Article

Genetic continuity and change among the Indigenous peoples of California

https://doi.org/10.1038/s41586-023-06771-5

Received: 22 May 2023

Accepted: 20 October 2023

Published online: 22 November 2023

Check for updates

Nathan Nakatsuka[1,2]✉, Brian Holguin[3], Jakob Sedig[4], Paul E. Langenwalter II[5], John Carpenter[6], Brendan J. Culleton[7], Cristina García-Moreno[6], Thomas K. Harper[7], Debra Martin[8], Júpiter Martínez-Ramírez[6], Antonio Porcayo-Michelini[9], Vera Tiesler[10], M. Elisa Villapando-Canchola[6], Alejandro Valdes Herrera[11], Kim Callan[1,12], Elizabeth Curtis[1,12], Aisling Kearns[1], Lora Iliev[1,12], Ann Marie Lawson[1,12], Matthew Mah[1,12], Swapan Mallick[1,12], Adam Micco[1,12], Megan Michel[1,12], J. Noah Workman[1,12], Jonas Oppenheimer[1,12], Lijun Qiu[1,12], Fatma Zalzala[1,12], Nadin Rohland[1], Jose Luis Punzo Diaz[11], John R. Johnson[13,15]✉ & David Reich[1,4,12,14,15]✉

Before the colonial period, California harboured more language variation than all of Europe, and linguistic and archaeological analyses have led to many hypotheses to explain this diversity[1]. We report genome-wide data from 79 ancient individuals from California and 40 ancient individuals from Northern Mexico dating to 7,400–200 years before present (BP). Our analyses document long-term genetic continuity between people living on the Northern Channel Islands of California and the adjacent Santa Barbara mainland coast from 7,400 years BP to modern Chumash groups represented by individuals who lived around 200 years BP. The distinctive genetic lineages that characterize present-day and ancient people from Northwest Mexico increased in frequency in Southern and Central California by 5,200 years BP, providing evidence for northward migrations that are candidates for spreading Uto-Aztecan languages before the dispersal of maize agriculture from Mexico[2–4]. Individuals from Baja California share more alleles with the earliest individual from Central California in the dataset than with later individuals from Central California, potentially reflecting an earlier linguistic substrate, whose impact on local ancestry was diluted by later migrations from inland regions[1,5]. After 1,600 years BP, ancient individuals from the Channel Islands lived in communities with effective sizes similar to those in pre-agricultural Caribbean and Patagonia, and smaller than those on the California mainland and in sampled regions of Mexico.

People have lived in California since at least 13,000 years BP (all dates are calibrated in years BP in the remainder of the paper) based on archaeological evidence from the Northern Channel Islands off Southern California[6] and the mainland, which were occupied by peoples speaking Chumashan languages at the time of European contact[1]. California also harbours some of the highest linguistic diversity of any region in the Americas, which is relevant to understanding human population relationships as language often correlates with movements of people[7]. Linguistic diversity among Native Californians includes multiple major groupings, which have varying divergence time estimates, some exceeding 6,000 years BP. In the north, language families include Algic (including Yurok, for example), Athabascan (including Hupa), and Yukian (named for the Yuki). In the central part of California, including the coast and interior valley, Utian languages occur, spoken by Miwok and Ohlone (Coastanoan) tribes[8]. Languages within the Chumashan family characterize the Northern Channel Islands and adjacent mainland in the Santa Barbara region. In Southern California, speakers of Uto-Aztecan languages predominate, including the Tongva (Gabrielino) and Payomkawichum (Luiseño). Last, there are smaller language families (for example, Yuman-Cochimí) and isolates (for example, Washo)[1]. How these divergent language families came to be in such close proximity needs to be understood in a larger context. This is because migration in one direction or another must have been responsible for some families, such as Algic and Uto-Aztecan, to have wide dispersals, in these cases extending far beyond California. The Uto-Aztecan language family in particular is one of the most geographically widespread families in the Americas, ranging from Shoshone in Idaho to Pipil in Costa Rica and covering the central and west coast of Mexico and the American

[1]Department of Genetics, Harvard Medical School, Boston, MA, USA. [2]Harvard–MIT Division of Health Sciences and Technology, Boston, MA, USA. [3]Department of Anthropology, University of California at Santa Barbara, Santa Barbara, CA, USA. [4]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA. [5]Department of Anthropology, Biola University, La Mirada, CA, USA. [6]Instituto Nacional de Antropología e Historia, Sonora, Hermosillo, México. [7]Institute of Energy and the Environment, The Pennsylvania State University, University Park, PA, USA. [8]Department of Anthropology, University of Nevada, Las Vegas, NV, USA. [9]Instituto Nacional de Antropología e Historia, Mexicali, México. [10]Universidad Autónoma de Yucatán, Facultad de Ciencias Antropológicas, Mérida, México. [11]Instituto Nacional de Antropología e Historia, Morelia, México. [12]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, USA. [13]Santa Barbara Museum of Natural History, Santa Barbara, CA, USA. [14]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [15]These authors jointly supervised this work: John R. Johnson and David Reich. ✉e-mail: Nathan_nakatsuka@hms.harvard.edu; JohnJ@sbnature2.org; reich@genetics.med.harvard.edu

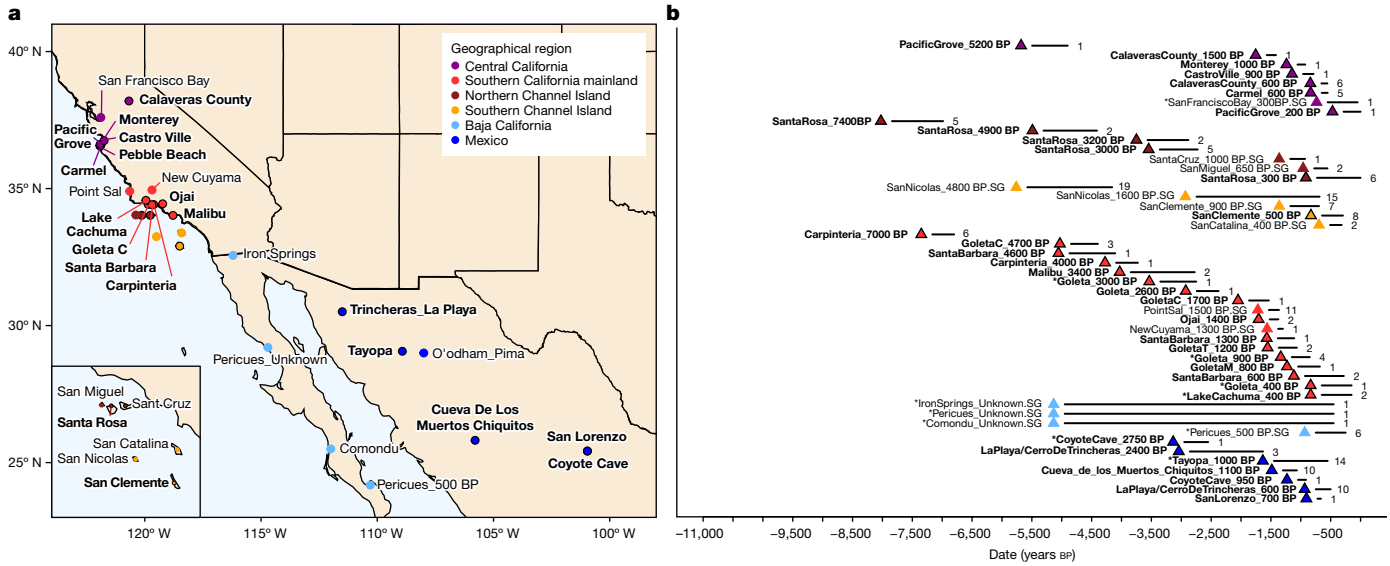


Fig. 1 | Summary of data. a, Locations of analysed ancient individuals. Newly reported locations are in bold. All data are ancient except O'odham. Map was made using open-source data and software using the R packages maps, sf, rnaturalearth, ggplot2 and ggrepel. **b**, Dates of analysed individuals. Width is the full range of all radiocarbon dates for the group (marine reservoir calibrated; Methods). Asterisk indicates no radiocarbon date. The number of individuals per site is indicated on the right.

Southwest. There are many proposed homelands, including the Great Basin[9], California's central valley[10], the Sonora desert[11], Central Mexico (from which it has been suggested to have spread with maize farming)[12], and Southern Arizona to Northern Mexico[13], with different types of linguistic and archaeological evidence adduced for each model.

We report data from 119 individuals, including 79 individuals from 7,400 to 200 years BP in Central and Southern California, and 40 individuals from 2,900 to 500 years BP in Northwest and North Central Mexico (Fig. 1 and Supplementary Data 1) (in this paper, we frequently refer to present-day political entities such as US or Mexican states, but caution that modern boundaries artificially divide Indigenous culture areas). To obtain these data, we extracted DNA, generated single-stranded and double-stranded DNA libraries (treated to remove characteristic damage signatures associated with ancient DNA) and enriched for mitochondrial DNA and about 1.2 million single nucleotide polymorphisms (SNPs) across the genome. We sequenced the enriched products using Illumina instruments and evaluated the authenticity of the data, which led us to restrict analyses to 112 newly reported individuals with no evidence of substantial modern contamination (Methods and Supplementary Information). We combined these data with previously published ancient and present-day data.

Consistent with previous work[14], the earliest DNA sequenced from the Chumash region in California going back to at least around 7,400 years BP was most closely related to modern people from South America and the Clovis culture-associated Anzick individual dated to approximately 12,800 years BP from Montana (Southern Native American (SNA) ancestry)[15]. Genetic clustering of the individuals correlated with geography, and correlated with language, with branches enriched in groups likely to have spoken Chumashan, Uto-Aztecan and Utian. We found evidence for large-scale movement of genetic lineages characteristic of ancient and modern individuals from Northwest Mexico into both Southern and Central California by at least around 5,200 years BP. This result raises the possibility that this movement was responsible for spreading Uto-Aztecan languages, documenting a period of major migration from the south not associated with farming, and undermining the argument that agriculture needed to have introduced these languages[12]. Finally, there was a strong genetic relationship between the earliest individual from Pacific Grove, Central California, dated to around 5,200 years BP to ancient individuals from Baja California. This result provides support for the theory that people speaking languages from an earlier linguistic substrate were once dispersed across large parts of California, and that the populations of the region were later transformed by new migrants who changed both the genetic and linguistic landscapes.

Ethics and inclusion statement

This research was carried out in consultation with Indigenous communities and other stakeholders in California and Mexico, with multiple engagements occurring before sampling, as well as return-of-results meetings before submission of the paper. Co-authors, including different subsets of N.N., J.R.J., B.H., P.L., J.L.P.D. and D.R., participated in consultations with members of the Chumash, Tongva, Ohlone communities and several communities in Mexico, with the goal of ensuring that the paper reflected community perspectives (additional details in the Methods and Supplementary Data 1). Our study includes co-authors who not only contributed to the scientific work but are also members of communities with connections to ancient individuals. The final paper addresses topics that were emphasized by community members as being of particular interest, including understanding how ancient peoples on the Channel Islands and mainland related to each other, understanding the relationships between ancient peoples in California and those of nearby regions, and understanding the processes that produced the historical distribution of Indigenous languages. We emphasized in these presentations that scientific discovery is a dynamic and iterative process that builds on itself, and that this study is not the final word even on a scientific level, as additional studies will refine and improve the models and interpretations here. We also emphasized that genetic ancestry is distinct from identity, which is often based on social relationships rather than biological ties; genetic findings should never be seen as challenging cultural identity.

Overview of genetic data

We grouped individuals on the basis of their archaeological site and age. Individuals in these groups were genetically homogeneous relative to other groups using qpWave (Supplementary Data 2). To cluster groups, we created a neighbour-joining tree (Fig. 2) and performed

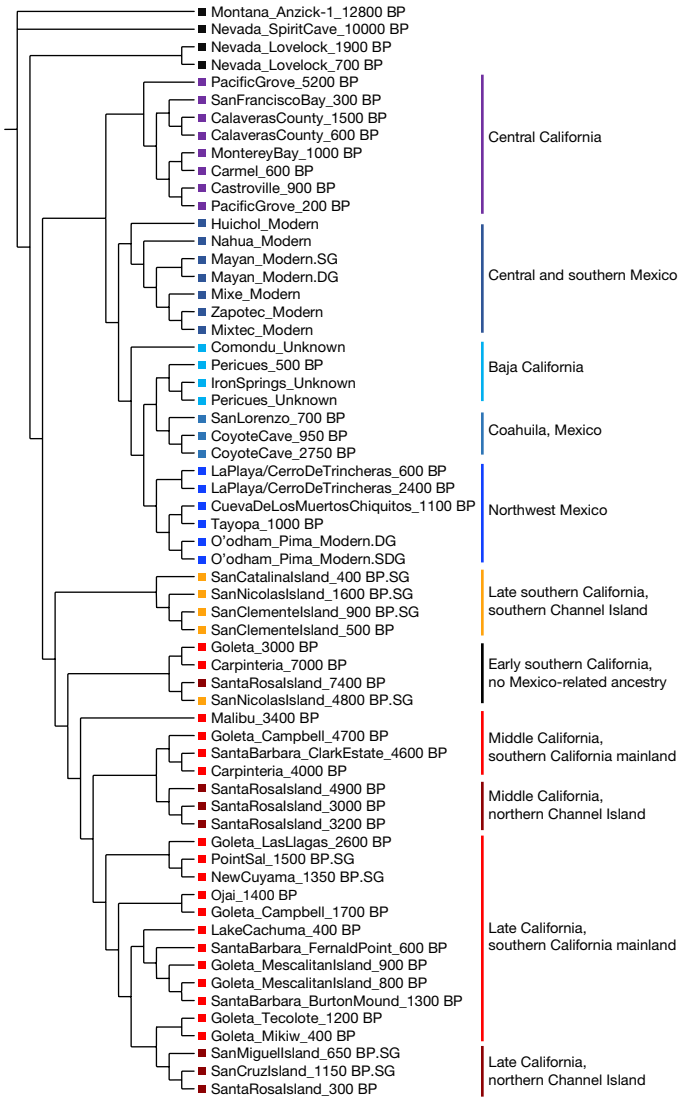


Fig. 2 | Neighbour-joining tree of groups. Tree created using a matrix of inverted outgroup-f_3 statistics (distances = $1/f_3$(Mbuti; Pop1, Pop2)) with USA-MT Anzick-1 12,800 BP as the outgroup. Designed using Itol.

an unsupervised analysis using the ADMIXTURE software (Extended Data Fig. 1). We also used multidimensional scaling (MDS) (Extended Data Fig. 2) based on an outgroup-f_3 matrix of statistics that measures the amount of shared genetic drift between two groups (Pop1 and Pop2) as follows: f_3(Mbuti; Pop1, Pop2). Pairwise F_{ST} values were also computed (Extended Data Fig. 3 and Supplementary Data 3). The clusters obtained using these procedures correlated with geography and time. From a geographical perspective, we observed differentiation of Northwest Mexico, Baja California, two Southern California regions (roughly the Northern Channel Islands, which was genetically similar to the Southern California mainland, and the Southern Channel Islands), and Central California. Geography-based clustering can be an indicator of local population continuity, with later populations descended in substantial part from earlier ones. In some cases, genetic clustering was more correlated with time than with location, a pattern that can be indicative of cross-regional migrations.

Mitochondrial haplogroup frequencies in California showed variability over time. Of the individuals dated to 3,500 years BP with sufficient data to make a determination, 30 out of 36 had an A2 haplotype, nearly all from the Santa Barbara Channel Islands and adjacent mainland, with the 6 non-A2 haplotypes all from mainland Southern California (Supplementary Data 1). After 3,500 years BP, only 35 out of 91 individuals carried A2 haplotypes, with B2, C1b, C1c, C5b, D1 and D4h3a all represented, a result consistent with whole genome evidence of movement of lineages into California from outside. All of the individuals from Mexico were younger than 3,500 years BP, and only 5 out of 44 had A2 haplotypes, with B2, C1b, C1c, C5b and D4h3a also represented.

All Y chromosome samples were Q1b1a except for one from the North Mexican site of Cueva de los Muertos Chiquitos. This result differs from the much higher rate (about 1 out of 3) of Q1a2a in very ancient (>5,000 years BP) individuals from South America[16] (Supplementary Data 1).

Spread of lineages characteristic of Mexico into California before the advent of agriculture

The time dependency of some of the clustering in the neighbour-joining tree and the changes in mitochondrial haplogroups indicated ancestry change over time. To quantify these patterns, we compared the most ancient individuals (around 7,400 years BP) from the Northern Channel Island of Santa Rosa (called *wima'* in the Samala or Ineseño Chumash language) to the more recent ones using statistics of the form f_4(Mbuti, *X*; Santa Rosa Island 7,400 years BP, Santa Rosa Island < 7,400 years BP), for groups from Santa Rosa Island in each time period. Almost all statistics were consistent with 0 ($|Z| < 3$) for populations *X* outside California (Supplementary Data 4). The only exceptions were a significant genetic affinity of the younger group of individuals from Santa Rosa Island to several groups in the south. This affinity included individuals dated to around 500 years BP from the southern tip of the peninsula of Baja California. Individuals dated to about 1,000 years BP from Sonora in Northwest Mexico (LaPlaya, Cerro De Trincheras and Tayopa) had similar results, as did a group from Northern Durango Mexico (Cueva de los Muertos Chiquitos). The MX_LaPlaya/CerroDeTrincheras_600 BP individuals gave the most consistent signals of extra affinity to the later individuals from Santa Rosa Island, a result that is in keeping with them being the most geographically proximal to California. Significant signals were also present when comparing individuals dated to 7,000 years BP from Carpinteria (on the California mainland coast across the Santa Barbara channel from Santa Rosa Island) relative to individuals dated to 600 years BP from the same area, and comparing the earliest individuals from the Southern Channel Islands (around 4,800 years BP) to the latest ones (about 900 years BP), with MX_LaPlaya/CerroDeTrincheras_600 BP individuals again showing greater affinity to the later groups relative to the earlier ones.

We used qpAdm to estimate ancestry in individuals from California as a mixture of sources related to the following two proxies highlighted by the preceding analyses: USA-CA_SantaRosa_7400 BP and MX_LaPlaya/CerroDeTrincheras_600 BP (Fig. 3). qpAdm is designed to give unbiased estimates even if the relationships are distant. We estimated Mexico-related ancestry of 20 ± 8% in USA-CA_SantaRosa_4900 BP (± is one standard error), 22 ± 6% in USA-CA_SantaRosa_3200 BP, 23 ± 6% in USA-CA_SantaRosa_3000 BP and 37 ± 5% in USA-CA_SantaRosa_300 BP, with all models fitting at $P > 0.05$ (Supplementary Data 5). By contrast, USA-CA_Carpinteria_7000 BP, USA-CA_SanNicolasIsland_4800 BP and USA-CA_Goleta_3000 BP were consistent with being directly descended without admixture from earlier groups in the same region (the estimates of Mexico-related ancestry were not significantly different from zero: −1 ± 6%, 5 ± 6% and −3 ± 11%, respectively; $P > 0.05$ for all). This result is consistent with them being a clade and clustering with the oldest individuals from Santa Rosa Island, and suggests that the migration was south to north (Supplementary Data 5). Note that the USA-CA_Goleta_3000 BP individual has not been radiocarbon dated and could therefore be older than 3,000 years BP. When SanNicolas_4800 BP or Carpinteria_7000 BP were used as sources

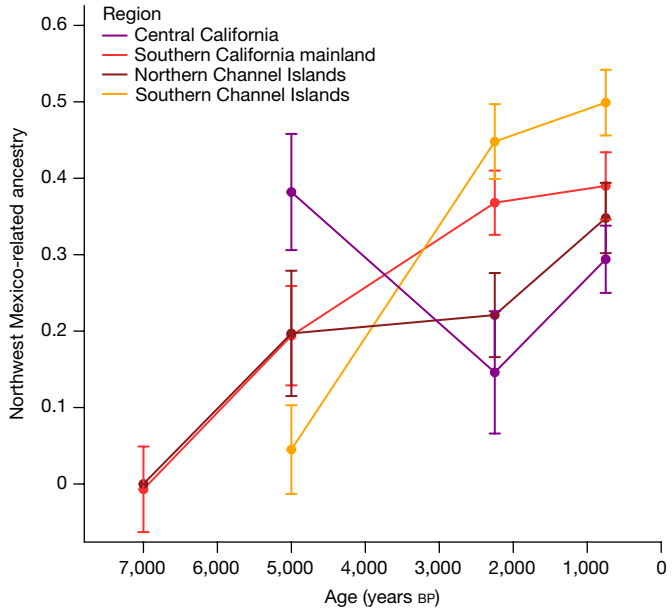


Fig. 3 | Northwest Mexico-related ancestry at different regions over time from qpAdm. Each data point represents the mean MX LaPlaya/CerroDeTrincheras 600 BP-related ancestry in a bin of time (8,000–6000 BP, 6,000–4,000 BP, 4,000–1,500 BP and <1,500 BP) with the number of individuals for Central California in each time bin respectively as (0, 1, 1, 15), Southern California mainland (6, 5, 6, 30), Northern Channel Islands (5, 2, 7, 5), and Southern Channel Islands (0, 19, 15, 17). Bars represent ±1 standard error, derived from a weighted block jackknife over 5-Mb blocks.

instead of SantaRosa_7400 BP, the estimates were not significantly different (Supplementary Data 5).

These statistics show that there was gene exchange among people in California and peoples related to those of Northwest Mexico in the second half of the Holocene beginning at least by the 4,900 years BP date of USA-CA_SantaRosa_4900 BP. O'odham speakers (Pima) currently occupy the region in Mexico where the individuals who maximize these affinities reside, and the modern O'odham in our dataset have very similar ancestry as the ancient Northwest Mexicans (Fig. 2, Extended Data Fig. 2 and Supplementary Data 4). The O'odham language belongs to the Uto-Aztecan family[10], and based on the degree of language diversity in the region, the geographical distribution of languages and knowledge about the rate at which languages evolve, linguists have argued that the majority of ancient Northwest Mexicans had almost certainly spoken Uto-Aztecan languages by 2,900–500 years BP, the time period of the ancient individuals from Mexico in our dataset[2–4,17]. Mexico-related ancestry increased over time, reaching the highest levels in the later populations of the Southern Channel Islands (Fig. 3). The highest proportion (44–51%) of Mexico-related ancestry occurred in later individuals from the Southern Channel Islands (San Nicolas, San Clemente and Santa Catalina), a result consistent with the observation that Indigenous people in that area at the time of colonial contact spoke an Uto-Aztecan language (Nicoleño). The present-day people in the area, the Tongva (some refer to themselves as Gabrieliños), also speak a closely related Uto-Aztecan language variety. By contrast, Chumash people from the Northern Channel Islands speak a language from an unrelated family, and show correspondingly less genetic affinity to people who speak a Northwest Mexican Uto-Aztecan language[1].

A notable aspect of the spread of ancestry related to individuals dated to 5,200–2,000 years BP from Mexico is its geographical extent, with evidence appearing from Central to Southern California and potentially even Baja California. Baja California could not be rigorously tested given the absence of a time transect, but when modelled as mixture of USA-CA_SantaRosa_7400 BP and MX_LaPlaya/CerroDeTrincheras_600 BP, the groups from Baja California all were inferred to have more than 60% Mexico-related ancestry. Further evidence that this spread must have been mediated in large part by migration into California is that we observed no evidence of an increase in California-related ancestry in Mexico. Modelling MX_LaPlaya/CerroDeTrincheras_600 BP as a mixture of groups related to MX_LaPlaya/CerroDeTrincheras_2400 BP and USA-CA_Carpinteria_7000 BP, we obtained a well-fitting model of LaPlaya/CerroDeTrincheras_600 BP as descending only from MX_LaPlaya/CerroDeTrincheras_2400 BP (−0.5 ± 2.3% USA-CA_Carpinteria_7000 BP ancestry; Supplementary Data 5). Taken together, our results document more than 5,000 years of movement of ancestry from people related to modern Northwest Mexican Uto-Aztecan speakers into the Channel Islands and mainland.

A notable exception to this pattern of monotonically increasing affinity to Northwest Mexico over time is an individual dated to approximately 5,200 years BP from Central California (USA-CA_PacificGrove_5200 BP), who could be well modelled as having 38 ± 8% Mexico-related ancestry (Supplementary Data 5), with later individuals from Central California in the dataset having similar or lower ancestry proportions. Although based on a sample size of one, these results demonstrate the presence of Mexico-related ancestry by this time. A non-monotonic pattern is what would be expected from a south-to-north migration by about 5,200 years BP followed by mixture with other (northern) groups without recent Mexican relatedness.

These findings show that ancestry related to that common in ancient and present-day people in Northwest Mexico began spreading at least as far north as Central California beginning at least 5,000 years ago, documenting demographically significant mid-Holocene gene flow between the two regions before the spread of agriculture (starting about 4,100 years BP)[9,18]. Further evidence for the presence of Mexican-related ancestry in California before the spread of maize-based agriculture northward from Mexico into the US Southwest comes from the presence of an individual dated to about 4,900 years BP from Santa Rosa Island and three individuals dated to around 4,700 years BP from Goleta with significantly increased (20 ± 8% and 19 ± 7%, respectively) Mexico-related ancestry. A group of individuals dated to approximately 1,900 years BP from Lovelock Cave, Nevada (Great Basin region)[19] showed one significant signal in comparison to more recent individuals from California and more ancient individuals ($Z = 3.3$ for affinity to SantaBarbara_900 BP relative to SantaBarbara_4600 BP). Although this could potentially reflect a random statistical fluctuation due to multiple hypothesis testing (Supplementary Data 4), it is a sufficiently strong signal to provide tentative evidence of population movement. More individuals from groups from the Great Basin would be necessary to test the extent to which such migration affected the Great Basin and California, and to test for evidence that the effects on California might have been mediated through the Great Basin.

The strongest case for agriculture being the vector for a northward spread of Uto-Aztecan languages from Central Mexico to the western USA has been the argument that the expansion of farming is the only process that could have been sufficiently demographically transformative to propel language change[20]. In this argument, agriculture led to an increase in population size in the US Southwest, which propelled individuals north into California even though agriculture never spread there. However, our analysis shows that major migration from south to north affected Central California by about 5,200 years BP, a time before agriculture began spreading and coinciding with the period when linguists have argued that 'Old Uto-Aztecan' had reached the San Joaquin Valley before being displaced by Yokutsan languages[1,10]. This result increases the weight of evidecne supporting the theory that the migrations before 5,200 years BP could have been the events that brought Uto-Aztecan languages to the region. However, our data also document a later increase in Mexico-related ancestry in the Southern California mainland and Southern Channel Islands between

5,000 and 3,000 years BP. This is a notable finding as the individuals from California in our dataset are primarily in the region occupied by speakers of the Takic subgroup of the Northern Uto-Aztecan branch, whereas the individuals from Northwest Mexico are primarily in the region occupied by speakers of the Piman subgroup of the Southern Uto-Aztecan branch. The date of the split of these two branches is estimated to be older than 7,000 years BP by some reconstructions[9], and 5,000–4,000 years BP in other reconstructions[11,13]. Our findings of south-to-north migrations into California both before 5,200 years BP and from 5,000 to 3,000 years BP could be consistent with both reconstructed split times.

Our genetic findings inform the debate about the likely homeland of Uto-Aztecan languages, beyond undermining the strongest argument in favour of the theory of an origin among agriculturalists from Central Mexico. This is because our results prove that movements of people associated with the spread of agriculture cannot have been the only demographically significant south-to-north migration. One group of linguistic reconstructions has suggested that Proto-Uto-Aztecan languages were spoken by hunter–gatherers living between Southern Arizona and Northern Mexico (immediately to the northeast of the MX_LaPlaya/CerroDeTrincheras individuals), in a woodland–grassland homeland in proximity to montane forests. This proposal is based on evidence that the reconstructed proto-language contained words for animals and plants from this region (for example, agave, long-needled pine, hawk and owl)[13]. Our genetic findings of south-to-north migrations into California beginning before 5,200 years BP and continuing until at least 3,000 years BP—alongside archaeological evidence of material culture exchange between Mexico and California (for example, the spread of the contracting stem dart point[21] and turquoise[22]) at least 4,000 years ago[23]—increases the weight of evidence for this theory. Conversely, the fact that we do not observe a significant increase in Great Basin (Lovelock Cave-related) or California Central Valley-related ancestry (Supplementary Data 5) in Northern Mexico decreases the weight of evidence for either a Great Basin[9] or a California Central Valley origin[10].

Genetic continuity and immigration in Central California

We assessed the ancestry of the oldest individual from Central California by comparing her to the oldest individuals from Santa Rosa Island with the statistic f_4(Mbuti,Test; USA_CA_PacificGrove_5200 BP, USA_CA_SantaRosa_7400 BP). The most significant attraction to PacificGrove_5200 BP is with MX_CA_Pericues_500 BP.SG ($Z = 6.46$, $f_4 = 0.00413$), a relatively isolated group that lived at the southern tip of Baja California in present-day Mexico. This signal was similarly strong when comparing PacificGrove_5200 BP to USA_CA_CalaverasCounty_1500 BP ($Z = 6.47$, $f_4 = 0.00384$, for attraction to Pericues_500 BP), a group about 290 kilometres to the east of Pacific Grove and probably within the territory seasonally occupied by speakers of Washo[1,24,25]. This signal is driven by components of ancestry other than Northwest Mexican, as Northwest Mexico-related ancestry in Pericues is of similar proportions in the later individuals from Central California (PacificGrove_200 BP, MontereyBay_1000 BP, Carmel_600 BP and Castroville_900 BP had between 24 and 42% Northwest Mexico-related ancestry), yet PacificGrove_5200 BP still had a significant attraction to Pericues_500 BP ($2.1 < Z < 4.3$, $0.00209 < f_4 < 0.00427$) relative to these groups (Extended Data Fig. 4 and Supplementary Data 4). Technical differences between shotgun and capture ancient DNA methods are also not likely to be an artefact and causing this effect. This is because the groups being compared for their affinity to the shotgun-sequenced groups from Pericues both had 1.24 million capture processing. Consequently, there should not be a differential affinity to shotgun-sequenced groups (as would be expected if one group had shotgun processing and the other had capture). This result raises the possibility that Pericues-related ancestry decreased over time, with the caveat that the approximately 5,200 years BP data point is from only one individual and more individuals are needed to understand the distribution of these ancestries over time.

The Baja California-related signal in the individual dated to 5,200 years BP from Central California is potentially consistent with a previous hypothesis of an earlier linguistic substrate widespread in California and Baja California (Fig. 1a), conjectured to have been broadly replaced later in time in Central California by speakers of Utian languages coming from inland to the coast approximately 4,000 years BP[1,5,26,27]. This would plausibly have been accompanied by migrations into the region within this time period, as anthropological evidence shows that language changes are often mediated by movements of people[28]. One possible source for this migration is Calaveras County in the Eastern Central Valley, as we observed a genetic affinity of CalaverasCounty_600 BP for Carmel_600 BP relative to PacificGrove_5200 BP ($Z = 3.7$, $f_4 = 0.0016$), consistent with migration between these regions from 5,200 years BP to 600 years BP (Carmel_600 BP was used for comparison because it was geographically the closest to Pacific Grove and had multiple high-coverage individuals sequenced). However, the individuals from Calaveras County were thought to have also spoken a non-Utian language (Washo)[24,25], so migration from this region does not neatly fit into the Utian migration hypothesis. Denser sampling from 5,000 to 3,000 years BP would be necessary to determine with more confidence the geographic origin of the source population that moved into the Central California coast and provide a clearer picture of the history of this region.

Movement of people to Central California did not fully displace the original ancestry in the region, as later populations from Central California have ancestry related to the PacificGrove_5200 BP individual. This result is consistent with previous evidence of a degree of local continuity[29] (f_4-statistics show significant affinity of PacificGrove_5200 BP to younger individuals from the coast of Central California relative to SantaRosa_7400 BP, though the statistics are non-significant or only marginally significant when compared with groups from Calaveras County ($2.1 < Z < 3.0$), possibly due to lack of power (Supplementary Data 4)). When we modelled later individuals from Central California as a mixture of PacificGrove_5200 BP and CalaverasCounty_600 BP, we found well-fitting models with between 55 ± 14% and 76 ± 9% related ancestry to PacificGrove_5200 BP (Supplementary Data 5). This result shows that the largest fraction of ancestry is consistent with having deep local roots, similar to the pattern in Southern California.

Relationship to the earliest sequenced Native Americans

Early Holocene individuals from Brazil, Chile and Nevada (Brazil_LapaDoSanto_9600 BP, Chile_LosRieles_12000 BP and USA_NV_SpiritCave_10000 BP) share more alleles with USA-MT_Anzick_12800 BP associated with the Clovis culture than with later populations in the same regions[16,19]. Our analysis showed that this specific affinity to an individual from the Clovis culture persisted for many more millennia in the Chumash region of California than it did in any other sampled regions of the Americas. Symmetry f_4-statistics and outgroup–f_3 statistics assessing the rate of allele sharing with Anzick relative to an outgroup such as USA-AK_USR1_11500 BP showed that the earliest individuals from California (USA_CA_SantaRosa_7400 BP and USA_CA_Carpinteria_7000 BP) had affinity similar to those of the earliest individuals from Brazil, Chile and Nevada. Moreover, these individuals had significantly more affinity to Anzick relative to Peru_Lauricocha_8600 BP ($Z > 3.3$) and Peru_Cuncaicha_9000 BP ($Z > 2.4$) (Supplementary Data 4). This result suggests that the ancient individuals from California descend from an early spread of people with affinity to USA-MT_Anzick_12800 BP and have more affinity to this lineage than the earliest individuals of similar age from the Central Andes[14,16].

We assessed whether the ancient individuals from California and Mexico in our dataset, particularly the younger ones, showed any

evidence of ancestry from the other main branch of Northern Native American (NNA) ancestry, related to ancient individuals from Southern Ontario (Canada_Lucier_4800-500 BP). We computed a statistic sensitive to this, f_4(Mbuti.DG, Canada_Lucier_4800-500 BP; Test, Chile_LosRieles_12000 BP), and found that it was consistent with zero for all ancient groups from California and Mexico, thereby providing no evidence for NNA ancestry (Supplementary Data 4). We also created an admixture graph using qpGraph with Canada_Lucier_4800-500 BP as an outgroup. We found admixture graphs with plausible fits (all Z-scores of <3.0), with almost all ancient groups from California not requiring additional ancestry from branches related to Canada_Lucier_4800-500 BP (Extended Data Fig. 5 and Supplementary Data 4). The exceptions were USA_CA_SanClemente_500 BP and USA_CA_SantaCatalina_400 BP.SG, who had poor fits ($3.05 < Z < 3.25$), although not due to an attraction of them with Canada_Lucier. We view these as probable artefacts given that the SanClemente_900 BP group had a good fit and that the SantaCatalina group could be a poor fit owing to technical biases that differ between shotgun sequencing and capture data. Overall, we did not find consistent evidence of NNA ancestry in the ancient individuals from California and Mexico. This result is in contrast to a previous study[14] that modelled intermediate proportions in all the individuals from California they reported, all of which we reanalysed here (Supplementary Data 4).

Ancient Mexicans harboured ancestry from non-Clovis-associated southern expansions

When we modelled ancient individuals from Northwest Mexico, in all fitting admixture graphs ($Z < 3.0$ for the worst residual), the predominant ancestry of this group of individuals was more basal (early splitting) than Chile_LosRieles_12000 BP and the ancient individuals from California, although still less basal than NNA (the best graph we found is presented in Extended Data Fig. 5b). This is due to the groups from Mexico being on an SNA lineage that is basal to Anzick and does not have the same affinity to Anzick that the individuals from Los Rieles and California have. This finding is also supported by f_4-statistics, which showed a significant affinity of USA-MT_Anzick_12800 BP for USA_CA_SantaRosa_7400 BP and USA_CA_Carpinteria_7000 BP relative to MX_Tayopa_1000 BP ($Z = 5.6$), MX_Cueva de los Muertos Chiquitos_1100 BP ($Z = 4.4$) and MX_LaPlaya/CerroDeTrincheras_600 BP ($Z = 5.0$) (Supplementary Data 4). Relative to ancient people from Mexico, there was also a significant affinity based on f_4-statistics between USA-MT_Anzick_12800 BP and Brazil_LapaDoSanto_9600 BP ($Z = 4.8$), Chile_LosRieles_12000 BP ($Z = 3.8$) and USA-NV_SpiritCave_10000 BP ($Z = 4.1$), but they were consistent with zero in comparisons with Peru_Cuncaicha_9000 BP (Z = 1.1) and Peru_Lauricocha_8600 BP ($Z = 0.2$) (the results were also qualitatively the same when using only transversion SNPs; Supplementary Data 4). This result suggests that the earliest individuals from California might have shared ancestry with the Anzick-related individuals found in Chile, Brazil and Nevada[16,19], whereas the ancient people from Mexico in our dataset might have shared ancestry with the earliest people from Peru. These findings appear superficially similar to those from a study[30] that found a contribution from a lineage basal to Anzick in Aridoamerican and some Mesoamerican Mexicans (all of our ancient individuals from Mexico were from Aridoamerica). However, the divergent SNA lineage we infer for the groups from Mexico is different from the previous findings of UPopA1 or UPopA2 lineages contributing to Mexicans[19,30,31]. This is because both UPopA populations were inferred to be lineages more basal to that of both SNA and NNA, whereas our deep Mexican lineage is consistent with being SNA.

Relationship to people of other regions of the world

We tested for evidence of Polynesian ancestry based on suggestions that the *tomol* (plank canoe) of the Chumash and Tongva might have had influence from Polynesia[32]. We used f_4-statistics to test for genetic affinity between individuals from Polynesia (a modern Native Hawaiian, an ancient individual from Tonga or another ancient sample of Polynesian ancestry) and individuals from California from 7,100 years BP to 300 years BP relative to SantaRosa_7400 BP. We also used qpAdm to test for Polynesian ancestry in the individuals from California. We did not find evidence of Polynesian ancestry in any of the individuals (Supplementary Data 4 and 5), a result consistent with arguments against Polynesian contribution to *tomol* development[33]. We also tested for excess relatedness to Australasians (population Y) using f_4(Mbuti, Onge or Papuan; Test, Mixe) and found no evidence for it in any of the ancient individuals from California and Northwest Mexico (Supplementary Data 4).

We tested the claim of an ancient migration into the Central Andes after about 4,200 years BP by people distinctively related to ancient individuals from Southern California[16]. We confirmed the previously reported signal, and made a new observation, namely that the signal can only be perceived when outgroups with Mexican-related ancestry are used in qpAdm. Thus, when USA-CA_SantaRosa_7400 BP, USA-CA_Carpinteria_7000 BP or USA-CA_SanNicolas_4800 BP are used as outgroups with no evidence of Mexican ancestry, there is no signal of extra migration into Peru after 4,200 years BP ($P > 0.05$; Supplementary Data 6). However, when groups from California with Northwest Mexico-related ancestry are used as outgroups, the signal is present ($P < 0.005$). In the publications reporting this finding[16,34], groups from California with Mexico-related ancestry were used as outgroups. This result indicates that the signal, also found by an independent analysis that used Mixe in Southern Mexico among the outgroups[19], might have been due to a migration of Mexico-related ancestry simultaneously into both the Central Andes and the California Channel Islands after about 4,200 years BP. The signal might also be due to Central Andes-related south-to-north migration affecting Mexico after about 4,200 years BP without new migration into the Central Andes[35].

Community sizes in the Channel Islands were smaller than on the mainland

We analysed runs of homozygosity (ROHs) to estimate effective community sizes of the ancient groups from California and Mexico, referring to the size of the mate pool in the last handful of generations. For this purpose we used the software hapROH, analysing 85 ancient individuals with data at over 400,000 SNPs. The most notable patterns were evident in small (4–8 centimorgans (cM)) and mid-size (8–20 cM) ROHs, which occurred at higher rates in the Channel Islands than the mainland of Southern and Central California (Fig. 4a, Extended Data Fig. 6 and Supplementary Data 7). This result indicates that mothers and fathers of individuals often descended from the same ancestors in the last handful of generations. We estimated effective community size (N_e) using the length distribution of ROHs at all spatial scales 4–20 cM, which arise from shared ancestry at different time depths in the last 50 generations and make it possible to detect signals of size change and migratory rates with neighbouring communities over this temporal scale. When analysing individuals younger than 1,600 years BP, and after filtering out individuals with evidence of recent close-kin unions (those with ROH fragments larger than 20 cM that total more than 50 cM), Northern and Southern Channel islands had an estimated N_e of 388 ± 42 and 175 ± 13, respectively, similar to pre-agriculture Archaic Caribbean sites[36] (232 ± 8) and ancient groups from Patagonia (171 ± 7), Guam (333 ± 11) and Saipan (375 ± 16) (Table 1). Southern California mainland and Central California had N_e values of 519 ± 49 and 418 ± 39, respectively (Fig. 4b and Supplementary Data 7). By contrast, ancient Northwest Mexico had N_e of 839 ± 74, which is in line with estimates using modern genomes (Figure 3a from ref. 30) and similar to estimates for ceramic-associated Caribbean sites with agriculture[36] (681 ± 21) and similar-aged Peruvian groups (817 ± 51). Effective community sizes in groups from Southern Mexico were larger than

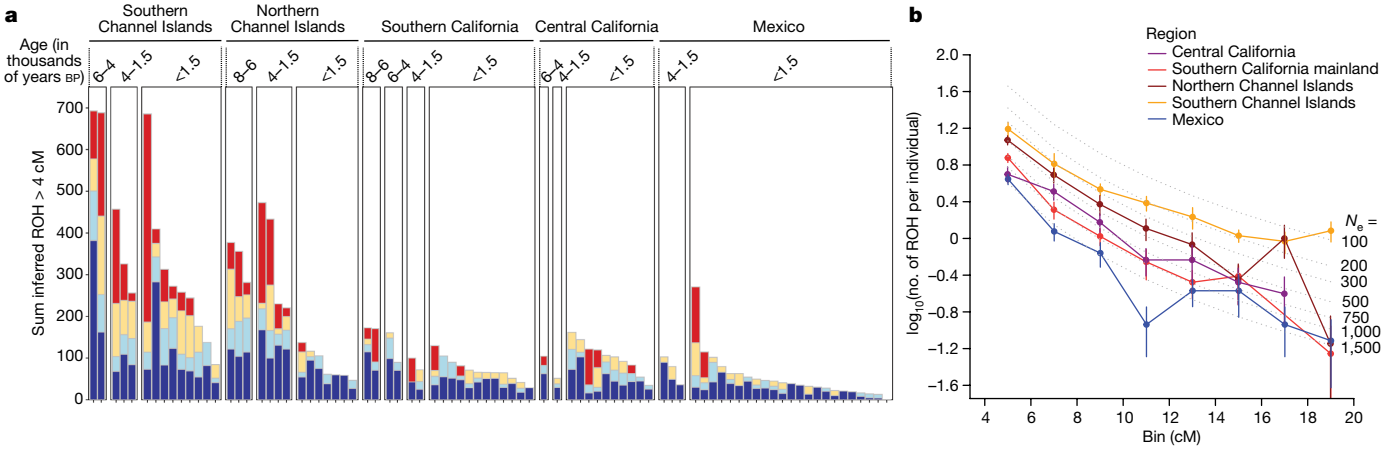


Fig. 4 | ROHs in all ancient peoples from California and Mexico. a, ROHs in all individuals with sufficient coverage. Dark blue indicates sum ROHs of 4–8 cM fragments; light blue 8–12 cM; tan 12–20 cM; and red 20–300 cM. Different regions grouped by age. Numbers above the bars are the age of the individuals in thousands of years BP. **b**, Average rate of ROH in different length bins for all ancient individuals. Points with no ROH fragments present in those bins are filtered out. No filtering was performed for these data; results after filtering out consanguineous individuals or individuals >16,00 years BP are in Extended Data Fig. 6.

the more northern ones (Cueva De Los Muertos Chiquitos = 1,105 ± 181, Tayopa = 895 ± 142, LaPlaya/Cerro De Trincheras = 605 ± 94, in order from south to north). This result could reflect southern settlements having more access to water and fertile land for agriculture, thereby allowing larger communities to develop. Alternatively, these patterns could reflect more frequent exchange of mates between southern villages than between northern villages, without implying that villages in the two regions were different in size from each other. The findings are further supported by analyses of conditional heterozygosity—rates of variation at sites polymorphic in an outgroup (Yoruba from Africa)—as we found lower heterozygosity in ancient people from Californian islands than in any other group, as expected if ancestral variation was lost owing to persistently small community sizes (Extended Data Fig. 7). However, the sizes of the mate pools in the Channel Islands and Southern California increased over time, as indicated by decreasing ROHs (Extended Data Fig. 6 and Supplementary Data 7).

Conclusion

The history of Indigenous peoples in California reflects late Pleistocene migrations into the region followed by mid-Holocene south-to-north migrations of people related to Uto-Aztecan speaking groups of Northwest Mexico. There was independent migration that affected the coast of Central California, correlated with ancestry found in inland Central California Valley populations. Our data and analyses demonstrate that the earliest sequenced people in the Chumash region were unusually closely related to the Clovis culture-associated Anzick individual of late Pleistocene age. In-place genetic continuity can be documented through the millennia down to modern Chumash as represented by sequences from 200 years BP. There has been substantial debate about whether early speakers of Uto-Aztecan languages originated as hunter–gatherers from the southwestern USA–northwestern Mexican border area, as maize farmers in Central Mexico, or as hunter–gatherers of the Great Basin region of the present-day USA who spread southward[9,10]. Our results show that ancestry related to present-day Mexican Uto-Aztecan speakers was present in admixed form in Central California at least 5,200 years BP and in Southern California at least 4,900 years BP, and provide no evidence for a spread of Central California-related or Great Basin-related ancestry southward into Mexico. This finding fits best with the scenario of hunter–gatherers moving both northwest into California and south into Mexico. Our results provide an alternative vector for the spread of Mexican ancestry to California than the spread of maize agriculture. This was the previous best argument for a south-to-north movement of Uto-Aztecan being associated with agriculture because the earliest evidence for maize expansion into the southwest is only about 4,100 years BP[9,37]. The finding of this ancestry in Central California at 5,200 years BP is also consistent with linguistic theories that Uto-Aztecan languages were already spoken in the Central Valley by the mid-Holocene[1,10].

It is possible that currently unsampled ancient groups from outside Southern California and Northwest Mexico (for example, the western Great Basin[10]) mixed into both of these regions, which could have produced some of the genetic signals. For example, there is evidence for drought conditions between about 6,300 and 4,800 years BP[38], particularly on the coasts and the northwest Mexico desert, whereas the Great Basin was wetter in this period. Some linguists have proposed that these conditions led to the diversification of Proto-Uto-Aztecan into its north and south branches, with the northern branch spreading up to the Great Basin region during this time[38,39]. There is also archaeological evidence for cultural exchange between the Great Basin and Southern and Central California between 5,900 and 4,700 years BP based on the distribution of *Olivella* grooved rectangle beads produced on the Southern Channel Islands and the adjacent Southern California

Table 1 | Effective population size estimates from hapROH

Group	Population size
Patagonia <1,500 years BP	171±7
Southern Channel Islands <1,600 years BP	175±13
Caribbean archaic <3,200 years BP	232±8
Venezuela ceramic 2,000–3,000 years BP	274±25
Guam <800 years BP	333±11
Saipan <800 years BP	375±16
Northern Channel Islands <1,500 years BP	388±42
Central California <1,600 years BP	418±39
Southern California mainland <1,500 years BP	519±51
Bolivia <1,700 years BP	663±62
Caribbean ceramic <1,700 years BP	681±21
Peru <1,800 years BP	817±51
Mexico <1,600 years BP	839±74

See Supplementary Data 7 for additional results. Error bars represent ±1 standard error.

coast[40] as well as in Central California[41], and the spread of obsidian throughout these areas[42].

Our analyses do not provide information about the geographical origin of the migration into Central California that we show began at least by 5,200 years BP. Collection of genetic data from present-day Indigenous groups from California, and analysis with ancient data in California and beyond, would provide additional insights. It is important to carry out such research in a way that is engaged with present-day Indigenous descendant groups, following approaches such as those taken in this and previous studies, and informed by recent discussions and recommendations concerning ethical analysis of DNA from ancient Indigenous individuals[14,16,19,29,36].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06771-5.

1. Golla, V. *California Indian Languages* (Univ. California Press, 2011).
2. Carpenter, J. P., Sanchez, G. & Villalpando, M. E. in *Traditions, Transitions, and Technologies: Themes in Southwestern Archaeology* (ed. Schlanger, S. H.) 245–258 (Univ. Press of Colorado, 2002).
3. LeBlanc, S. A. in *Archaeology without Borders: Contact, Commerce, and Change in the US Southwest and Northwestern Mexico* (eds Webster, L. D. & McBrinn, M. E.) 107–142 (Univ. Press of Colorado, 2008).
4. Mabry, J. B., Carpenter, J. P. & Sanchez, G. in *Archaeology without Borders: Contact, Commerce, and Change in the US Southwest and Northwestern Mexico* (eds Webster, L. D. & McBrinn, M. E.) 155–183 (Univ. Press of Colorado, 2008).
5. Breschini, G. S. *Models of Population Movement in Central California Prehistory* (Coyoye Press, 1984).
6. Erlandson, J. M. et al. Paleoindian seafaring, maritime technologies, and coastal foraging on California's Channel Islands. *Science* **331**, 1181–1185 (2011).
7. Johnson, J. R. & Lorenz, J. G. Genetics, linguistics, and prehistoric migrations: An analysis of California Indian mitochondrial DNA lineages. *J. Calif. Gt. Basin Antrhopol.* **26**, 33–64 (2006).
8. DeLancey, S. & Golla, V. The Penutian hypothesis: retrospect and prospect. *Int. J. Am. Linguist.* **63**, 171–202 (1997).
9. Merrill, W. L. et al. The diffusion of maize to the southwestern United States and its impact. *Proc. Natl Acad. Sci. USA* **106**, 21019–21026 (2009).
10. Shaul, D. L. *A PreHistory of Western North America: The Impact of Uto-Aztecan Languages*. (Univ. of New Mexico Press, 2014).
11. Greenhill, S. J. et al. A recent northern origin for the Uto-Aztecan family. Preprint at *SocArXiv* https://doi.org/10.31235/osf.io/k598j (2023).
12. Hill, J. H. Proto-Uto-Aztecan: a community of cultivators in Central Mexico? *Am. Anthropol.* **103**, 913–934 (2001).
13. Fowler, C. S. Some lexical clues to Uto-Aztecan prehistory. *Int. J. Am. Linguist.* **49**, 224–257 (1983).
14. Scheib, C. L. et al. Ancient human parallel lineages within North America contributed to a coastal expansion. *Science* **360**, 1024–1027 (2018).
15. Rasmussen, M. et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014).
16. Posth, C. et al. Reconstructing the deep population history of Central and South America. *Cell* **175**, 1185–1197.e22 (2018).
17. Hill, J. in *Examining the Farming/Language Dispersal Hypothesis* (eds Bellwood, P. & Renfew, C.) 331–340 (MacDonald Institute for Archaeological Research, 2003).
18. Vellanoweth, R. L. AMS radiocarbon dating and shell bead chronologies: Middle Holocene trade and interaction in western North America. *J. Archaeolog. Sci.* **28**, 941–950 (2001).
19. Moreno-Mayar, J. V. et al. Early human dispersals within the Americas. *Science* **362**, eaav2621 (2018).
20. Bellwood, P. et al. First farmers: the origins of agricultural societies. *Camb. Archaeol. J.* **17**, 87–109 (2007).
21. Glassow, M. A. et al. in *California Prehistory: Colonization, Culture, and Complexity* (eds Jones, T. L. & Klar, K. A.) 191–213 (Altamira Press, 2007).
22. Kelley, J. C. & Reyman, J. in *The Gran Chichimeca: Essays on the Archaeology and Ethnohistory of Northern Mesoamerica* (ed. Reyman, J. E.) 103–172 (Avebury Ashgate, 1995).
23. Coulam, N. J. The appearance of contracting stem dart points in the Western United States, diffusion or migration? *KIVA* **88**, 355–371 (2022).
24. Barrett, S. A. The Washo Indians. *Bull. Pub. Mus. Milwaukee* **2**, 1–13 (1917).
25. d'Azevedo, W. L. *Handbook of North American Indians. Great Basin.* Vol. 11 (Smithsonian Institution, 1986).
26. Eshleman, J. A. et al. Mitochondrial DNA and prehistoric settlements: native migrations on the western edge of North America. *Hum. Biol.* **76**, 55–75 (2004).
27. Kaestle, F. A. & Smith, D. G. Ancient mitochondrial DNA evidence for prehistoric population movement: the Numic expansion. *Am. J. Phys. Anthropol.* **115**, 1–12 (2001).
28. Siegel, J. S. in *Demographic and Socioeconomic Basis of Ethnolinguistics* 427–484 (Springer, 2018).
29. Severson, A. L. et al. Ancient and modern genomics of the Ohlone Indigenous population of California. *Proc. Natl Acad. Sci. USA* **119**, e2111533119 (2022).
30. García-Ortiz, H. et al. The genomic landscape of Mexican Indigenous populations brings insights into the peopling of the Americas. *Nat. Commun.* **12**, 5942 (2021).
31. Villa-Islas, V. et al. Demographic history and genetic structure in pre-Hispanic Central Mexico. *Science* **380**, eadd6142 (2023).
32. Jones, T. L. & Klar, K. A. Diffusionism reconsidered: linguistic and archaeological evidence for prehistoric Polynesian contact with southern California. *Am. Antiq.* **70**, 457–484 (2005).
33. Arnold, J. E. Credit where credit is due: the history of the Chumash oceangoing plank canoe. *Am. Antiq.* **72**, 196–209 (2007).
34. Nakatsuka, N. et al. A Paleogenomic reconstruction of the deep population history of the Andes. *Cell* **181**, 1131–1145.e21 (2020).
35. Kennett, D. J. et al. South-to-north migration preceded the advent of intensive farming in the Maya region. *Nat. Commun.* **13**, 1530 (2022).
36. Fernandes, D. M. et al. A genetic history of the pre-contact Caribbean. *Nature* **590**, 103–110 (2021).
37. da Fonseca, R. R. et al. The origin and evolution of maize in the southwestern United States. *Nat. Plants* **1**, 14003 (2015).
38. Carpenter, J., Sánchez, G., Sánchez, I. & Vierra, B. J. in *The Archaic Southwest: Foragers in an Arid Land* (ed. Vierra, B. J.) 98–118 (Univ. of Utah Press, 2018).
39. Carpenter Slavens, J. & Sánchez, G. Los cambios ambientales del Holoceno medio/Holoceno tardío en el desierto de Sonora y sus implicaciones en la diversificación del Yuto-Aztecano y la difusión del maíz. *Diálogo Andino* https://doi.org/10.4067/S0719-26812013000100013 (2013).
40. Kennett, D. J., Kennett, J. P., Erlandson, J. M. & Cannariato, K. G. Human responses to Middle Holocene climate change on California's Channel Islands. *Quat. Sci. Rev.* **26**, 351–367 (2007).
41. Fitzgerald, R. T., Rosenthal, J. S., Eerkens, J. W., Nicholson, D. & Spero, H. J. The distribution of Olivella grooved rectangular beads in the Far West. *J. Calif. Gt. Basin Anthropol.* **38**, 241–252 (2018).
42. Hughes, R. E. Obsidian studies in California archaeology. *Quat. Int.* **482**, 67–82 (2018).

Methods

Ethics approval

We acknowledge the Indigenous peoples of California and Mexico who supported this study and the ancient individuals whose skeletal remains we analysed. Studies of DNA from ancient individuals can have deep and important implications for present-day groups because they can reveal information about their ancestors, including their history and interactions with others, and because the physical handling of the skeletal materials can be sensitive to descendant communities. We performed this study in strong engagement and with participation from local Indigenous communities with closest ties to the ancient individuals we studied. We also performed this study according to ethical guidelines for working with human remains, treating the Indigenous ancient individuals with the respect owed to deceased people.

For the ancient individuals from California, the ancient skeletal remains we analysed were curated primarily at the Santa Barbara Museum of Natural History. The newly sequenced individuals from San Clemente were curated at the Peabody Museum of Archaeology and Ethnology. All ancient skeletal remains from California were repatriated to the tribes residing in the region where the ancient individuals originally lived, and the skeletal remains were reburied by the tribes (additional details provided in Supplementary Data 1). The exception to this were the newly sequenced individuals from San Clemente, whose skeletal remains were deemed culturally unidentifiable and for which an official federal register notice was posted, with discussions currently ongoing to determine the best approach for the repatriation of these ancient individuals. Co-authors J.R.J., N.N., B.H., P.L. and D.R. participated in multiple engagements with several Chumash groups in Southern California (B.H. is a tribal descendant of the Santa Ynez Band and meetings occurred with permission granted from the Santa Ynez, Barbareño and Barbareño-Ventureño bands), as well as with the Tongva in Southern California and Ohlone and Esselen groups in Central California. M. Armenta, an elder of the Santa Ynez Band of Chumash Indians and NAGPRA representative for the tribe, gave permission for DNA sequencing. M. Armenta and his colleague R. Saint-Onge, met with J.R.J. to formulate research goals. Several of the ancient individuals in Central California were sequenced and studied as part of long-term engagements by the late G. Breschini who obtained support for DNA testing by Ohlone tribal members, and by the late E. Rodriguez who was designated by the State Native American Commission as Most Likely Descendant (MLD) for the Monterey Bay area.

In Mexico, all legal authorizations were obtained for this work, sanctioned by the Consejo de Arqueología from the Instituto Nacional de Antropología e Historia. The research followed their guidance and was directed by archaeologists from Mexico (J.C., C.G.-M., J.M.-R., A.P.-M., M.E.V-C. and J.L.P.D.). Individuals from Trincheras and La Playa were part of the PIPANOM (Proyecto de investigación de poblaciones antiguas en el norte y occidente de México) project and curated at different centres of the Instituto Nacional de Antropología e Historia in West and North Mexico. Individuals from San Lorenzo, Tayopa and Coyote Cave were approved for research by collections committees at the Peabody Museum of Anthropology and Ethnology and the American Museum of Natural History. Individuals from Cueva de los Chiquitos were curated at the Anthropology Department of University of Nevada, Las Vegas. In Mexico, consultation occurred through the Mexican government cultural agencies by J.L.P.D. and J.S., including with groups in Northwest Mexico with closer connections to the Indigenous cultures. Information for repatriation of the ancient individuals from Mexico held at American institutions to their homelands was provided to the INAH such that repatriation efforts for these individuals are being guided by their cultural agencies.

During all community engagements, results were shared and support was obtained for the data to be made public, with possible implications of this also discussed. With the help of Indigenous community members, we prepared a frequently asked questions document (Supplementary Data 8) to assist the general public with understanding the findings in this study. This project involved components in both the USA and in Mexico providing Indigenous tribal and local community members training in genetics, archaeology and ancient DNA research techniques, as well as career advice and mentoring. Community members provided feedback on the paper before final publication, with the goal of ensuring sensitivity of the final paper to community perspectives.

We emphasized in these presentations that scientific discovery is a dynamic and iterative process that builds on itself, and that this study is not the final word even on a scientific level, as additional studies will refine and improve the models and interpretations here. We also emphasized that genetic ancestry is different from identity, which is often based on social relationships rather than biological ties; genetic findings should never be seen as challenging cultural identity.

Direct accelerator mass spectrometry ^{14}C bone dates

We generated 54 new direct accelerator mass spectrometry ^{14}C dates for 54 ancient individuals, which we added to previously reported ^{14}C dates for other individuals as well as archaeological context information to provide information on chronology (Supplementary Data 1).

Calibration of radiocarbon dates

All calibrated ^{14}C ages were calculated using OxCal (v.4.4) with different mixtures of the Northern Hemisphere terrestrial (IntCal20)[43] and marine (Marine20)[44] calibration curves. Marine dietary contribution was estimated using stable carbon and nitrogen isotope measurements from collagen (Supplementary Data 1). Nitrogen is sensitive to the relative importance of marine dietary resources, with δ^{15}N values of about 11.5‰ expected for a wholly terrestrial diet and around 22.0‰ expected for a predominately (about 90%) marine diet. We used nine categories of calibration curve mixing defined by 10% increments (10–90%), each with an applied uncertainty value of ±10%. For individuals from the Santa Barbara Basin, we used a variable marine ΔR model based on the variable reservoir ages for this region from paired organic and planktonic marine foraminiferal carbonate in laminated varves and linear regression[45]. For individuals from the Monterey Bay area, we used the nearest published ΔR values from a previous study[46] (based on modern molluscs). In both cases, ΔR values were recalculated according to the Marine20 calibration curve.

Ancient DNA laboratory work

We extracted DNA using a method that is optimized to retain small DNA fragments[47–49]. We converted the DNA into a form that could be sequenced using a double-stranded library preparation protocol, usually pretreating with the enzyme uracil-DNA glycosylase (UDG) to reduce the characteristic cytosine-to-thymine errors in ancient DNA[50]. For some libraries, we substituted the MinElute columns used for cleaning up reactions with magnetic beads, and the MinElute column-based PCR cleanup at the end of library preparation with SPRI beads[51,52]. We enriched the libraries both for sequences overlapping mtDNA[53] and for sequences overlapping about 1.24 million nuclear targets after two rounds of enrichment[54–56]. We sequenced the enriched products on an Illumina NextSeq500 instrument using v.2 150 cycle kits for 2 × 76 cycles and 2 × 7 cycles, or on an Illumina HiSeq X10 instrument using 2 × 101 cycles and 2 × 8 cycles, and sequenced up to the point that the expected number of new SNPs covered per 100 additional read pairs sequenced was approximately less than 1.

Computational processing of initial sequence data

We merged paired reads that overlapped by at least 15 nucleotides using SeqPrep (https://github.com/jstjohn/SeqPrep), taking the highest quality base to represent each nucleotide, and then mapped the sequences to the human genome reference sequence (GRCh37 from the 1000 Genomes Project) using the samse command

of the Burrows–Wheeler aligner (v.0.6.1)[57]. We removed duplicate sequences using Picard (v.2.23.0; http://broadinstitute.github.io/picard/). We trimmed two nucleotides from the end of each sequence and then randomly selected a single sequence at each site covered by at least one sequence in each individual to represent their genotype at that position (pseudo-haploid genotyping).

Contamination estimation

We assessed evidence for ancient DNA authenticity by measuring the rate of damage in the first nucleotide (flagging individuals as potentially contaminated if they had a less than 3% cytosine-to-thymine substitution rate in the first nucleotide for a UDG-treated library and less than 10% substitution rate for a non-UDG-treated library). We used contamMix to determine evidence of contamination based on polymorphism in mtDNA[58] and used ANGSD to determine evidence of contamination based on polymorphism on the X chromosome in males[59]. We also used ContamLD[60] to estimate the rate of contamination in autosomal DNA. We removed (but still report) 7 individuals from analyses with point estimates of more than 7% contamination according to ContamLD, 5% from ANGSD or 10% from contamMix applied to mtDNA.

Kinship analyses

We analysed all pairs of individuals to determine whether any of them had evidence of close genetic relatedness. In these analyses, we examined all non-CpG autosomal sites and calculated an average mismatch rate at all SNPs covered by at least one sequence for both individuals. We then compared these rates to the rate of difference between the two chromosomes in each individual, which was assumed for this analysis to come from individuals not closely related to each other[61]. We removed from group analyses all individuals inferred to have a first cousin or closer relationship with another person in the dataset (retainin the individual with higher coverage data) but analysed them in individual-level analyses.

Analyses of uniparental haplogroups

We determined the mtDNA haplogroups for all individuals by analysing the .bam files, restricting to reads with MAPQ ≥ 30 and base quality ≥ 20. We created consensus sequences using samtools and bcftools (v.1.31) with majority rule and then inferred the haplogroup using HaploGrep2 with Phylotree (v.17). We determined Y chromosome haplogroups with the same filtering as for mtDNA reads. We called Y chromosome haplogroups on the basis of the most derived mutation using the nomenclature of the International Society of Genetic Genealogy (http://www.isogg.org; v.14.76, April 2019) and using a previously reported method[62] using YFull YTree (v.8.09) phylogeny (https://github.com/YFullTeam/YTree/blob/master/ytree/tree_8.09.0.json).

Admixture clustering analysis

Using PLINK2 (ref. 63), we first removed SNPs in high linkage disequilibrium using the command –indep-pairwise 50 5 0.5. We removed individuals and genetic variants with high missingness and variants with low minor allele frequency using the command –mind 0.9 –geno 0.5 –maf 0.01. We ran ADMIXTURE[64] with 10 replicates, reporting the replicate with the highest likelihood and stopping at $K = 7$ owing to the significantly higher cross-validation errors that occur after this point (the cross-validation errors from 2 until 9 are, in order: 0.807, 0.822, 0.824, 0.850, 0.870, 0.878, 0.947 and 0.962). We therefore show results for $K = 2$ to 7 in Extended Data Fig. 1.

Testing of group homogeneity using qpWave

We used the qpWave methodology[56] in the ADMIXTOOLS package (v.6.0) to test for genetic homogeneity within groups. We tested all pairs of individuals within each group with three outgroups chosen to be in close geographical proximity and age to the test group. Pairs of individuals were considered to be consistent with being genetically homogeneous relative to the outgroups if their P values were greater than 0.01.

f-statistics

We used the qp3pop and qpDstat packages in ADMIXTOOLS (v.6.0) to compute f_3-statistics and f_4-statistics (using the f4Mode: Yes parameter). We computed standard errors using a weighted block jackknife over 5-Mb blocks. We computed outgroup f_3-statistics of the form f_3(Mbuti; Pop1, Pop2), which measures the shared genetic drift between population 1 and population 2. We used these statistics to create a MDS plot and neighbour-joining tree by creating a matrix of outgroup–f_3 statistics values between all pairs of populations and converting to distances by either taking the inverse of the values for the neighbour-joining tree or subtracting the values from 1 for the MDS plot. We generated the MDS plot using R, and the neighbour-joining tree using the PHYLIP (v.3.696)[65] neighbour function setting USA_MT_Anzick-1_12800 BP as the outgroup. We plotted the tree using Itol[66] with all lengths set to ignore.

F_{ST} analyses

We used smartpca (v.5.0)[67] to compute F_{ST} values between all groups with at least two individuals. We used fstonly: Yes and inbreed: Yes with all other settings left at default. We then used this matrix to create a heatmap using a hierarchical clustering-based dendrogram in R with symm=T.

Admixture graph analyses

We used the qpGraph package[68] in ADMIXTOOLS (v.6.0) to fit models of population splitting and mixture to the allele frequency correlation statistics (*f*-statistics) relating the different groups. We used a basic graph for Native Americans[45,69] and then successively added additional populations in all combinations, allowing up to one admixture from the previously fit groups into the graph. We took the graph with the lowest maximum Z-score and then repeated the process, adding another population until all populations of interest were added. Our process for choosing the added populations was to start with the oldest populations and those known to have the most divergent ancestries and then add the younger populations. We also explored choosing alternative orders of populations to determine whether the final graphs were affected by the order in which populations were added (they were not).

Quantifying mixture using qpAdm and qpWave

We used the qpAdm methodology[56] in the ADMIXTOOLS package (v.6.0) to estimate the proportions of ancestry of populations deriving from a mixture of reference populations by assessing the relative shared genetic drift with a set of 'outgroup' populations. We set the parameters as details: Yes, which reports a normally distributed Z-score for fit (estimated with a block jackknife), and Allsnps: Yes to maximize information content in the context of the relatively low coverage of many of the individuals. We computed P values through block jackknife resampling and using a likelihood ratio test (two-sided). We considered a model to be a plausible fit if $P > 0.01$. For qpWave analyses, we analysed all triplets with Brazil_LapaDoSanto_9600 BP or Chile_LosRieles_12000 BP as pop1, Peru_Lauricocha_8600 BP or Peru_Cuncaicha_9000 BP as pop2, and Peruvian, Chilean or Bolivian groups after 4,200 years BP as pop3, as previously described[16,34]. We used the following outgroups to test for Anzick-1 relatedness in Lapa Do Santo and Los Rieles and the California Channel Island groups without evidence of Mexico-related ancestry: USA-CA_SanNicolas_4800 BP.SG, USA-CA_Carpinteria_7000 BP, USA-CA_SantaRosa_7400 BP, USA-NV_SpiritCave_10000 BP.SG, USA-MT_Anzick_12800 BP.SG, Russia_MA1_2400 BP.SG, Papuan.DG and Karelia_HG.SG. To study individuals from the California Channel Islands with evidence of Mexico-related ancestry, we replaced Carpinteria_7000 BP and SanNicolas_4800 BP in these analyses with USA-CA_SantaBarbara_600 BP and USA-CA_SantaBarbara_1500 BP.

We used the Allsnps: No parameter to decrease biases for these analyses and left all other settings as default.

Conditional heterozygosity analyses

We estimated conditional heterozygosity to infer the cumulative effect of bottlenecks in the history of a population over millennia by examining polymorphisms between two randomly chosen Yoruba chromosomes. We performed these analyses at transversion variants on all groups with at least two individuals per site using PopStats (https://github.com/pontussk/popstats) with the 26 September 2018 default settings. We computed this on individuals from this study, ancient individuals from Peru[16,34], Brazil[16], the Caribbean[36] and Patagonia[46], as well as on sequencing data from present-day Native Americans[70,71]. We assessed statistical significance for differences between groups using two-sided Student's *t*-tests.

Analyses of ROH

We used hapROH (v.0.1a8; https://pypi.org/project/hapROH/) to identify ROHs[72].We used the 1000 Genomes Project haplotype panel as the reference panel with 5,008 global haplotypes. We analysed the ancient and present-day data of individuals with at least 400,000 SNPs covered to identify ROHs longer than 4 cM. We also estimated N_e using a maximum-likelihood inference framework for a ROH size range of 4–20 cM[36]. We estimated the confidence interval using the curvature of the likelihood (Fisher information matrix). We used the default settings of hapROH for all analyses. The individuals analysed are shown in Supplementary Data 1 and include groups from Guam and Saipan[73], Patagonia[46], Peru and Bolivia[34], and the Caribbean and Venezuela[36].

Map plotting

Fig. 1a was made using the open-source R packages maps (v.3.4.1), sf (v.1.14)[74], rnaturalearth (v.0.3.4)[75], ggplot2 (v.3.4.3)[76] and ggrepel (v.0.9.3)[77]. Extended Data Fig. 4 was generated in R using ggplot2 (v.3.4.3)[76], fields (v.15.2)[78] and RcolorBrewer (v.1.13).

Reporting summary

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

Data availability

All sequencing data newly generated in this study are available from the European Nucleotide Archive (ENA) under accession number PRJEB66319. Genotype data obtained by random sampling of sequences at approximately 1.24 million analysed positions are available from Harvard Dataverse under accession number Z2JD58. The data we are publishing in this study are the DNA libraries for each of the ancient individuals we analysed, which are molecular copies of the original molecules extracted from the ancient individuals whose remains in many cases may no longer be available for scientific study. The data we report are therefore not only stored after publication in digital form (the sequences we uploaded) but in molecular form for as long as the libraries are maintained in freezers. This means that more sequences may be generated by those who can support generating a higher quality digital readout of the library, with permission to generate such sequences covered by the current publication. These libraries can only be requested for scholarly use and cannot be used for commercial purposes. If the relevant Indigenous communities request them to be repatriated or reburied, they will no longer be available. In addition, we used the following publicly available datasets: ref. 14 (ENA: PRJEB25445); ref. 34 (ENA: PRJEB37446 and PRJEB39010); ref. 16 (ENA: PRJEB28961); ref. 36 (ENA: PRJEB3555); ref. 70 (ENA: PRJNA470966); ref. 71 (ENA: PRJEB9586 and ERP010710); ref. 79 (NCBI Sequence Read Archive database identifier: SRP029640); and ref. 19 (ENA: PRJEB29074). The hg19 human genome reference sequence was used for all analyses, available at https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.25/.

43. Reimer, P. J. et al. The IntCal20 Northern Hemisphere radiocarbon age calibration curve (0–55 cal kBP). *Radiocarbon* **62**, 725–757 (2020).
44. Heaton, T. J. et al. Marine20—the marine radiocarbon age calibration curve (0–55,000 cal BP). *Radiocarbon* **62**, 779–820 (2020).
45. Hendy, I. et al. Resolving varve and radiocarbon chronology differences during the last 2000 years in the Santa Barbara Basin sedimentary record, California. *Quat. Int.* **310**, 155–168 (2013).
46. Nakatsuka, N. et al. Ancient genomes in South Patagonia reveal population movements associated with technological shifts and geography. *Nat. Commun.* **11**, 3868 (2020).
47. Dabney, J. et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl Acad. Sci. USA* **110**, 15758–15763 (2013).
48. Korlevic, P. et al. Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *BioTechniques* **59**, 87–93 (2015).
49. Sirak, K. A. et al. A minimally-invasive method for sampling human petrous bones from the cranial base for ancient DNA analysis. *BioTechniques* **62**, 283–289 (2017).
50. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20130624 (2015).
51. DeAngelis, M. M., Wang, D. G. & Hawkins, T. L. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* **23**, 4742–4743 (1995).
52. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
53. Maricic, T., Whitten, M. & Paabo, S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE* **5**, e14004 (2010).
54. Fu, Q. et al. An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
55. Mathieson, I. et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
56. Haak, W. et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
57. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
58. Fu, Q. et al. A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23**, 553–559 (2013).
59. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
60. Nakatsuka, N. et al. ContamLD: estimation of ancient nuclear DNA contamination using breakdown of linkage disequilibrium. *Genome Biol.* **21**, 199 (2020).
61. Kennett, D. J. et al. Archaeogenomic evidence reveals prehistoric matrilineal dynasty. *Nat. Commun.* **8**, 14115 (2017).
62. Lazaridis, I. et al. The genetic history of the Southern Arc: a bridge between West Asia and Europe. *Science* **377**, eabm4247 (2022).
63. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
64. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
65. Felsenstein, J. PHYLIP—phylogeny inference package (Version 3.2). *Cladistics* **5**, 164–166 (1989).
66. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
67. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
68. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
69. Skoglund, P. et al. Genetic evidence for two founding populations of the Americas. *Nature* **525**, 104–108 (2015).
70. Lindo, J. et al. The genetic prehistory of the Andean highlands 7000 years BP though European contact. *Sci. Adv.* **4**, eaau4921 (2018).
71. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
72. Ringbauer, H., Novembre, J. & Steinrücken, M. Parental relatedness through time revealed by runs of homozygosity in ancient DNA. *Nat. Commun.* **12**, 5425 (2021).
73. Liu, Y.-C. et al. Ancient DNA reveals five streams of migration into Micronesia and matrilocality in early Pacific seafarers. *Science* **377**, 72–79 (2022).
74. Pebesma, E. J. Simple features for R: standardized support for spatial vector data. *R J.* **10**, 439 (2018).
75. Massicotte, P., South, A. rnaturalearth: World Map Data from Natural Earth. R package version 0.3.2 (2023).
76. Wickham, H. & Wickham, H. *Data Analysis* (Springer, 2016).
77. Slowikowski, K. et al. ggrepel: Automatically Position Non-overlapping Text Labels with 'ggplot2'. R package version 4.3.1 (2018).

78. Nychka, D., Furrer, R., Paige, J. & Sain, S. fields: Tools for Spatial Data. R package version 9, https://doi.org/10.5065/D6W957CT (2017).
79. Raghavan, M. et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014).

Acknowledgements We acknowledge and thank the ancient individuals whose remains were analysed in this study, and the present-day Indigenous communities who supported this study and provided cultural contextualization that we sought to reflect in the final paper, particularly representatives from the Chumash, Tongva, Ohlone and Esselen groups; staff at the Consejo de Arqueología and the Instituto Nacional de Antropología e Historia for the permits and facilities granted for the study of the samples from Mexico; expert linguists J. Yee (member of the Barbareño Chumash tribe), L. Campbell, M. Mithun, M. Walworth, J. W. Powell, P. Munro and D. Shaul for their comments and suggestions regarding how best to discuss the implications of our genetic results for linguistic debates; H. Ringbauer for advice on the ROH analyses; I. Lazaridis for determining Y chromosome haplogroups; M. Armenta, I. Lazaridis, M. Lipson, I. Olalde and N. Patterson for critical comments and helpful discussions; N. Adamski, R. Bernardos, M. Ferry, G. Fisher, I. Greenslade, K. Mable, K. Stewardson, Z. Zhang, staff at the American Museum of Natural History and the Peabody Museum of Archaeology and Ethnology for support with wet laboratory work or bioinformatics or sample management; and archaeologist G. S. Breschini, who would have been an author on this paper had he not passed away in 2018. Support for analysis of DNA from ancient individuals from Monterey County was provided by the late E. Rodriguez (Most Likely Descendant; Rumsen tribe) and provided for this study by the late G. Breschini. N.N. was supported by a National Institutes of General Medical Sciences fellowship. The PIPANOM Project, sampling individuals from Northern Mexico, was supported by a grant from the National Geographic Society to J.L.P.D. The ancient DNA data collection and statistical analyses were supported by a grant from the National Human Genome Research Institute (R01-HG012287), the John Templeton Foundation (grant 61220), by a private gift from Jean-Francois Clin, by the Allen Discovery Center programme, a Paul G. Allen Frontiers Group advised programme of the Paul G. Allen Family Foundation, and by the Howard Hughes Medical Institute (D.R.).

Author contributions N.N. performed population genetics analyses. N.N., B.H., J.S., P.E.L., J.C., C.G.-M., D.M., J.M.-R., A.P.-M., V.T., M.E.V.-C., A.V.H., J.L.P.D., J.R.J. and D.R. interpreted the data. J.S., P.E.L., J.C., C.G.-M., D.M., J.M.-R., A.P.-M., V.T., M.E.V.-C., A.V.H., J.L.P.D. and J.R.J. collected and described archaeological material and site contexts. K.C., E.C., A.K., L.I., A.M.L., M.M., J.N.W., J.O., L.Q., F.Z. and N.R. performed or supervised sample preparations. N.R. and A.M.L. generated genetic data. N.N., J.R.J., J.L.P.D., J.R.J. and D.R. conceived the study. T.K.H. and B.J.C. performed or supervised accelerator mass spectrometry radiocarbon dating analyses and marine correction. N.R., S.M., M. Mah, A.M. and D.R. performed bioinformatics analyses. N.N., B.H., J.S., J.L.P.D., J.R.J. and D.R. wrote the paper with the assistance of the other co-authors. J.R.J. and D.R. directed the study together.

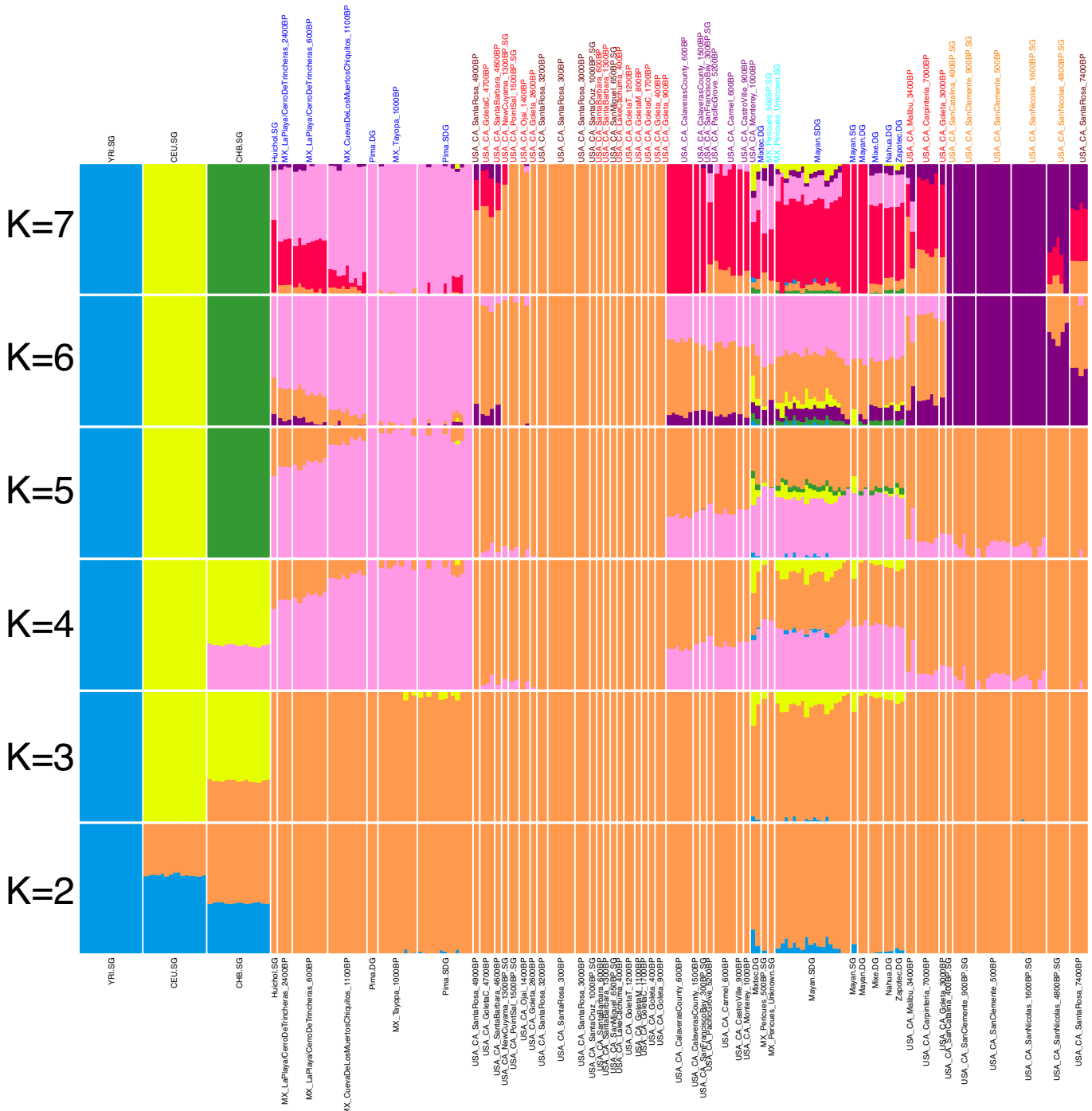
Competing interests The authors declare no competing interests.

Additional information

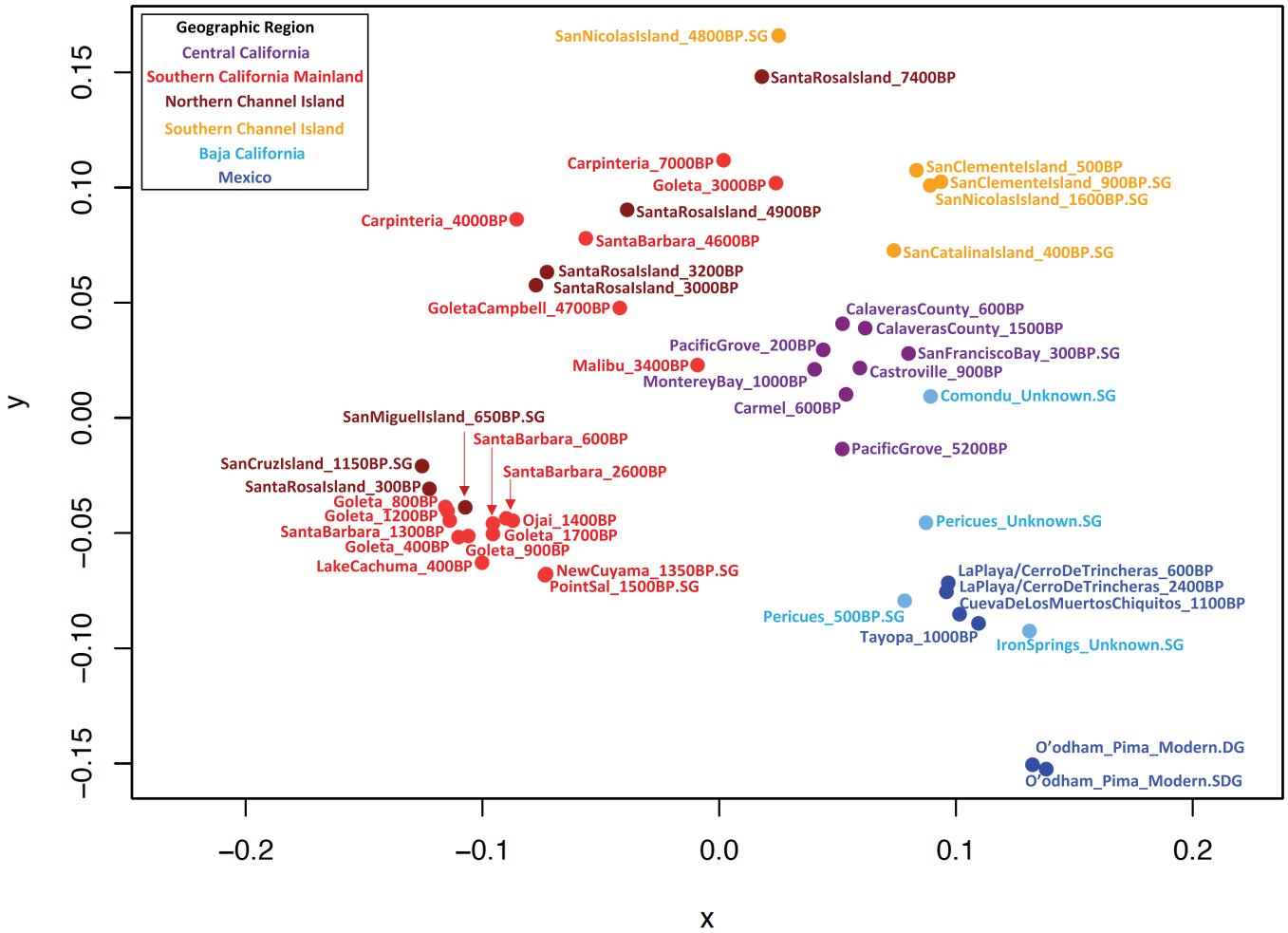
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06771-5.

Correspondence and requests for materials should be addressed to Nathan Nakatsuka, John R. Johnson or David Reich.

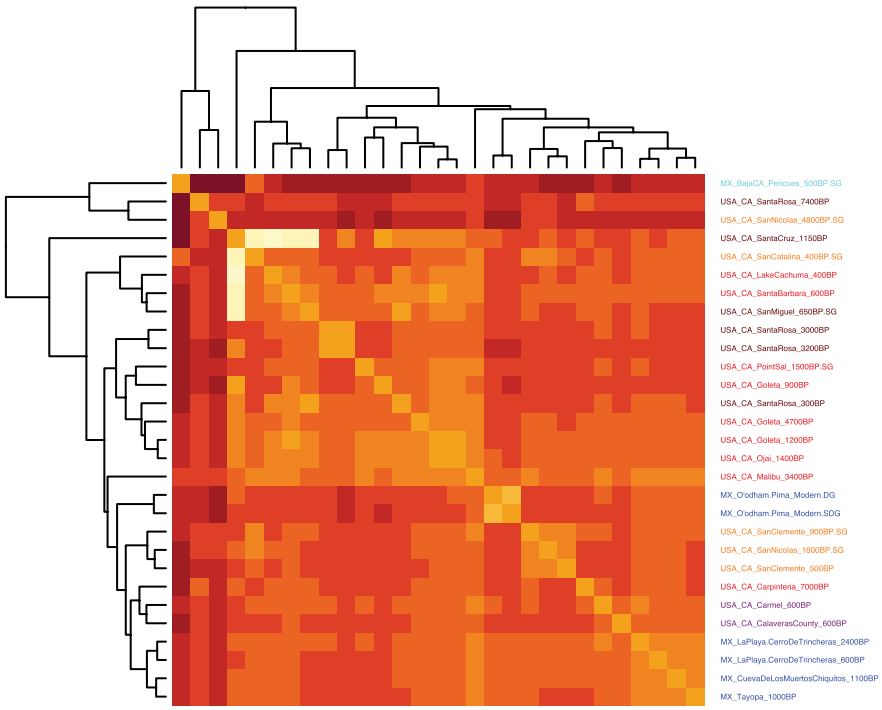
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.



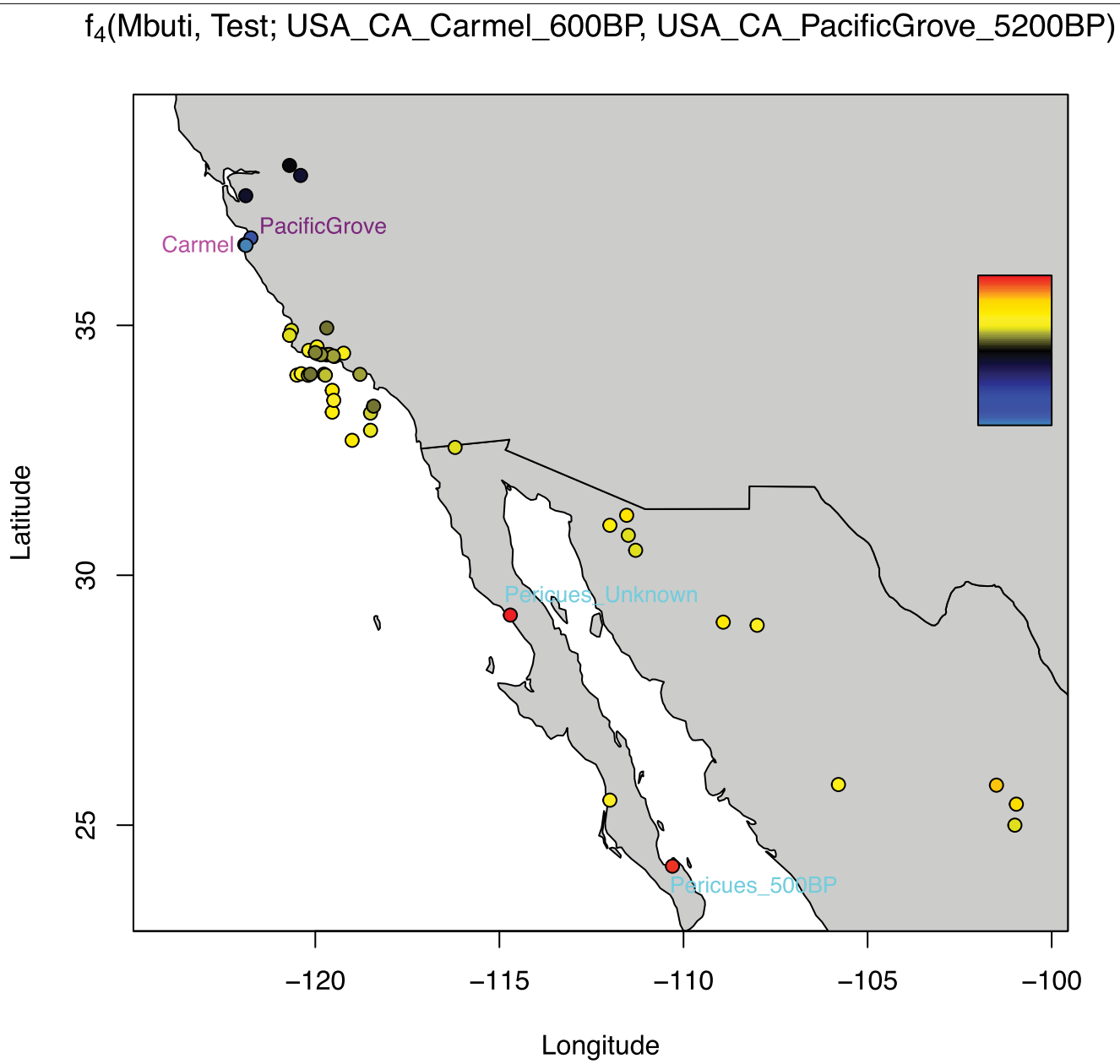
Extended Data Fig. 1 | ADMIXTURE plot at different K values. Purple=Central California, red=Southern California mainland, dark red=Northern Channel Islands, orange=Southern Channel Islands and nearby mainland, light blue=Baja California, blue=Mexico excluding Baja California.



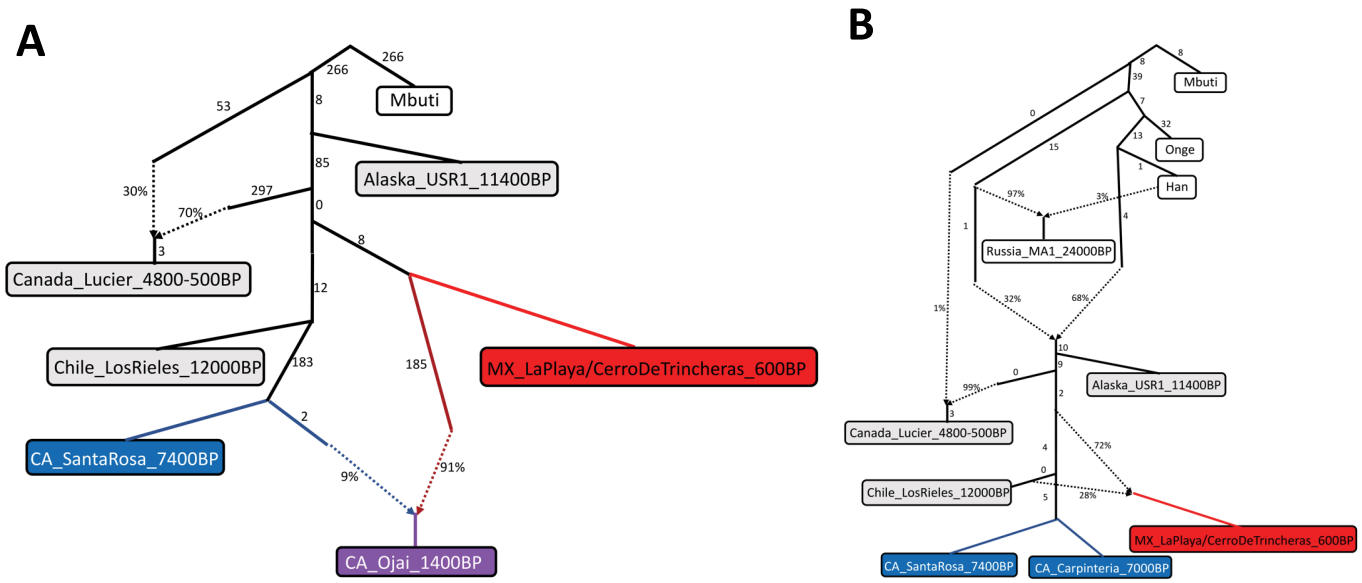
Extended Data Fig. 2 | MDS plot of groups created using a matrix of inverted outgroup-f_3 statistics (distances = 1-f_3(*Mbuti; Group1, Group2*)). Purple=Central California, red=Southern California mainland, dark red=Northern Channel Islands, orange=Southern Channel Islands and nearby mainland, light blue=Baja California, blue=Mexico excluding Baja California.



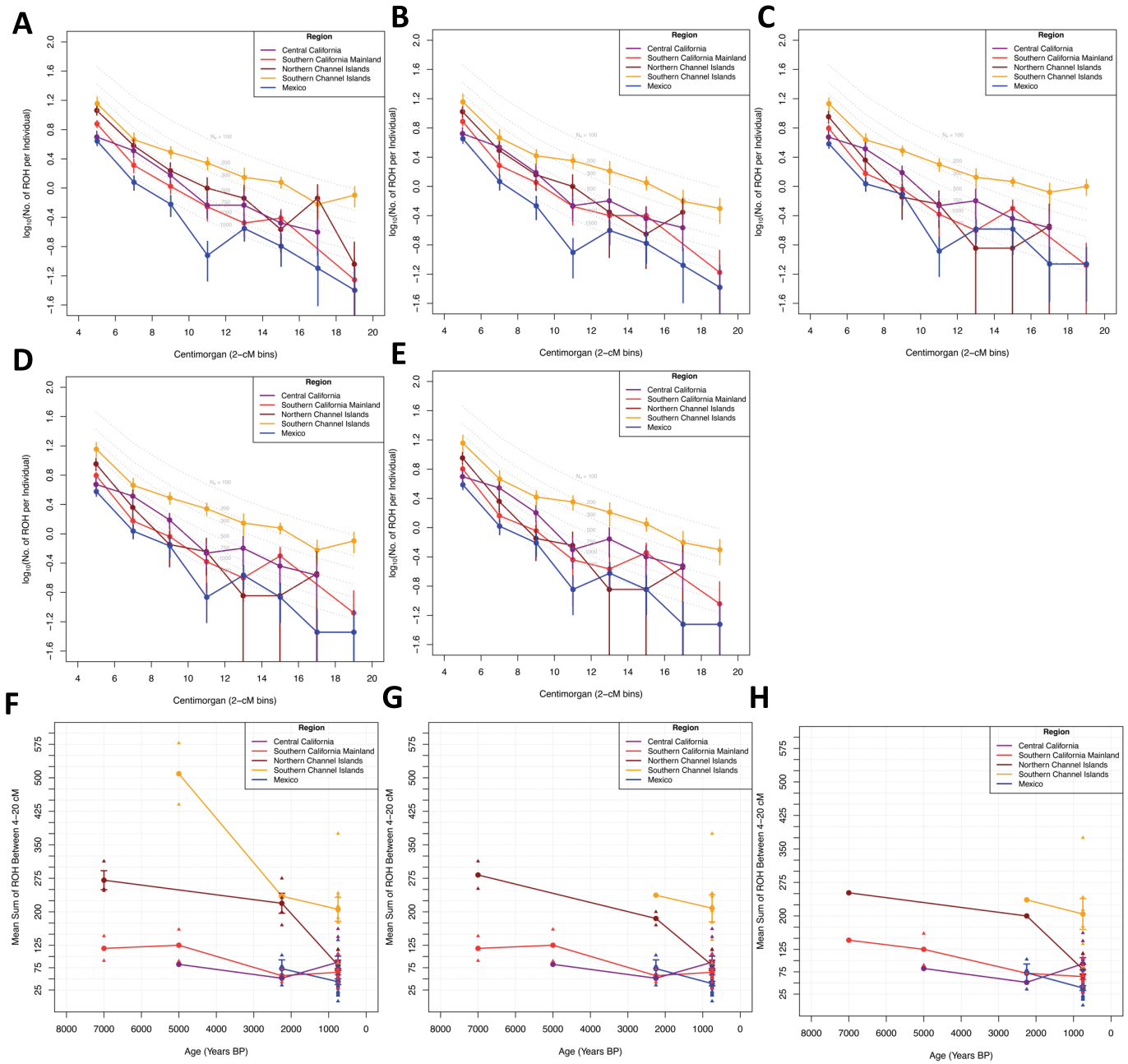
Extended Data Fig. 3 | Heatmap of pairwise F_{ST}. F_{ST} between groups was estimated using smartpca. Only groups with at least 2 individuals of greater than 100,000 SNP coverage were used. Heatmap and dendrogram were created in R with symm=T. Supplementary Data File 3 shows F_{ST} values. *Purple=Central California, red=Southern California mainland, dark red=Northern Channel Islands, orange=Southern Channel Islands and nearby mainland, light blue=Baja California, blue=Mexico excluding Baja California.*



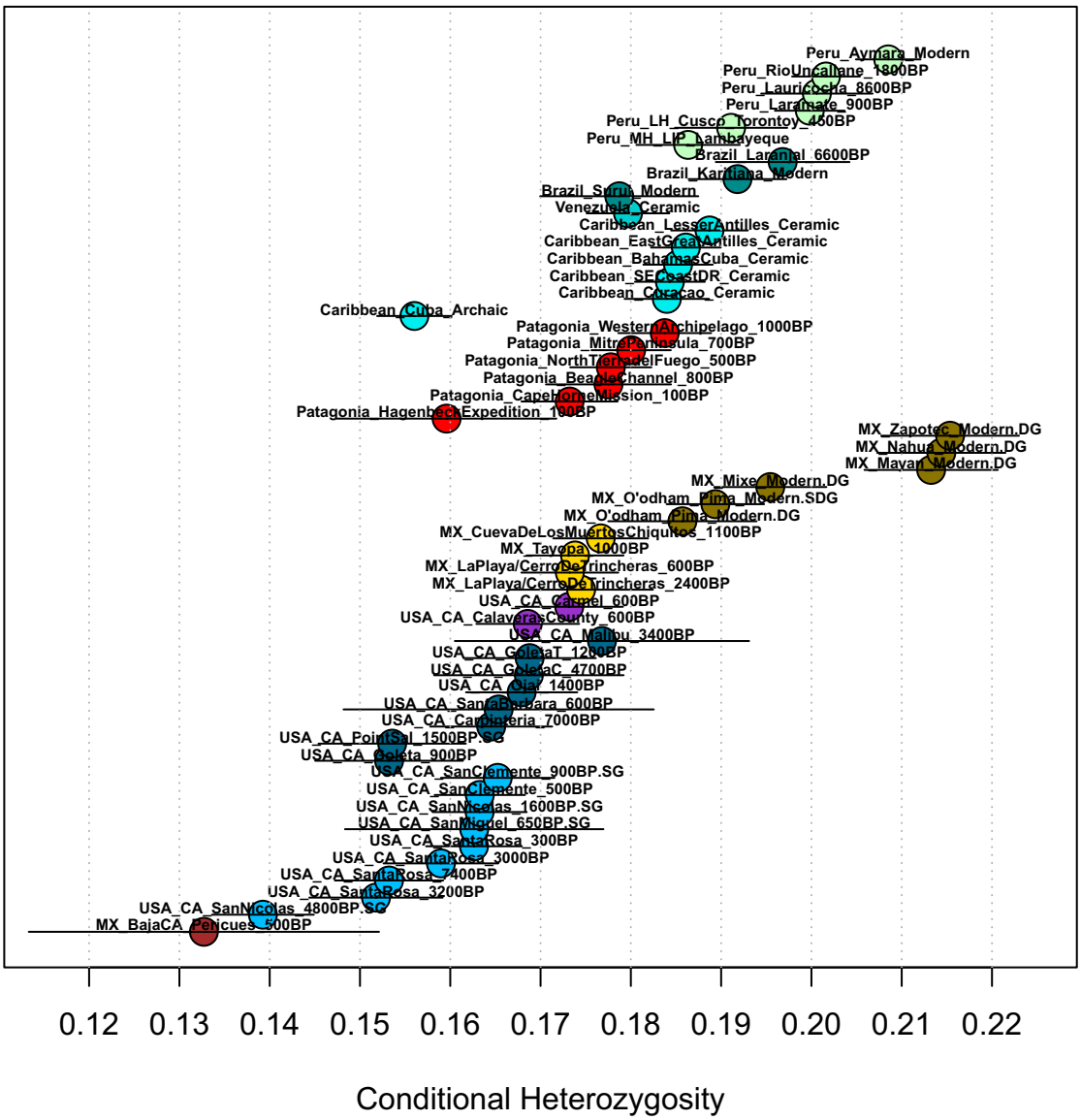
Extended Data Fig. 4 | Map of statistics of the form *f*$_4$*(Mbuti, Test; USA-CA_Carmel_600BP, USA-CA_PacificGrove_5200BP). Dots in red show greater genetic affinity to *PacificGrove_5200BP* relative to *Carmel_600BP*, while dots in black and blue have greater affinity to *Carmel_600BP*. Points are jittered to allow better visualization. Figure is generated with open source data and software in R with ggplot2 and the 'fields' and 'RcolorBrewer' libraries.



Extended Data Fig. 5 | Admixture graphs. A) Example admixture graph testing for attraction to *Canada_Lucier_4800-500 BP*. This graph fits with a maximum |Z-score| of 2.81. We tested all subsequent graphs replacing CA_Ojai_1400BP with another ancient California group (Supplementary Data File 4). **B**) Admixture graph consistent with relationships between ancient California and Mexico groups. This graph fits with a maximum |Z-score| of 2.98. All graphs we explore require a lineage more basal than that of *Chile_LosRieles_12000 BP* to fit the Mexico individuals, although we caution that the total space of admixture graph topologies is too large to explore exhaustively so we are making no claim that these particular graphs are correct (only that they are plausible and not ruled out by the data). The basal ancestry into *Canada_Lucier_4800-500 BP* is present to account for known European contamination.



Extended Data Fig. 6 | ROH in California and Mexico. A) Average rate of ROH segments in different length bins after filtering out individuals with a sum of ROH segments of ≥20 cM of 100 cM or more and **B**) after filtering out individuals using a lower stringency threshold of 50 cM or more. Points with no ROH fragments present in those bins were filtered out. **C**) Average rate of ROH segments in different length bins after filtering out individuals over 1600 BP and **D**) after filtering out individuals with summed 20 cM over 100 cM or **E**) over 50 cM. **F**) ROH over time where each data point represents the average sum of ROH between 4–20 cM of individuals in a bin of its corresponding time-period (8000-6000 BP, 6000-4000 BP, 4000-1500 BP, and <1500 BP); the number of individuals for each of these time bins is (0,1,1,10) for Central California, (2,2,2,12) for Southern California Mainland, (3,0,4,7) for Northern Channel Islands, (0,2,3,9) for Southern Channel Islands), and (0,0,3,23) for Mexico. **G**) ROH over time after filtering out individuals with a sum of ROH segments of ≥20 cM of 100 cM or more; the number of individuals for each of these time bins is (0,1,1,10) for Central California, (2,2,2,12) for Southern California Mainland, (2,0,2,7) for Northern Channel Islands, (0,0,2,8) for Southern Channel Islands), and (0,0,3,22) for Mexico. **H**) ROH over time after filtering out individuals with a sum of ROH segments of ≥20 cM of 50 cM or more; the number of individuals for each of these time bins is (0,1,1,9) for Central California, (1,2,1,11) for Southern California Mainland, (1,0,1,7) for Northern Channel Islands, (0,0,1,7) for Southern Channel Islands), and (0,0,3,21) for Mexico. For all figures, data are presented as mean values ± 1 standard error (no standard errors are presented for points with fewer than 3 individuals).



Extended Data Fig. 7 | Conditional heterozygosity of groups. Ancient Californian, Mexican, Peruvian, Brazilian, Caribbean, and Patagonian groups and present-day Mexican, Brazilian and Peruvian groups are shown. Only groups with at least 2 individuals could be included in these analyses.

nature portfolio

Corresponding author(s): Nathan Nakatsuka, David Reich, John Johnson

Last updated by author(s): Sep 26, 2023

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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	Illumina NextSeq500 v.2 or Illumina HiSeq X10 was used to sequence the DNA. Accelerator Mass Spectrometry was used to determine carbon 14 dates for the ancient individuals.
Data analysis	SeqPrep version 1.2 (https://github.com/jstjohn/SeqPrep) or custom software (https://github.com/DReichLab/ADNA-Tools) was used to merge paired forward and reverse reads. BWA version 0.6.1 (bio-bwa.sourceforge.net) was used to align the reads to the hg19 human genome reference sequence. Duplicates were removed with Picard version 2.23.0 (http://broadinstitute.github.io/picard/). OxCal version 4.4 (https://c14.arch.ox.ac.uk/oxcal.html) was used to calculate 14C ages. ContamMix version 1.0-12 (https://github.com/DReichLab/ADNA-Tools) was used to measure mitochondrial contamination. ANGSD version 0.930 (https://github.com/ANGSD/angsd) was used to measure X-chromosome contamination. Autosomal contamination was measured with ContamLD version 1.0 (https://github.com/nathan-nakatsuka/ContamLD). HaploGrep2 with Phyotree version 17 was used to determine mtDNA haplogroups. Y chromosome haplotypes were determined using the nomenclature of the International Society of Genetic Genealogy version 14.76 using YFull YTree v. 8.09 phylogeny. Samtools version 1.10 (http://samtools.sourceforge.net/) was used to generate VCFs from BAM files and find genotype information. Haplogrep version 17 (http://haplogrep.uibk.ac.at/index.html) was used to obtain mitochondrial haplogroup assignments. ADMIXTURE version 1.3.0 (https://www.genetics.ucla.edu/software/admixture/download.html) was used to do unsupervised clustering analysis. smartPCA in EIGENSOFT version 5.0 (https://github.com/DReichLab/EIG) was used to perform FST analyses. f-statistics, qpWave, qpAdm, and qpGraph were calculated with ADMIXTOOLS version 6.0 (https://github.com/DReichLab/AdmixTools). Conditional Heterozygosity was done using POPSTATS September 26, 2018 version (https://github.com/pontussk/popstats). PLINK2 version 2.0 alpha (https://www.cog-genomics.org/plink/2.0/) was used to determine correlation between SNPs. Damage rates were calculated with PMDTools version 0.60 (https://github.com/pontussk/PMDtools). Kinship was determined using custom software based on mismatch rates described in Kennett et al., 2017 (available upon request but not yet ready for broader distribution). hapROH version 0.1a8 was used to identify runs of homozygosity. maps (version 3.4.1), sf (version 1.14),

rnaturalearth (version 0.3.4), ggplot2 (version 3.4.3), ggrepl (version 0.9.3), fields (version 15.2), and RcolorBrewer (version 1.13) were used to make maps. Scripts for making outgroup-f3 neighbor-joining trees and MDS plots as well as counting alleles are available upon request.

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All sequencing data are available from the European Nucleotide Archive, accession number PRJEB66319. Genotype data obtained by random sampling of sequences at approximately 1.24 million analyzed positions are available from Harvard Dataverse at accession number Z2JD58. See Data availability statement for additional details.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	The genetic sex of ancient individuals was inferred by analyses of sex chromosomes, but given there was no data on the gender expression of the individuals, we make no mention of gender in the manuscript. We lacked sufficient power to determine whether there were sex differences in our findings.
Reporting on race, ethnicity, or other socially relevant groupings	Throughout the manuscript, individuals older than 200 years old were referenced based on the location that they were found and their age. However, individuals younger than 200 years old were referred to either with their self-expressed ethnic grouping or with the ethnic grouping of the groups living in the region where the ancient individual currently existed (e.g. Chumash, provided by the ethnic groups in the region today). We made clear to specify that modern political boundaries do not necessarily reflect past boundaries.
Population characteristics	N/A
Recruitment	Skeletal material from ancient individuals were excavated
Ethics oversight	For the California ancient individuals, ethical oversight was provided by the UC Santa Barbara Museum and Peabody Museum of Archaeology and Ethnology as well as by the local Indigenous groups (Chumash groups, Tongva, and Ohlone, as detailed in the Ethical Approval section). In Mexico, all legal authorizations were obtained for this work, sanctioned by the Consejo de Arqueología from the Instituto Nacional de Antropología e History; the research followed their guidance, and was directed by Mexican archaeologists (authors JC, CGM, JMR, APM, EVC, and JLPD).

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

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

Ecological, evolutionary & environmental sciences study design

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

Study description	DNA analyses of 79 ancient individuals from California and 40 ancient individuals from Mexico were performed to determine the population genetic structure of California and Mexico over time and how this structure was changed by admixture.
Research sample	79 ancient California individuals were chosen based on availability of skeletal material and also to represent the different Channel Islands, the Southern California and Central California area at different time points. 60 additional ancient California individuals were analyzed from Scheib et al., Science 2018. 40 ancient individuals from NW Mexico were also analyzed to represent a region near California. Additional previously published ancient individuals were analyzed as detailed in Supplementary Data File 1.
Sampling strategy	Sample size was not pre-determined beforehand and was based on availability of skeletal material and ability to generate DNA from the material. The sufficiency of these sample sizes depends on the analyses, but for the analyses performed in this study, usually a few individuals per group (or even one in some cases) is sufficient for obtaining some insight about the genetic make-up of the group (more is useful for learning about outliers and genetic variation within the group).

Data collection	Extensive detail of the different data collection methods (with wide variety based on the different ancient individuals) is provided in Supplementary Information "Description of Archaeological Sites".
Timing and spatial scale	Extensive detail of the timing and spatial scale of the excavations (with wide variety based on the different ancient individuals) is provided in Supplementary Information "Description of Archaeological Sites".
Data exclusions	We removed (but still report) 7 individuals from analyses with point estimates of more than 7% contamination from ContamLD, 5% from ANGSD or 10% from contamMix applied to mitochondrial DNA.
Reproducibility	Reproducibility was ensured by analyzing the data across different sequencing platforms and treatment types (with only transversions as well as all sites) and analyzing them as individuals and in groups. The analyses are all qualitatively equivalent across these differences.
Randomization	No randomization of the individuals was done because the analyses were not based on experimental treatment of different groups looking forward in time. The analyses were based on comparisons of the genetics of the different groups (looking backwards in time) where knowledge of the groups is necessary to attain the results.
Blinding	Blinding was not relevant to this study for the same reason as stated above for randomization.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	Extensive descriptions for all sites are provided in the Supplementary Information "Description of Archaeological Sites".
Location	Extensive descriptions for all sites are provided in the Supplementary Information "Description of Archaeological Sites".
Access & import/export	Excavations of the human skeletal material was done with permissions from the local Indigenous communities and additional permissions from the relevant museums. Additional detail is provided in the Supplementary Information "Description of Archaeological Sites".
Disturbance	Excavations were done, where possible, with oversight from local Indigenous groups and with rescue archaeology methods to minimize disturbances (as detailed in Supplementary Information). Most ancient individuals were repatriated and reburied by local Indigenous groups.

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 and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Palaeontology and Archaeology

Specimen provenance	Excavations of the human skeletal material was done with permissions from the local Indigenous communities and additional permissions from the relevant museums. Additional detail is provided in the Supplementary Information "Description of Archaeological Sites".
Specimen deposition	Most of the skeletal material has been repatriated and reburied by local Indigenous groups except as detailed in Supplementary Data File 1 with the relevant museums including the Peabody Museum at Harvard University, the Santa Barbara Museum of Natural History, the Museo Nacional de Antropología, the Anthropology Department at UNLV, and the American Museum of National History.
Dating methods	We generated new direct Accelerator Mass Spectrometry (AMS) 14C dates for ancient individuals from PSUAMS, Poz and UCIAMS

Dating methods	(Supplementary Data File 1). All calibrated 14C ages were calculated using OxCal version 4.4 using different mixtures of the northern hemisphere terrestrial (IntCal20) 88 and marine (Marine20) 89 calibration curves. Marine dietary contribution was estimated using stable carbon and nitrogen isotope measurements from collagen (Supplementary Data File 1).

☒ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	For the California ancient individuals, ethical oversight was provided by the UC Santa Barbara Museum and Peabody Museum of Archaeology and Ethnology as well as by the local Indigenous groups (Chumash groups, Tongva, and Ohlone, as detailed in the Ethical Approval section). In Mexico, all legal authorizations were obtained for this work, sanctioned by the Consejo de Arqueología from the Instituto Nacional de Antropología e History; the research followed their guidance, and was directed by Mexican archaeologists (authors JC, CGM, JMR, APM, EVC, and JLPD).

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