Article

The evolutionary history and unique genetic diversity of Indigenous Americans

https://doi.org/10.1038/s41586-026-10406-w

Received: 15 May 2025

Accepted: 11 March 2026

Published online: 22 April 2026

Open access

Check for updates

Marcos Araújo Castro e Silva[1,2,3,27], Kelly Nunes[3,27], Maíra R. Ribeiro[3,27], Hemanoel Passarelli-Araujo[4], Renan Barbosa Lemes[3], Lilian Kimura[3], Putira Sacuena[5], Carlos Eduardo G. Amorim[6], Maria Cátira Bortolini[7], José Geraldo Mill[8], João Farias Guerreiro[5], Chiara Barbieri[9,10], Diana Iraíz Hernández-Zaragoza[11,12], Antonia Walter[11], Trija Nag Chowdhury[11], Daniela Macías-Herrera[12], Julio César Lara-Riegos[13], Oana Del Castillo-Chávez[14], Camilo Zurita[15], Ana María Tito-Álvarez[16], Emilia Vásquez-Domínguez[17], María Ermila Moo-Mezeta[18], Julio César Torres-Romero[13], Abraham Aguilar-Campos[19], Ricardo Serrano-Osuna[20], Maria Laura Parolín[21], Claudio M. Bravi[22], Virgínia Ramallo[23], Graciela Bailliet[22], Susana Revollo[24], José R. Sandoval[25], Ricardo Fujita[25], Rodrigo Barquera[11], Fabrício R. Santos[26], David Comas[1,2] & Tábita Hünemeier[1,3] ✉

Indigenous peoples of America represent the last principal expansion of humans across the globe[1], yet their genetic history remains one of the least explored[2]. Although these populations have inhabited the continent for thousands of years[3], their evolutionary history remains largely unresolved[4,5], owing to the limited availability of genomic data. Here we present data on 128 high-coverage Indigenous American genomes and show they harbour extensive and previously uncharacterized genetic diversity, reflecting at least three dispersals into South America, followed by regional differentiation and long-term continuity. We identified widespread natural selection signals in genes associated with immunity, metabolism, reproduction and development, which were shaped by adaptation to diverse environmental conditions. Notably, several genomic regions exhibit a remarkable allele sharing with Australasian populations, probably originating from an ancient admixture event and partly maintained by selection for more than 10,000 years. We also detected distinct contributions from archaic humans with adaptive introgression affecting key biological functions. The limited overlap between the regions of Australasian affinity and archaic ancestry indicates independent evolutionary origins of these signals. These findings challenge simplified models of continental settlements and show a more dynamic and complex evolutionary history for the Indigenous peoples in America.

America was the final frontier of early human global dispersal[1], marking the last continent to be populated not only by *Homo sapiens* but also by any hominin species. The American continent and its populations have unique characteristics crucial for understanding human evolution. Human history on the American continent began at least 15,000 years ago[3]; however, the ancestry of Indigenous American populations traces back further in time, with their formation beginning about 25,000 years ago in and around the Beringian land bridge[6]. This period was crucial for the emergence of Indigenous American diversity[7], and after millennia of partial isolation in Beringia, the end of the Last Glacial Maximum triggered rapid and large-scale dispersal into America[4,5].

As the only continent spanning both northern and southern global extremes, America encompasses a vast range of climates and biomes, further shaped by the most extensive mountain ranges in the world, including the two American cordilleras, the Andes in South America and the Rocky Mountains in North America. This environmental diversity imposed diverse selective pressures on human populations[8–10]. After the initial settlement, except for the northernmost regions, the first American inhabitants remained virtually isolated from the other continental groups for nearly 15,000 years, giving rise to diverse and numerous autochthonous populations after the initial wave of dispersal from Beringia and, in the case of South America, a few dispersals from North America[4,5]. Over the past five centuries, European-driven colonization has profoundly transformed the continent, leading to the widespread extermination of Indigenous populations and giving rise to a complex mosaic of diverse and multiethnic surviving Indigenous and admixed groups[11].

Despite this profound and impactful history, Indigenous populations from America remain the least studied from a genomic perspective among the principal human continental groups[2]. At present, only a few high-coverage genomes are available for evolutionary and biomedical research[12,13]. Over the past decade, studies of ancient and present-day individuals based primarily on genotyping arrays and, to a lesser extent, whole-genome sequencing have provided important insights into the initial peopling of the continent and the factors shaping genetic diversity among Indigenous populations[3–5,11,14,15]. However, inherent biases

A list of affiliations appears at the end of the paper.

in genotyping arrays and the limited number of studied populations have left many gaps unfilled.

In this context, questions that have been extensively explored using high-quality genomic data from other continental populations remain largely unexamined in Indigenous Americans[16,17]. For instance, the full extent of genomic variability among Indigenous American populations, represented by diverse and geographically broad sampling, and its effect on health and evolution remain unknown. The fine-scale geographic structure and comprehensive picture of the relationships between present-day and ancient populations across the continent have yet to be thoroughly explored through genomic studies. Moreover, the effects of natural selection, the role of archaic introgressions—whether adaptive or not—and the influence of several ancient dispersals on shaping genetic diversity over time remain compelling and unresolved questions.

Here we generated a diverse and high-quality dataset of Indigenous American genomes to address these unresolved issues and shed light on how demographic history and natural selection have shaped the genetic diversity of Indigenous Americans, providing crucial insights into their unique genomic variability, history, relatedness, archaic introgressions and genetic adaptations.

Data overview

We present the Indigenous American Genomic Diversity Project, comprising 128 newly generated high-coverage (approximately 44 times) whole genomes from Indigenous individuals across 8 Latin American countries (Argentina, Bolivia, Brazil, Colombia, Ecuador, Mexico, Paraguay and Peru), representing 45 populations and 28 language families (Fig. 1a, Extended Data Fig. 1 and Supplementary Table 1). This dataset expands Indigenous representation in genomics, with emphasis on geographically and linguistically diverse populations, particularly from the South American lowlands. For most analyses, we integrated these data with published ancient and present-day Indigenous American genomes (Supplementary Tables 2 and 3) after quality filtering (Methods). The final dataset included 199 contemporary Indigenous individuals from 53 populations and 31 language families. Global ancestry inference (Methods) showed no detectable post-European admixture in most newly sequenced individuals (Supplementary Note 3). When required, we inferred local ancestry and masked non-Indigenous genomic segments (Methods and Supplementary Note 3).

All data were collected, generated and analysed following international ethical standards for research with Indigenous populations (Methods). The study was conducted in collaboration with Indigenous communities, with results returned in accessible formats. Community feedback was constructive and positive, emphasizing how genomic findings could complement traditional knowledge and historical narratives.

Global genomics miss Indigenous diversity

Owing to an evolutionary history marked by both ancient (out-of-Beringia[6]) and recent (European colonization[11]) demographic shifts, Indigenous populations from America have traditionally been considered and expected to be less genetically diverse than other continental human groups[1,12,13]. By analysing 128 high-coverage genomes against a dataset of more than 270,000 individuals from various genomic databases (1000 Genomes Project (1KGP)[18], Human Genome Diversity Project (HGDP)[13], Simons Genome Diversity Project (SGDP)[12], Genome Aggregation Database (gnomAD)[19] and Single Nucleotide Polymorphism Database (dbSNP)[20]), we identified 12,493,650 single-nucleotide variants (SNVs), of which 11.4% were new (1,426,511; Fig. 1b). Of these, 4.7% were polymorphic (allele frequency > 0.01) and 0.02% were common (allele frequency > 0.05; Fig. 1c–e). This corresponds to about 11,100 new variants per individual, fewer than reported for Africans (about 27,800 per individual in 180 genomes)[16] and similar to Oceanians (about 12,600 per individual in 159 genomes)[17], on the basis of comparable reference datasets.

Thus, despite lower overall diversity, Indigenous Americans carry many previously unreported variants. This reflects both substantial genetic variation and long-standing underrepresentation in genomic resources, underscoring the need for broader inclusion to improve evolutionary and biomedical inference and to ensure benefits for Indigenous and descendant communities.

Indigenous American genomic variation

We assessed present-day genetic structure in Indigenous Americans using principal component analysis (PCA) and ADMIXTURE (Methods). Global PCA showed no batch effects, with clustering by geography rather than dataset source (Supplementary Note 1). Analysis of 160 unrelated masked genomes showed clear differentiation between North and South America and substructure within regions (Fig. 2 and Extended Data Fig. 2). Because PCA and ADMIXTURE are sensitive to population-specific drift[21], several groups, including Karitiana, Suruí, Amahuaca and Yaminahua, show distinctive profiles. This pattern is partly explained by elevated inbreeding, reflected by increased runs of homozygosity (ROHs) and higher F_{ROH}, an ROH-based inbreeding coefficient (Fig. 2, Extended Data Fig. 3 and Supplementary Table 4). These results underscore the limitations of using the Suruí and Karitiana as representative proxies for lowland South American genetic diversity (a common practice in population genomic studies), given their pronounced genetic differentiation, even relative to geographically nearby populations.

Genetic similarity mirrors geography when assessed with outgroup f_3 statistics, which are unbiased by population-specific drift. Four principal South American genetic clusters that closely follow geography were identified by applying a neighbour-joining tree to 1 − outgroup f_3 and multidimensional scaling to 1/outgroup f_3 (Methods, Extended Data Figs. 2 and 8a,b and Supplementary Table 5). These clusters are the Southern Cone, Eastern South America, Western South America (mostly lowlands) and Chaco. Southern Cone populations, including Mapuche-Tehuelche, were the most differentiated, yet Andean Quechua grouped within the same clade, indicating long-distance affinity. Chaco populations formed the sister branch, followed by the unclustered Tsimané, probably reflecting cultural and linguistic isolation[22]. The remaining split separates Eastern and Western lowland populations.

In North America, the earliest split separates Pima and Yaqui (both in northern Mexico), followed by divergence of the Nahua-Xochitlán group (central Mexico/Mexican Atlantic coast). Mixe, Mixtec and Zapotec then branch off (in southern Mexico), followed by Maya (on the Yucatán Peninsula) and finally the remaining continental diversity. This pattern mirrors the archaeological division between Aridoamerica and Mesoamerica and corresponding linguistic differences, with Pima and Yaqui belonging to Uto-Aztecan languages. On the basis of this concordance, we group these populations into two genetic clusters, Aridoamerica (northern Mexico) and Mesoamerica (central and southern Mexico, and Extended Data Figs. 2 and 8a). Although broadly consistent with previous studies[11,23–26], our results show finer substructure and reinforce the link between genetic differentiation and geography, in line with PCA and ADMIXTURE results (Fig. 2).

Analysis of molecular variance showed that genetic clusters explained a modest proportion of genetic variation (9%; $P = 0.04$), whereas ethnolinguistic groupings explained none ($P = 0.57$; Supplementary Table 6). Genetic distances, estimated as 1 − outgroup f_3(Mbuti; X, Y), were correlated with geographic distance among Indigenous American populations, both continent-wide (Spearman's $r = 0.4401$, $P \cong 0$) and within South America (Spearman's $r = 0.148$, $P \cong 0$; Supplementary Table 7 and Supplementary Note 4). However, isolation by distance alone does not account for the observed structure. Estimated effective migration

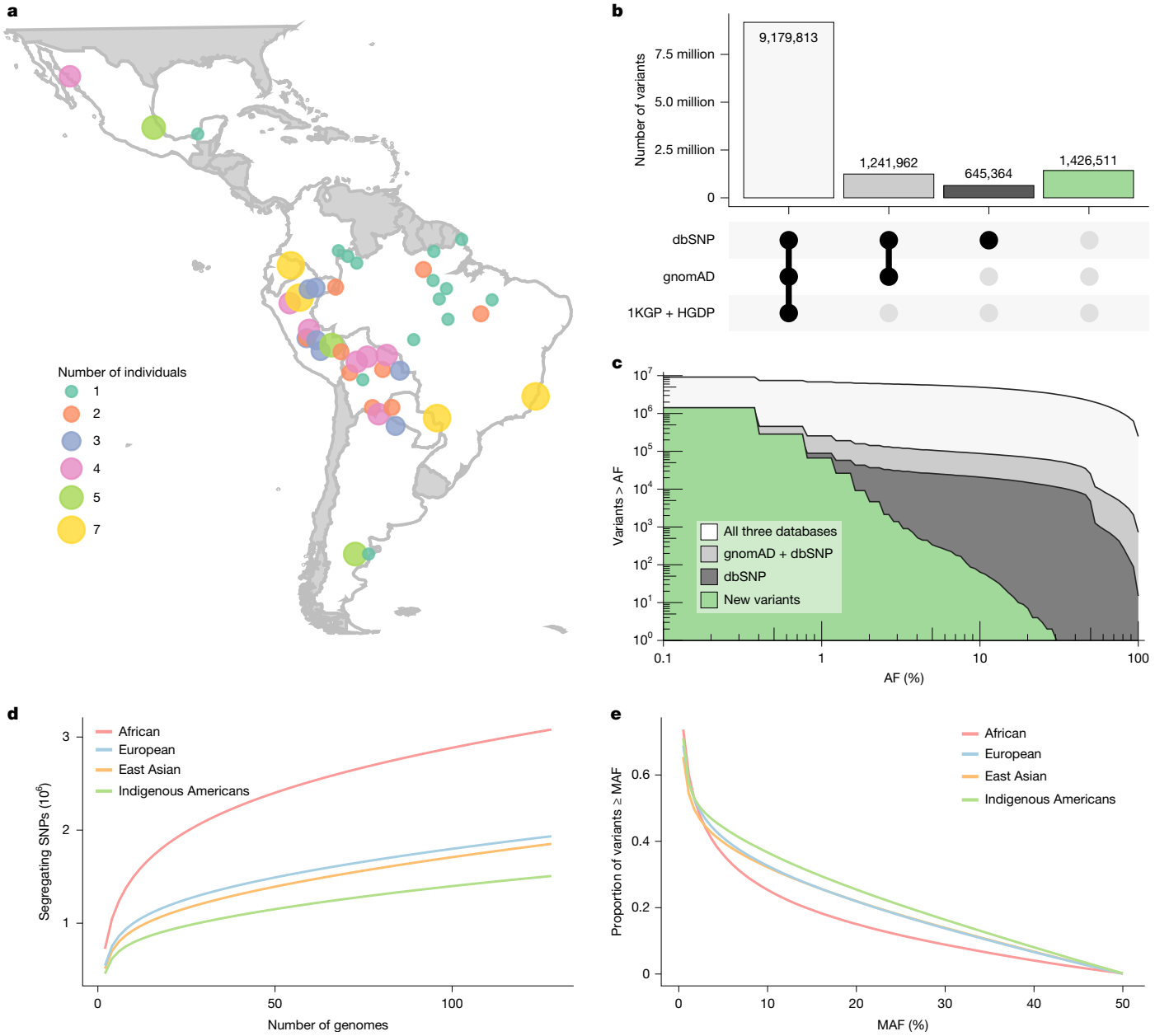


Fig. 1 | Genetic diversity of present-day Indigenous American populations. **a**, Geographic distribution of Indigenous American individuals included in the study. Each point represents a sampling location and its size is proportional to the number of individuals from that population. **b**, Stepwise filtering of SNVs to identify new variants. The UpSet plot shows the number of variants in different combinations of public databases: 1KGP + HGDP, gnomAD and dbSNP. Total variants = 12,493,650. **c**, Cumulative allele frequency spectrum: new variants (n = 1,426,511). The area plot displays the number of variants with allele frequency equal to or higher than a given threshold (greater than or equal to allele frequency (AF) in the x axis), separated by database overlap groups. **d**, Rarefaction curves showing the number of segregating SNPs as a function of the number of sampled genomes across African, European, East Asian and Indigenous American ancestry groups. **e**, Distribution of minor allele frequencies (MAFs) for SNPs in each population. Lines represent the proportion of SNPs with MAF greater than or equal to a given threshold.

surface (EEMS) showed regions of greater-than-expected genetic differentiation under an isolation-by-distance model, including southeastern Peru, northern Bolivia and southwestern Amazonia (Extended Data Fig. 4), consistent with previous work[24]. Reduced migration was inferred in Aridoamerica, the northern Amazon, the Chaco and eastern South America, particularly north of the Amazon River. These patterns indicate that demographic history and geographic and cultural factors have shaped genetic structure. By contrast, elevated migration between Mesoamerica and South America indicates greater connectivity than expected from geographic distance alone, consistent with previous evidence of North to South America gene flow[4,27].

Genomic legacies of colonization

Compared to other continental groups, Indigenous Americans show a distinctive ROH length distribution, with more segments in all categories except the shortest (less than 2 cM), where Oceanians exceed them (Extended Data Fig. 5), reflecting ancient bottlenecks and recent inbreeding. Mesoamerican and Southern Cone populations have the lowest ROH levels (Extended Data Figs. 3 and 5), indicating higher genetic diversity. By contrast, Moseten (Tsimané), Pano-Takana (Amahuaca, Yaminahua), Zamuco (Ayoreo) and Tupi (Sirionó, Suruí, Karitiana) speakers have the highest ROH counts and total length

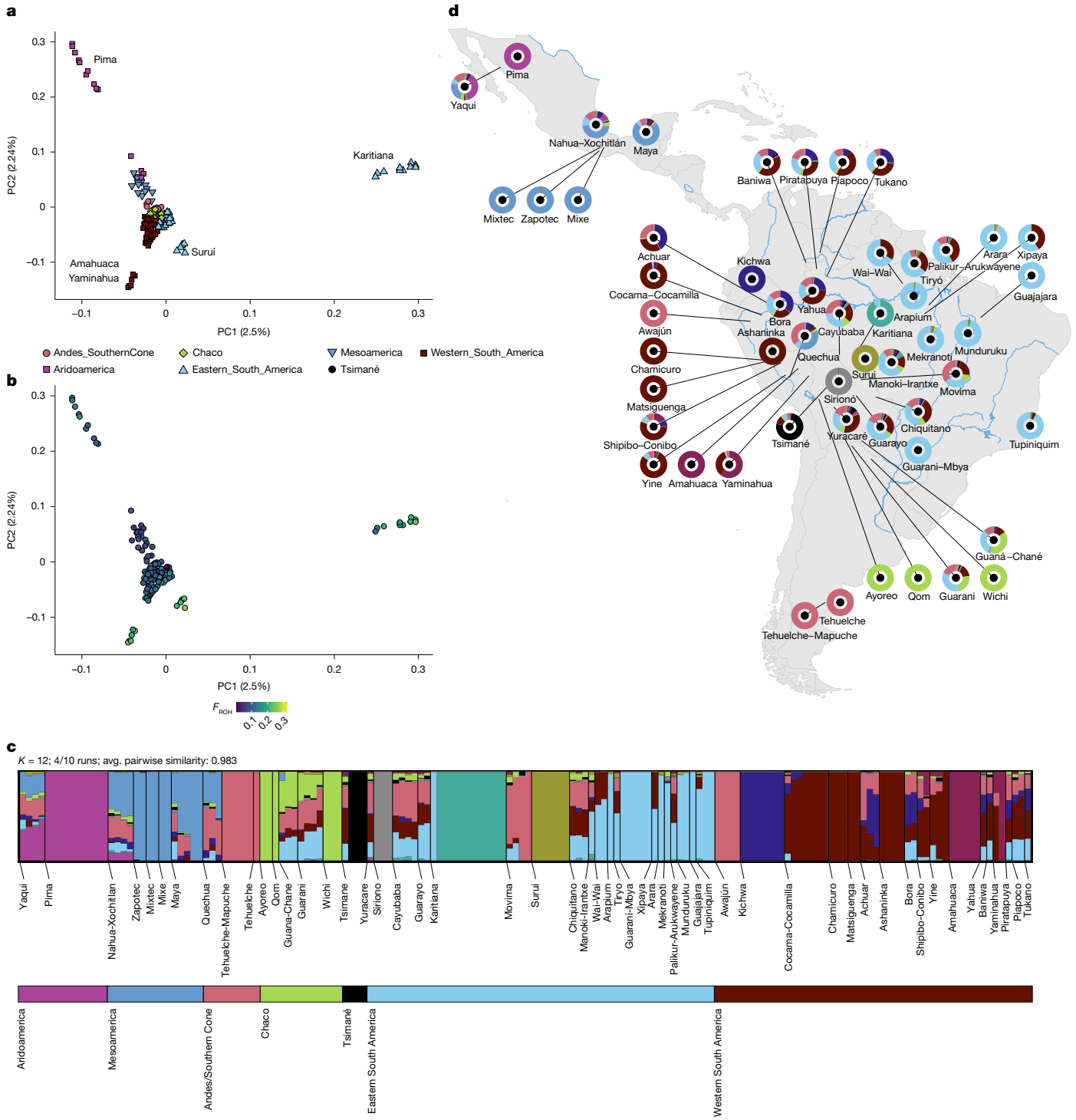


Fig. 2 | Genetic structure of present-day Indigenous American populations. **a**, The first two PCs inferred from linkage-disequilibrium-pruned masked genome-wide data. Individuals are colour-coded by genetic cluster, as defined in the analysis presented in Extended Data Fig. 2a. **b**, PCA visualization of the same data, colour-scaled by individual ROH-based inbreeding coefficients (F_{ROH}). **c**, Ancestry proportions inferred by means of unsupervised ADMIXTURE analysis at $K = 12$ (the highest K for which a consensus was reached), ordered by genetic cluster (as defined in Extended Data Fig. 2a) and plotted with PONG. **d**, Geographic distribution of population-averaged ancestry proportions at $K = 12$ overlaid on a partial map of America. Both analyses were restricted to the maximum set of individuals unrelated at first degree.

(Extended Data Figs. 3, 5 and Supplementary Table 4), with an excess of long segments (Extended Data Fig. 5) probably due to population collapse, fragmentation and isolation after European colonization[11,28], driven by epidemics, enslavement, warfare, displacement, instigation of intergroup violence, habitat destruction and disruption of subsistence and traditional knowledge[29,30].

Indigenous American ancestry was positively correlated with ROH number (Spearman's $r = 0.69$; $P = 3.16 \times 10^{-22}$), whereas African ($r = -0.45$; $P = 1.97 \times 10^{-11}$) and European ancestries ($r = -0.68$; $P = 4.20 \times 10^{-22}$) were negatively correlated, highlighting elevated ROH in Indigenous Americans and corroborating previous studies[31]. The Arapium individual from Northeastern Brazilian Amazon showed low

ROH consistent with their admixed profile (26.92% European, 11.25% African). This individual speaks Nheengatu, a Tupi-Guarani-derived creole historically used as a Brazilian lingua franca, reflecting increased contact and admixture with non-Indigenous populations.

Contrary to expectations under serial bottlenecks following entry into the Americas, population-averaged F_{ROH} values (Extended Data Fig. 3) showed no significant correlation with geographic distance from Alaska ($r = 0.22$; $P = 0.1091$), longitude ($r = 0.16$; $P = 0.2506$) or latitude ($r = -0.25$; $P = 0.0734$). Clusters of high F_{ROH} are observed in southern Amazonia, northern Bolivia and southeastern Peru, though uneven sampling could influence these patterns.

Identity by descent (IBD) analyses showed that segments greater than 8 cM were mostly shared within populations (Extended Data Fig. 6 and Supplementary Table 8), indicating limited postcolonial gene flow. Small, isolated populations probably experienced increased inbreeding, explaining the high prevalence of long ROHs (Extended Data Fig. 5) and reflecting population fragmentation and reduced intergroup contact following European colonization.

Effective population size (N_e) histories inferred from IBD (Methods) showed a widespread postcontact bottleneck across Indigenous American populations, with recent recovery in Chaco and western South America (Extended Data Fig. 2d). Precolonial expansions of roughly an order of magnitude (from 10^4 to 10^5) occurred in eastern and western South America (Extended Data Fig. 2d), consistent with Late Holocene growth[32,33]. Mesoamerican populations maintained high N_e, although eastern and western South America exceeded them about 1,500–500 years ago, with western South America and Chaco surpassing all in the most recent period. Aridoamerica had persistently low N_e and the strongest postcontact bottleneck (Extended Data Fig. 2d), matching the high number of long IBD segments in this cluster (Extended Data Fig. 6).

Cross-coalescence rate estimates (Methods) show that Indigenous American populations, grouped by genetic cluster, diverged from other continental groups between about 70,000 and 15,000 years ago (Extended Data Fig. 7), consistent with Out-of-Africa dispersal and the initial peopling of the Americas, when their ancestors were already genetically distinct from Asian ancestors[4,5]. Within the continent, differentiation among Indigenous American clusters increased during the Middle Holocene and became pronounced in the Late Holocene (Fig. 3c), coinciding with the emergence of distinct ancestral components and regionally structured genetic profiles (Fig. 3c). Estimates of cross-coalescence rates between South American clusters (regions) rise between 5,000 and 3,000 years ago for most pairs (Extended Data Fig. 7). These dates contrast with an independent study estimating divergence of principal South American groups (Amazonia, Andes, Chaco, Patagonia/Southern Cone) at 13,900–10,000 years ago[34]. Our results better match the timing of dispersals into South America, particularly the second main wave, which is inferred to have contributed most to present-day populations and to have started arriving around 9,000 years ago[4,5]. The difference probably reflects our broader, more representative sampling across ethnolinguistic and geographic diversity.

Our inferred timeframe matches the diversification of several language families[35] and supports the increasing population structure during this period. The Middle Holocene brought major climatic shifts, weakening the South American monsoon and causing instability[36]. Combined with megafauna extinction[37], these changes disrupted ecosystems and reduced resources. Archaeological evidence indicates demographic decline and retreat into refugia, probably driving the genetic isolation and differentiation we observe[38]. Population recovery followed, aided by early plant cultivation and regionally structured societies[38].

We inferred the N_e history of genetic clusters from within-population coalescence rates (Methods) and found that over the past 10,000 years, Indigenous American effective population sizes declined by 40–90%, consistent with previous reports[34]. Historically, North American populations had slightly higher N_e than South American groups (Extended Data Fig. 2e), but in the past 1,000 years, South American lowland populations (Western and Eastern clusters) show marginally higher N_e, matching IBD-based inferences of recent expansion (Extended Data Fig. 2d,e). IBD-based methods capture recent events (past 50–100 generations) but miss deeper history, whereas coalescent approaches resolve ancient events but have limited recent resolution. Combining both provides a fuller view across timescales. These results highlight the complex demographic history of Indigenous Americans, from divergence from Asian ancestors to continental differentiation, and show the lasting effect of colonization on genetic diversity.

Continental dispersals and continuity

To assess shared ancestry across time and space, we estimated genetic affinities among all Indigenous Americans, including ancient genomes from the Allen Ancient DNA Resource[39] and Brazilian sambaqui mound builders[15] (Methods, Supplementary Note 3 and Supplementary Table 9). A neighbour-joining tree based on unbiased genetic distances showed a strong link between genetic similarity and geography, with some regions displaying long-term continuity (Fig. 3 and Extended Data Fig. 8c,d). Clustering mirrored ADMIXTURE results (Fig. 3 and Extended Data Figs. 8 and 9), forming geographically restricted clades indicative of at least partial continuity and limited mobility. Genetic and geographic distances were moderately correlated (Spearman's $r = 0.46$; $P \cong 0$), similar to when only present-day populations were considered (Supplementary Note 4). Genetic distance and temporal separation between individuals showed a weak but significant correlation ($r = 0.29$; $P \cong 0$). Inferred correlations were stronger in North America and weaker in South America and across regions (Supplementary Note 4).

The neighbour-joining tree indicates at least three main dispersals into South America. The earliest includes Southern Native American (SNA) individuals older than about 9,000 years. They cluster together (Fig. 3 and Extended Data Fig. 8c,d) and show mixed ancestry components (Fig. 3 and Extended Data Fig. 9), reflecting an early undifferentiated SNA ancestry that later diverged into distinct lineages[4,5]. Key representatives include Anzick-1, Montana, USA (about 12,700 years; Clovis culture[40]); a Spirit Cave, Nevada, USA individual (about 11,000 years)[4,5]; a Los Rieles, Chile individual (about 12,000 years)[5]; and Sumidouro and Lapa do Santo, Lagoa Santa, Brazil individuals (more than 9,000 years)[4,5].

About 9,000 years ago, a distinct genetic lineage spread through Central America into southern Mexico, Belize and Panama. This ancestry also appears in the Caribbean, especially in Archaic-period (preceramic) Cuban individuals (about 5,000–2,500 years ago) and in northern South America, including present-day Venezuela. We reference the archaeological period for Caribbean individuals because genetic patterns in the region align closely with shifts in material culture, as previously demonstrated[33] and as we further corroborated. This lineage extended into southern South America, where groups such as the Quechua in Peru and the Tehuelche in Argentina show continuity with it today. Genetic evidence supports this later dispersal and partial replacement of earlier populations[4,5], consistent with ancient DNA studies showing a population turnover in the Americas beginning at least about 9,000 years ago.

Notably, nearly all present-day Indigenous South American and Caribbean individuals from the Ceramic period (about 2,500–500 years ago) show a distinct genetic affinity, showing a previously unrecognized third dispersal into South America. This ancestry peaks in the Ceramic-period Caribbean group (Extended Data Fig. 9; $K = 5$–8), and these populations form a separate branch in the neighbour-joining tree (Fig. 3 and Extended Data Fig. 8c,d). The dispersal probably occurred at least 1,300 years ago, on the basis of the earliest individuals in this cluster.

Our analysis of ancient and modern genomes improves resolution of population relationships, especially for the second dispersal. Archaic

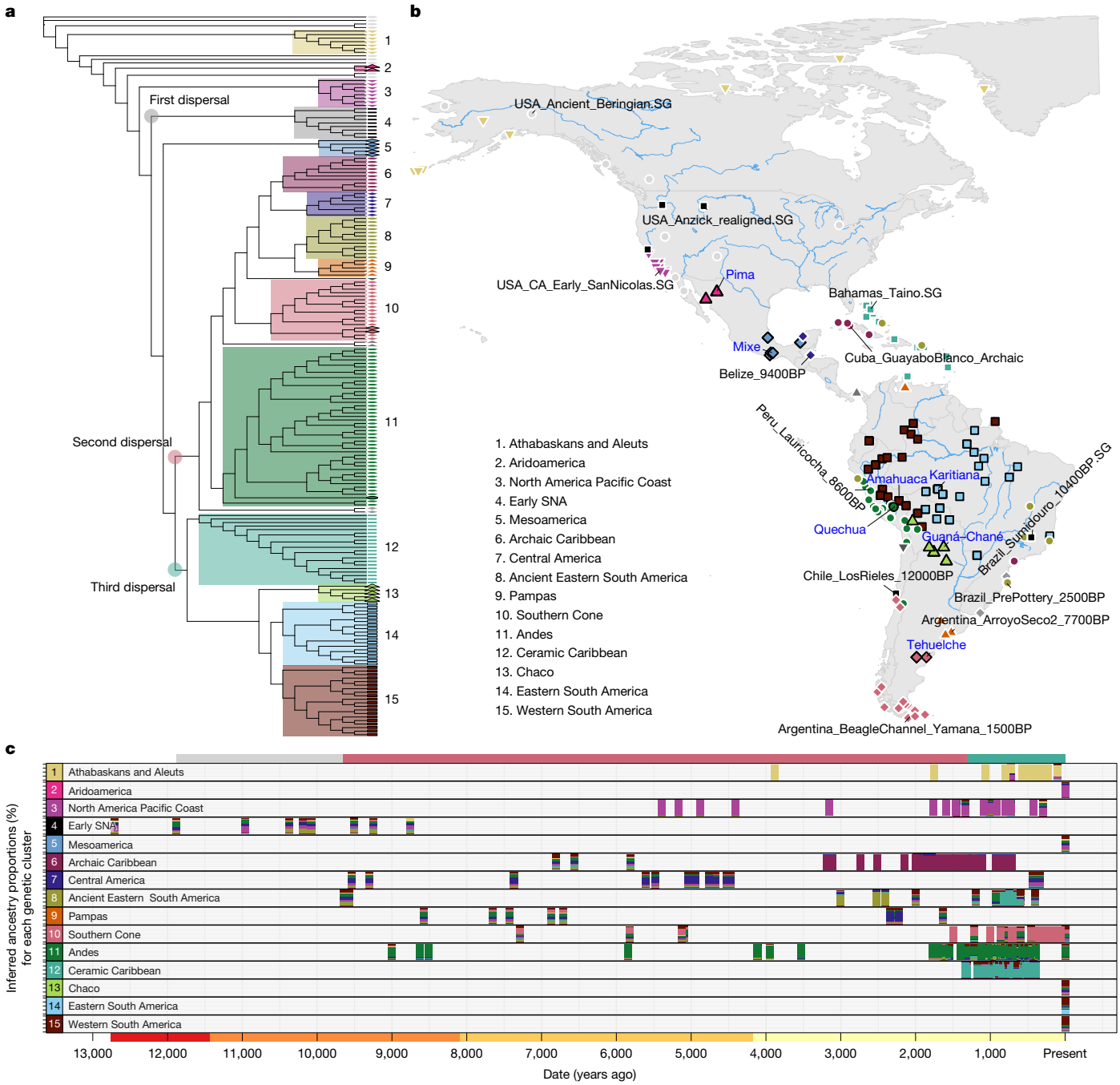


Fig. 3 | Spatiotemporal patterns of genetic similarity among Indigenous Americans. a, Neighbour-joining tree constructed using genetic distances calculated as $1 - f_3$(Mbuti; X, Y), where X and Y represent pairs of Indigenous American groups. Colour-coded clades define genetic clusters, with branch lengths omitted to emphasize topology, and clades encompassing individuals from the three proposed dispersals are highlighted by coloured circles. Present-day (black stroke) and ancient (white stroke) individuals are distinguished. **b**, The geographic distribution of genetic clusters shown in **a**, with labels for selected present-day (blue text) and ancient (black text) groups. **c**, Ancestry proportions inferred by unsupervised ADMIXTURE analysis at $K = 10$. For each genetic cluster (panels), individual ancestry estimates are shown along a temporal axis spanning the present to 15,000 years ago (x axis), with proportions ranging from 0 to 100% (y axis). The top bar indicates the earliest occurrence of individuals associated with the first (grey), second (red) and third (green) dispersals in South America and the Caribbean. The bottom bar denotes the geological timescale: Late Pleistocene (red) and Early (orange), Middle (gold) and Late Holocene (yellow).

Caribbean and Middle Holocene sambaqui individuals from Brazil (Laranjal and Moraes) show high genetic similarity (Archaic Caribbean cluster; Fig. 3 and Extended Data Fig. 8c,d). Late Holocene sambaqui people are closer to ancient Caribbean Ceramic groups (Ancient Eastern South America cluster). The Archaic Caribbean cluster aligns with ancient Central Americans, whereas ancient Eastern South Americans group with Argentinean Pampas populations. Both diverge first from Southern Cone and then from Andes populations, indicating that the second dispersal might have moved south along the Pacific Coast before expanding north.

Several South American Indigenous groups cluster with ancient individuals from the second dispersal. The Quechua align with ancient Andean populations, whereas the Mapuche-Tehuelche and Tehuelche cluster with ancient Southern Cone individuals, reflecting continuity

in their current regions. Indigenous North American populations are the most genetically distinct from South American groups. Aridoamerican populations diverged before all lineages that later expanded into South America, whereas Mesoamerican groups split after the initial dispersal but before the second-dispersal clusters formed (Fig. 3 and Extended Data Fig. 8c,d).

A Late Holocene third dispersal

We inferred population history by integrating genomic data from ancient and present-day Indigenous Americans. Representative individuals from all genetic clusters were selected (Fig. 3 and Extended Data Fig. 8c,d), focusing on the most differentiated ancestry components, and admixture graph modelling was applied (Methods). Late Pleistocene individuals were included to reconstruct Indigenous American ancestry. Starting from tree models without admixture, we progressively added admixture events using an unsupervised, genome-based approach.

Admixture graph analyses placed present-day Indigenous South Americans in a clade with Ceramic-period ancient Caribbeans (Extended Data Fig. 10a–c), whereas Central and South American ancient individuals, including Archaic-period Caribbeans, formed a separate clade, indicating distinct dispersals. This mirrors earlier genetic affinities (Fig. 3 and Extended Data Figs. 8 and 9), with present-day Indigenous South Americans closer to Ceramic-period Caribbeans and second-dispersal individuals more similar to each other. The 11,000-year-old Spirit Cave individual shows affinity to the first-dispersal clade (Fig. 3) but splits earliest in the second-dispersal clade (Extended Data Fig. 10a–c).

Single-admixture models identify the ancestral mixture forming Indigenous Americans from East Asians (Han) and Ancient North Siberians (Russia_MA1_HG.SG), with Ancient Beringians diverging first (Extended Data Fig. 10b). Including a second admixture into the model, third-dispersal individuals, including present-day Indigenous South Americans and Ceramic-period Caribbeans, derive from a mixture of branches related to Mixe and Sumidouro (Brazil_Sumidouro_10400BP.SG, and Extended Data Fig. 10c).

The third dispersal can be modelled as a mixture of early SNA and Mesoamerican ancestry, represented by Lagoa Santa (Brazil_Sumidouro_10400BP.SG) and Mixe. Late survival of Early SNA lineages in South America has been suggested for ancient individuals from Eastern South America[15]. The second and third dispersals remained genetically distinct, with no early admixture between the second dispersal and Mesoamericans. Over time, genetic affinity with Mesoamericans increases (f_3(Mbuti; Mixe, X), Spearman's $r = 0.7028$, $P < 2.2 \times 10^{-16}$; Fig. 4b). A similar trend is seen with North American Pacific Coast groups (f_3(Mbuti; USA_CA_Early_SanNicolas.SG, X), Spearman's $r = 0.4198$, $P < 2.2 \times 10^{-16}$), reflecting their phylogenetic proximity to Mesoamericans (Fig. 4a). Some later individuals in the second-dispersal clade (Fig. 3a and Extended Data Fig. 8c) also show increasing Mesoamerican affinity (Fig. 4b), indicating a gradual influx rather than rapid replacement. This progressive gene flow would have eventually reached a threshold sufficient to generate the genetic differentiation observed between the third- and second-dispersal populations.

Present-day Tehuelche and Quechua, showing strong continuity with second-dispersal ancestors (Fig. 3 and Extended Data Fig. 8c,d), were included in admixture graph models to assess ancestry. Without admixture, both clustered with other contemporary populations and Ceramic-period Caribbeans (third dispersal), whereas other Central and South Americans formed a separate second-dispersal branch (Extended Data Fig. 10d). Accounting for admixture, Tehuelche retain continuity with Southern Cone second-dispersal ancestors (Argentine Pampas) but derive most ancestry from a Chaco sister branch (third dispersal). Quechua mostly descend from an early-diverging third-dispersal branch while maintaining continuity with second-dispersal ancient Andeans (Extended Data Fig. 10f,g).

In automated models including Tehuelche and Quechua, the third dispersal is not modelled as early SNA-Mesoamerican admixture. The second-dispersal clade forms from admixture between early SNA lineages in South America (Brazil_Sumidouro_10400BP.SG) and North America (post-Anzick-1), indicating distinct early SNA contributions (Extended Data Fig. 10f,g).

We built a comprehensive admixture graph integrating previous knowledge about Indigenous American population history (Methods; Fig. 4). The summary model includes ancestry from an unsampled population A (PopA; Fig. 4a), contributing to present-day Mixe[4]. Gene flow into third-dispersal populations was modelled using both the Mixe and ancient California Channel Island individuals as a sister branch. We also modelled ancestry from an unsampled population Y (Ypikuéra, 'ancestor' in Tupi), explaining excess affinity between some Indigenous Americans and Australasians[41] (PopY; Fig. 4a). The model also accommodates PopA contributions to Mesoamericans and ancient Californians (Fig. 4a). Observed and predicted F-statistics show no significant deviation (maximum $|Z| < 3$; Fig. 4a). Adding Quechua and Tehuelche reduced model fit (maximum $|Z| = 4.43$), indicating that their ancestry is not fully captured.

We tested the minimum number of ancestral sources explaining genetic diversity in third-dispersal populations using qpWave, with first- and second-dispersal populations and ancient North American Pacific Coast individuals as candidate sources (right populations; Methods). Models with rank 0 and rank 1 were rejected ($P = 0$; $P = 1.8 \times 10^{-25}$), whereas rank 2 was not ($P = 1$; Supplementary Table 10). Thus, one or two ancestry streams are insufficient, and at least three distinct dispersal events from genetically differentiated sources are required to explain the observed patterns of genetic diversity.

Using qpAdm, we inferred the ancestral sources of present-day Indigenous Americans, initially excluding populations linked to the third dispersal (Methods). Most South American groups show substantial ancestry related to the North American Pacific Coast (Extended Data Fig. 11a and Supplementary Table 10). In the summary admixture graph (Fig. 4a), this source is modelled as sister to present-day Mesoamericans (Mixe), indicating gene flow from Mesoamerica into South America that probably shaped third-dispersal populations. Further ancestry derives from first-dispersal SNA groups and, to a lesser extent, second-dispersal populations, mainly ancient Eastern South Americans. These results are consistent across top-ranked models (Extended Data Fig. 11a and Supplementary Table 10).

Including all cluster representatives as sources (Methods) yields concordant patterns (Figs. 2–4), with most present-day Indigenous South American ancestry attributed to third-dispersal sources (Extended Data Fig. 11b and Supplementary Table 10). Limited continuity with the first and second dispersals is detected in a few groups, particularly Early SNA and ancient Eastern South American clusters. Tehuelche and Quechua are notable exceptions, deriving most ancestry from Early SNA or a mix of Early SNA and North American Pacific Coast sources, a pattern stable across high-probability models (Extended Data Fig. 11b and Supplementary Table 10).

Our results indicate that present-day Indigenous South Americans and Ceramic-period Caribbean populations were shaped by an unidentified third main dispersal, probably from Mesoamerican-related groups. This Late Holocene dispersal predates the arrival of ceramic-associated groups in the Caribbean and is at least 1,300 years old, on the basis of the earliest individuals assigned to it. Distinctive ancestry and affinity patterns (Fig. 3 and Extended Data Fig. 8c,d), inferred Mesoamerican-related contributions (Fig. 4a and Extended Data Figs. 10 and 11) and elevated Mesoamerican–South American affinity (Extended Data Fig. 5), together with increasing affinity through time (Fig. 4b), collectively support this model.

These findings are consistent with previously inferred Mesoamerican admixture in some Indigenous South American populations[4] and with a 'non-Anzick' SNA lineage related to ancient California Channel

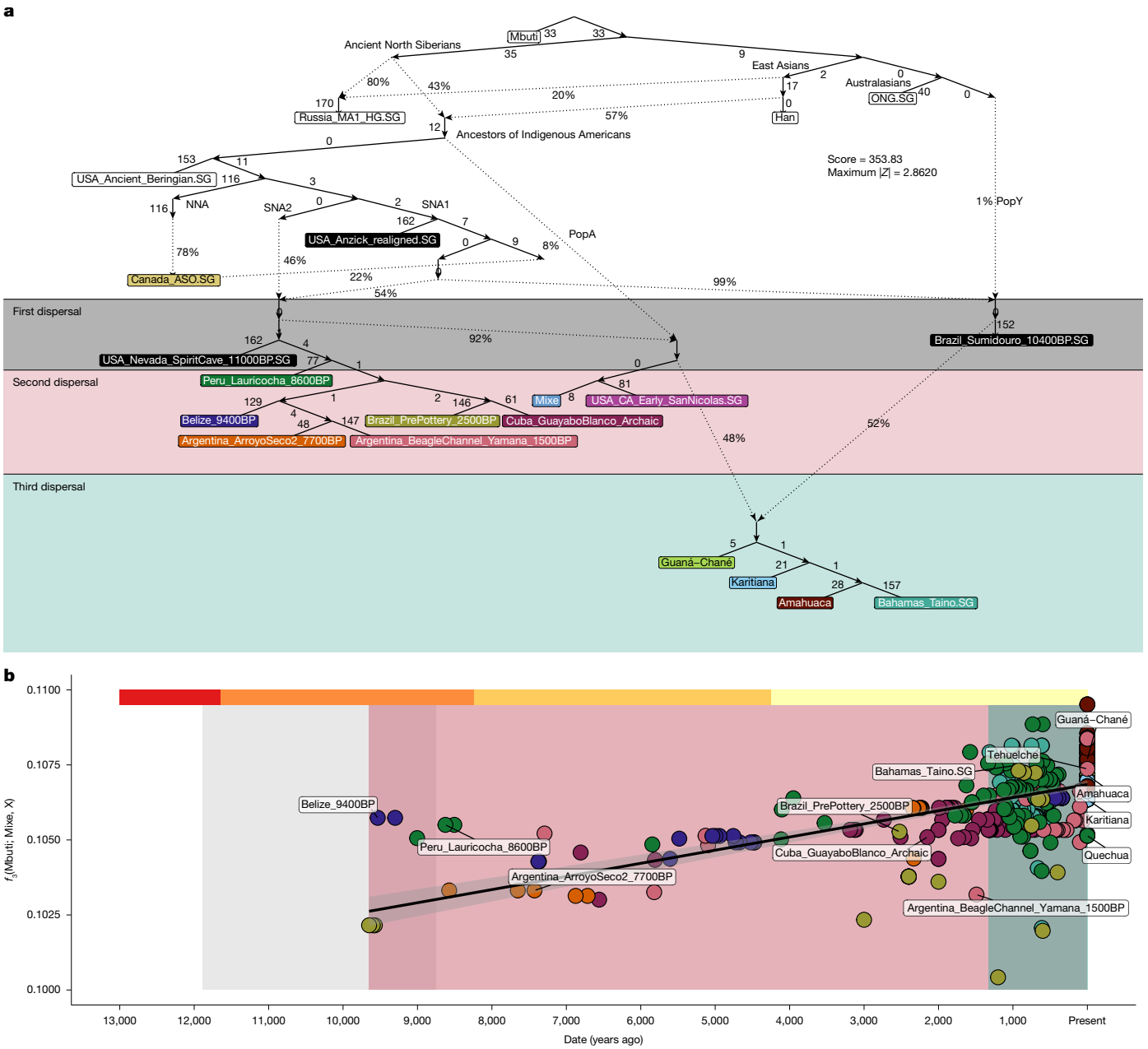


Fig. 4 | Population history modelling and genetic affinity to Mesoamericans. **a**, Admixture graph summarizing divergence and admixture events among ancient and contemporary populations. Models were inferred by means of an automated graph-search algorithm ('find_graphs' function from 'admixtools' R package), iteratively adding admixture events (0–5) to a base tree. Best-fit graphs (highest score and Extended Data Fig. 10) were refined manually, incorporating evidence from the literature, and then summarized and validated using 'qpgraph' ('admixtools' R package). Confidence intervals for drift lengths and admixture weights were calculated by means of SNP-block resampling. Analyses used transversions only to minimize biases. **b**, Genetic affinity of individuals to Mesoamericans (*x* axis) versus sample dates (*y* axis). Shaded backgrounds mark the earliest presence of individuals linked to the first (grey), second (red) and third (green) dispersals in South America and the Caribbean. The top bar denotes the geological timescale: Late Pleistocene (red) and Early (orange), Middle (gold) and Late Holocene (yellow).

Islands individuals who largely replaced local South American groups, especially in the central Andes[5]. Our results further indicate that these components were introduced by the dispersal of closely related populations.

Affinity with Australasians and archaics

Some Indigenous American populations show elevated genetic affinity to present-day Australasians relative to other groups[41,42], contradicting a single non-Arctic Indigenous American clade. This affinity is best explained by admixture between the ancestors of Indigenous Americans and an unsampled ancient Asian population, termed Ypykuéra[41,42] (here referred to as Ypykuéra ancestry), partially related to a sister clade of present-day Australasians.

We assessed genetic affinity between ancient and modern Indigenous Americans and present-day Australasians, the closest living proxies for Ypykuéra ancestry (Methods). We applied *F*-statistics to modern Indigenous American pairwise comparisons (Supplementary Table 11) and to comparisons including ancient individuals (Supplementary Table 12).

Several Indigenous groups, including the Awajún, Ayoreo, Guarani, Karitiana, Sirionó, Suruí and Tsimané, show significant excess genetic affinity to Australasians relative to other present-day populations

($Z > 3$; Supplementary Table 11). These groups span eastern and western South America and the Chaco, with the strongest enrichment in the southwestern Amazon, where five of these seven populations are located.

A second analysis detected at least one individual with significant affinity in all examined clusters (Extended Data Fig. 12 and Supplementary Table 12), except Arctic and northern North American groups, which were excluded from this analysis because of partial or complete ancestry from independent Siberian dispersals[43,44]. The earliest signal occurs in the 10,400-year-old Sumidouro individual (Extended Data Fig. 12). Signals persist from the Early Holocene to the present, increasing in frequency during the Late Holocene, especially in the Andes, Pacific Coast and western South America (Extended Data Fig. 12). The partially discontinuous spatiotemporal pattern probably reflects variation in prevalence within and among populations[42]. Taken together, these findings indicate that this ancestry was present during the initial peopling of America and that it may have contributed more strongly to Late Holocene and present-day genetic diversity.

We tested whether Ypykuéra-related ancestry in Indigenous Americans reflects shared ancestry with Australasians by means of archaic hominins (Neanderthals and/or Denisovans). We compared D(Mbuti, Onge; Mixe, X) with D(Mbuti, Neanderthal or Denisova; Mixe, X), where X denotes Indigenous American groups (Supplementary Table 13). Mixe served as a Mesoamerican reference to match earlier studies reporting Australasian affinity[41,42]. No correlation was detected between Australasian and Neanderthal (Spearman's $r = -0.006$, $P = 0.971$) or Denisovan affinity ($r = -0.1002$, $P = 0.5372$). By contrast, Neanderthal and Denisovan affinities were strongly correlated ($r = 0.6572$, $P = 7.2 \times 10^{-6}$), consistent with homogeneous archaic ancestry in the founding populations.

An alternative hypothesis proposes that Australasian affinity reflects retention of the Ypykuéra component in isolated groups with high internal genetic similarity[14]. Such populations and genomic regions, characterized by elevated ROH, would be less affected by admixture that could dilute signals of ancient Population Y ancestry. This hypothesis is not supported by our data, which show no correlation between Australasian affinity and inbreeding (F_{ROH}) (Spearman's $r = 0.2503$; $P = 0.1192$). Moreover, ROH hotspots, defined as regions with ROH density greater than three standard deviations above the mean, show little overlap with loci of Australasian affinity, with only about 6% of such positions coinciding (Supplementary Note 5).

We tested whether Indigenous American affinity to present-day Australasians also includes ancient Hòabìnhian individuals, proposed ancestors of mainland Southeast Asian hunter-gatherers, including the Onge[45]. Using D(Mbuti, Y; Mixe, X), with Y as Onge or Hòabìnhian individuals (La368, La364) and X as Indigenous American populations, we evaluated correlations in affinity to Onge and Hòabìnhian individuals. La368 forms a sister branch to Onge, whereas La364 is modelled as Australasian-related plus Austronesian ancestry, sister to Ami[45]. We observed significant correlations for La368 (Spearman's $r = 0.6444$; $P = 1.1856 \times 10^{-5}$) and La364 (Spearman's $r = 0.6208$; $P = 2.8848 \times 10^{-5}$; Supplementary Table 14). These results support a shared ancestry component between Indigenous Americans and Australasians that extends deep into the past.

Genetic adaptation landscape

To investigate shared selective pressures shaping Indigenous American genomes, we performed a genome-wide scan for natural selection using complementary approaches based on population differentiation and extended haplotype homozygosity (Supplementary Note 7). We identified 12 candidate genomic regions (Fig. 5a), predominantly associated with immune response, cardiometabolic traits, fertility and anthropometric features (Supplementary Table 15).

Among the candidate genes, some have previously exhibited selection signals in Indigenous American or Latin American populations; however, these signals were reported to be restricted to specific ecoregions or populations rather than being broadly distributed across the American continent, as observed here. The strongest signal was in *LINC00871*, associated with the fertility rate[46,47]. Other signals were detected in *GALNT13*, involved in fetal growth[48] and malaria protection[49] and in *IFH1*, associated with viral responses and autoimmune diseases[50]. Signals in FADS (*FADS2* and *FADS3*) and *SLC25A17* genes[10] were also detected above the 99th percentile threshold.

Our analyses also identified new candidate genes under selection, including *CMKLR1*, a chemokine receptor involved in early immune response, adipocyte development and reproductive regulation[51]. *CCDC134* (coiled-coil domain containing 134) encodes a secreted protein with cytokine-like activity that modulates $CD8^+$ T cell activation[52]. *DOCK2* (dedicator of cytokinesis 2), expressed in hematopoietic cells, regulates immune cell activation and has a central role in inflammatory responses[53]. Together, these candidate genes show diverse biological processes shaped by natural selection in Indigenous American populations, highlighting the evolutionary forces that have structured their genetic landscape (Supplementary Table 15).

Evolutionary effect of Ypykuéra ancestry

Ypykuéra ancestry in present-day Indigenous American populations remains consistently low (1–3%)[41,42], comparable to estimates for the 10,400-year-old Sumidouro individual. Its persistence at similar levels over millennia raises the possibility that specific haplotypes were maintained by positive selection. To test this, we identified genomic regions in modern Indigenous Americans with excess affinity to present-day Australasians using two complementary approaches: a local ancestry deviation test and a genome-wide D-statistic in an overlapping sliding-window framework. Regions with Z-scores greater than 3 in both analyses were classified as candidate Ypykuéra ancestry regions. We then assessed evidence of selection using population branch statistics (PBS), evaluating whether these regions fall in the top 1% and 5% of the genome-wide PBS distribution (Supplementary Note 8).

This integrative approach identified 2,297 positions across 24 genomic regions spanning 17 chromosomes (Fig. 5b), of which 17 ranked in the 99th percentile of the PBS distribution (Supplementary Table 17). The strongest affinity to present-day Australasians mapped to *LINC00871*, associated with fertility[46], and was also the top hit in the genome-wide selection scan (Fig. 5a). Further prominent signals included *ALX4*, a transcription factor involved in craniofacial and hair follicle development[54], *HMGA1*, associated with immune response and insulin signalling[55], and *MTDH*, implicated in angiogenesis and cancer progression[56]. A complete list of candidate regions is provided in Supplementary Table 17.

Together, these findings indicate that several genomic regions sharing alleles with present-day Australasians were probably targets of positive selection. The candidate genes in these loci participate in critical biological processes that may have shaped the health and adaptive history of Indigenous American populations. However, positive selection seems to have acted on only a subset of Ypykuéra ancestry loci, whereas the remaining regions show no evidence of selection and are probably shaped by neutral evolutionary processes.

Adaptive introgression in America

Although archaic introgression from Neanderthals and Denisovans has been reported in the Americas[13,57], its adaptive significance in Indigenous American populations remains largely unexplored. We therefore conducted a genome-wide scan using SPrime on local-ancestry-masked genomes from 128 newly sequenced Indigenous

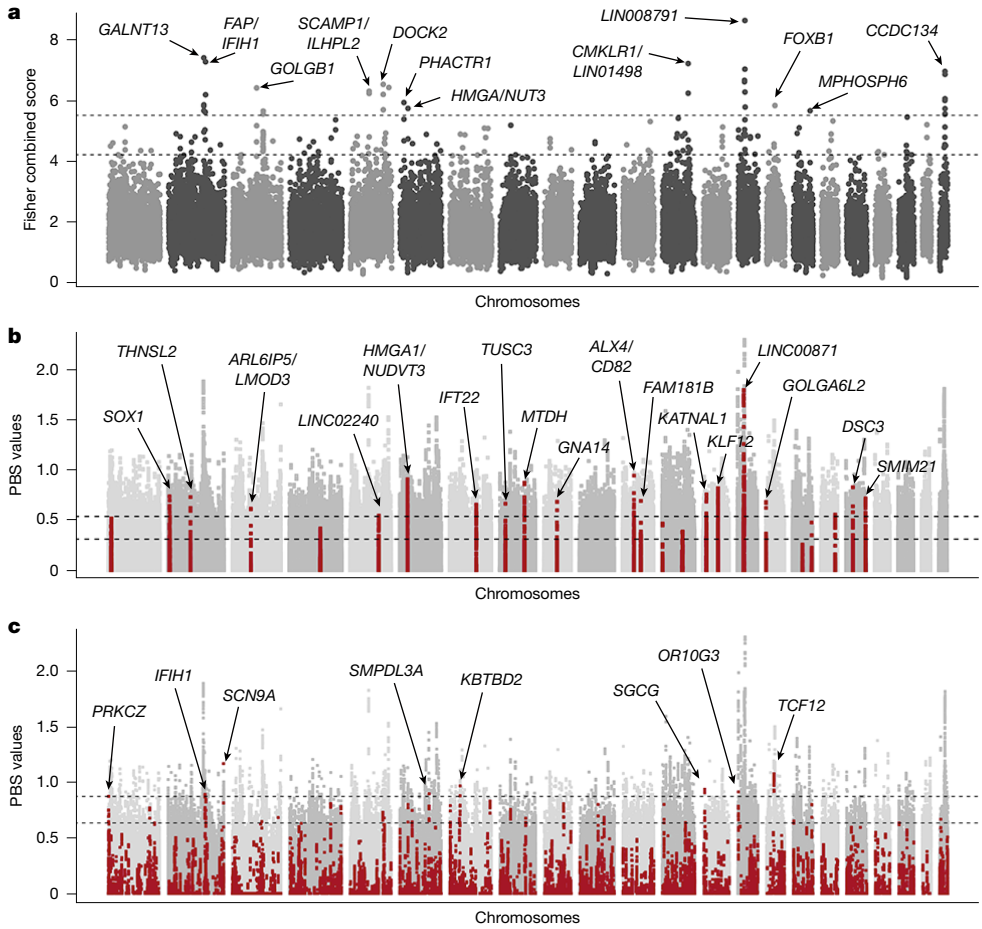


Fig. 5 | Genome-wide selection scan. a, Manhattan plot of the Fisher combined score of Indigenous Americans. The *x* axis shows the autosomal chromosomes, and the *y* axis represents the average of Fisher combined score values in a window of 200 SNPs, using a step size of 50 SNPs. The 99.5th and 99.9th percentiles of the empirical distribution are indicated by black dashed horizontal lines. **b**, Manhattan plot of PBS values and regions with excess genetic affinity to Australasians. Grey dots represent PBS values across the genome, and red dots mark positions with inferred excess genetic affinity to Australasians. The 95th and 99th percentiles of the empirical distribution are indicated by black dashed horizontal lines. **c**, Manhattan plot of PBS values and introgressed archaic alleles. Grey dots represent PBS values across the genome, and red dots highlight introgressed archaic alleles identified in Indigenous Americans. The 99.5th and 99.9th percentiles of the empirical distribution are indicated by black dashed horizontal lines.

American individuals (Methods). After quality filtering (Supplementary Note 9), high-confidence archaic alleles comprised about 1.2% of the genome.

Overrepresentation analysis (ORA) on the basis of Gene Ontology categories identified significant enrichment for genes related to epidermal structure ($P = 5.01 \times 10^{-4}$), keratin filament ($P = 1.19 \times 10^{-4}$) and intermediate filament organization and keratinization ($P = 6.07 \times 10^{-8}$ and 2.22×10^{-6}) (Supplementary Table 18). These findings support previous evidence that keratin-related Neanderthal alleles contributed to adaptation outside Africa[58,59], consistent with an adaptive role in the Americas.

Similar patterns were observed for Denisovan-exclusive introgressed variants, with further enrichment in genes associated with calcium channel activity ($P = 6.03 \times 10^{-5}$) (Supplementary Table 19). By contrast, Neanderthal-exclusive introgressed variants were enriched for C–C chemokine receptor activity and structural constituents of the skin epidermis ($P = 4.51 \times 10^{-7}$ and 1.44×10^{-5}). Notably, Neanderthal introgression of C–C chemokine receptor genes has been widely associated with immune responses and human health[60–63].

Among the putative archaic alleles shared with Neanderthals and Denisovans and observed exclusively in Indigenous American populations, 488 alleles were identified across 84 genes. ORA showed functional enrichment for phenotypes associated with cyanosis ($P = 5.68 \times 10^{-4}$), limitation of joint mobility ($P = 7.77 \times 10^{-4}$) and agenesis of molars ($P = 1.15 \times 10^{-3}$) (Supplementary Table 21).

To identify adaptive archaic introgression in the Americas, we conducted a genome-wide PBS scan using Indigenous Americans as the focal population, Siberians as the sister group and East Asians as the outgroup (Fig. 5c). The list of introgressed archaic alleles in the 99.9th percentile is provided in Supplementary Table 22. Further selection signals were observed in functionally enriched loci such as *IFIH1*, *PRDM16* and *SCN9A* (Supplementary Tables 21 and 22). Notably, several of these genes were also reported under selection in admixed Latin American populations, although previous analyses could not disentangle Indigenous American from European ancestry components[64].

Signals in *PRKCZ*, *KBTBD2* and *OR10G3* have not been reported in previous studies. *PRKCZ* encodes a serine/threonine kinase involved in immune regulation, memory formation and glucose metabolism[65]. *KBTBD2* (BTB domain containing 2) has a key role in insulin signalling[66] and bone development[67], and *OR10G3* encodes an olfactory receptor linked to fruit consumption[68]. Together, these findings highlight the contribution of introgressed archaic variants to immune function and energy metabolism in Indigenous American populations.

Finally, to determine whether the observed affinity between some Indigenous Americans and Australasians could be attributed to shared

archaic ancestry, we examined the genomic overlap between regions with excess Ypykuéra ancestry and those with archaic introgression. This analysis showed a minimal overlap of 0.4%, corresponding to 11 genes/genomic regions (Supplementary Table 23). This minimal overlap indicates that the Ypykuéra ancestry is unlikely to have been derived from a known archaic hominin.

Conclusion

Our comprehensive genomic analysis showed a complex and dynamic demographic and evolutionary history that improved our understanding of human diversity and adaptation in the American continent. We present evidence of at least three principal population dispersals into South America, with subsequent regional differentiation and long-term genetic continuity in many areas, challenging oversimplified models of continental peopling. We demonstrated that Indigenous American genomes harbour extensive genetic diversity, highlighting the need to better represent these populations using global genomic datasets. This diversity reflects both deep-time evolutionary processes, such as the Beringian standstill and postglacial expansions, and more recent demographic shifts, including severe bottlenecks following European colonization and subsequent population isolation. Although structural variation was not assessed, these genomes probably contain substantial uncharacterized diversity, including copy number variation and insertions and deletions. Our analyses showed widespread signatures of natural selection across loci related to immune function, metabolism, fertility and development, indicating that Indigenous Americans underwent strong and diverse selective pressures shaped by the continent's vast environmental heterogeneity. Particularly remarkable are genomic regions with excess allele sharing with Australasian populations, the Ypykuéra signal, many of which we show are probably targets of positive selection. This indicates that the persistence of this genetic signal, present at low but consistent levels for more than 10,000 years, is probably not just a signal of ancient admixture but may also reflect adaptive advantages in specific environmental contexts. Furthermore, we identified candidate regions of adaptive archaic introgression from Neanderthals and Denisovans that contribute to functions related to immunity, metabolism and epidermal integrity, thereby reinforcing the role of archaic alleles in shaping the evolutionary trajectory of non-African populations. Importantly, our data indicate minimal overlap between genomic regions with Australasian affinity and those introgressed from archaic hominins, supporting the interpretation that these signals represent distinct evolutionary phenomena. Together, these findings underscore the unique evolutionary history of Indigenous Americans, shaped by complex interactions between ancient population structure, migration, natural selection and introgression. They also emphasized the importance of expanding genomic research among underrepresented populations to capture the full scope of human genomic diversity.

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/s41586-026-10406-w.

1. Henn, B. M., Cavalli-Sforza, L. L. & Feldman, M. W. The great human expansion. *Proc. Natl Acad. Sci. USA* **109**, 17758–17764 (2012).
2. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* https://doi.org/10.1038/538161a (2016).
3. Reich, D. et al. Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
4. Moreno-Mayar, J. V. et al. Early human dispersals within the Americas. *Science* **362**, eaav2621 (2018).
5. Posth, C. et al. Reconstructing the deep population history of Central and South America. *Cell* **175**, 1185–1197 (2018).
6. Hoffecker, J. F., Elias, S. A. & O'Rourke, D. H. Anthropology. Out of Beringia? *Science* **343**, 979–980 (2014).
7. Niedbalski, S. D. & Long, J. C. Novel alleles gained during the Beringian isolation period. *Sci. Rep.* **12**, 4289 (2022).
8. Couto-Silva, C. M. et al. Indigenous people from Amazon show genetic signatures of pathogen-driven selection. *Sci. Adv.* **9**, eabo0234 (2023).
9. Schlebusch, C. M. et al. Human adaptation to arsenic-rich environments. *Mol. Biol. Evol.* **32**, 1544–1555 (2015).
10. Amorim, C. E. G. et al. Genetic signature of natural selection in first Americans. *Proc. Natl Acad. Sci. USA* **114**, 2195–2199 (2017).
11. Castro E Silva, M. A. et al. Population histories and genomic diversity of South American natives. *Mol. Biol. Evol.* **39**, msab339 (2022).
12. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
13. Bergström, A. et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
14. Capodiferro, M. R. et al. Archaeogenomic distinctiveness of the Isthmo-Colombian area. *Cell* **184**, 1706–1723 (2021).
15. Ferraz, T. et al. Genomic history of coastal societies from eastern South America. *Nat. Ecol. Evol.* **7**, 1315–1330 (2023).
16. Fan, S. et al. Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell* **186**, 923–939 (2023).
17. Silcocks, M. et al. Indigenous Australian genomes show deep structure and rich novel variation. *Nature* **624**, 593–601 (2023).
18. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440 (2022).
19. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
20. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
21. Moorjani, P. & Hellenthal, G. Methods for assessing population relationships and history using genomic data. *Annu. Rev. Genomics Hum. Genet.* **24**, 305–332 (2023).
22. Lea, A. J. et al. Natural selection of immune and metabolic genes associated with health in two lowland Bolivian populations. *Proc. Natl Acad. Sci. USA* **120**, e2207544120 (2023).
23. Barbieri, C. et al. The current genomic landscape of western South America: Andes, Amazonia, and Pacific Coast. *Mol. Biol. Evol.* **36**, 2698–2713 (2019).
24. Harris, D. N. et al. Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proc. Natl Acad. Sci. USA* **115**, E6526–E6535 (2018).
25. Arango-Isaza, E. et al. The genetic history of the Southern Andes from present-day Mapuche ancestry. *Curr. Biol.* **33**, 2602–2615 (2023).
26. 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).
27. Krettek, K.-L. et al. A 6000-year-long genomic transect from the Bogotá Altiplano reveals multiple genetic shifts in the demographic history of Colombia. *Sci. Adv.* **11**, eads6284 (2025).
28. Adhikari, K., Chacón-Duque, J. C., Mendoza-Revilla, J., Fuentes-Guajardo, M. & Ruiz-Linares, A. The genetic diversity of the Americas. *Annu. Rev. Genomics Hum. Genet.* **18**, 277–296 (2017).
29. Collen, E. J., Johar, A. S., Teixeira, J. C. & Llamas, B. The immunogenetic impact of European colonization in the Americas. *Front. Genet.* **13**, 918227 (2022).
30. Thornton, R. *American Indian Holocaust and Survival: A Population History Since 1492* (Univ. Oklahoma Press, 1987).
31. Nunes, K. et al. Admixture's impact on Brazilian population evolution and health. *Science* **388**, eadl3564 (2025).
32. Castro E Silva, M. A. et al. Genomic insight into the origins and dispersal of the Brazilian coastal natives. *Proc. Natl Acad. Sci. USA* **117**, 2372–2377 (2020).
33. Nägele, K. et al. Genomic insights into the early peopling of the Caribbean. *Science* **369**, 456–460 (2020).
34. Gusareva, E. S. et al. From North Asia to South America: tracing the longest human migration through genomic sequencing. *Science* **388**, eadk5081 (2025).
35. Campbell, L. in *The Indigenous Languages of South America: A Comprehensive Guide* (eds Campbell, L. & Grondona, V.) Ch. 2 (De Gruyter Mouton, 2012).
36. Prado, L. F., Wainer, I., Chiessi, C. M., Ledru, M.-P. & Turcq, B. A mid-Holocene climate reconstruction for eastern South America. *Clim. Past* **9**, 2117–2133 (2013).
37. Borrero, L. A. in *American Megafaunal Extinctions at the End of the Pleistocene* (ed. Haynes, G.) Ch. 8 (Springer, 2009).
38. Riris, P. & Arroyo-Kalin, M. Widespread population decline in South America correlates with mid-Holocene climate change. *Sci. Rep.* **9**, 1–10 (2019).
39. Mallick, S. et al. The Allen Ancient DNA Resource (AADR): a curated compendium of ancient human genomes. *Sci. Data* **11**, 182 (2024).
40. Rasmussen, M. et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014).
41. Skoglund, P. et al. Genetic evidence for two founding populations of the Americas. *Nature* **525**, 104–108 (2015).
42. Castro E Silva, M. A., Ferraz, T., Bortolini, M. C., Comas, D. & Hünemeier, T. Deep genetic affinity between coastal Pacific and Amazonian natives evidenced by Australasian ancestry. *Proc. Natl Acad. Sci. USA* **118**, e2025739118 (2021).
43. Flegontov, P. et al. Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America. *Nature* **570**, 236–240 (2019).
44. Raghavan, M. et al. The genetic prehistory of the New World Arctic. *Science* **345**, 1255832 (2014).
45. McColl, H. et al. The prehistoric peopling of Southeast Asia. *Science* **361**, 88–92 (2018).
46. Barban, N. et al. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat. Genet.* **48**, 1462–1472 (2016).

47. Cheng, J. Y., Stern, A. J., Racimo, F. & Nielsen, R. Detecting selection in multiple populations by modeling ancestral admixture components. *Mol. Biol. Evol.* **39**, msab294 (2022).
48. Tekola-Ayele, F. et al. Admixture mapping identifies African and Amerindigenous local ancestry loci associated with fetal growth. *Hum. Genet.* **140**, 985–997 (2021).
49. Mackinnon, M. J. et al. Environmental correlation analysis for genes associated with protection against malaria. *Mol. Biol. Evol.* **33**, 1188–1204 (2016).
50. Onengut-Gumuscu, S. et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).
51. Goralski, K. B. et al. Chemerin, a novel adipokine that regulates adipogenesis and adipocyte metabolism. *J. Biol. Chem.* **282**, 28175–28188 (2007).
52. Huang, J. et al. Cytokine-like molecule CCDC134 contributes to $CD8^+$ T-cell effector functions in cancer immunotherapy. *Cancer Res.* **74**, 5734–5745 (2014).
53. Chen, Y. et al. Dock2 in the development of inflammation and cancer. *Eur. J. Immunol.* **48**, 915–922 (2018).
54. Kayserili, H. et al. ALX4 dysfunction disrupts craniofacial and epidermal development. *Hum. Mol. Genet.* **18**, 4357–4366 (2009).
55. Bianchi, M. E. & Agresti, A. HMG proteins: dynamic players in gene regulation and differentiation. *Curr. Opin. Genet. Dev.* **15**, 496–506 (2005).
56. Dhiman, G. et al. Metadherin: a therapeutic target in multiple cancers. *Front. Oncol.* **9**, 349 (2019).
57. Qin, P. & Stoneking, M. Denisovan ancestry in East Eurasian and Native American populations. *Mol. Biol. Evol.* **32**, 2665–2674 (2015).
58. Sankararaman, S. et al. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014).
59. Simonti, C. N. et al. The phenotypic legacy of admixture between modern humans and Neandertals. *Science* **351**, 737–741 (2016).
60. Taskent, R. O. et al. Variation and functional impact of Neanderthal ancestry in Western Asia. *Genome Biol. Evol.* **9**, 3516–3524 (2017).
61. Hoover, K. C. Intragenus (Homo) variation in a chemokine receptor gene (CCR5). *PLoS ONE* **13**, e0204989 (2018).
62. Jagoda, E. et al. Detection of Neanderthal adaptively introgressed genetic variants that modulate reporter gene expression in human immune cells. *Mol. Biol. Evol.* **39**, msab304 (2022).
63. Jagoda, E. et al. Regulatory dissection of the severe COVID-19 risk locus introgressed by Neanderthals. *eLife* **12**, e71235 (2023).
64. Witt, K. E., Funk, A., Añorve-Garibay, V., Fang, L. L. & Huerta-Sánchez, E. The impact of modern admixture on archaic human ancestry in human populations. *Genome Biol. Evol.* **15**, evad066 (2023).
65. Hirai, T. & Chida, K. Protein kinase Czeta (PKCzeta): activation mechanisms and cellular functions. *J. Biochem.* **133**, 1–7 (2003).
66. Zhang, Z. et al. Insulin resistance and diabetes caused by genetic or diet-induced KBTBD2 deficiency in mice. *Proc. Natl Acad. Sci. USA* **113**, E6418–E6426 (2016).
67. Xun, Y. et al. KBTBD2 controls bone development by regulating IGF-1 signaling during osteoblast differentiation. *Cell Death Differ.* https://doi.org/10.1038/s41418-024-01416-0 (2024).
68. Cole, J. B., Florez, J. C. & Hirschhorn, J. N. Comprehensive genomic analysis of dietary habits in UK Biobank identifies hundreds of genetic associations. *Nat. Commun.* **11**, 1467 (2020).

[1]Institut de Biologia Evolutiva, CSIC/Universitat Pompeu Fabra, Barcelona, Spain. [2]Departament de Medicina i Ciències de la Vida, Universitat Pompeu Fabra, Barcelona, Spain. [3]Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil. [4]Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [5]Laboratório de Genética Humana e Médica, Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Brazil. [6]Institute of Human Origins and School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA. [7]Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil. [8]Departamento de Fisiologia, Universidade Federal do Espírito Santo, Espírito Santo, Brazil. [9]Department of Life and Environmental Sciences, University of Cagliari, Cagliari, Italy. [10]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland. [11]Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology (MPI-EVA), Leipzig, Germany. [12]Escuela Nacional de Antropología e Historia (ENAH), Mexico City, Mexico. [13]Chemistry Faculty, Universidad Autónoma de Yucatán (UADY), Mérida, Mexico. [14]Centro INAH Yucatán, Instituto Nacional de Antropología e Historia (INAH), Mérida, Mexico. [15]Biomedicine Research Unit, Zurita & Zurita Laboratorios, Quito, Ecuador. [16]Nursing School, Faculty of Health Sciences, Universidad de las Américas (UDLA), Quito, Ecuador. [17]Department of Biotechnology, Faculty of Engineering and Applied Sciences, Universidad de las Américas (UDLA), Quito, Ecuador. [18]Nursing Faculty, Universidad Autónoma de Yucatán (UADY), Mérida, Mexico. [19]Blood Bank Laboratory, Hospital General de Zona #30, Instituto Mexicano del Seguro Social (IMSS), Mexicali, Mexico. [20]Clinical Laboratory, Unidad Médica de Alta Especialidad (UMAE) #2, Instituto Mexicano del Seguro Social (IMSS), Ciudad Obregón, Mexico. [21]Instituto de Diversidad y Evolución Austral (IDEAus, CCT CONICET-CENPAT), Puerto Madryn, Argentina. [22]Laboratorio de Genética Molecular Poblacional, Instituto Multidisciplinario de Biología Celular (IMBICE), Buenos Aires, Argentina. [23]Instituto Patagónico de Ciencias Sociales y Humanas, Centro Nacional Patagónico, CONICET, Puerto Madryn, Argentina. [24]Universidad Mayor de San Andrés (UMSA), La Paz, Bolivia. [25]Centro de Genética y Biología Molecular (CGBM), Instituto de Investigación, Facultad de Medicina Humana, Universidad de San Martín de Porres, Lima, Perú. [26]Laboratório de Biodiversidade e Evolução Molecular (LBEM), Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [27]These authors contributed equally: Marcos Araújo Castro e Silva, Kelly Nunes, Maíra R. Ribeiro. ✉e-mail: hunemeier@usp.br

Methods

Sample description and DNA sequencing

In this study, we generated new whole-genome sequencing (WGS) data from 128 Indigenous American individuals representing 45 populations and 28 language families across 8 Latin American countries (Extended Data Fig. 1 and Supplementary Note 1). Whole genomes were sequenced at the Beijing Genomics Institute (China) and Dasa Genômica (Brazil) using BGISEQ-500 and Illumina NovaSeq 6000, with an average sequencing depth of about 44×.

Ethical approval for sample collection was granted by local ethics committees in each country: Argentina (Puerto Madryn Zonal Hospital, resolution no. 009/2015; San Carlos de Bariloche Zonal Hospital, resolution no. 1510/2015), Brazil (CONEP, resolution nos. 763, 4599, 3828655, 7107656 and 8273857), Bolivia (Universidad Mayor de San Andrés), Ecuador (Universidad de Las Américas; Consejo Nacional de Ciencia y Tecnología—CONACyT, grant no. 69856; Instituto Nacional de Ciencias Médicas y de la Nutrición Salvador Zubirán, refs. 15,18), Mexico (CONACyT grant no. 69856; Instituto Nacional de Ciencias Médicas y de la Nutrición Salvador Zubirán, refs. 15,18; CNIC Salud 2013-01-201471; Committee of Ethics and Research, UADY, notice F-FENC-SAC-14/REV: 04, registry no. 09/17) and Peru (Universidad San Martín de Porres). Written informed consent was obtained from all participants before sample collection. Logistical support in Brazil was provided by the Fundação Nacional do Índio. All sampling adhered to the Declaration of Helsinki and the relevant national laws and regulations at the time (Supplementary Note 10).

Read mapping, variant calling and annotation

Whole-genome sequence data preprocessing, variant calling and annotation were performed using the Sarek v.3.5.0 pipeline[69]. Specifically, sequence data in FASTQ format were aligned to the GRCh38 reference genome and preprocessed according to the GATK Best Practices for germline variant discovery and joint variant calling. Variants in the joint cohort variant call format were then normalized and annotated using the Ensembl Variant Effect Predictor (VEP)[70] v.113, incorporating several annotation sources. Annotations from dbSNP, ClinVar and more custom annotations were retrieved using SnpSift and VEP plugins. Ancestral alleles were filtered using the VEP Ancestral Allele plugin to improve the specificity of downstream population genetic analyses. New SNVs were identified by comparing their positions and alleles with those of the variants in public databases (1KGP[18], HGDP[13], gnomAD[19] and dbSNP[20]). Variants absent from the dataset of more than 270,000 individuals were classified as new (Supplementary Note 2).

Site frequency spectra

We estimated the number of segregating single-nucleotide polymorphisms (SNPs) as a function of sample size using a rarefaction approach on the basis of the site frequency spectrum (SFS). Alternate allele counts were computed for biallelic sites and were used to construct the SFS using Scikit-Allel v.1.3.13. To account for different sample sizes and normalize the comparisons, the folded SFS was projected onto smaller sample sizes using a hypergeometric downsampling method.

Dataset assembly and quality control

Genomic coordinates of the newly sequenced individuals were mapped to the hg38 reference genome. The 128 newly generated genomes were merged with the following publicly available WGS databases (Supplementary Note 1): (1) 1KGP High Coverage, (2) HGDP and (3) SGDP. Sites and individuals with more than 5% missing data were eliminated, biallelic SNVs were selected, and positions with significant deviations ($P < 10^{-8}$) from the Hardy–Weinberg equilibrium expectations were excluded. Ambiguous positions (A-T and C-G) were also removed. The resulting dataset comprised 199 Indigenous American individuals from 31 language families and 53 ethnic groups. It includes 5,308,880 biallelic SNPs and 3,710 individuals from 201 populations worldwide.

After the initial quality assessment, linkage disequilibrium pruning was performed with 'SNPRelate' v.1.28.0 R package[71] to exclude markers exhibiting a pairwise correlation greater than 20% ($r^2 > 0.2$) in a 50-kb sliding window, advancing in 10-kb steps. This procedure yielded an linkage-disequilibrium-pruned dataset for downstream analyses that required an independent set of markers (for example, PCA and ADMIXTURE).

PCA

PCA was performed using the 'SNPRelate' v.1.28.0 R package[71] on both the complete dataset and a subset comprising only local ancestry-masked Indigenous American populations to assess potential biases introduced during data merging and quality control, as well as to explore broad patterns of ancestry and genetic differentiation. In the case of the second analysis, positions with more than 10% missing data were eliminated, as well as those with a minor allele frequency below 5%.

Global ancestry inference

We analysed the Indigenous American genomes using the supervised mode of ADMIXTURE[72] v.1.3.0 and three putative ancestry components ($K = 3$) using a reference panel of diverse African (Bantu from Kenya, Bantu from South Africa, Biaka, Dinka, Khomani-San, Mandenka, Mbuti, San and Yoruba), European (Basque, Bergamo Italian, French, Orcadian, Sardinian and Tuscan) and Indigenous American (Karitiana, Surui, Colombian and Pima) populations without evidence of recent admixture with other continental groups, running 10 independent iterations with 100 bootstrap replicates per run. The consensus results of independent runs were obtained using CLUMPP[73] v.1.1.2.

We also investigated the genetic structure of present-day Indigenous Americans and their relationship with ancient Indigenous individuals through unsupervised ADMIXTURE analyses, considering 2 to 12 putative ancestry components ($K = 2$ to $K = 12$) and visualized the results using PONG[74] v.1.5. In the first analysis, we incorporated a reference panel of African, European and East Asian populations to model the non-Indigenous American ancestry. In the second analysis, we masked non-Indigenous American ancestry (following the approach detailed below) in present-day Indigenous American individuals and then combined them with ancient ones.

Relatedness analysis and sample selection

Subsequently, we performed kinship analysis to identify and remove closely related individuals from the dataset, minimizing the bias introduced by close relatives in downstream analyses. Using PLINK v.1.9 (ref. 75), we estimated the IBD between all pairs of individuals, calculated as $\mathrm{PI\_HAT} = P(\mathrm{IBD} = 2) + 0.5 \times P(\mathrm{IBD} = 1)$. On the basis of these estimates, we identified the largest set of unrelated individuals by applying a first-degree kinship cut-off. Filtering was conducted using PRIMUS[76] v.1.9.0.

Haplotypic phase, local ancestry inference and masking

The haplotypic phase of the genomic data was statistically inferred using ShapeIT4 (ref. 77) v.4.2.2, with the 1KGP dataset as the reference panel. The parameters were adjusted for sequencing data using the '--sequencing' option, with the following settings: 15 burn-in iterations, 15 pruning iterations and 100 main iterations. Local ancestry inference was conducted using RFMix[78] v.1.5.4, applying a window size of 0.2 cM and a minimum of five reference haplotypes per tree node, using a reference panel of unadmixed (that is, with no evidence of recent admixture) Indigenous Americans, Sub-Saharan Africans and Western Europeans. We then used the inferred local ancestry tracts to mask genomic sites (code to perform local ancestry masking is available at https://github.com/macscastro/lamask), assigning segments

with a posterior probability of being ancestrally Indigenous Americans below 99% as missing data (Supplementary Note 3).

Two further datasets were generated: (1) the first dataset was generated by combining the local ancestry-masked dataset with ten Indigenous Andamanese, of which six were Onge and four were Jarawa[79]; and (2) the second dataset was created by combining the local ancestry-masked dataset with the Allen Ancient DNA Resource[39] and ancient DNA data from sambaqui mound builders found in Brazil[15].

Effective migration surface modelling

EEMS modelling[80] (https://github.com/dipetkov/eems) was applied to subsets of 116 unadmixed and 160 ancestry-masked Indigenous Americans. The model used 1,200 demes and ran for 4×10^6 iterations with a 2×10^5 burn-in period (Supplementary Note 4). Migration and diversity rates were visualized using scripts from EEMS developers.

Patterns of allele sharing

Using the 'admixtools' v.2.0.10 R package, we calculated population pairwise f_3(Mbuti; X, Y) to investigate genetic similarity patterns among contemporary Indigenous American populations. These patterns were visualized using a neighbour-joining tree and multidimensional scaling. The same method was used to assess spatial and temporal genetic variation, although genetic similarity was estimated for all pairs of contemporary and ancient X and Y groups. Clusters with genetic similarities were identified as clades in a neighbour-joining tree. The 1 − outgroup f_3 distances were also used to infer the existence of a correlation with geographic distances, estimated as great circle distances with the R package 'geosphere' v.1.5.18.

Admixture graph and ancestry modelling

We used the 'find_graphs' function from the 'admixtools' v.2.0.10 R package to identify plausible population history models for contemporary Indigenous American populations (Supplementary Note 6). The algorithm ran ten times with 200 iterations for each inferred number of admixture events. The best-fitting admixture graph for each scenario (ranging from zero to five admixture events) was selected on the basis of its highest score. These graphs were constructed using f_2 statistics restricted to transversions, including one representative of each genetic cluster identified in the previous step.

To summarize our findings alongside existing evidence from the literature, we manually constructed admixture graphs and tested their fit to the data using the 'qpgraph' function from 'admixtools' v.2.0.10 R package. Confidence intervals for drift lengths and admixture weights were computed with the 'qpgraph_resample_snps' function, which fits the graph several times using random SNP subsets. The results were summarized using the 'summarize_fits' function to generate a data frame with parameter estimates. As in previous analyses, these models were inferred using only transversions, thus avoiding systematic errors in ancient DNA data caused by postmortem damage, which induces C-to-T transitions at methylated CpG sites.

We used the qpWave method to estimate the minimum number of independent ancestral sources contributing to third-dispersal populations and the rotating qpAdm approach to model the sources and their proportional ancestry contributions to present-day Indigenous South Americans (Supplementary Note 6), both from the 'admixtools' v.2.0.10 R package. Analyses were restricted to transversions with a maximum per-site missing rate of 10%. Two complementary analyses were performed: (1) including representatives from all genetic clusters and (2) focusing on individuals from the first and second dispersals, alongside ancient genomes from the North American Pacific Coast (that is, excluding present-day Indigenous Americans and Ceramic-period Caribbeans). We identified feasible models in which source contributions summed to 100% and plotted their probabilities and admixture proportions (Extended Data Fig. 11). Full model statistics are reported in Supplementary Table 10.

Effective population size history and IBD sharing

We used IBDNe[81] v.07May18.6a4 to infer the effective population size (N_e) histories of present-day Indigenous Americans. IBD segments were identified using the Refined IBD[82] v.12Jul18.a0b by merging those with short gaps (maximum gap = 0.6, maximum discordant homozygotes = 1). N_e was inferred by pooling individuals by language families or genetic clusters (minimum ten individuals; populations from Mesoamerica and Aridoamerica were pooled together), as previous studies indicate that historical N_e trajectories remain robust with smaller sample sizes (Supplementary Note 5). Segments greater than 2 cM were analysed using the default IBDNe parameters.

IBD sharing patterns were assessed by categorizing segments by length, which reflected the time since a shared ancestor, to examine shared IBD within and between populations over time. The IBD networks were estimated using the 'as_tbl_graph' function (directed = FALSE) and visualized with the 'ggraph' function (layout = 'fr') from the 'tidygraph' v.1.3.1 and 'ggraph' v.2.2.1 R packages, respectively.

Effective population size, coalescence rates and divergence times

We applied the coalescent-based method Relate[83] v.1.2.2 to the phased WGS data to infer historical changes in N_e and estimate divergence times between contemporary populations. Input files were converted from variant call format to the haps/sample format using the RelateFileFormats script, and haplotypes were flipped according to the ancestral genome using the PrepareInputFiles script, which also filtered SNPs and adjusted the distances on the basis of the genomic mappability mask. Input preparation was performed using the GRCh38 ancestral genome (human_ancestor_GRCh38), the GRCh38 genome mask (20160622_genome_mask_GRCh38) and GRCh38 recombination maps provided by the developer.

Ancestral recombination graphs were inferred using the parameters $-m = 1.25 \times 10^{-8}$ (mutation rate) and $-N = 30{,}000$ (effective population size), and the coalescence rate trajectories for each population were estimated using the EstimatePopulationSize script. N_e estimates were obtained by calculating the inverted coalescence rate in the form of 0.5/(coalescence rate), and putative divergence times between groups were identified as the time points at which the inverted coalescence rate values in each population started to diverge from one another, whereas the inverted cross-coalescence rate values between populations increased. This indicates a decrease in cross-coalescence rates between populations, indicating genetic separation and increasing divergence over time. We also inferred the N_e history by inferring the ancestral recombination graph, while considering all individuals as a single population.

ROHs

ROHs were inferred for unrelated and unadmixed Indigenous Americans using PLINK v.1.9 with a sliding window of 50 SNPs, allowing for up to one heterozygous site and five missing calls, a minimum density of one SNP per 50 kb, a maximum gap of 100 kb and a minimum ROH length of 500 kb. We compared the total and average ROH lengths per individual across global populations and Indigenous American clusters of genetic similarity. For global populations, we visualized individual total ROH counts and lengths, whereas for Indigenous Americans, we additionally plotted population-wise averages. We estimated the inbreeding coefficient (F_{ROH} – ROH-base inbreeding coefficient) and average inbreeding per population and tested the correlations between ROH counts and ancestry proportions. Additionally, ROH hotspots, defined as regions with an above-average ROH occurrence (more than three standard deviations), were identified (Supplementary Note 5).

Excess affinity with Australasian populations

We analysed the excess genetic affinity between Indigenous American and Australasian populations by computing D(Mbuti, Australasian; X, Y), where X and Y are Indigenous American groups, and Australasians are represented by Australian, Bougainville, Jarawa, Onge and Papuan ('PapuanHighlands' and 'PapuanSepik') populations. This included comparisons between present-day and ancient Indigenous American populations to trace the Ypykuéra ancestry across space and time. To minimize bias from missing data, we used a subset of unrelated and unadmixed Indigenous American individuals to ensure more reliable results when integrating present-day and ancient genomes, the latter generally having high missingness levels. We also investigated the signatures of natural selection in genomic regions with excess genetic affinity for Australasian populations (Ypykuéra ancestry) and their potential functional effects (Supplementary Note 7).

Selection scans

Natural selection analysis was performed on a subset of unrelated individuals, masking segments of non-Indigenous American ancestry. To detect positive selection signals, we used two approaches based on population differentiation (d_i statistic[84] and PBS[85]) and extended haplotype homozygosity (iHS[86] and xpEHH[87]). For all four statistics, we conducted sliding-window analysis using 200 SNPs per window with a step size of 50 SNPs. We then combined the genome-wide ranks of the four statistics for each window using Fisher's combined score[88]. This score is calculated as the sum, over the four statistics, of −ln(rank of the statistic/number of windows tested). Outlier regions were defined as windows with Fisher's combined score scores in the 99.9th percentile (Supplementary Note 8).

Archaic introgression inference

To identify genomic regions exhibiting signals of archaic introgression in ancestry-masked Indigenous Americans, we used segments detected by the SPrime method[89], using Indigenous Americans as targets and African Mbuti from the HGDP as unadmixed outgroups. To enhance robustness, we applied filtering steps following ref. 64, retaining only (1) high-confidence archaic sites and (2) those found at low frequency in Africa (less than 0.01) but present at greater than or equal to 0.01 in at least one non-African population (Supplementary Note 9).

For sites passing through these filters, we classified a match when the Archaic genotype contained the putative Archaic-specific allele (present in both Neanderthals and Denisovans). Additionally, we identified Neanderthal-specific (matching Neanderthals but differing from Denisovans) and Denisovan-specific (matching Denisovans but differing from Neanderthals) sites.

ORA

ORA was performed using the WEB-based GEne SeT AnaLysis Toolkit (WEB-GESTALT)[90], focusing on phenotypes and Gene Ontology categories, including Biological Processes, Cellular Components and Molecular Functions. To address redundancy, we applied a weighted set cover approach, which identified the minimum subset of gene sets that covered all genes from the enriched sets. The weight or cost of adding a gene set was based on P.

Reporting summary

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

Data availability

The corresponding author (T.H.) can be contacted at tabita.hunemeier@csic.es or hunemeier@usp.br. All data needed to evaluate the conclusions of this study are presented in the paper and/or Supplementary Information. The datasets generated in this study have been deposited in the European Genome-Phenome Archive (EGA) and are available under accession number EGAD50000002396. Access to individual-level genomic data is controlled and granted through the EGA Data Access Committee for research in human evolutionary genetics and related fields, in accordance with the ethical approvals, informed consent and community agreements governing the use of Indigenous genomic data. Requests for access will be evaluated by the Data Access Committee, and applicants can expect an initial response in about 2–4 weeks. Allele frequencies are freely accessible through the Indigenous American Genomic Diversity Project Variant Browser (IAGDP Browser).

Code availability

Newly generated code essential for reproducing the results has been deposited on GitHub (https://github.com/macscastro/lamask). All remaining data processing and analyses were performed using previously published software, as described and referenced in the Methods.

69. Hanssen, F. et al. Scalable and efficient DNA sequencing analysis on different compute infrastructures aiding variant discovery. *NAR Genom. Bioinform.* **6**, lqae031 (2024).
70. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
71. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
72. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
73. Jakobsson, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 (2007).
74. Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P. & Ramachandran, S. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* **32**, 2817–2823 (2016).
75. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
76. Staples, J., Nickerson, D. A. & Below, J. E. Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genet. Epidemiol.* **37**, 136–141 (2013).
77. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
78. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
79. Mondal, M. et al. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nat. Genet.* **48**, 1066–1070 (2016).
80. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).
81. Browning, S. R. et al. Ancestry-specific recent effective population size in the Americas. *PLoS Genet.* **14**, e1007385 (2018).
82. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
83. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
84. Akey, J. M. et al. Tracking footprints of artificial selection in the dog genome. *Proc. Natl Acad. Sci. USA* **107**, 1160–1165 (2010).
85. Yi, X. et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
86. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
87. Sabeti, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
88. Deschamps, M. et al. Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *Am. J. Hum. Genet.* **98**, 5–21 (2016).
89. Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell* **173**, 53–61 (2018).
90. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019).

Acknowledgements We thank all individuals and Indigenous communities who participated in this study, as well as V. Acuña Alonzo and G. Macín for their valuable assistance in working with Indigenous populations from Mexico. We also thank Google Cloud Brazil for providing support with cloud computing. Funding was provided by grant nos. FAPESP 21/06860-8 and 15/26875-9 (T.H.), grant no. CSIC 20223AT005 (T.H.), European Union's Horizon 2020 Research and Innovation Programme/Marie Skłodowska-Curie COFUND/EUTOPIA grant no. 945380 (M.A.C.S.), grant no. JDC2022-049175-I funded by MICIU/AEI/10.13039/501100011033 and 'European Union NextGenerationEU/PRTR' (M.A.C.S.), CAPES grant no. 88887.505029/2020 (K.N.), grant no. CNPq INCT 406864/2022-5 (F.R.S.), URPP 'Evolution in Action' of the University of Zurich (C.B.), the NCCR Evolving Language Swiss National Science Foundation Agreement grant no. 51NF40_180888 (C.B.), the National Institute of General Medical Sciences of the

National Institutes of Health under grant no. R35GM142939 (C.E.G.A.) and the Max Planck Society (D.I.H.-Z., A.W., T.N.C. and R.B.).

Author contributions Conceptualization: T.H. Methodology: M.A.C.S., K.N., M.R.R., H.P.-A., R.B.L and L.K. Investigation: M.A.C.S., K.N., M.R.R., D.C. and T.H. Data curation: P.S., C.E.G.A., M.C.B., J.G.M., J.F.G., C.B., D.I.H.-Z., A.W., T.N.C., D.H.-M., J.C.L.R., O.D.C.-C., C.Z., A.M.T.-A., E.V.-D., M.E.M.-M., J.C.T.-R., A.A.-C., R.S.-O., M.L.P., C.M.B., V.R., G.B., S.R., J.R.S., R.F. and F.R.S. Funding acquisition: T.H. Project administration: T.H. Supervision: T.H. Writing—original draft: M.A.C.S., K.N. and T.H. Writing—review and editing: T.H. All authors reviewed the final version of the manuscript.

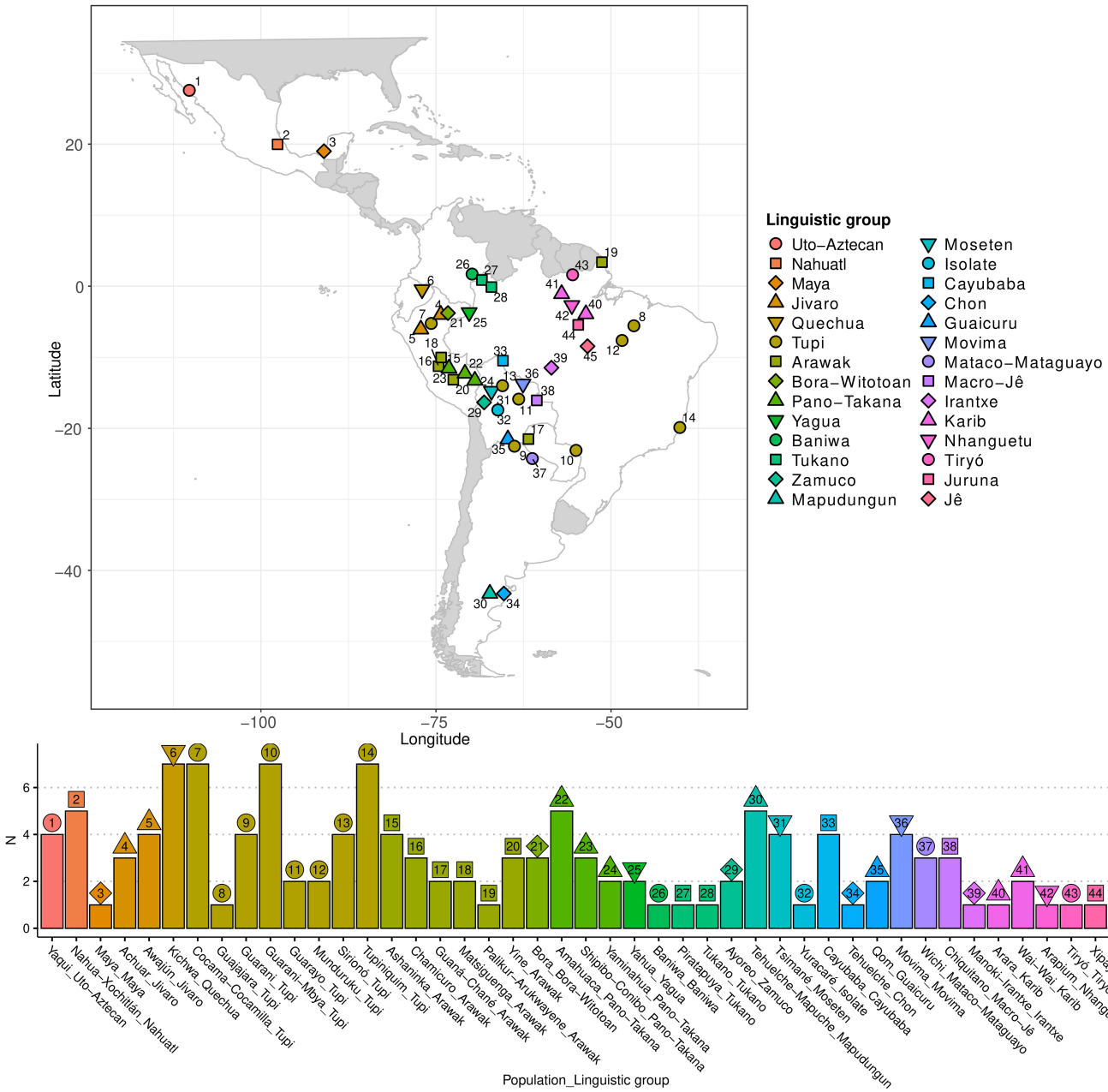
Competing interests The authors declare no competing interests.

Additional information

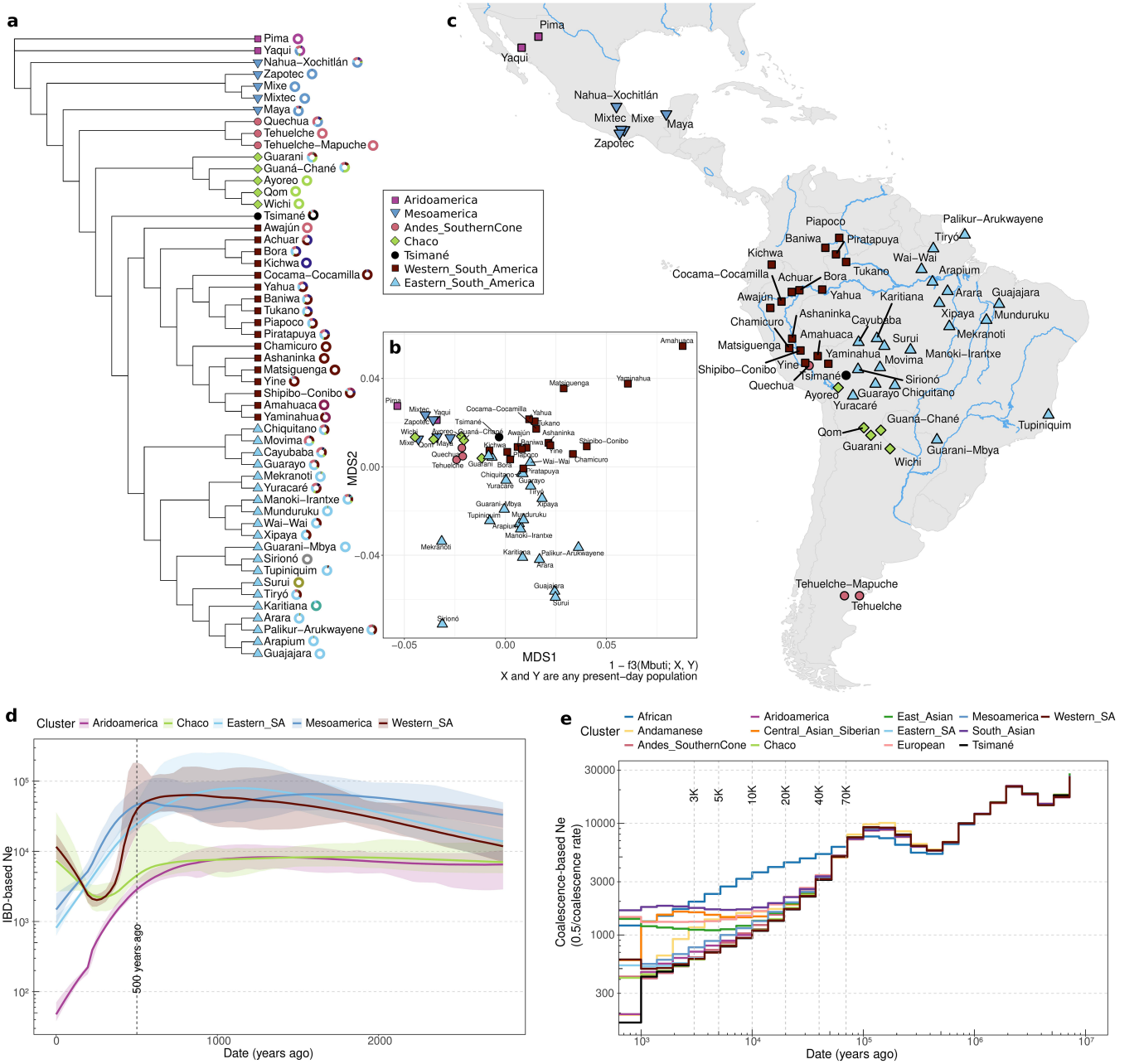
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10406-w.

Correspondence and requests for materials should be addressed to Tábita Hünemeier.

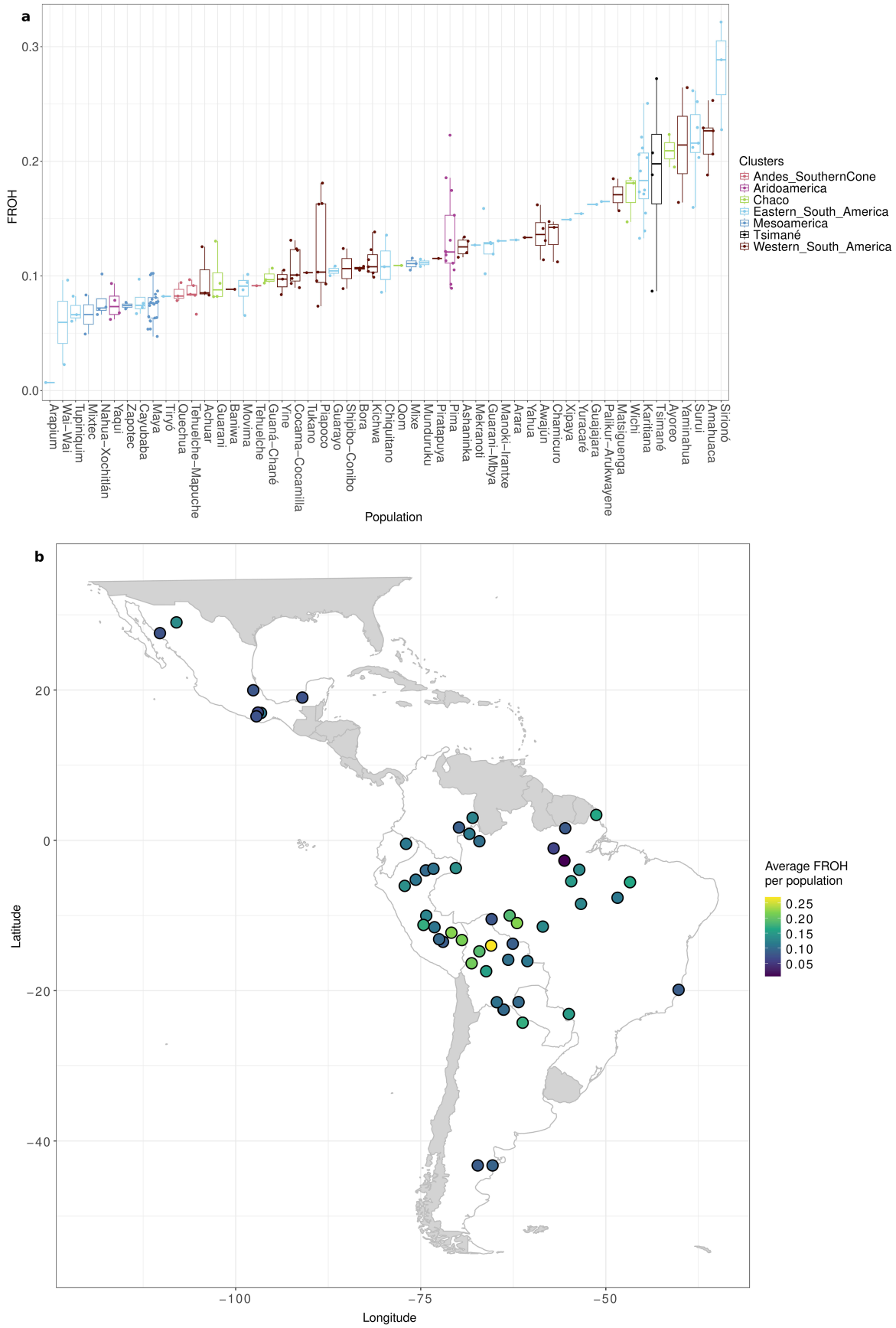
Peer review information *Nature* thanks María Ávila-Arcos, Ryan Hernandez and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



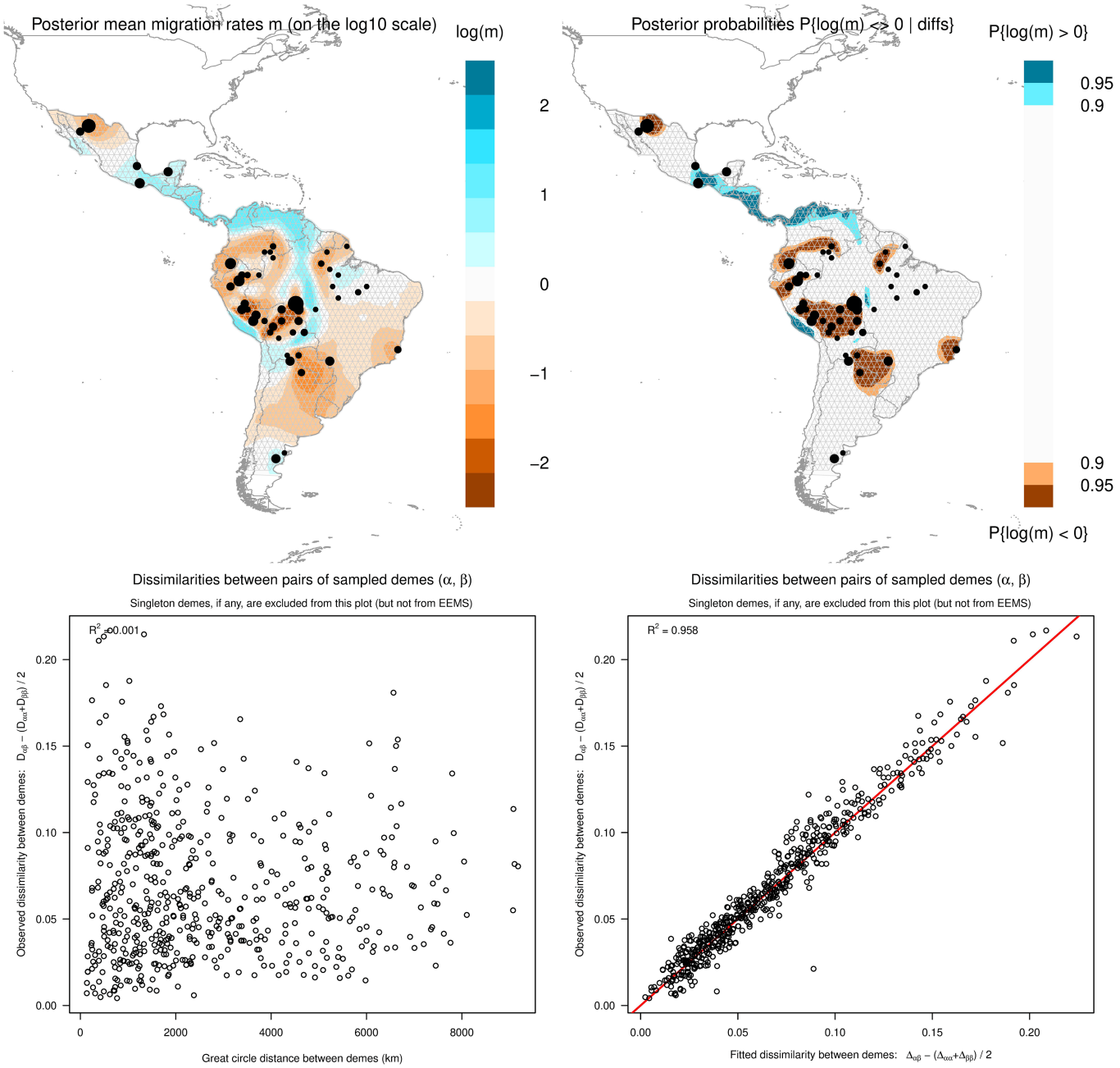
Extended Data Fig. 1 | Geographic distribution, linguistic affiliation, and sample sizes of newly-generated whole genome sequences. The approximate locations of newly sampled Indigenous American communities are shown on a partial map of the Americas, with language families represented by different symbol shapes and colors. A bar plot below lists ethnic groups and language families (x-axis) along with the number of sequenced individuals per community (y-axis), with numbers linking communities between the map and bar plot. Countries with at least one sampled Indigenous community are highlighted in white.



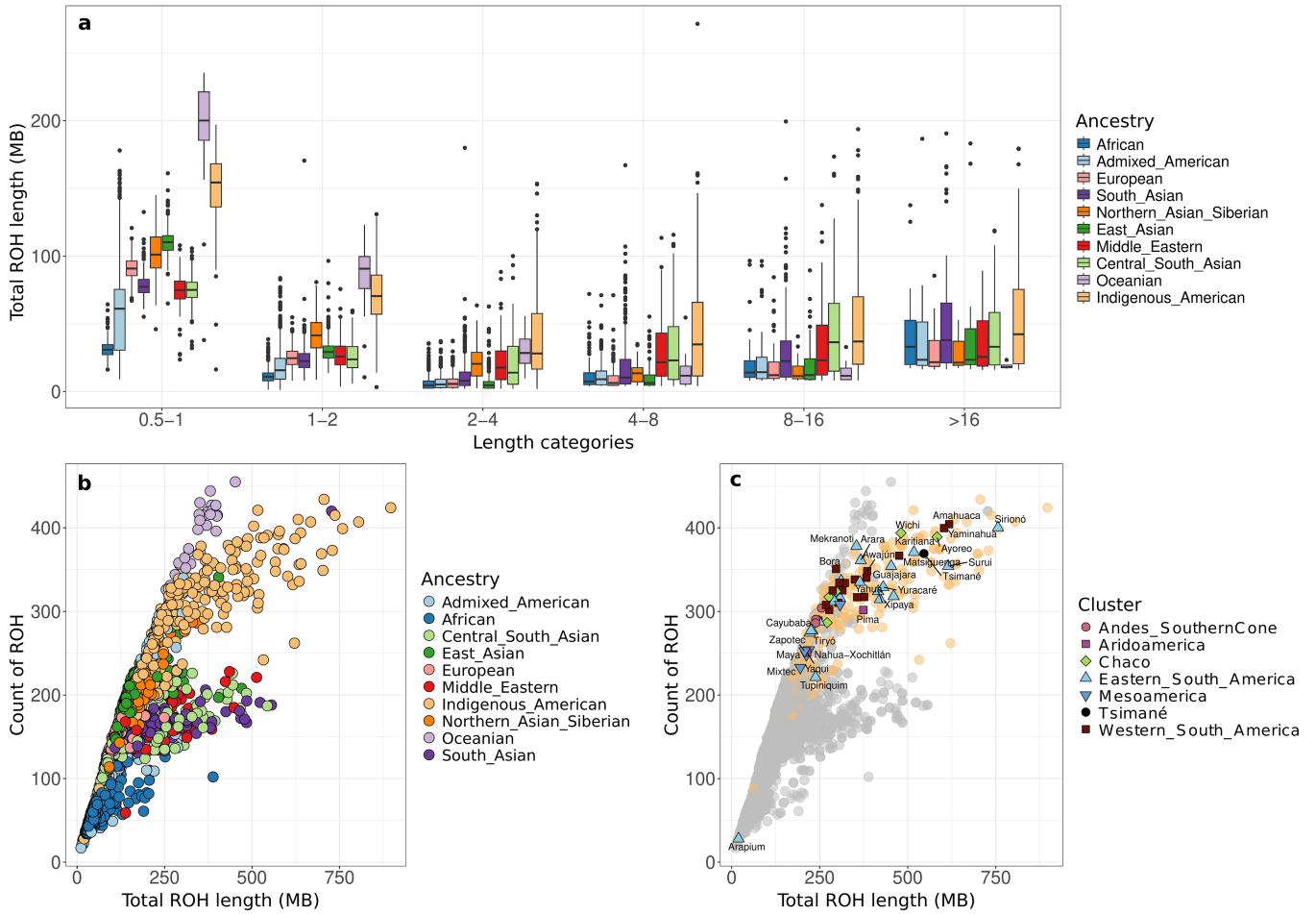
Extended Data Fig. 2 | Present-day patterns of genetic similarity and demographic history. (**a**) Neighbour-Joining (NJ) tree constructed from genetic distances calculated as $1 - f_3$(Mbuti; X, Y), where X and Y represent pairs of contemporary Indigenous American populations, masked for non-Indigenous American ancestry. Genetic clusters are color-coded and branch lengths are omitted to emphasize topology. Population-averaged ancestry proportions (from Fig. 2d) are shown to the right of each tree tip. (**b**) Multidimensional scaling (MDS) plot based on the same genetic distance matrix, illustrating genetic similarities in reduced dimensional space. (**c**) Geographic distribution of populations, color-coded by genetic cluster. Effective population size (N_e) trajectories for contemporary Indigenous Americans were inferred (**d**) using IBDNe, based on identity-by-descent (IBD) segments exceeding 2 cM (y-axis $\log_{10}$-transformed), and (**e**) using Relate, based on coalescence rates (axes $\log_{10}$-transformed), with individuals pooled and color-coded according to the genetic clusters defined in (**a**), assuming a generation time of 28 years.



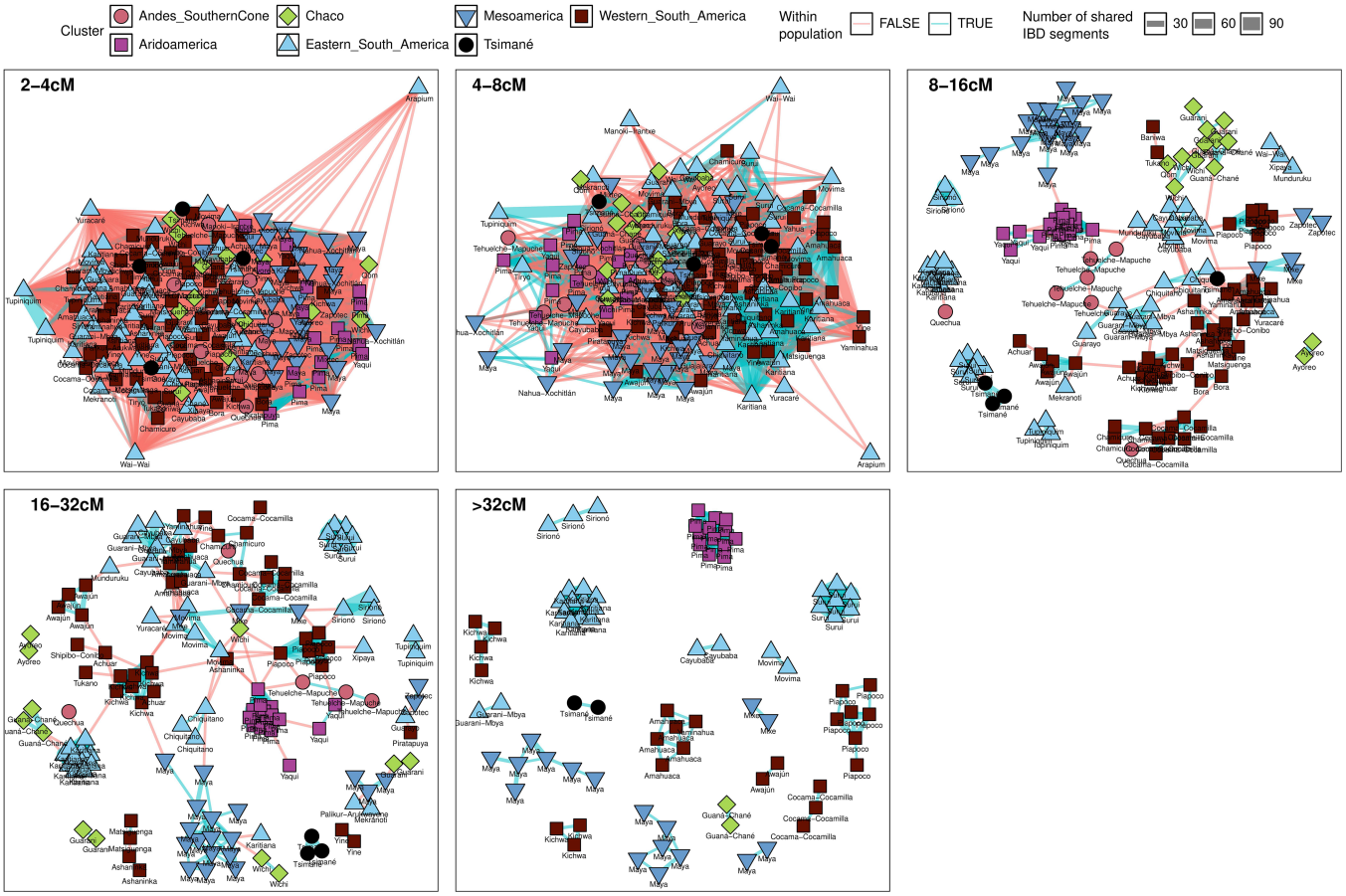
Extended Data Fig. 3 | Indigenous American individual and population-wise F_{ROH} distribution. (**a**) Boxplot showing individual-wise ROH-based inbreeding coefficients (F_{ROH}) for each contemporary Indigenous American population, color-coded by genetic cluster (see Extended Data Fig. 2). (**b**) A partial map of the Americas displaying the approximate geographic location of each population, color-scaled by their average F_{ROH}.



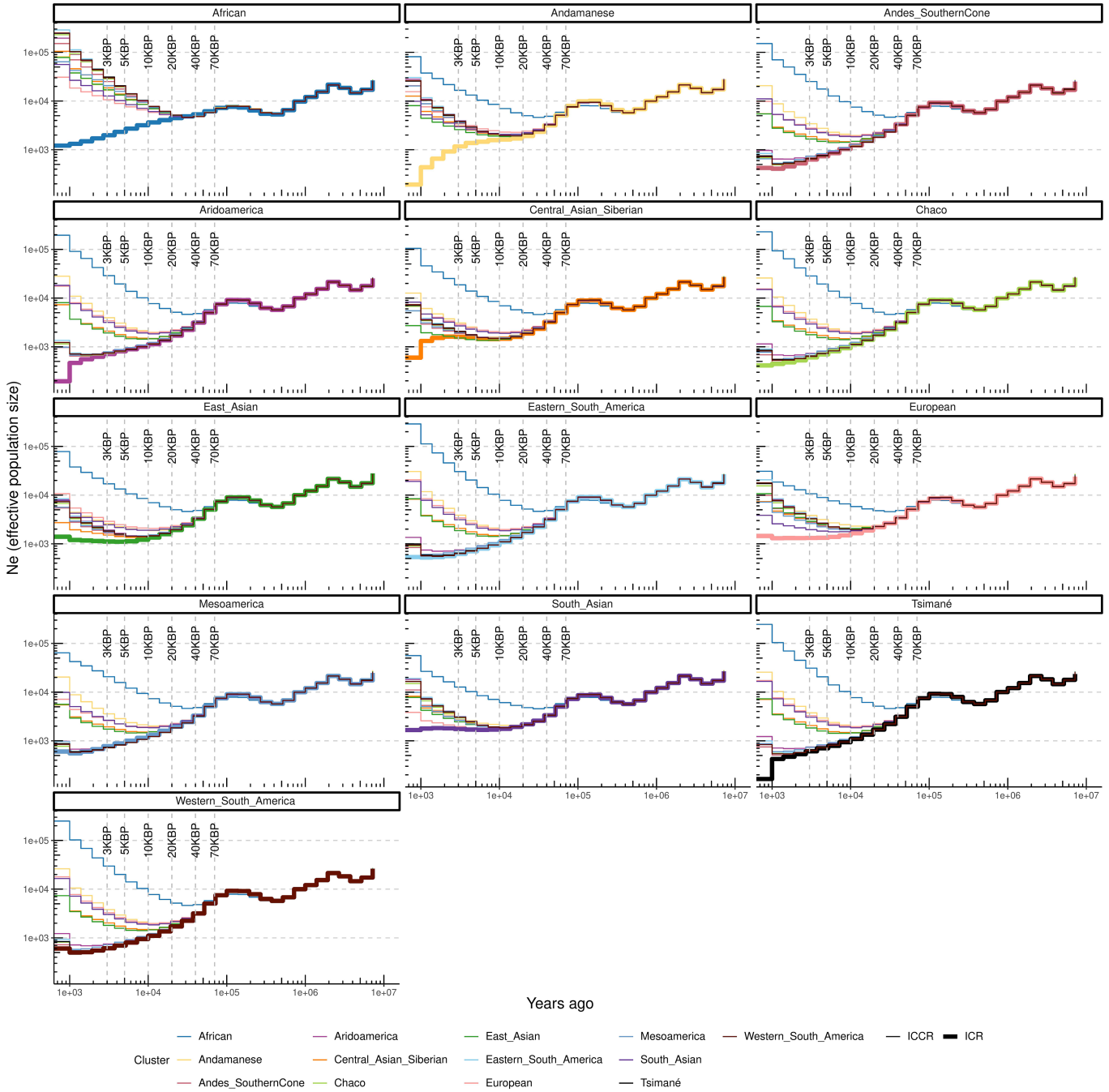
Extended Data Fig. 4 | Effective migration rates in Indigenous American populations. Effective migration surface modelling was applied to the ancestry-masked Indigenous Americans. Blue shades indicate increased migration rates, while orange shades indicate migration barriers. The model used 1,200 demes and ran for 4×10^6 iterations with a 2×10^5 burn-in period. Migration rates and model fit were visualized using scripts provided by the developers. The estimated effective migration rates are shown (top left), and posterior probabilities >0.9 (top right) highlight regions with rates significantly higher (blue) or lower (orange) than expected under an isolation-by-distance model. The bottom panels show the correlation between observed genetic dissimilarity and geographic distance (bottom left), and between observed and fitted dissimilarities under the EEMS model (bottom right).



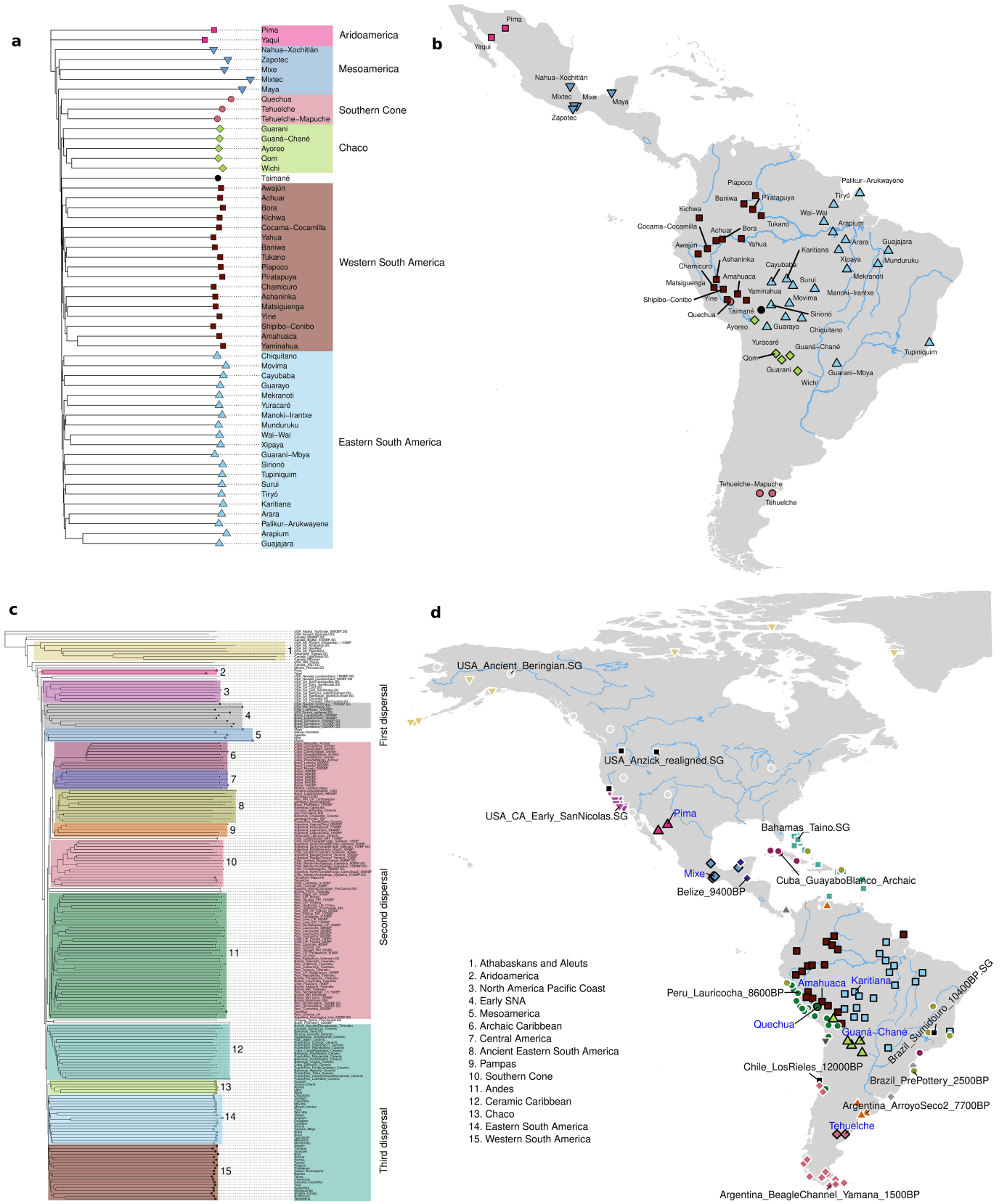
Extended Data Fig. 5 | Worldwide individual and population-wise ROH distribution. (**a**) Boxplot of individual-wise total ROH length (y-axis) across different segment length categories (x-axis), with individuals from different continental regions separated and color-coded as shown in the legend. (**b**) Scatterplot of individual-wise total ROH length (x-axis) versus count (y-axis), with points color-coded by their continental region of origin. (**c**) Population average total ROH length (x-axis) and count (y-axis), overlaid on individual-wise values shown in panel B, with points color-coded according to genetic cluster (Extended Data Fig. 2).



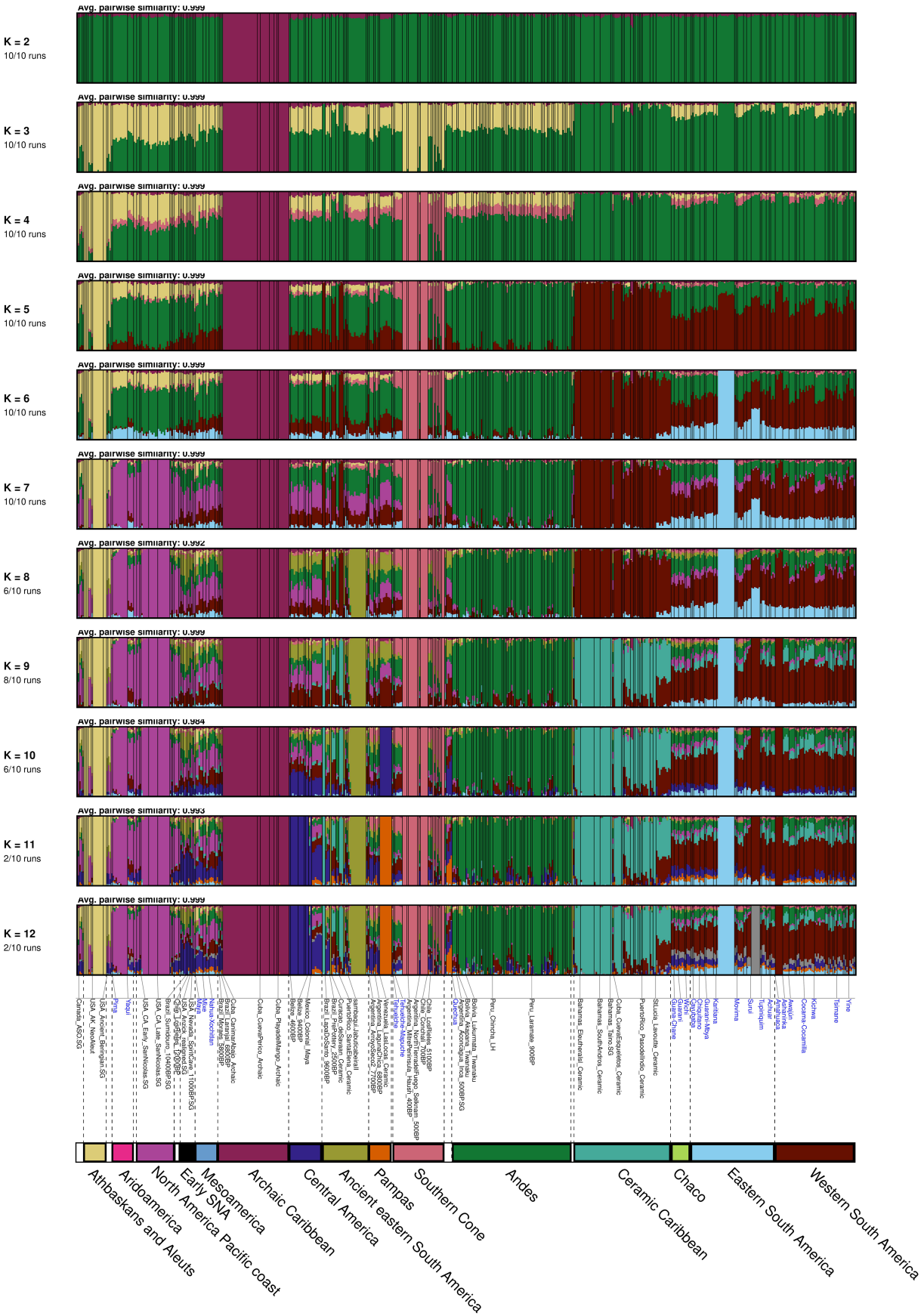
Extended Data Fig. 6 | Patterns of IBD sharing between individuals. Identity by descent (IBD) segments were identified between all pairs of contemporary Indigenous Americans. These segments were categorized by length and independently analysed in network analysis. IBD segments shared among individuals from the same population are color-coded in blue, while those shared between individuals from different populations are color-coded in red. The number of segments shared between pairs of individuals is represented by the thickness of the connecting lines. Individuals are color-coded according to their genetic cluster (Extended Data Fig. 2).



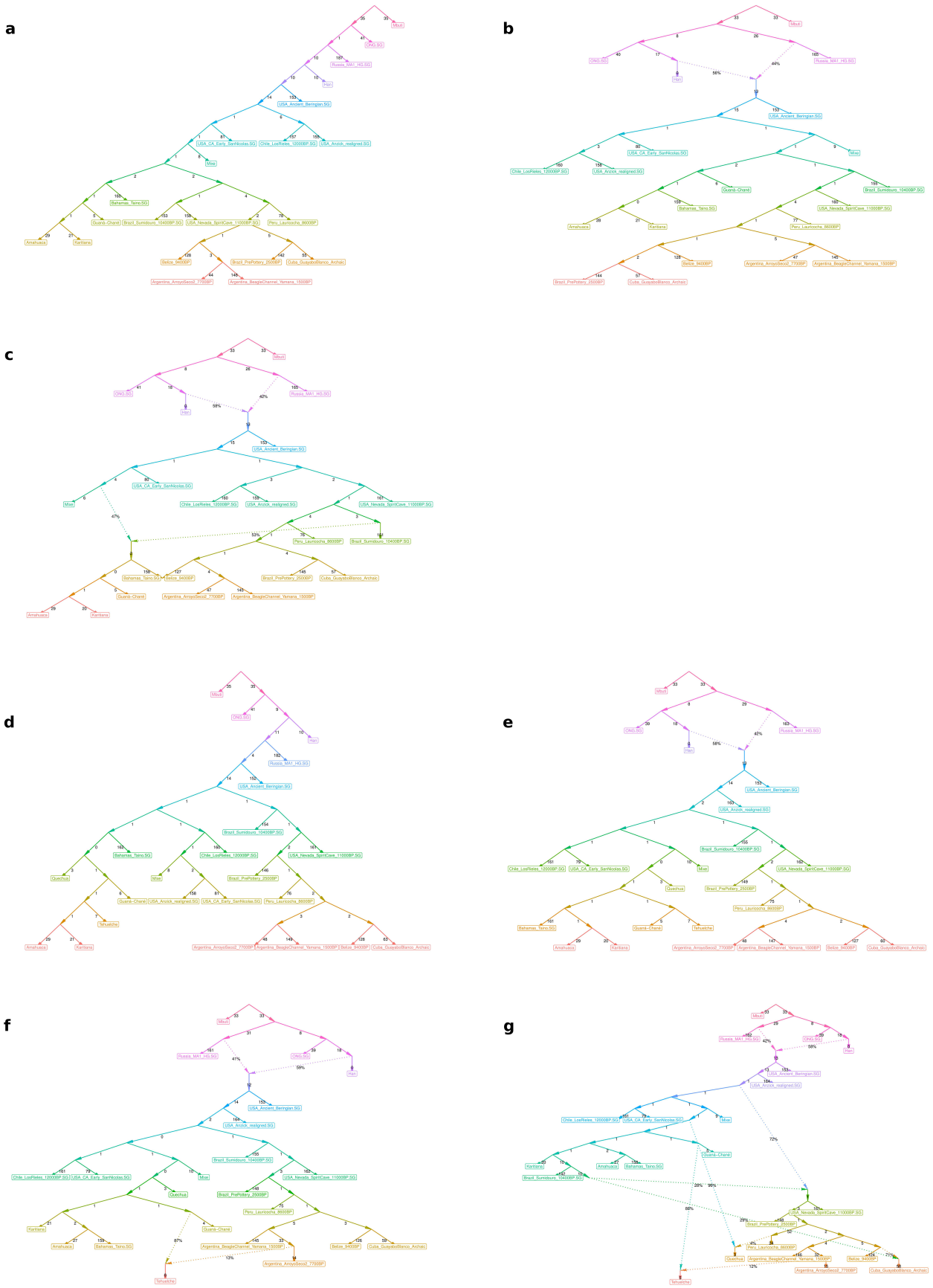
Extended Data Fig. 7 | Coalescence and cross-coalescence rates in contemporary Indigenous Americans and worldwide populations. Inverse coalescence rates (ICR estimated within groups) and inverse cross-coalescence rates (ICCR; estimated between groups) were inferred using Relate (0.5/coalescence rate in both cases) to estimate divergence times, grouping individuals by genetic clusters (Extended Data Fig. 2). Each panel displays the ICR and ICCR for a given group, with ICCR represented by color-coded lines corresponding to comparisons with other groups. The ICR for each group is depicted by a thicker line color-coded with the same group as the panel's name. Both axes are $\log_{10}$-transformed, and key periods marking the onset of group divergence are highlighted. Other continental regions are represented by the following populations: YRI (African); IBS (European); CHB (East Asian); GIH (South Asian); Yakut (Central Asian/Siberian); Jarawa and Onge (Andamanese).



Extended Data Fig. 8 | Continental-wide long-term patterns of genetic similarity. Neighbour-joining (NJ) trees based on genetic distances calculated as $1-f_3$(Mbuti; X, Y), where X and Y are (**a**) pairs of present-day Indigenous American groups or (**c**) combinations of present-day and ancient groups. color-coded clades define genetic clusters, with branch lengths representing genetic distances. (**b**,**d**) Geographic distribution of the genetic clusters shown in the corresponding NJ trees (**a**,**c**). In (**b**), all populations are labelled; in (**d**), selected present-day (blue text) and ancient (black text) groups are labelled. In (**c**) and (**d**), present-day individuals are outlined in black and ancient individuals in white. In (**c**), numbers to the right of tree branches indicate genetic clusters as listed in the figure, and shades below group names denote the associated dispersal events.

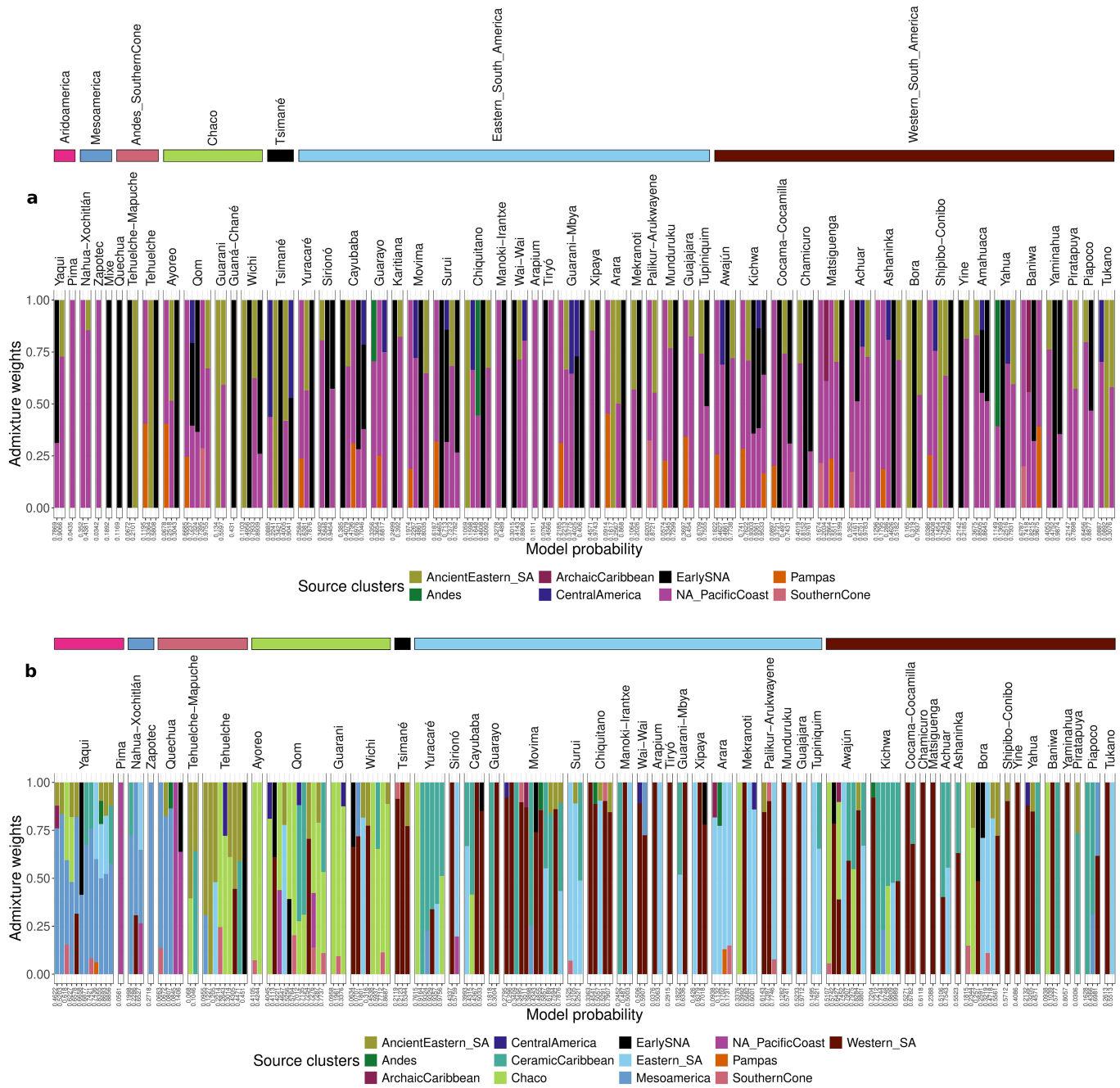


Extended Data Fig. 9 | Global ancestry inference of contemporary and ancient Indigenous Americans. Ancestry proportions inferred from unsupervised ADMIXTURE analysis for K = 2 to 12. Individual estimates are shown for each K, with individuals ordered according to the NJ tree clustering (Fig. 3; Extended Data Fig. 8). Genetic clusters are indicated in the bottom bar, with selected present-day and ancient individuals distinguished in blue and black, respectively.

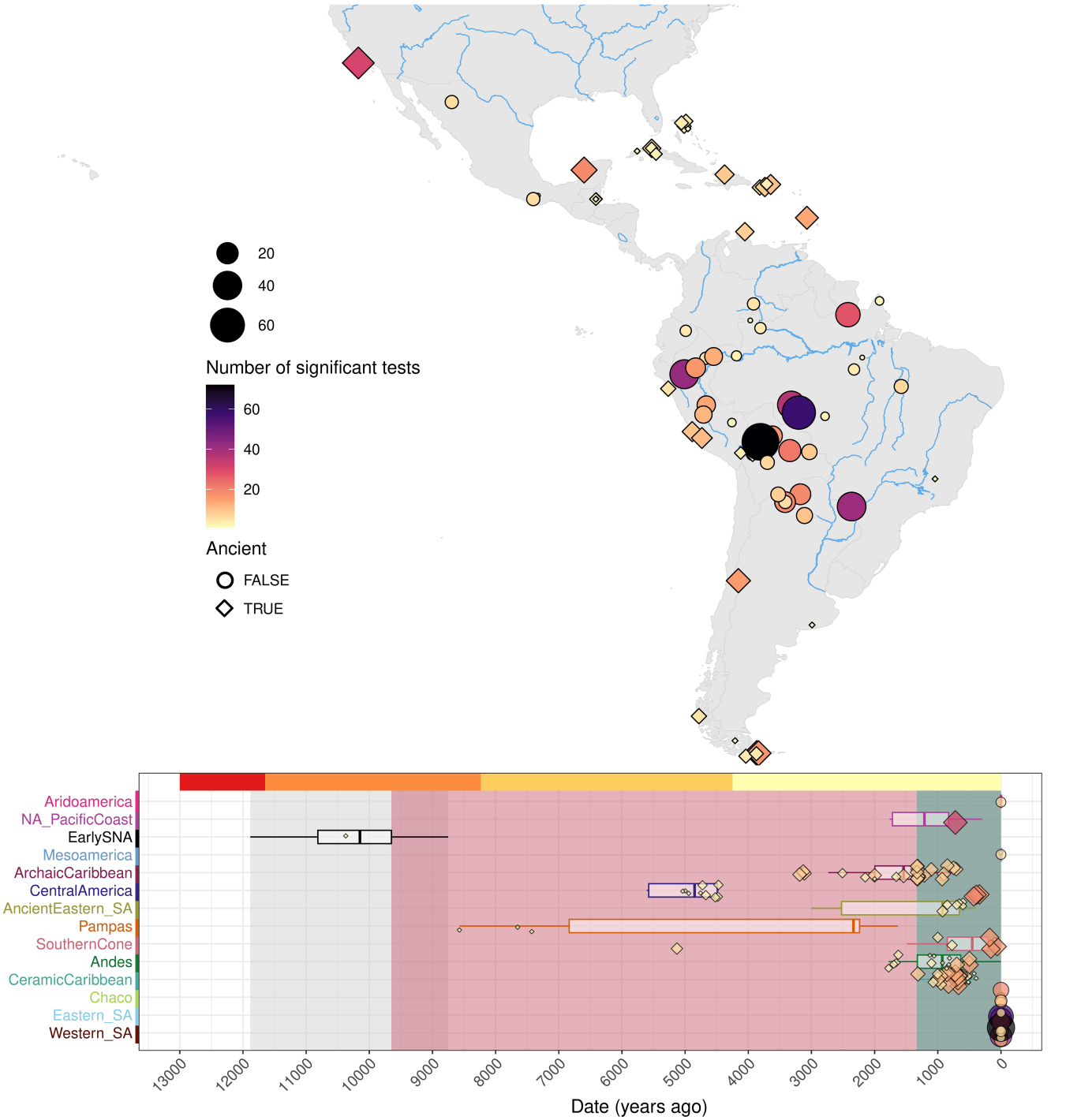


Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Admixture graph modelling of Indigenous American populations. Population history models were inferred using 'find_graphs' from 'admixtools' in R, with 10 replicates and 200 iterations per admixture event. The best-fitting graph for each scenario (0–5 admixture events) was selected based on the highest score using f_2 statistics with transversions. One representative per genetic cluster (Fig. 3) was included, alongside Late Pleistocene individuals and outgroups necessary to model the ancestry of Indigenous Americans. (**a**–**c**) Models were refined by progressively adding admixture events (examples of models with 0 to 2 admixture events are shown), (**d**–**g**) then expanded to include present-day populations with genetic continuity to ancient Southern Cone and Andean individuals (Tehuelche and Quechua, respectively; examples of models with 0, 1, 2 and 5 admixture events are shown).



Extended Data Fig. 11 | Ancestry modelling of contemporary Indigenous American populations. Ancestry sources and proportions were estimated using qpADM with a rotating strategy, based on transversions and a maximum per-site missing rate of 10%. (**a**) Models restricted to source populations representing the first and second dispersals. (**b**) Models incorporating source populations from all defined genetic clusters. For each target population, the top 50% of feasible models are shown, ordered from lowest to highest model probability (left to right). Bar plots indicate inferred ancestry proportions, and model probabilities are shown on the y-axis. Full model statistics are available in Supplementary Table 10.



Extended Data Fig. 12 | Spatiotemporal distribution of relative excess genetic affinity of Indigenous Americans to Australasians. Excess Australasian affinity in Indigenous Americans was assessed using D(Mbuti, Australasian; X, Y), where X and Y represent Indigenous American groups, and Australasian affinity (Ypykuéra ancestry) was modelled using Australian, Bougainville, Jarawa, Onge, and Papuan (PapuanHighlands and PapuanSepik) populations. To minimize biases from missing data, only groups with less than 50% missing data were analysed, and unadmixed individuals were used instead of the masked dataset. The number of significant tests ($Z > 3$) for each Y population is represented by point size and color on a partial map of the Americas and a timeline of sample dates (x-axis) by genetic cluster (y-axis). Ancient and present-day individuals are indicated by diamonds and circles, respectively. In the bottom panel, shaded areas denote the earliest and latest presence of individuals associated with the first (gray), second (red), and third (green) dispersals in South America and the Caribbean, while the top bar indicates the geological timescale: Late Pleistocene (red), Early (orange), Middle (gold), and Late Holocene (yellow).

nature portfolio

Corresponding author(s): Tábita Hünemeier

Last updated by author(s): Jun 4, 2025

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	No software was used.
Data analysis	Sarek v.3.5.0, GATK Best Practices, Ensembl Variant Effect Predictor (VEP) v.113, Scikit-Allel v1.3.13, 'SNPRelate' v1.28.0 R package, ADMIXTURE v1.3.0, CLUMPP v.1.1.2, PLINK 1.9, PRIMUS v1.9.0, ShapeIT4 v.4.2.2, RFMix v.1.5.4, 'admixtools' v.2.0.10 R package, IBDNe v.07May18.6a4, Relate v1.2.2, selscan v.2.0, SPrime method, WEB-based GEne SeT AnaLysis Toolkit (WebGestalt 2024). Newly generated code essential for reproducing the results has been deposited on GitHub (https://github.com/macscastro/lamask).

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 data needed to evaluate the conclusions of this study are presented in the paper and/or Supplementary Materials. The datasets used in this study have been

deposited in the European Genome-Phenome Archive and are available for download under accession number EGAXXXXXXX. They can also be freely accessed using the IAGDP Variant Browser.

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	No analysis in this study was conducted using biological or social information related to sex or gender.
Reporting on race, ethnicity, or other socially relevant groupings	The present study included individuals who self-identify and are concomitantly officially recognized by their countries of origin as Indigenous Americans, and whose mother tongue is a native language of the American continent.
Population characteristics	No demographic or population-related characteristics were used in the study.
Recruitment	Participants were recruited using a stratified linguistic approach to maximize Native American genetic diversity by sampling a small number of unrelated speakers per Indigenous language. This strategy prioritizes breadth of representation rather than proportional sampling. We acknowledge potential self-selection bias due to voluntary participation and limited sample sizes per group. However, as our primary analyses focus on inter-population structure and broad patterns of genetic diversity, these factors are unlikely to substantially affect the main conclusions.
Ethics oversight	Ethical approval for sample collection was provided by the respective local ethics committees in each country: Argentina (Puerto Madryn Zonal Hospital, Resolution No 009/2015; San Carlos de Bariloche Zonal Hospital, Resolution No 1510/2015), Brazil (CONEP, Resolution No. 123, 763, and 4599), Bolivia (Universidad Mayor de San Andrés), Ecuador (Universidad de Las Américas; Consejo Nacional de Ciencia y Tecnología—CONACyT, grant # 69856; Instituto Nacional de Ciencias Médicas y de la Nutrición Salvador Zubirán Ref.: 1518), Mexico (Consejo Nacional de Ciencia y Tecnología—CONACyT, grant # 69856; Instituto Nacional de Ciencias Médicas y de la Nutrición Salvador Zubirán Ref.: 1518, CNIC Salud 2013-01-201471; Committee of Ethics and Research, UADY, official notice no. F-FENC-SAC-14/REV: 04; registry no. 09/17) and Peru (Universidad San Martín de Porres). Individual and/or tribal informed oral consent was obtained from participants who were unable to read or write. The ethics committees also approved the oral consent procedure. Logistical support for sample collection in Brazil was provided by the Fundação Nacional do Índio. All sampling was conducted in accordance with the Declaration of Helsinki and the applicable laws and regulations of each country at the time of sampling.

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

Life sciences study design

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

Sample size	No formal a priori statistical power calculation was performed, as the primary objective was not to estimate within-population allele frequencies with high precision, but rather to maximize representation of Native American genetic diversity across ethnolinguistic groups. Sample sizes were therefore determined based on the inclusion of a small number of unrelated speakers per sampled language, prioritizing breadth of linguistic and geographic coverage. Given the current scarcity of available Indigenous genomes and the limited representation of many populations in public datasets, this strategy substantially expands existing diversity and is sufficient for analyses focused on inter-population structure and large-scale evolutionary patterns.
Data exclusions	No data exclusions.
Replication	Methods are published and freely available, and all data is available for replication purposes.
Randomization	Individuals were grouped at different levels according to the analyses, such as by language and geographic region.
Blinding	Blinding was not performed in this study. Sample collection was conducted with prior knowledge of the populations, as cultural and linguistic information was fundamental to the sampling strategy and study design. Blinding was not relevant to the primary analyses, which rely on genome-wide data processed through standardized bioinformatic pipelines, thereby minimizing the potential for investigator-driven bias.

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

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.