited the manuscript.

naturereserch

Corresponding author(s): Andres Moreno-Estrada, Alexander Ioannidis

Last updated by author(s): Mar 14, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	DNA samples were genotyped by using either Affymetrix (Mountain View, CA) Axiom LAT-1 arrays or by using Illumina (San Diego, CA) Multi-Ethnic Genotyping Array (MEGA) arrays. Genotype calling was performed following default parameters using Affymetrix's Genotyping Console software and Illumina's GenomeStudio 2.0 application, respectively. The average call rate was 98.5% for all newly genotyped samples. Before filtering and merging, the total number of SNPs called on the Axiom LAT-1 and Illumina MEGA platforms were 813,036 and 1,738,289 respectively. To remove genotyping errors, all samples genotyped on the same array were filtered together using Plink 1.9, eliminating the following: individuals missing more than 1% of genotypes sites (mind .01), SNPs missing in more than 1% of individuals (geno .01), and SNPs out of Hardy-Weinberg equilibrium with a p-value below 10e-110.
Data analysis	PLINK 1.9, EIGENSOFT 7.2.1, ADMIXTURE 1.3.0, R 3.5.2, ggplot2 3.1.0, Pophelper 2.2.9, Compositions 1.4.0, SHAPEITv2.837, RFMixv1.5.4, GERMLINE 1.5.3, qgraph, Tracts 1.0, ALDER 1.0, MALDER 1.0, badmixture 1.0, Affymetrix's Genotyping Console 4.0, GenomeStudio 2.0, UCSC liftover 1.0, fourpop from TreeMix 1.1, fastsimcoal 2.6, custom matrix completion code is available at https://github.com/AlexIoannidis/completion

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Individual-level genotypes for new data presented in this study are available through a data access agreement to respect the privacy of the participants for the

transfer of genetic data from the European Genome Archive (EGA).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   ☐ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. The sample size was selected before analysis was begun based on available samples and budgetary constraints for genotyping. We sought to include sufficient sample to power statistical comparisons.
Data exclusions	Standard pre-established genotype quality controls were applied to remove all individuals missing more than 1% of genotypes sites (mind .01), SNPs missing in more than 1% of individuals (geno .01), and SNPs out of Hardy-Weinberg equilibrium with a p-value below 10e-110.
Replication	No experiments were conducted, so replication of experimental results is not relevant.
Randomization	This is not relevant as our study does not consider variable assignments or categories.
Blinding	Blinding was not relevant to this study, which is standard in the population genetics field.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	No covariate-relevant population characteristics were used as recruitment criteria for any of the samplings, other than the geographic location of participants throughout the various sampling sites.
Recruitment	The samples used for this analysis were collected by the University of Chile, the University of Oxford, and Stanford University during various expeditions across Latin America and the Pacific. No self-selection bias was introduced as recruitment procedures were inclusive and addressed to the general population at each sampling site.
Ethics oversight	Written informed consent was obtained from all participants and research/ethics approval and permits were obtained from the following institutions: Stanford University Institutional Review Board (IRB approval No. 20839), Oxford University Tropical Research Ethics Committee (reference No. 537-14), and the Scientific Ethics Committee of the Catholic University of Chile (reference No. 1971092).

Note that full information on the approval of the study protocol must also be provided in the manuscript.