Article

Genomic insights into the formation of human populations in East Asia

https://doi.org/10.1038/s41586-021-03336-2

Received: 19 March 2020

Accepted: 5 February 2021

Published online: 22 February 2021

Check for updates

The deep population history of East Asia remains poorly understood owing to a lack of ancient DNA data and sparse sampling of present-day people[1,2]. Here we report genome-wide data from 166 East Asian individuals dating to between 6000 BC and AD 1000 and 46 present-day groups. Hunter-gatherers from Japan, the Amur River Basin, and people of Neolithic and Iron Age Taiwan and the Tibetan Plateau are linked by a deeply splitting lineage that probably reflects a coastal migration during the Late Pleistocene epoch. We also follow expansions during the subsequent Holocene epoch from four regions. First, hunter-gatherers from Mongolia and the Amur River Basin have ancestry shared by individuals who speak Mongolic and Tungusic languages, but do not carry ancestry characteristic of farmers from the West Liao River region (around 3000 BC), which contradicts theories that the expansion of these farmers spread the Mongolic and Tungusic proto-languages. Second, farmers from the Yellow River Basin (around 3000 BC) probably spread Sino-Tibetan languages, as their ancestry dispersed both to Tibet—where it forms approximately 84% of the gene pool in some groups—and to the Central Plain, where it has contributed around 59–84% to modern Han Chinese groups. Third, people from Taiwan from around 1300 BC to AD 800 derived approximately 75% of their ancestry from a lineage that is widespread in modern individuals who speak Austronesian, Tai–Kadai and Austroasiatic languages, and that we hypothesize derives from farmers of the Yangtze River Valley. Ancient people from Taiwan also derived about 25% of their ancestry from a northern lineage that is related to, but different from, farmers of the Yellow River Basin, which suggests an additional north-to-south expansion. Fourth, ancestry from Yamnaya Steppe pastoralists arrived in western Mongolia after around 3000 BC but was displaced by previously established lineages even while it persisted in western China, as would be expected if this ancestry was associated with the spread of proto-Tocharian Indo-European languages. Two later gene flows affected western Mongolia: migrants after around 2000 BC with Yamnaya and European farmer ancestry, and episodic influences of later groups with ancestry from Turan.

East Asia was one of the earliest centres of animal and plant domestication, and harbours an extraordinary diversity of language families including Sino-Tibetan, Tai–Kadai, Austronesian, Austroasiatic, Hmong–Mien, Indo-European, Mongolic, Turkic, Tungusic, Koreanic, Japonic, Yukaghiric and Chukotko–Kamchatkan[1]. Our current understanding of the human population history in the region remains poor because of the minimal sampling of genetic diversity of present-day people on the Tibetan Plateau and southern China[2], and a paucity of ancient DNA data compared with West Eurasia[3–6].

We collected DNA from 383 people from 46 populations from China (n = 337) and Nepal (n = 46) who provided informed consent for broad studies of population history; we carried out community consultation with minority group leaders as an integral part of the consent process (see Methods, 'Ethics statement'). We genotyped DNA using the Affymetrix Human Origins array at about 600,000 single-nucleotide polymorphisms (SNPs) (Extended Data Table 1 and Supplementary Information section 1).

For ancient individuals, we obtained permission for analysis from sample custodians, following protocols to minimize damage to skeletal material and including members of local minority groups as part of our study team when there was a plausible cultural connection between modern communities and ancient individuals (see Methods, 'Ethics statement'). We prepared powder from bones and teeth, extracted DNA, and prepared double- or single-stranded libraries for sequencing on Illumina instruments (Methods). For most samples, we enriched the DNA for a set of about 1.2 million SNPs[3,7]; for the Chinese samples, we used exome enrichment (Methods, Supplementary Information section 1 and Supplementary Table 1). We sequenced the DNA and processed the data using one of two nearly identical bioinformatics procedures (Methods and Supplementary Table 2), for which we found indistinguishable results from the perspective of analyses of population history (Supplementary Table 3). We considered samples to fail screening if they had fewer than 5,000 of the targeted SNPs covered at least once; if they had a too-low rate of

A list of authors and their affiliations appears online. ✉e-mail: wang@xmu.edu.cn; krause@shh.mpg.de; ron.pinhasi@univie.ac.at; reich@genetics.med.harvard.edu

cytosine-to-thymine substitution in the terminal nucleotide; or if they showed evidence of major contamination based on polymorphisms in mitochondrial DNA sequences[8] or the X chromosome in male individuals[9] or a ratio of Y-to-X chromosomes that would be unexpected for a male or female individual (Supplementary Tables 1, 2). We newly report data from 166 individuals (Fig. 1 and Supplementary Table 1): 82 individuals from Mongolia from between around 5700 BC and AD 1400, 11 individuals from the Chinese Mainland from a site dating to approximately 3000 BC in the Yellow River Basin, 7 individuals from Japan comprising Jomon hunter-gatherers dating to around 2500–800 BC, 18 individuals from the Russian Far East interred in the Boisman-2 cemetery dating to 5400–3600 BC as well as an individual dating to around 900 BC and another dating to around AD 1100, and 46 individuals from 2 sites in Taiwan dating to between 1300 BC and AD 800 (Supplementary Table 1). For analysis we focused on 130 individuals after excluding 16 individuals with evidence of low but non-zero contamination, 10 individuals in which 5,000–15,000 SNPs were covered and 11 individuals who were close relatives of another higher-coverage individual in the dataset (Extended Data Table 2). We merged our dataset with published data: 1,079 ancient individuals reported in 30 publications (Supplementary Table 4a) and 3,265 present-day individuals reported in 16 publications (Supplementary Table 4b). We grouped individuals by geography, time (aided by 108 newly reported direct dates; Supplementary Table 5), archaeological context and genetic cluster (Supplementary Table 1).

We carried out principal component analysis[10], projecting ancient individuals onto axes computed using present-day people. The population structure is correlated with geography ($R^2 = 0.261$; $P < 0.0001$) and language ($R^2 = 0.087$; $P < 0.0001$) (Supplementary Table 6), with some exceptions. Groups in northwest China, Nepal and Siberia deviate towards West Eurasian populations (Supplementary Information section 2), reflecting admixture that occurred, on average, between 5 and 70 generations ago[11] (Supplementary Tables 7, 8). Differentiation was much higher in East Asian individuals living in the early Holocene (fixation index (F_{ST}) = 0.067) compared to present-day populations ($F_{ST} = 0.013$) (Supplementary Table 9), reflecting mixture between deep East Asian lineages. Present-day East Asian individuals with minimal West-Eurasian-related ancestry grade between three poles. The 'Amur Basin cluster' correlates with ancient and present-day people in the Amur River Basin, and linguistically with speakers of Tungusic languages and the Nivkh. The 'Tibetan Plateau cluster' is most strongly represented in ancient people from Nepal and Indigenous Tibetan peoples. The 'Southeast Asian cluster' is maximized in ancient Taiwan and in East Asian individuals speaking Tai–Kadai, Austroasiatic and Austronesian languages (Extended Data Figs. 1–3). Automated clustering[12] provides similar results (Extended Data Fig. 4 and Supplementary Information section 2).

We organize our findings around themes. First, we considered deep time and determined the early branching lineages contributing to East Asian populations. Then, we shed light on how population structure came to be how it is today by testing three hypotheses about language expansions and their possible connection to farming spreads. Finally, we document how West and East Eurasian groups mixed along their geographical contact zone.

A Late Pleistocene coastal expansion

Only two pre-Ice Age genomes are available from East Asia: the approximately 40,000-year-old individual from Tianyuan Cave in northern China[13] and the around 35,000-year-old Salkhit individual from Mongolia[14]. Nevertheless, important insights can be gleaned from analysis of post-Ice Age genomes. One question concerns the extent to which the peopling of East Asia by modern humans occurred via a coastal or interior route. Suggestive genetic evidence for a coastal route comes from Y chromosome data as Tibetan populations have a high frequency (around 50%) of the deeply branching haplogroup D-M174, which is shared with modern Japanese groups (and ancient Jomon hunter-gatherers of Japan) along with Indigenous Andaman islanders of the Bay of Bengal[15].

We used qpGraph[16] to explore scenarios of population splits and gene flow that are consistent with the data and to therefore identify a parsimonious working model for the deep history of key lineages that contribute to ancestry extremes in our principal component analysis (Extended Data Fig. 5 and Supplementary Information section 3). Our fit (Fig. 2 and Extended Data Fig. 6) suggests that much of the ancestry of East Asian individuals can be derived from mixtures in different proportions of two ancient populations: one from the same lineage as the approximately 40,000-year-old Tianyuan individual[10,13] and the other from the same lineage as Indigenous Andaman Islanders (Onge).

We infer that a Tianyuan-related lineage with a northern geographical distribution contributed 98% of the ancestry of Neolithic people from Mongolia and 90% to Neolithic farmers from the Upper Yellow River. (The Upper Yellow River farmer lineage then mixed with an Onge-related branch, which we speculate is related to Tibetan hunter-gatherers to form modern Tibetan populations.) We infer that another Tianyuan-related lineage with a more southern geographical distribution contributed 73% of the ancestry of a hunter-gatherer from the Liangdao site on an island off the southeast coast of China[17] and 56% to Jomon hunter-gatherers from Japan. Japan was occupied by humans before and after the Ice Age and southern and northern Jomon were morphologically distinct[18], which may relate to the admixture that we detect. The northerly Tianyuan-related lineage also contributed to farmers from the West Liao River (67%) and from Taiwan (25%) with the rest of the ancestry of these latter groups being related to Liangdao southern hunter-gatherers. The fact that this northern Tianyuan-related lineage is different from (albeit related to) the lineage that contributed to farmers from the Upper Yellow River suggests that there was probably an expansion of northern farmers to Taiwan that was not linked to the expansion of Yellow River farmers.

The contributions of the Onge-related lineage are concentrated in coastal groups: we estimate 100% in Andamanese, 44% in Jomon and 20% in ancient Taiwan farmers, consistent with the coastal route expansion hypothesized based on the Y-chromosomal haplogroup D-M174 that is found in both Andamanese and Japanese populations[15]. Although Tibet is not coastal, the relatively high inferred contribution of this lineage to ancient Tibetan populations (16%) and the presence of D-M174 with a frequency of around 50% in modern Tibetan individuals, provides a link between this Y-chromosomal haplogroup and Onge-related ancestry. We hypothesize that Tibetan hunter-gatherers represent an early splitting branch of this Late Pleistocene coastal expansion that spread inland and occupied the high plateau.

Refining the trans-Eurasian hypothesis

The farming-and-language-dispersal hypothesis[19] suggests that increases in population densities in and around centres of domestication were important in propelling movements of people that spread languages. However, in East Asia there have been limited data available to test this theory. We searched for genetic correlates of the 'trans-Eurasian hypothesis'[20], which proposes a macrofamily that includes Mongolic, Turkic, Tungusic, Koreanic and Japonic languages based on reconstructed features including shared agricultural terms. The trans-Eurasian hypothesis proposes that languages of these families descend from a proto-language that was associated with the expansion of early millet farmers around the West Liao River in northeast China who spread west towards Mongolia, north towards Siberia and east towards Korea and Japan.

To obtain insight into possible genetic correlates of this language spread, we studied our time transect in the Amur River Basin[21]. From the early Neolithic individuals (around 5500 BC) and Boisman individuals (about 5000 BC) until the Iron Age Yankovsky culture (around 900 BC) and Xianbei culture (AD 50–250), individuals from the Amur

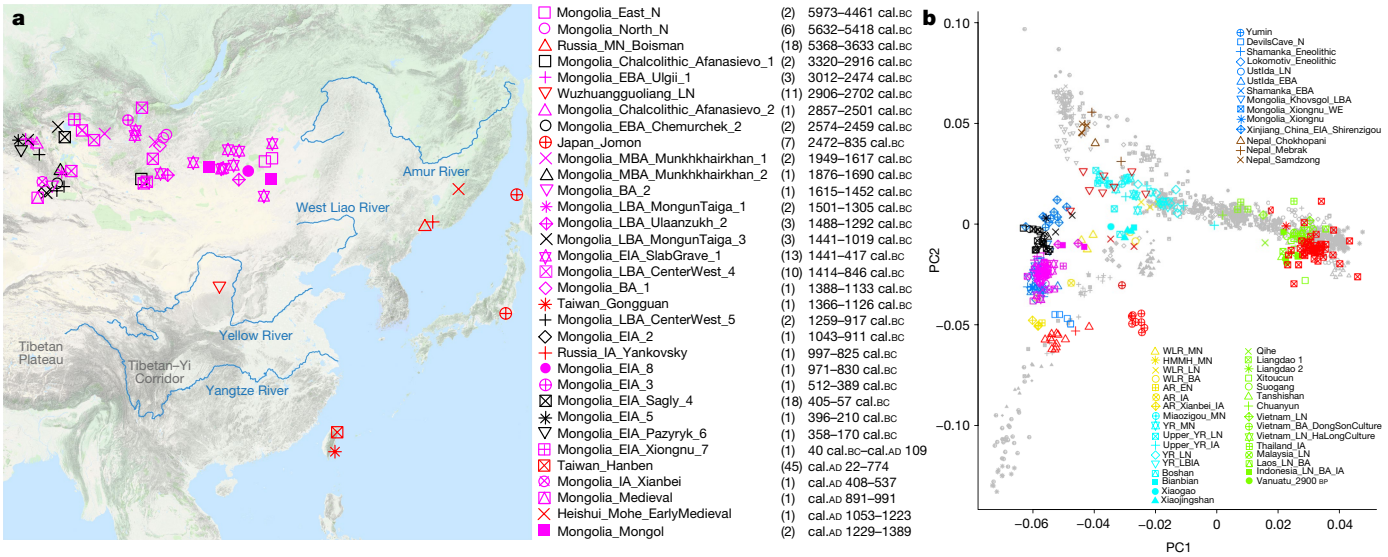


Fig. 1 | Overview. a, Locations, sample size (in brackets) and temporal distribution of newly reported ancient individuals, plotted using the 'Google Map Layer' from ArcGIS Online Basemaps (map data ©2020 Google). **b**, Plot of the first and second principal components (PCs) defined in an analysis of East Asian individuals with minimal West Eurasian-related mixture. cal.AD, calibrated years AD; cal.BC, calibrated years BC. N, Neolithic; BA, Bronze Age; IA, Iron Age; E, Early; M, Middle; L, Late.

River Basin are consistent with being a clade according to qpWave (Supplementary Table 10). This locally continuous population also contributed to later populations, as reflected in the Y-chromosomal haplogroup C2b-F1396 and mitochondrial haplogroups D4 and C5 of Boisman individuals—which are predominant in present-day speakers of Tungusic, Mongolic and some Turkic languages—and in an individual from the Heishui Mohe culture (around AD 1100) who had an estimated 43 ± 15% ancestry from the Amur River Basin lineage (the remaining ancestry was well-modelled as Han Chinese ancestry, which could be expected if there was an immigration from the south in historical times) (Supplementary Table 10). This anciently established Amur River Basin lineage was part of a cline of more Jomon-relatedness in the east and most Mongolian Neolithic-related ancestry in the west. We infer 77–94% Mongolian Neolithic-related ancestry in Baikal hunter-gatherers[5] (the remainder from Ancient North Eurasian populations comprising a deeply splitting West Eurasian-related lineage that was established in the Baikal region during the Ice Age) (Supplementary Table 11). We infer around 87% Mongolian Neolithic-related ancestry in Amur River Basin hunter-gatherers such as Boisman (the remaining ancestry is Jomon-related). Native American individuals share more alleles with Boisman and the Mongolian Neolithic individuals than with most other East Asian populations, suggesting that an early branch of this lineage—reflecting the northern distribution of the Tianyuan-related branch in Fig. 2—was the source for the East-Asian-related ancestry in Native American peoples (Supplementary Table 12).

The trans-Eurasian hypothesis is that the Mongolic, Turkic, Tungusic, Koreanic and Japonic proto-languages were spread by agriculturalists from the West Liao River region, who had a mixture of ancestries related to individuals from the Upper Yellow River (around 67%) and Liangdao (~33%) (Fig. 2). Notably, we observe that this characteristic mixture of ancestries is absent from the time transects of Mongolia and the Amur River Basin in our study (Fig. 3), which is not what is expected on the basis of the hypothesis that expansions of West Liao River farmers spread Mongolic and Tungusic languages. By contrast, the ancestry of West Liao River farmers did plausibly have an influence further east. For example, we can model present-day Japanese populations as two-way mixtures of around 92% West Liao River farmer-related ancestry from the Bronze Age and about 8% Jomon-related ancestry, with a negligible contribution from sources related to Yellow River farmers. We confirmed this by including the Yellow River farming groups in the outgroup set of the qpAdm analysis of Japanese populations and finding that the models continued to fit (Supplementary Tables 13, 14). The West Liao River ancestry is consistent with having been transmitted through Korea, as Japanese populations can be modelled as deriving from Korean (91%) and Jomon (9%) groups (Supplementary Tables 13, 14). None of the six Jomon individuals reported here carried the derived allele in the gene encoding the EDAR(V370A) variant of the human ectodysplasin A receptor, which affects hair, sweat and mammary glands (Supplementary Table 15). This variant has been estimated to have arisen in mainland East Asia around 30,000 years ago[22] and that then reached a high frequency in nearly all Holocene individuals from mainland East Asia and the Americas. The fact that it is nearly absent from the Jomon people highlights the genetic distinctiveness of this population compared with mainland groups.

Northern origin of Sino-Tibetan languages

The Tibetan Plateau has been occupied by modern humans since 40,000–30,000 years ago[23], but it is only since around 1600 BC, with the advent of agriculture, that there is evidence for permanent occupation[24]. Indigenous Tibetan peoples speak Sino-Tibetan languages linked to languages in the coastal plain of China. The northern origins hypothesis for the origin of these closely related languages suggests that farmers who cultivated foxtail millet in the Upper and Middle Yellow River Basin expanded southwest to the Tibetan Plateau and spread present-day Tibeto-Burman languages, and east and south to the Central Plains and eastern coast, spreading Sinitic languages including the linguistic ancestor of Han Chinese[25]. The southern origins hypothesis suggests that the proto-language arose in the Tibetan–Yi Corridor connecting the highlands to the lowlands, and then expanded in the early Holocene[26].

To shed light on Tibetan ancestry, we grouped 17 present-day populations into three genetic clusters (Extended Data Fig. 7): 'Core Tibetan individuals'; 'northern Tibetan individuals' who are admixed between lineages related to Core Tibetan and West Eurasian individuals; and 'Tibeto–Yi Corridor' populations who we estimate using qpAdm[3,16] have 30–70% ancestry related to Southeast Asian populations (Supplementary Table 16) and include not only speakers of Tibetan languages but also speakers of Qiang and Lolo-Burmese languages. Ancient farmers

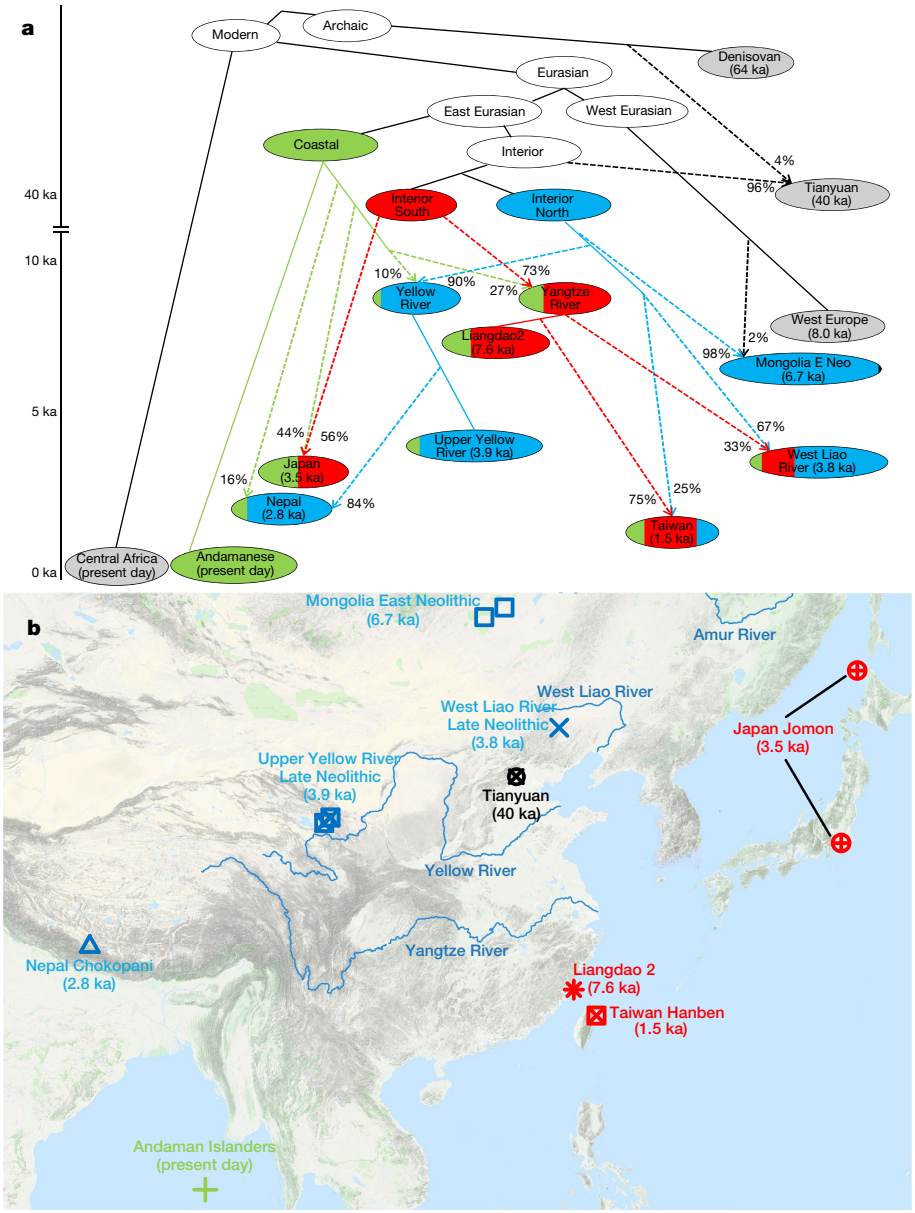


Fig. 2 | Model of deep population relationships. We started with a skeleton tree with one admixture event that fits the data for Denisovan, Mbuti, Onge, Tianyuan and Loschbour according to qpGraph. We grafted on Mongolia East Neolithic (E Neo), Late Neolithic farmers from the Upper Yellow River, Liangdao 2, Japan Jomon, Nepal Chokhopani, Taiwan Hanben and Late Neolithic farmers from the West Liao River, adding them consecutively to all possible edges and retaining only graphs that provided no differences of $|Z| < 3$ between fitted and estimated statistics (maximum $|Z| = 2.95$ here). We used relative population split time estimates from the multiple sequentially Markovian coalescent (MSMC) and MSMC2 analyses[48,49] to constrain models. **a**, We colour lineages modelled as derived from the hypothesized coastal expansion (green), interior southern expansion (red) or interior northern expansion (blue), and populations according to ancestry proportions. Dashed lines represent admixture (proportions are indicated). The grey circles represent sampled populations and white circles represent unsampled hypothesized nodes. **b**, Locations and dates of East Asian individuals used in model fitting, with colours indicating the majority ancestry source, are plotted using the 'Google Map Layer' from ArcGIS Online Basemaps (map data ©2020 Google).

from the Yellow River and present-day Han and Qiang individuals share the most drift with Core Tibetan individuals (Supplementary Table 17), consistent with the hypothesis that Tibetan, Han and Qiang peoples all harbour ancestry from a population related to Neolithic farmers from the Yellow River. We confirm large-scale admixture related to Yellow River farmers (minimum 22% but plausibly a much higher percentage, which is consistent with the 84% estimate in Fig. 2) in Core Tibetan individuals through the decay of admixture linkage disequilibrium[11]. This provides independent evidence that Core Tibetan populations and their genetically almost indistinguishable relatives in ancient Nepal are unlikely to represent continuous descendants of Tibetan hunter-gatherers[75]. We estimate that mixture occurred between, on average, around 290 BC and AD 270 using models of a single pulse of admixture (Supplementary Table 18). The start of admixture could plausibly be as long ago as around 1600 BC, the inferred date for the spread of agriculture onto the Tibetan plateau.

Han Chinese populations are characterized by a north–south genetic cline[27,28]. Farmers from the Upper and Middle Yellow River and Tibetan individuals share more alleles with Han Chinese populations compared with the Southeast Asian cluster, whereas the Southeast Asian cluster groups share more alleles with most Han Chinese groups when compared with Yellow River farmers (Supplementary Tables 19, 20). Using qpWave[3,29], we determined that two sources are consistent with contributing all of the ancestry of most Han Chinese individuals (Supplementary Table 21), with the exception of the northern Han populations

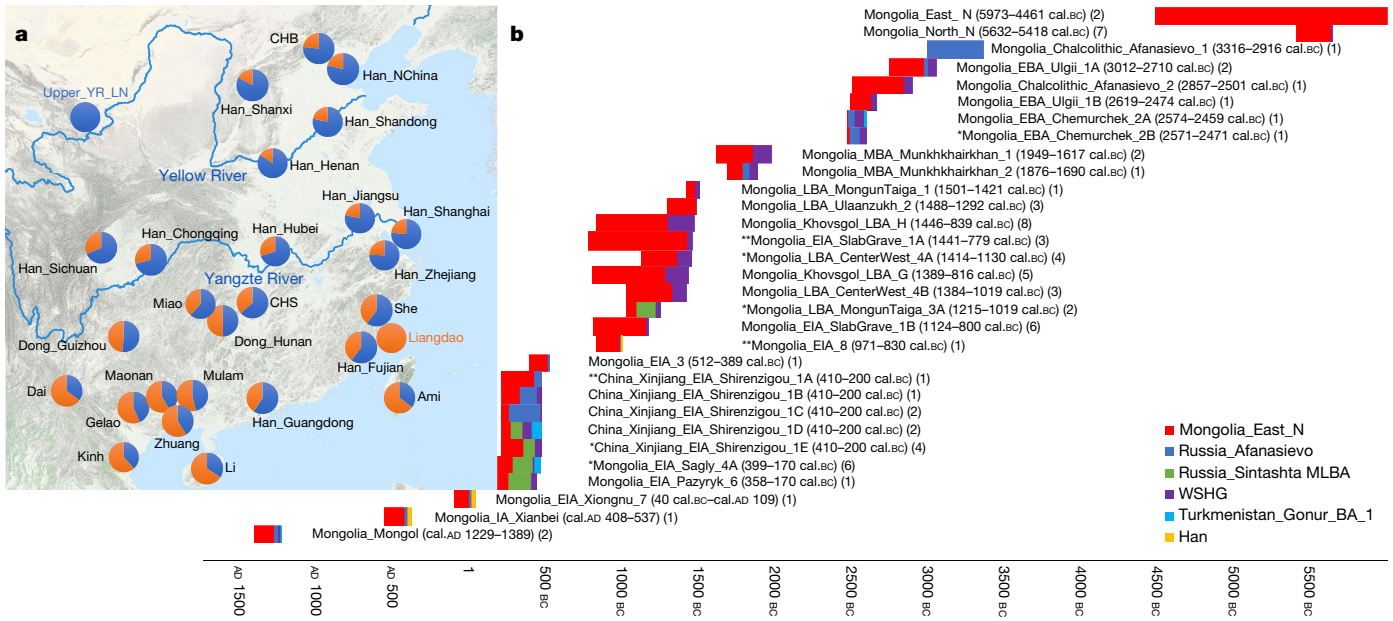


Fig. 3 | Estimates of mixture proportions using qpAdm. a, qpAdm modelling of ancestry related to Yellow River farmers (blue) and Liangdao (orange) in present-day East Asian populations. Proportions are described in Supplementary Table 22 and the map was plotted using the 'Google Map Layer' from ArcGIS Online Basemaps (map data ©2020 Google). CHB, Han Chinese in Beijing; CHS, Han Chinese South; Upper_YR_LN, Upper Yellow River Late Neolithic. **b**, Mongolian and Xinjiang populations. As sources we explored all possible subsets of Mongolia_East_N, Afanasievo, west Siberian hunter-gatherers (WSHG), Sintashta_MLBA, Turkmenistan_Gonur_BA_1 and Han Chinese individuals, adding all groups to the reference set when not used as sources, and identifying parsimonious models (smallest numbers of sources) that fit at $P > 0.05$ based on the Hotelling T^2 test implemented in qpAdm (Supplementary Table 25). These P values do not incorporate any correction for multiple-hypothesis testing. *Parsimonious models pass at only $P > 0.01$. **Multiple equally parsimonious models pass at $P > 0.05$, so we cannot determine whether the West-Eurasian-related source was Afanasievo, west Siberian hunter-gatherers or Sintashta_MLBA (we plot the model with the largest P value). Bars show ancestry proportions, and time spans are unions of all samples. We do not visualize results from singleton outliers. N, Neolithic; BA, Bronze Age; IA, Iron Age; E, Early; M, Middle; L, Late.

for whom we infer West-Eurasian-related admixture of 2–4% (Supplementary Tables 7, 8). We estimate this mixture occurred, on average, 32–45 generations ago, which overlaps the Tang (AD 618–907) and Song (AD 960–1279) dynasties for which historical records of integration of Han Chinese and western ethnic groups are available. For all other Han Chinese groups, we estimate that 59–84% of ancestry is related to farmers from the Upper and Middle Yellow River, and the remainder from a population related to the ancient Liangdao hunter-gatherers. This latter group possibly corresponds to rice farmers of the Yangtze River Basin, an inference that gains strength from the fact that it comprises the primary ancestry of many Austronesian speakers, Tai–Kadai speakers on Hainan Island (Li, around 66%), Southeast Asian individuals from the Bronze Age and around two-thirds of the ancestry of some Austroasiatic speakers[30,31] (Fig. 3 and Supplementary Table 22).

Our results support the northern origins hypothesis for Sino-Tibetan languages, as we detect a specific link between present-day individuals who speak Sino-Tibetan languages and farmers from the Upper and Middle Yellow River. A timing that coincides with the archaeologically attested expansions of farming from this region is also supported by the Y-chromosome evidence of a shared haplogroup (Oα-F5) between Han Chinese and Tibetan peoples that derives from a single male ancestor around 3800 BC[32]. The cline of increasing Liangdao-related ancestry in present-day southern Han Chinese people is plausibly due to expanded mixing of Han Chinese individuals with southern groups as they spread into southern China as recorded in the historical literature[33]. However, this was not the first southward migration, as southern Chinese populations are genetically closer to Late Neolithic farmers from the Yellow River than to earlier Middle Neolithic ones[34] and because we also observe about 25% northern ancestry in ancient farmers from Taiwan (Fig. 2).

Rice farming expansions linked by shared ancestry

Previous ancient DNA analysis in Southeast Asia has shown that the earliest farmers of Southeast Asia had about two-thirds ancestry from East Asian populations that were plausibly related to southern Chinese agriculturalists, and about one-third ancestry from a deeply diverged hunter-gatherer lineage, a pattern that is most-strongly evident in Austroasiatic speakers, which suggests that there is an association with the spread of these languages[30,31]. By capitalizing on our time series, which spans about 2,000 years from ancient Taiwan, we confirm that this was part of a broader pattern. The ancient individuals from Taiwan show strong genetic links to modern Austronesian speakers, a connection that is further supported by the fact that the dominant haplogroups in these ancient individuals are the Y-chromosome lineage O3a2c2-N6 and maternal mitochondrial DNA lineages E1a, B4a1a, F3b1 and F4b[35,36]. These Y-chromosome and mitochondrial lineages are shared by modern Indigenous Taiwanese peoples, and mitochondrial lineages are also present in individuals of the Lapita culture from Vanuatu who were plausibly part of the first spread of Austronesian languages into the southwest Pacific region[37] (Supplementary Table 12). Ancient Taiwan groups and modern Indigenous Taiwanese peoples who speak Austronesian languages share significantly more alleles with speakers of Tai–Kadai languages in southern Chinese Mainland and in Hainan Island[38] than with other East Asian populations (Supplementary Table 12), which is consistent with the hypothesis that ancient populations that were related to present-day speakers of Tai–Kadai languages and descended more anciently from farmers of the Yangtze River (for whom ancient DNA samples have not yet been analysed) spread agriculture to Taiwan around 3000 BC[39]. A surprising finding is our observation that ancient North Chinese individuals are more closely related to

ancient individuals of our Taiwan time transect than to early Holocene hunter-gatherers on the mainland side of the Straits of Taiwan (Supplementary Table 23). This suggests gene flow from Neolithic northern Chinese Mainland into Taiwan, which we estimate to be around 25% if we model it as derived from one of the two source lineages of Yellow River farmers (Fig. 2). This ancestry does not fit as coming from Yellow River farmers themselves, suggesting a north-to-south migration that is not associated with expansions of these farmers. A speculative possibility is that this ancestry was carried by cultivators of foxtail millet—which was domesticated in the north by around 8000 BC[40] and which, in the south, appears relatively early in the Neolithic Tapenkeng culture (around 3000–2500 BC) of Taiwan.

Admixture of West and East Eurasian populations

Mongolia falls near the eastern extreme of the Eurasian Steppe, and archaeological evidence shows that throughout the Holocene this region was a conduit for cultural exchanges between East and West Eurasia. For example, the Afanasievo culture—an eastward extension of the Yamnaya steppe pastoralist culture—brought the first dairying to the region[41] and had a cultural influence on subsequent phenomena such as Chemurchek.

Our Mongolian time transect overwhelmingly derives ancestry from four sources from 6000 to 600 BC. The earliest-established source—and the only source that is primarily East-Asian-associated—is represented at essentially 100% frequency in the two East Mongolian hunter-gatherer individuals from the Neolithic (6000–5000 BC) who are some of the earliest individuals in our dataset (Fig. 3 and Supplementary Tables 24, 25). The second source appears the earliest in seven Neolithic hunter-gatherers from northern Mongolia from 5700 to 5400 BC who can be modelled as having around 5% of ancestry related to previously reported west Siberian hunter-gatherers[6] (Supplementary Table 25). The third source appears the earliest in individuals from the Afanasievo culture (around 3100 BC), who are genetically extremely similar to Yamnaya steppe pastoralists which is consistent with the pattern in individuals of the Afanasievo culture from Russia[4,6]. The fourth source appears by around 1400 BC and is well-modelled as deriving from people with ancestry similar to the pastoralists of the Sintashta culture who derive from a mixture of the Yamnaya culture (around two-thirds) and European farmers (approximately one-third).

To quantify the admixture history in Mongolia, we used qpAdm[3,16] (Supplementary Table 25). Many eastern Mongolian individuals can be modelled as simple two-way admixtures of Neolithic eastern Mongolian populations as one source (65–100%) and the remainder of the ancestry deriving from west Siberian hunter-gatherers (Fig. 3). The individuals who fit this model were not only from Neolithic groups (0–5% west Siberian hunter-gatherers), but also a child from the Early Bronze Age from the Afanasievo Kurgak govi site (15%), the Ulgii group (21%), the main grouping from the Middle Bronze Age Munkhkhairkhan culture (31–36%) and, in the Late Bronze Age, a combined group from the Centre–West region (24–31%), as well as individuals of the Mongun Taiga type (35%). The fact that the child from Kurgak govi has no evidence of Yamnaya-related ancestry despite his clear Afanasievo cultural association and chronology makes him the first case of an individual buried with Afanasievo traditions who has no evidence of Yamnaya ancestry. The legacy of the spread during the Yamnaya era into Mongolia continued in two individuals from the Chemurchek culture whose ancestry can only be modelled using Yamnaya–Afanasievo ancestry as a source (around 33–51%) (Supplementary Table 25). This fits even when ancient European farmers are included in the outgroups, providing no evidence for the theory that long-distance movement of people spread West European megalithic cultural traditions to people of the Chemurchek culture[42].

The one instance before 600 BC for which our four source model does not fit occurs in a Chemurchek individual ($P = 3.7 \times 10^{-4}$ from qpAdm), but we can successfully model the ancestry of this individual by adding 15% additional ancestry from populations related to the Turan region far to the south (Fig. 3). A parallel study[43] models a Chemurchek-associated individual as a mixture of Turan and early Kazakhstan pastoralists from the site of Botai, without any of the other three ancestries that we detect in all Chemurchek individuals in our study. As our best-fit model passes when Botai is in the reference set ($P > 0.63$) (Supplementary Table 25), the two findings would indicate an extremely complex origin for Chemurchek if both were correct, with one migration stream carrying Botai-related ancestry and the other not carrying it.

From the Middle Bronze Age, there is no compelling evidence in the Mongolian time transect data for the persistence of the Yamnaya-derived lineages that spread with the Afanasievo culture. Instead, the Yamnaya-related ancestry can only be modelled as deriving from a later spread related to people of the Sintashta and Andronovo horizons of the Middle to Late Bronze Age who were themselves a mixture of around two-thirds Yamnaya-related and one-third European farmer-related ancestry[4–6]. The Sintashta-related ancestry is detected in proportions of 0–57% in groups from this time onward, with substantial proportions of Sintashta-related ancestry only in western Mongolia (Fig. 3 and Supplementary Table 25). For all of these groups, qpAdm ancestry models pass with Afanasievo groups in the outgroups whereas models with the Afanasievo-associated peoples as the source and Sintashta-related groups in the outgroups are all rejected (Fig. 3 and Supplementary Table 25).

New ancestry began reaching Mongolia in large proportions starting in the Late Bronze Age, with qpAdm models failing when using Neolithic eastern Mongolian populations as a single East Asian source in some individuals from the Late Bronze Age of Khovsgol, Ulaanzukh and the Centre–West region, two individuals from the Early Iron Age associated with Slab Grave culture, and for Xiongnu, Xianbei and Mongol peoples. However, when we include Han Chinese populations as a source, we estimate Han-related ancestry proportions of 9–80% in the aforementioned individuals (Supplementary Table 25). Turan-derived ancestry spread into the region again by the sixth to fourth century BC as we detect it in multiple individuals from the Iron Age Sagly culture. We find that alleles with two polymorphisms (rs1426654 and rs16891982) that are associated with light skin pigmentation and one (rs12913832) associated with blue eyes in European individuals occur frequently in the Sagly samples, but that the rs4988235 allele associated with lactose tolerance is nearly absent in all East Asian individuals that we analysed (Supplementary Table 15).

Although the Yamnaya–Afanasievo-associated lineages are consistent with having largely disappeared in Mongolia by the Middle to Late Bronze Age, we confirm and strengthen previous ancient DNA analysis that suggested that the legacy of this expansion persisted in western China into the time of the Iron Age Shirenzigou culture (410–190 BC)[44]. Considering many of the Shirenzigou individuals separately as well as three of the five genetically homogeneous subclusters, the only parsimonious models derive all of their West-Eurasian-related ancestry from groups related to the Afanasievo culture, confirming that Afanasievo ancestry without the characteristic European farmer-related mixture, which appeared later in Central Asia and Mongolia, persisted in Xinjiang. For example, for the two individuals with the most West-Eurasian-related ancestry (Xinjiang_EIA_Shirenzigou_1C), all three-way models that fit include Russian Afanasievo ancestry (71–77%) (Fig. 3 and Supplementary Table 25). Moreover, the total ancestry from the two other West-Eurasian-related groups that can fit in small proportions in such models is always less than 9% (Supplementary Table 25). In pre-state societies, languages are thought to spread primarily through the movements of people[45], and these results therefore add weight to the theory that the Tocharian languages of the Tarim Basin spread through the migration of Yamnaya descendants to the Altai Mountains and Mongolia (in the guise of the Afanasievo culture), from

whence they spread further to Xinjiang[4–6,44,46,47]. These results are important for theories of the diversification of Indo-European languages, as they increase the evidence in favour of the hypothesis that the split of the second-oldest branch in the Indo-European language tree occurred at the end of the fourth millennium BC[44,46,47].

Conclusion

While this study marks considerable progress in understanding the population history of East Asia, the findings raise as many questions as answers, motivating the collection of additional ancient DNA data. A particular priority should be to generate an ancient DNA time transect through southern China, including early farmers of the Yangtze River region—the putative source for the ancestry prevalent in the Southeast Asian Cluster of present-day groups—which would make it possible to test and extend the model presented in this study, and to better understand how dispersals of languages in Southeast Asia do or do not correlate to ancient movements of people. Another priority should be to generate data on many additional pre-Ice Age individuals from East Asia, which will make it possible to test the model of deep population relationships presented in Fig. 2 and to better understand the origins, migrations, and mixtures of the diverse modern human populations that have lived in East Asia for more than 40,000 years.

Online content

Any methods, additional references, Nature Research 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/s41586-021-03336-2.

1. Cavalli-Sforza, L. L. The Chinese human genome diversity project. *Proc. Natl Acad. Sci. USA* **95**, 11501–11503 (1998).
2. HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. *Science* **326**, 1541–1545 (2009).
3. Haak, W. et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
4. Allentoft, M. E. et al. Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
5. Damgaard, P. B. et al. 137 ancient human genomes from across the Eurasian steppes. *Nature* **557**, 369–374 (2018).
6. Narasimhan, V. M. et al. The formation of human populations in South and Central Asia. *Science* **365**, eaat7487 (2019).
7. Fu, Q. et al. An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
8. Fu, Q. et al. DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl Acad. Sci. USA* **110**, 2223–2227 (2013).
9. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
10. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
11. Loh, P. R. et al. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233–1254 (2013).
12. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
13. Yang, M. A. et al. 40,000-year-old individual from Asia provides insight into early population structure in Eurasia. *Curr. Biol.* **27**, 3202–3208 (2017).
14. Massilani, D. et al. Denisovan ancestry and population history of early East Asians. *Science* **370**, 579–583 (2020).
15. Wang, C. C. & Li, H. Inferring human history in East Asia from Y chromosomes. *Investig. Genet.* **4**, 11 (2013).
16. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
17. Yang, M. A. et al. Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* **369**, 282–288 (2020).
18. Nakashima, A., Ishida, H., Shigematsu, M., Goto, M. & Hanihara, T. Nonmetric cranial variation of Jomon Japan: implications for the evolution of eastern Asian diversity. *Am. J. Hum. Biol.* **22**, 782–790 (2010).
19. Bellwood, P. & Renfrew, C. *Examining the Farming/Language Dispersal Hypothesis* (McDonald Institute for Archaeological Research, 2002).
20. Robbeets, M. & Savelyev, A. *The Oxford Guide to the Transeurasian Languages* (Oxford Univ. Press, 2020).
21. Siska, V. et al. Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. *Sci. Adv.* **3**, e1601877 (2017).
22. Kamberov, Y. G. et al. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* **152**, 691–702 (2013).
23. Zhang, X. L. et al. The earliest human occupation of the high-altitude Tibetan Plateau 40 thousand to 30 thousand years ago. *Science* **362**, 1049–1051 (2018).
24. Chen, F. H. et al. Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 B.P. *Science* **347**, 248–250 (2015).
25. Zhang, M., Yan, S., Pan, W. & Jin, L. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. *Nature* **569**, 112–115 (2019).
26. van Driem, G. in *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics* (eds Sagart, L. et al.) 81–106 (Routledge, 2005).
27. Liu, S. et al. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. *Cell* **175**, 347–359 (2018).
28. Chiang, C. W. K., Mangul, S., Robles, C. & Sankararaman, S. A comprehensive map of genetic variation in the world's largest ethnic group—Han Chinese. *Mol. Biol. Evol.* **35**, 2736–2750 (2018).
29. Reich, D. et al. Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
30. Lipson, M. et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* **361**, 92–95 (2018).
31. McColl, H. et al. The prehistoric peopling of Southeast Asia. *Science* **361**, 88–92 (2018).
32. Wang, L. X. et al. Reconstruction of Y-chromosome phylogeny reveals two neolithic expansions of Tibeto-Burman populations. *Mol. Genet. Genomics* **293**, 1293–1300 (2018).
33. Ge, J. X., Wu, S. D. & Chao, S. J. *Zhongguo yimin shi (The Migration History of China)* (Fujian People's Publishing House, 1997).
34. Ning, C. et al. Ancient genomes from northern China suggest links between subsistence changes and human migration. *Nat. Commun.* **11**, 2700 (2020).
35. Wei, L. H. et al. Phylogeography of Y-chromosome haplogroup O3a2b2-N6 reveals patrilineal traces of Austronesian populations on the eastern coastal regions of Asia. *PLoS ONE* **12**, e0175080 (2017).
36. Ko, A. M. et al. Early Austronesians: into and out of Taiwan. *Am. J. Hum. Genet.* **94**, 426–436 (2014).
37. Skoglund, P. et al. Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**, 510–513 (2016).
38. Lipson, M. et al. Reconstructing Austronesian population history in island Southeast Asia. *Nat. Commun.* **5**, 4689 (2014).
39. Bellwood, P. The checkered prehistory of rice movement southwards as a domesticated cereal—from the Yangzi to the equator. *Rice* **4**, 93–103 (2011).
40. Yang, X. et al. Early millet use in northern China. *Proc. Natl Acad. Sci. USA* **109**, 3726–3730 (2012).
41. Wilkin, S. et al. Dairy pastoralism sustained eastern Eurasian steppe populations for 5,000 years. *Nat. Ecol. Evol.* **4**, 346–355 (2020).
42. Kovalev, A. The great migration of the Chemurchek people from France to the Altai in the early 3rd millennium BCE. *Int. J. Eurasian Stud.* **1**, 1–58 (2011).
43. Jeong, C. et al. A dynamic 6,000-year genetic history of Eurasia's Eastern Steppe. *Cell* **183**, 890–904 (2020).
44. Ning, C. et al. Ancient genomes reveal Yamnaya-related ancestry and a potential source of Indo-European speakers in Iron Age Tianshan. *Curr. Biol.* **29**, 2526–2532 (2019).
45. Bellwood, P. in *The Encyclopedia of Global Human Migration* (Wiley-Blackwell, 2013).
46. Mallory, J. P. in *Search of the Indo-Europeans: Language, Archaeology and Myth* (Thames & Hudson, 1991).
47. Anthony, D. *The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World* (Princeton Univ. Press, 2007).
48. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
49. Wang, K., Mathieson, I., O'Connell, J. & Schiffels, S. Tracking human population structure through time from whole genome sequences. *PLoS Genet.* **16**, e1008552 (2020).
75. Jeong, C. et al. Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. *Proc. Natl Acad. Sci. USA.* **113**, 7485–7490 (2016).

Article

Chuan-Chao Wang[1,2,3,4,44]✉, Hui-Yuan Yeh[5,44], Alexander N. Popov[6,44], Hu-Qin Zhang[7,44], Hirofumi Matsumura[8], Kendra Sirak[2,9], Olivia Cheronet[10], Alexey Kovalev[11], Nadin Rohland[2], Alexander M. Kim[2,12], Swapan Mallick[2,9,13,14], Rebecca Bernardos[2], Dashtseveg Tumen[15], Jing Zhao[7], Yi-Chang Liu[16], Jiun-Yu Liu[17], Matthew Mah[2,13,14], Ke Wang[3], Zhao Zhang[2], Nicole Adamski[2,14], Nasreen Broomandkhoshbacht[2,14], Kimberly Callan[2,14], Francesca Candilio[10], Kellie Sara Duffett Carlson[10], Brendan J. Culleton[18], Laurie Eccles[19], Suzanne Freilich[10], Denise Keating[10], Ann Marie Lawson[2,14], Kirsten Mandl[10], Megan Michel[2,14], Jonas Oppenheimer[2,14], Kadir Toykan Özdoğan[10], Kristin Stewardson[2,14], Shaoqing Wen[20], Shi Yan[21], Fatma Zalzala[2,14], Richard Chuang[16], Ching-Jung Huang[16], Hana Looh[22], Chung-Ching Shiung[16], Yuri G. Nikitin[23], Andrei V. Tabarev[24], Alexey A. Tishkin[25], Song Lin[7], Zhou-Yong Sun[26], Xiao-Ming Wu[7], Tie-Lin Yang[7], Xi Hu[7], Liang Chen[27], Hua Du[28], Jamsranjav Bayarsaikhan[29], Enkhbayar Mijiddorj[30], Diimaajav Erdenebaatar[30], Tumur-Ochir Iderkhangai[30], Erdene Myagmar[15], Hideaki Kanzawa-Kiriyama[31], Masato Nishino[32], Ken-ichi Shinoda[31], Olga A. Shubina[33], Jianxin Guo[1], Wangwei Cai[34], Qiongying Deng[35], Longli Kang[36], Dawei Li[37], Dongna Li[38], Rong Lin[38], Nini[36], Rukesh Shrestha[4], Ling-Xiang Wang[4], Lanhai Wei[1], Guangmao Xie[39,40], Hongbing Yao[41], Manfei Zhang[4], Guanglin He[1], Xiaomin Yang[1], Rong Hu[1], Martine Robbeets[42], Stephan Schiffels[3], Douglas J. Kennett[43], Li Jin[4], Hui Li[4], Johannes Krause[3]✉, Ron Pinhasi[10]✉ & David Reich[2,9,13,14]✉

[1]Department of Anthropology and Ethnology, Institute of Anthropology, State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen, China. [2]Department of Genetics, Harvard Medical School, Boston, MA, USA. [3]Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany. [4]MOE Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai, China. [5]School of Humanities, Nanyang Technological University, Nanyang, Singapore. [6]Scientific Museum, Far Eastern Federal University, Vladivostok, Russia. [7]Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, China. [8]School of Health Science, Sapporo Medical University, Sapporo, Japan. [9]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA. [10]Department of Evolutionary Anthropology, University of Vienna, Vienna, Austria. [11]Institute of Archaeology, Russian Academy of Sciences, Moscow, Russia. [12]Department of Anthropology, Harvard University, Cambridge, MA, USA. [13]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [14]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, USA. [15]Department of Anthropology and Archaeology, National University of Mongolia, Ulaanbaatar, Mongolia. [16]Institute of Archaeology, National Cheng Kung University, Tainan, Taiwan. [17]Department of Anthropology, University of Washington, Seattle, WA, USA. [18]Institutes of Energy and the Environment, The Pennsylvania State University, University Park, PA, USA. [19]Department of Anthropology, Pennsylvania State University, University Park, PA, USA. [20]Institute of Archaeological Science, Fudan University, Shanghai, China. [21]School of Ethnology and Sociology, Minzu University of China, Beijing, China. [22]Institute of History and Philology, Institute of History and Philology, Academia Sinica, Taipei, Taiwan. [23]Museum of Archaeology and Ethnology, Institute of History, Archaeology and Ethnology, Far Eastern Branch of the Russian Academy of Sciences, Vladivostok, Russia. [24]Institute of Archaeology and Ethnography, Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia. [25]Department of Archeology, Ethnography and Museology, Altai State University, Barnaul, Russia. [26]Shaanxi Provincial Institute of Archaeology, Xi'an, China. [27]School of Cultural Heritage, Northwest University, Xi'an, China. [28]Xi'an AMS Center, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an, China. [29]Research Center at the National Museum of Mongolia, Ulaanbaatar, Mongolia. [30]Department of Archaeology, Ulaanbaatar State University, Ulaanbaatar, Mongolia. [31]Department of Anthropology, National Museum of Nature and Science, Tsukuba, Japan. [32]Archaeological Center of Chiba City, Chiba, Japan. [33]Department of Archeology, Sakhalin Regional Museum, Yuzhno-Sakhalinsk, Russia. [34]Department of Biochemistry and Molecular Biology, Hainan Medical University, Haikou, China. [35]Department of Human Anatomy and Center for Genomics and Personalized Medicine, Guangxi Medical University, Nanning, China. [36]Key Laboratory for Molecular Genetic Mechanisms and Intervention Research on High Altitude Disease of Tibet Autonomous Region, Ministry of Education, School of Medicine, Xizang Minzu University (Tibet University for Nationalities), Xianyang, China. [37]Institute for History and Culture of Science & Technology, Guangxi University for Nationalities, Nanning, China. [38]Department of Biology, Hainan Medical University, Haikou, China. [39]College of History, Culture and Tourism, Guangxi Normal University, Guilin, China. [40]Guangxi Institute of Cultural Relics Protection and Archaeology, Nanning, China. [41]Belt and Road Research Center for Forensic Molecular Anthropology, Key Laboratory of Evidence Science of Gansu Province, Gansu Institute of Political Science and Law, Lanzhou, China. [42]Eurasia3angle Research group, Max Planck Institute for the Science of Human History, Jena, Germany. [43]Department of Anthropology, University of California Santa Barbara, Santa Barbara, CA, USA. [44]These authors contributed equally: Chuan-Chao Wang, Hui-Yuan Yeh, Alexander N. Popov, Hu-Qin Zhang.
✉e-mail: wang@xmu.edu.cn; krause@shh.mpg.de; ron.pinhasi@univie.ac.at; reich@genetics.med.harvard.edu

Methods

Ethics statement

The collection of modern samples was carried out in 2014 in strict accordance with the ethical research principles of The Ministry of Science and Technology of the People's Republic of China (Interim Measures for the Administration of Human Genetic Resources, 10 June 1998). Our sample collection and genotyping protocol was further reviewed and approved by the Ethics Committee of the School of Life Sciences, Fudan University (22 October 2014). Study staff informed potential participants about the goals of the project, and individuals who chose to participate gave informed consent consistent with broad studies of population history and human variation and public posting of anonymized data. There were no rewards for participating and no negative consequences for not participating; all participants signed or affixed a thumbprint to the consent form reviewed by Fudan University. An important principle of our study was to ensure not only that the research was underpinned by individual informed consent, but also that it had support from community representatives sensitive to local perspectives, and we therefore carried out community consultation with minority group leaders or village leaders as an integral part of the consent process. For each minority group, community representatives affirmed community support for the study through a signature or thumbprint on a form summarizing the community consultation process (these forms were completed between 10 November 2014 and 10 December 2014). Co-authors of the manuscript who were culturally Indigenous and in some cases were legally registered as members of minority groups specifically reviewed the discussions of population history in this manuscript to increase sensitivity to local perspectives. Specifically, L.W. is a Tai–Kadai-speaking Zhuang person from Guangxi in southwest China; R.S. is from Nepal; and L.K. and N. are based at the Tibet University for Nationalities, and N. is an Indigenous Tibetan. We emphasize that Indigenous and community narratives co-exist with scientific ones and may or may not align with them. Indigenous ancestry should not be confused with identity, which is about self-perception and culture and cannot be defined by genetics alone.

The ancient samples newly reported in this study were collected with the permission of the custodians of the samples, who are the archaeologists or museums in each of the countries for which we analysed the data. We applied a case-by-case approach to obtaining permissions for each set of samples depending on the local expectations as these vary by region and cultural context. Every newly reported ancient sample in this study has permission for analysis from custodians of the samples who are co-authors and who affirm that ancient DNA analysis of these samples is appropriate. For most samples, we prepared formal collaboration agreements to explicitly list the ancient DNA work being performed by our team. In other instances, sample custodians who are co-authors determined that the generation and publication of ancient DNA data was covered under their existing permissions for sample analysis, and so new sampling agreements were not required. Going beyond what was formally required, we also sought to make the presentation of the scientific findings sensitive to local perspectives from the regions from which the skeletons were excavated. For some regions for which we obtained DNA such as the southern islands of Japan and the Russian Far East sites we are not aware of modern communities with traditions of biological or cultural connection to the ancient remains. For other regions, such as the Upper Yellow River Chinese or Mongolian regions, the modern nation-states in which the ancient individuals lived are modern inheritors of the cultural and genetic heritage of the ancient groups. In Taiwan, in addition to obtaining formal permission for sampling from government institutions, we sought to ensure that the presentation of our results was sensitive to the perspectives of Indigenous Taiwanese groups who plausibly descend thousands of years ago from groups related to those individuals whose data we report. The existence of at least 16 non-Han Chinese Indigenous groups in Taiwan makes it difficult to connect particular sites to specific modern ethnic groups for prehistoric sites older than 400 years, and it is rare for local communities to express connections with prehistoric sites. Nevertheless, two co-authors with Indigenous Taiwanese ancestry or cultural affiliation to these groups specifically reviewed the discussion of the results of Taiwanese groups to increase the sensitivity of our study to the perspectives of Indigenous groups. H.-Y.Y., who is co-first author of the study, has ancestry from the Paiwan Indigenous group. H.L. was the excavation leader for the Bilhun Hanben site and is the local community leader for the Ami group, whose present-day culture shows some similarities to the material culture of the site.

Ancient DNA laboratory work

All samples except those from Wuzhuangguoliang were prepared in dedicated clean room facilities at the Harvard Medical School, Boston, USA and in some cases also the University of Vienna, Vienna, Austria. Supplementary Table 2 lists experimental settings for each sample and library included in the dataset. Skeletal samples were surface-cleaned and drilled or sandblasted and milled to produce a fine powder for DNA extraction[50,51]. We either followed a previously published extraction protocol[52] replacing the extender-MinElute-column assembly with the columns from the Roche High Pure Viral Nucleic Acid Large Volume Kit[53] (manual extraction) or, for samples prepared later, used a DNA-extraction protocol based on silica beads instead of spin columns (and Dabney buffer) to enable automated DNA purification[54] (robotic extraction). We prepared individually barcoded double-stranded libraries for most samples using a protocol that included a DNA repair step with uracil-DNA glycosylase (UDG) to cut molecules at locations containing ancient DNA damage but that is inefficient at the terminal positions of DNA molecules, allowing the rate of damage at the final nucleotide to be used as a measure of authenticity (Supplementary Table 1, UDG: 'half')[55]. We also prepared some libraries without UDG pre-treatment (double-stranded minus). For a few extracts, single-stranded DNA libraries[56] were prepared with USER (NEB) addition in the dephosphorylation step, which results in inefficient uracil removal at the 5′ end of the DNA molecules, and does not affect deamination rates at the terminal 3′ end[57]. We performed target enrichment via hybridization with previously reported protocols[8]. We either enriched for the mitochondrial genome and 1.2 million SNPs in two separate experiments or together in a single experiment. If split over two experiments, the first enrichment was for sequences aligning to mitochondrial DNA (mtDNA)[55,58] with some baits overlapping nuclear targets spiked-in to screen libraries for nuclear DNA content. The second enrichment was for a targeted set of 1,237,207 SNPs that comprises a merge of two previously reported sets of 394,577 SNPs[3] and 842,630 SNPs[7]. We sequenced the enriched libraries on an Illumina NextSeq500 instrument for 2 × 76 cycles (and both indices) or on Hiseq X10 instruments at the Broad Institute of MIT and Harvard for 2 × 101 cycles. We also shotgun-sequenced a few hundred thousand molecules from each library to assess the fraction of human DNA.

Extractions of the Wuzhuangguoliang samples were performed in the clean room at the Xi'an Jiaotong University and Xiamen University following a previously published protocol[59]. Each extract was converted into double-stranded Illumina libraries following the manufacturer's protocol (Fast Library Prep Kit, iGeneTech). Sample-specific indexing barcodes were added to both sides of the fragments via amplification. Nuclear DNA capture was performed with the AIExome Enrichment Kit V1 (iGeneTech) according to the manufacturer's protocol and sequenced on an Illumina NovaSeq instrument with 150-base-pair paired-end reads.

Bioinformatics processing

We de-multiplexed the data and assigned sequences to samples based on the barcodes and/or indices, allowing up to one mismatch per barcode or index. We trimmed adapters and restricted to fragments for

which the two reads overlapped by at least 15 nucleotides. We merged sequences (allowing up to one mismatch) choosing bases in the merged region based on highest quality in case of a conflict, using either a modified version of Seqprep[60] (if we were using bioinformatics processing pipeline 1 as specified in Supplementary Table 2) or custom software (if we were using bioinformatics processing pipeline 2; https://github.com/DReichLab/ADNA-Tools). We aligned the merged sequences using bwa (v.0.6.1 for pipeline 1 and v.0.7.15 for pipeline 2)[61] to the mitochondrial genome RSRS[62] and to the human genome (GRCh37, https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/). We removed duplicate sequences with the same orientation, start and stop positions and barcodes. We determined haplogroups using HaploGrep2[63]. To assess authenticity, we estimated the rate of cytosine to thymine substitution in the final nucleotide, which is expected to be at least 3% at cytosines in libraries prepared with a partial UDG treatment protocol and at least 10% for untreated libraries (minus) and single-stranded libraries; all libraries that we analysed met this threshold. We also assessed authenticity using contamMix (v.1.0.9 for pipeline 1 and v.1.0.12 for pipeline 2)[8] to determine the fraction of mtDNA sequences in an ancient sample that matches the endogenous majority consensus more closely than a comparison set of 311 worldwide present-day human mtDNAs. For whole-genome analysis, we randomly selected a single sequence covering every SNP position of interest ('pseudo-haploid' data) using custom software, only using nucleotides that were a minimum distance from the ends of the sequences to avoid deamination artefacts (https://github.com/DReichLab/adna-workflow). The coverages and numbers of SNPs covered at least once on the autosomes (chromosomes 1–22) are included in Supplementary Table 1 for a merge of data from all libraries for each sample. Supplementary Table 2 shows the results by library.

To evaluate whether there was evidence that ancient DNA data processed using the same bioinformatics pipeline was artefactually biased to appear similar to each other in *f*-statistic analysis, we computed statistics of the form f_4(Group1Pipeline1, Group1Pipeline2; Group2Pipeline1, Group2Pipeline2) for all groups for which we had individuals in our main analysis dataset processed by both pipelines (Mongolia_EIA_Sagly_4, Mongolia_EIA_SlabGrave_1, Mongolia_LBA_CenterWest_4, Mongolia_LBA_MongunTaiga_3, Russia_MN_Boisman and Taiwan_Hanben). For all 15 possible pairwise comparisons, the *Z*-scores for deviation from zero as computed based on a block jackknife standard error had a magnitude of less than |2.7|, which is not significant after correcting for the 15 tests that we performed ($P = 0.11$ after applying a Bonferroni correction) (Supplementary Table 3).

Although these analyses reduce concerns about systematic differences in population genetic analysis driven by changes over time in the software that we used to carry out our bioinformatics processing steps, we caution that there are other inhomogeneities in our ancient DNA dataset that have the potential to affect inferences. Other sources of inhomogeneity include systematic differences in the chemical properties and preservation conditions of DNA from different archaeological sites, differences in wet laboratory protocols including differences between data from in-solution enrichment and direct shotgun sequencing, and differences in wet laboratory and bioinformatics processing protocols across research groups that published the various datasets co-analysed in our study. The fact that we can obtain fitting models of population history through admixture graph analysis (Fig. 2) even in the presence of these differences, and that the admixture graph model also fits when restricting to transversion polymorphisms (Supplementary Information section 3) and finally that our f_4-symmetry tests reveal no significant differences between data generated for this study using wet laboratory and bioinformatics protocols that changed over time (Supplementary Table 3) increases confidence that our inferences are valid even in the presence of inhomogeneities[64].

Customized damage restriction to address contamination in Wuzhuangguoliang

We explored authenticity metrics for different filtering strategies for the data from the Wuzhuangguoliang individuals: restricting only to damaged sequences, and merging damaged sequences with sequences that do not show damage in the final nucleotides but that are short (requiring a minimum of 30 bp, and increasing in 10-bp increments from there up to 180 bp). We considered data from an individual usable for analysis if it consisted of a minimum 5,000 SNPs, if the lower bound of its ANGSD[9] 95% confidence interval is less than 0.01, and if the upper bound of its contamMix 95% confidence interval is more than 0.98. We choose the version of each sample that has the most SNPs covered as long as it meets the criteria above (Supplementary Table 26).

Accelerator mass spectrometry radiocarbon dating

We generated 108 direct accelerator mass spectrometry (AMS) radiocarbon (^{14}C) dates; 70 at the Pennsylvania State University (PSUAMS), 32 through a collaboration of Pennsylvania State University (PSUAMS) and the University of California Irvine (UCIAMS), and 6 at Poznan Radiocarbon Laboratory. The methods used at Poznan are available at https://radiocarbon.pl/en/ and here we summarize the methods used for the samples measured at PSUAMS and UCIAMS. Bone collagen from petrous, phalanx or tooth (dentine) samples was extracted and purified using a modified Longin method with ultrafiltration (>30 kDa gelatin)[65]. If bone collagen was poorly preserved or contaminated, we hydrolysed the collagen and purified the amino acids using solid-phase extraction columns (XAD amino acids)[66]. Before extraction, we sequentially sonicated all samples in ACS-grade methanol, acetone and dichloromethane (30 min each) at room temperature to remove conservants or adhesives possibly used during curation. Extracted collagen or amino acid preservation was evaluated using crude gelatin yields (%wt), %C, %N and C/N ratios. Stable carbon and nitrogen isotopes were measured on a Thermo DeltaPlus instrument with a Costech elemental analyser at Yale University. C/N ratios between 3.06 and 3.45 indicate that all radiocarbon-dated samples are well-preserved. All samples were combusted and graphitized at PSUAMS and UCIAMS using methods described elsewhere[65]. ^{14}C measurements were made on a modified National Electronics Corporation 1.5SDH-1 compact accelerator mass spectrometer at either the PSUAMS facility or the Keck-Carbon Cycle AMS Facility at the University of California Irvine. All dates were calibrated using the IntCal20 curve[67] in OxCal v.4.4.2[68] and are presented in calibrated calendar years BC or AD.

Y-chromosomal haplogroup analysis

We determined Y-chromosomal haplogroups by examining the state of SNPs in ISOGG v.15.56 (https://isogg.org/tree/index.html) (Supplementary Information section 4).

X-chromosome contamination estimates

We performed an X-chromosomal contamination test for the male individuals following a previously described approach[69] and implemented in the ANGSD software[9]. We used the 'methods of moments' estimates. The estimates for some male individuals are not informative because of the limited number of X-chromosomal SNPs covered by at least two sequences (we only report results for individuals with at least 200 SNPs covered at least twice).

Procedure for combining new Affymetrix human origins genotyping data on modern individuals with previously published data

We merged the newly generated data with previously published datasets genotyped on Affymetrix Human Origins arrays[16], restricting to present-day individuals with more than 95% genotyping completeness. We manually curated the data using ADMIXTURE[12] and principal component analysis as implemented in EIGENSOFT[10] to identify individuals that were outliers compared with others from their own populations in cases in which a main cluster was identifiable. We removed seven present-day individuals as outliers from subsequent analysis; the population identifiers for these individuals are prefixed by the string

"Ignore_" in the dataset that we released (for analyses of ancient individuals, we do not remove outliers).

Principal component analysis

We used the smartpca program of EIGENSOFT[10], using default parameters and the lsqproject: YES and numoutlieriter: 0 options.

ADMIXTURE analysis

We carried out ADMIXTURE analysis in unsupervised mode[12] after pruning for linkage disequilibrium in PLINK[70] with parameters --indep-pairwise 200 25 0.4, which retained 256,427 SNPs. We ran ADMIXTURE with default fivefold cross-validation (--cv = 5), varying the number of ancestral populations between $K = 2$ and $K = 18$ in 100 bootstraps with different random seeds.

Clustering of ancient individuals

We clustered ancient individuals based on chronology and archaeological association, and then further based on both qualitative similarity (in principal component analysis (PCA), ADMIXTURE and outgroup f_3-statistics) and quantitative homogeneity (based on f_4-statistics and qpAdm results). In general, group names have the format '<country>_<additional geographical detail if any>_<time period>_<cultural association if any>_<genetic cluster>'. For the individuals in Mongolia and the Xinjiang Iron Age Shirenzigou group, we carried out finer clustering guided by qpWave to tests for homogeneity. We use an alphabetical suffix to designate the qpWave-based subcluster (for example, Mongolia_EBA_Chemurchek_2A).

f-Statistics

We computed *f*-statistics using ADMIXTOOLS[12] with default parameters, and standard errors using a block jackknife[71]. We use 'outgroup-f_3' statistics of the form f_3(African_outgroup; Test, Comparison) to measure allele sharing between a Test population and a Comparison panel. If we detect a significantly negative value for an admixture-f_3' statistic of the form f_3(Test; Source1, Source2) we have evidence that a Test population is mixed between at least two ancestral populations that are differentially related (perhaps anciently) to Source1 and Source2. If we detect a significantly non-zero value of a statistic of the form f_4(*A*, *B*; *C*, *D*) we can be confident that populations *A* and *B* (or *C* and *D*) are not consistent with being descended from a homogeneous ancestral population that split earlier in time from the ancestors of the other two groups. A significantly positive value of an f_4-statistic of the form f_4(*A*, *B*; *C*, *D*) implies an excess allele sharing between populations *A* and *C* or *B* and *D*, while a negative value implies sharing between populations *B* and *C*, or *A* and *D*.

F_{ST} computation

We estimated F_{ST} using the smartpca program of EIGENSOFT[10] with default parameters and fstonly: YES and inbreed: YES. The populations and groupings used in this analysis are shown in Supplementary Table 9.

Admixture graph modelling

We modelled population relationships and admixture with qpGraph in ADMIXTOOLS[16] using Mbuti as an outgroup. We computed f_2-, f_3- and f_4-statistics measuring allele sharing of two, three or four sets of populations and reported the maximum $|Z|$-score between predicted and observed values. We ranked models that passed according to this metric based on relative likelihood (Supplementary Information section 3).

Determining a minimum number of streams of ancestry

We used qpWave[3,29] as implemented in ADMIXTOOLS[16] to test whether a set of test populations is consistent with being related via *N* streams of ancestry from a set of outgroup populations. In qpWave, a test for rank *N*, implemented as a single hypothesis Hotelling T^2 test, means that we are evaluating whether the test populations are consistent with descending from as few as $N + 1$ sources of ancestry.

Inferring mixture proportions without an explicit phylogeny

We used qpAdm[3,29] as implemented in ADMIXTOOLS[16] to estimate mixture proportions for a Test population as a combination of *N* 'reference' populations by exploiting (but not explicitly modelling) shared genetic drift with a set of 'Outgroup' populations. We compute standard errors with a block jackknife and a *P* value for fit using a single hypothesis Hotelling T^2 test.

Weighted linkage disequilibrium analysis

Linkage disequilibrium decay was calculated using ALDER[11] to infer admixture parameters including dates and mixture proportions, with a standard error computed as a block jackknife over chromosomes.

MSMC and MCMC2

We used MSMC[48] as previously described[72] to infer cross-coalescence rates and population sizes among Ami and Atayal, Tibetan and Ulchi. We also ran MCMC2 as described previously[49].

Kinship analysis

We used the READ software[73] as well as a custom method[65] to determine genetic kinship between individual pairs.

Detecting runs of homozygosity

We detected runs of homozygosity in ancient DNA using the hapROH software as described previously[74].

Reporting summary

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

Data availability

The aligned sequences are available through the European Nucleotide Archive under accession number PRJEB42781. The newly generated genotype data of 383 modern East Asian individuals have been deposited in Zenodo (https://doi.org/10.5281/zenodo.4058532). The previously published data co-analysed with our newly reported data can be obtained as described in the original publications, which are all referenced in Supplementary Table 4; a compiled dataset that includes the merged genotypes used in this paper is available as the Allen Ancient DNA Resource at https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data. Any other relevant data are available from the corresponding authors upon reasonable request.

50. Pinhasi, R., Fernandes, D. M., Sirak, K. & Cheronet, O. Isolating the human cochlea to generate bone powder for ancient DNA analysis. *Nat. Protocols* **14**, 1194–1205 (2019).
51. Sirak, K. A. et al. A minimally-invasive method for sampling human petrous bones from the cranial base for ancient DNA analysis. *Biotechniques* **62**, 283–289 (2017).
52. Dabney, J. et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl Acad. Sci. USA* **110**, 15758–15763 (2013).
53. Korlević, P. et al. Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *Biotechniques* **59**, 87–93 (2015).
54. Rohland, N., Glocke, I., Aximu-Petri, A. & Meyer, M. Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. *Nat. Protocols* **13**, 2447–2461 (2018).
55. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Phil. Trans. R. Soc. Lond. B* **370**, 20130624 (2015).
56. Gansauge, M. T. & Meyer, M. Selective enrichment of damaged DNA molecules for ancient genome sequencing. *Genome Res.* **24**, 1543–1549 (2014).
57. Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
58. Maricic, T., Whitten, M. & Pääbo, S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE* **5**, e14004 (2010).

59. Rohland, N. & Hofreiter, M. Ancient DNA extraction from bones and teeth. *Nat. Protocols* **2**, 1756–1762 (2007).
60. John, J. S. SeqPrep. *GitHub* https://github.com/jstjohn/SeqPrep (2011).
61. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
62. Behar, D. M. et al. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90**, 675–684 (2012).
63. Weissensteiner, H. et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–W63 (2016).
64. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* **15**, e1008302 (2019).
65. Kennett, D. J. et al. Archaeogenomic evidence reveals prehistoric matrilineal dynasty. *Nat. Commun.* **8**, 14115 (2017).
66. Lohse, J. C., Madsen, D. B., Culleton, B. J. & Kennett, D. J. Isotope paleoecology of episodic mid-to-late Holocene bison population expansions in the southern Plains, U.S.A. *Quat. Sci. Rev.* **102**, 14–26 (2014).
67. Reimer, P. J. et al. The IntCal20 Northern Hemisphere radiocarbon age calibration curve (0–55 cal kBP). *Radiocarbon* **62**, 725–757 (2020).
68. Bronk Ramsey, C. Bayesian analysis of radiocarbon dates. *Radiocarbon* **51**, 337–360 (2009).
69. Rasmussen, M. et al. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**, 94–98 (2011).
70. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
71. Busing, F. T. A., Meijer, E. & van der Leeden, R. Delete-*m* jackknife for unequal *m*. *Stat. Comput.* **9**, 3–8 (1999).
72. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
73. Monroy Kuhn, J. M., Jakobsson, M. & Günther, T. Estimating genetic kin relationships in prehistoric populations. *PLoS ONE* **13**, e0195491 (2018).
74. Ringbauer, H., Novembre, J. & Steinruecken, M. Human parental relatedness through time — detecting runs of homozygosity in ancient DNA. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.31.126912 (2020).

Acknowledgements We thank D. Anthony, O. Bar-Yosef, K. Brunson, R. Flad, P. Flegontov, Q. Fu, W. Haak, I. Lazaridis, M. Lipson, I. Mathieson, R. Meadow, I. Olalde, N. Patterson, P. Skoglund, D. Xu, P. Bellwood and C. Chiang for comments; N. Saitou and the Asian DNA Repository Consortium for sharing genotype data from present-day Japanese groups; T. Nishimoto and T. Fujisawa from the Rebun Town Board of Education for sharing the Funadomari Jomon samples, and H. Tanaka and W. Nagahara from the Archeological Center of Chiba City, who are excavators of the Rokutsu Jomon site. The excavations at Boisman-2 site (Boisman culture), the Pospelovo-1 site (Yankovsky culture) and the Roshino-4 site (Heishui Mohe culture) were funded by the Far Eastern Federal University and the Institute of History, Archaeology and Ethnology Far Eastern Branch of the Russian Academy of Sciences; research on Pospelovo-1 is funded by RFBR project number 18-09-40101. C.-C.W. was funded by the Max Planck Society, the National Natural Science Foundation of China (NSFC 31801040), the Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), the Major project of National Social Science Foundation of China (20&ZD248), a European Research Council (ERC) grant to D. Xu (ERC-2019-ADG-883700-TRAM) and Fundamental Research Funds for the Central Universities (ZK1144). H.M. was supported by grant JSPS 16H02527. M.R. and C.-C.W. received funding from the ERC under the European Union's Horizon 2020 research and innovation program (grant no. 646612) to M.R. H. Li was funded NSFC (91731303, 31671297), B&R International Joint Laboratory of Eurasian Anthropology (18490750300). J.K. was funded by DFG grant KR 4015/1-1, the Baden Württemberg Foundation and the Max Planck Institute. Accelerator Mass Spectrometry radiocarbon dating work was supported by the National Science Foundation (NSF) (BCS-1460369) to D.J.K. and B.J.C. D.R. was funded by NSF grant BCS-1032255, NIH (NIGMS) grant GM100233, the Paul M. Allen Frontiers Group, John Templeton Foundation grant 61220, a gift from J.-F. Clin and the Howard Hughes Medical Institute.

Author contributions C.-C.W., H.-Y.Y., A.N.P., H.M., A.M.K., L.J., H. Li, J.K., R.P. and D.R. conceptualized the study. C.-C.W., R.B., M. Mah, S.M., Z.Z., B.J.C. and D.R. carried out the formal analysis; C.-C.W., K. Sirak, O.C., A.K., N.R., A.M.K., M. Mah, S.M., K.W., N.A., N.B., K.C., F.C., K.S.D.C., B.J.C., L.E., S.F., D.K., A.M.L., K.M., M. Michel, J.O., K.T.O., K. Stewardson, S.W., S.Y., F.Z., J.G., Q.D., L.K., Dawei Li, Dongna Li, R.L., W.C., N., R.S., L.-X.W., L.W., G.X., H.Y., M.Z., G.H., X.Y., R.H., S.S., D.J.K., L.J., H. Li, J.K., R.P. and D.R. carried out the investigation. H.-Y.Y., A.N.P., R.B., D.T., J.Z., Y.-C.L., J.-Y.L., M. Mah, S.M., Z.Z., R.C., H. Looh, C.-J.H., C.-C.S., Y.G.N., A.V.T., A.A.T., S.L., Z.-Y.S., X.-M.W., T.-L.Y., X.H., L.C., H.D., J.B., E. Mijiddorj, D.E., T.-O.I., E. Myagmar, H.K.-K., M.N., K.-i.S., O.A.S., D.J.K., R.P. and D.R. provided resources. C.-C.W., K. Sirak, O.C., A.K., N.R., R.B., M. Mah, S.M., B.J.C., L.E., A.A.T. and D.R. curated the data. C.-C.W., H.-Y.Y., A.N.P., H.M., A.K. and D.R. wrote the paper. C.-C.W., H.-Q.Z., N.R., M.R., S.S., D.J.K., L.J., H. Li, J.K., R.P. and D.R. supervised the study.

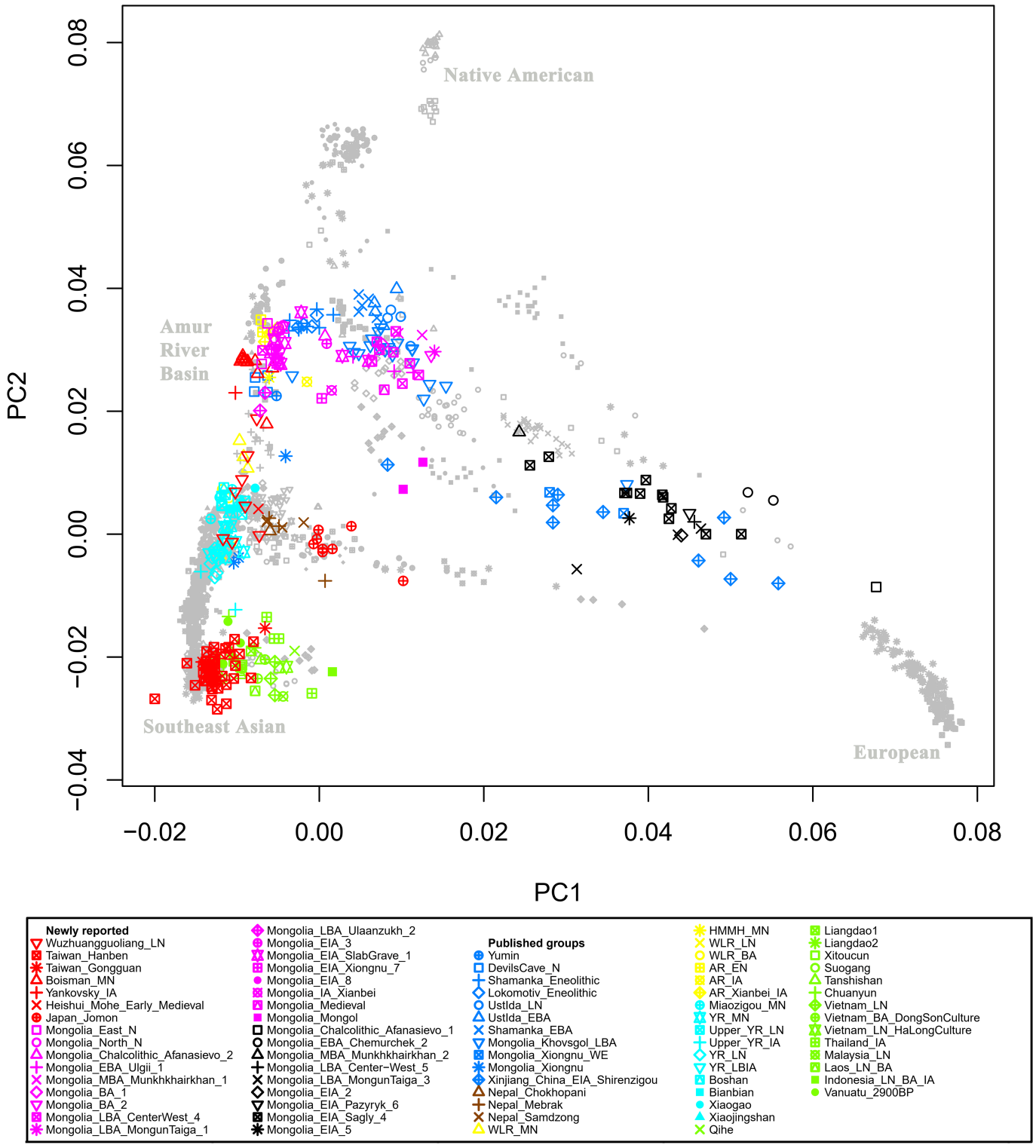
Competing interests The authors declare no competing interests.

Additional information

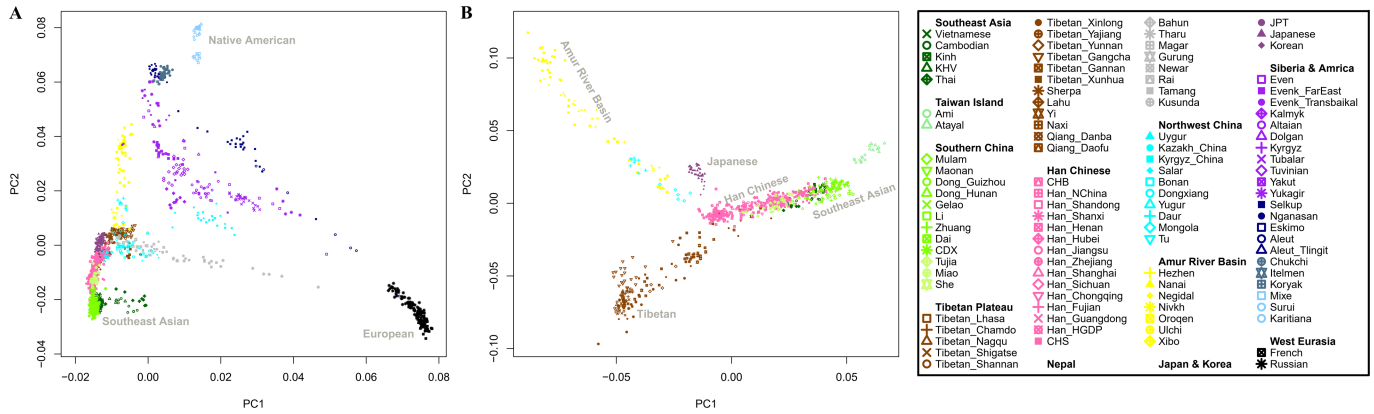
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03336-2.

Correspondence and requests for materials should be addressed to C.-C.W., J.K., R.P. or D.R.

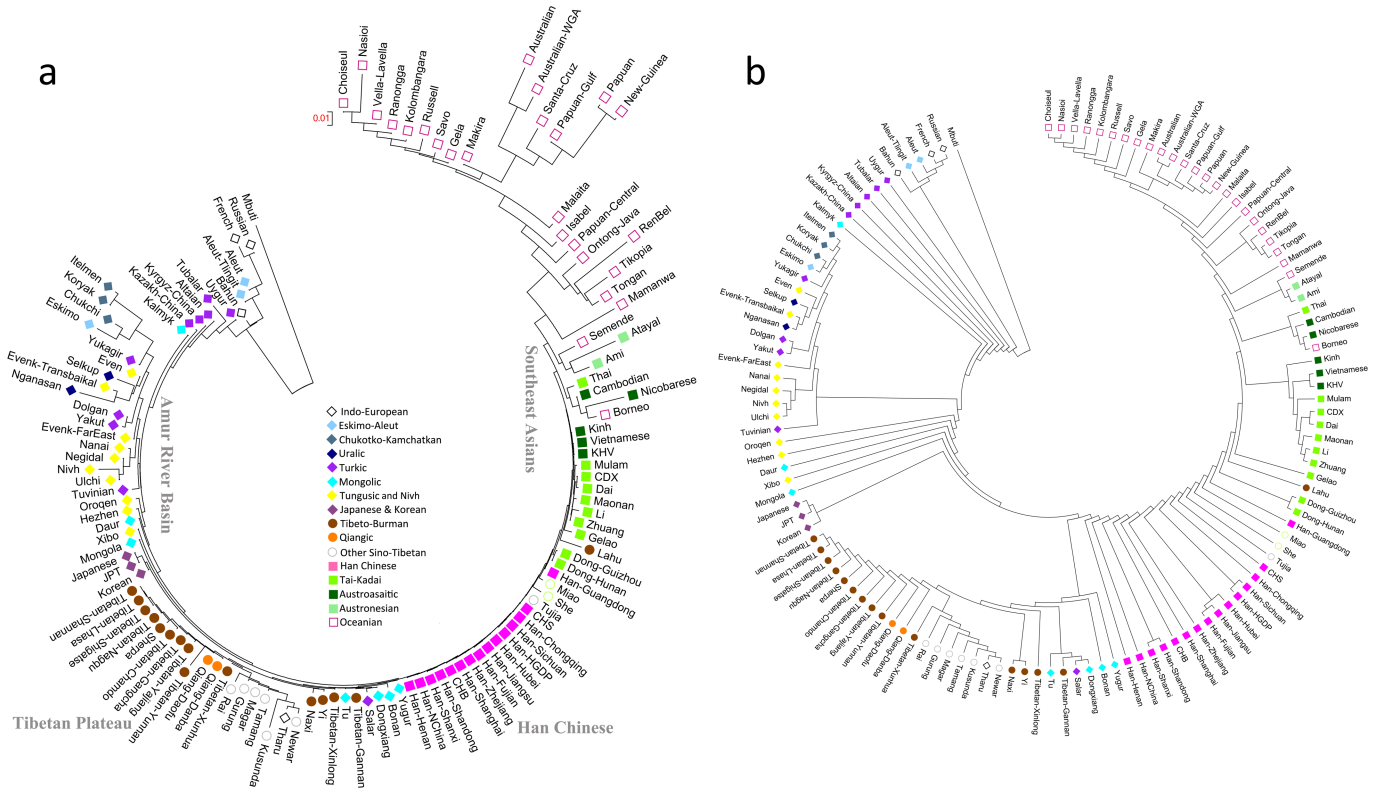
Peer review information *Nature* thanks Peter Bellwood, Charleston Chiang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



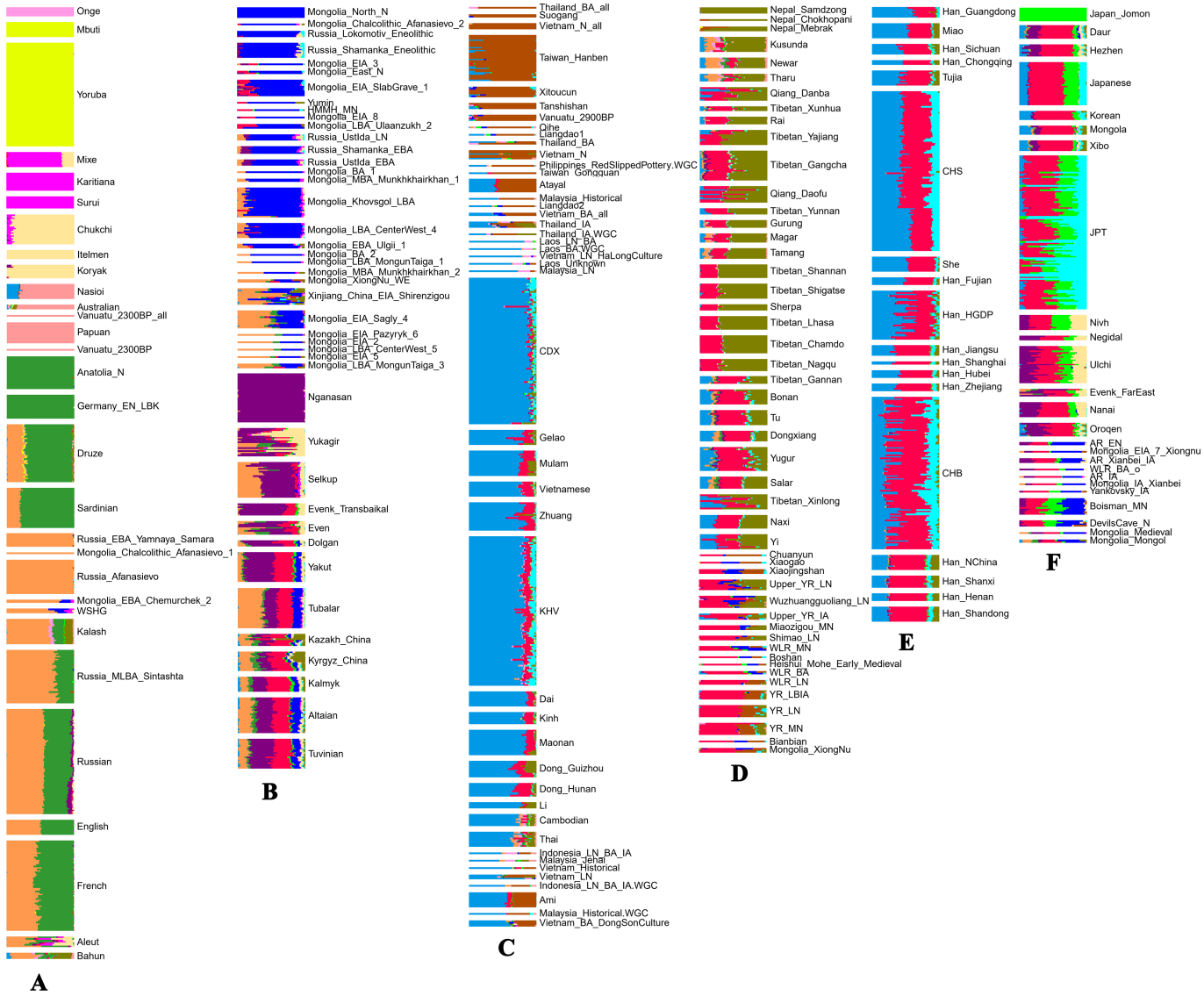
Extended Data Fig. 1 | PCA of ancient samples. Projection of ancient samples onto PCA dimensions 1 and 2 defined by East Asian, European, Siberian and Native American populations.



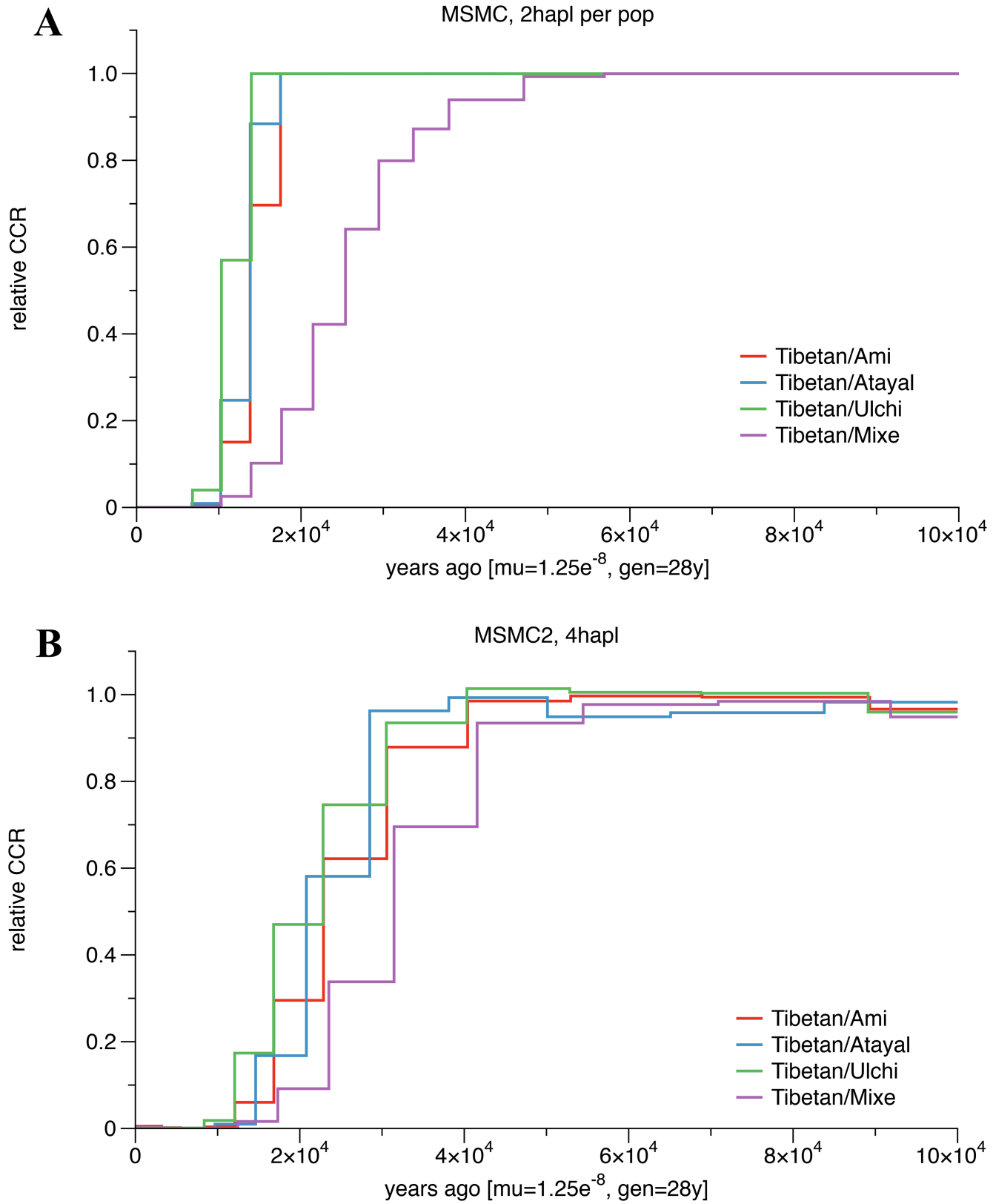
Extended Data Fig. 2 | PCA of present-day samples. a, PCA dimensions 1 and 2 defined by present-day East Asian, European, Siberian and Native American populations. **b**, PCA dimensions 1 and 2 defined by present-day East Asian groups with little West Eurasian mixture.



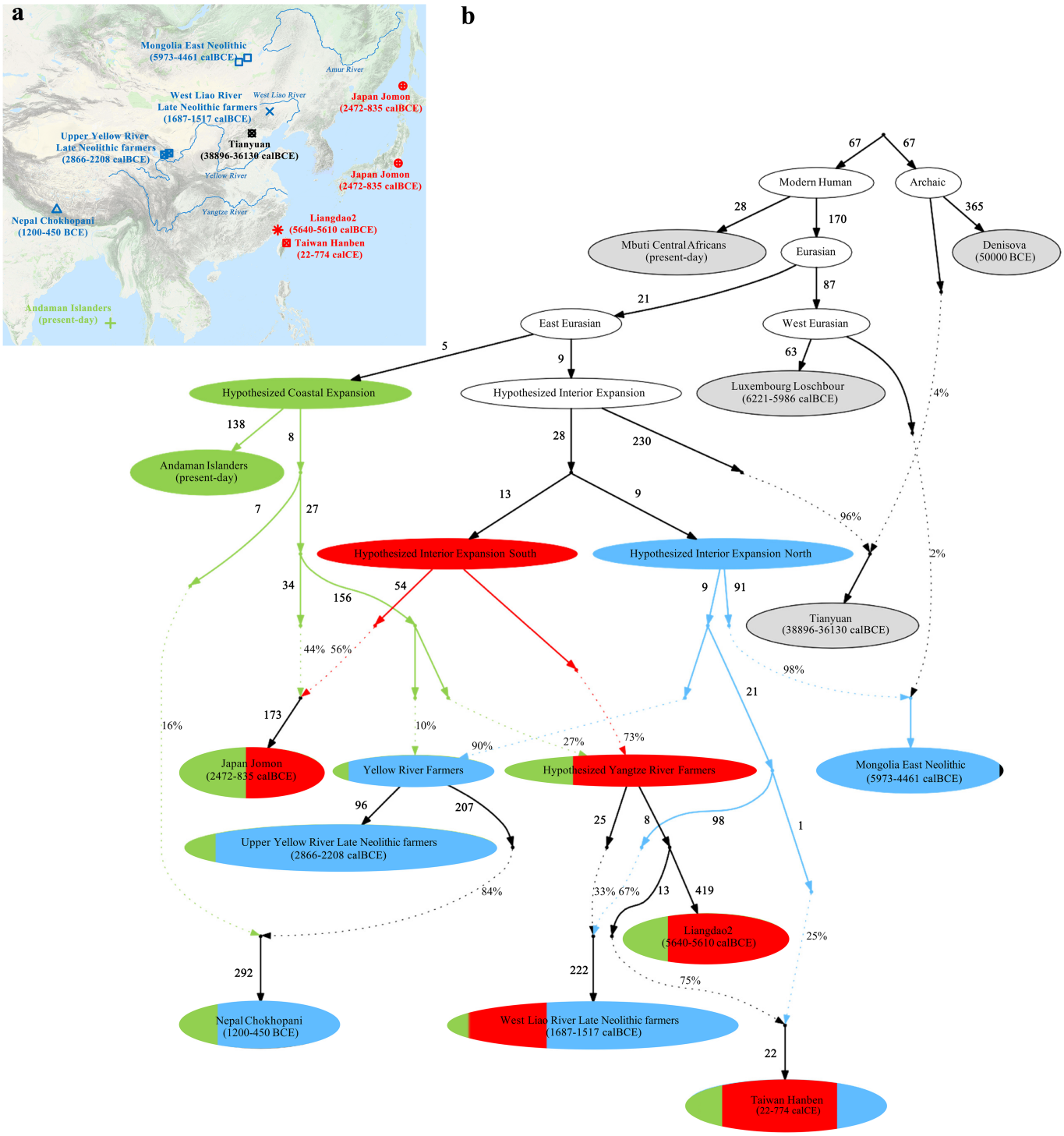
Extended Data Fig. 3 | Neighbour-joining tree of present-day East Eurasian individuals using the human origin dataset. a, Neighbour-joining tree of present-day East Eurasian individuals based on F_{ST} distances using the human origin dataset. The branch length is shown in F_{ST} distance. **b**, Neighbour-joining tree of present-day East Eurasian individuals in which internal branches are all shown with the same branch length for better visualization.



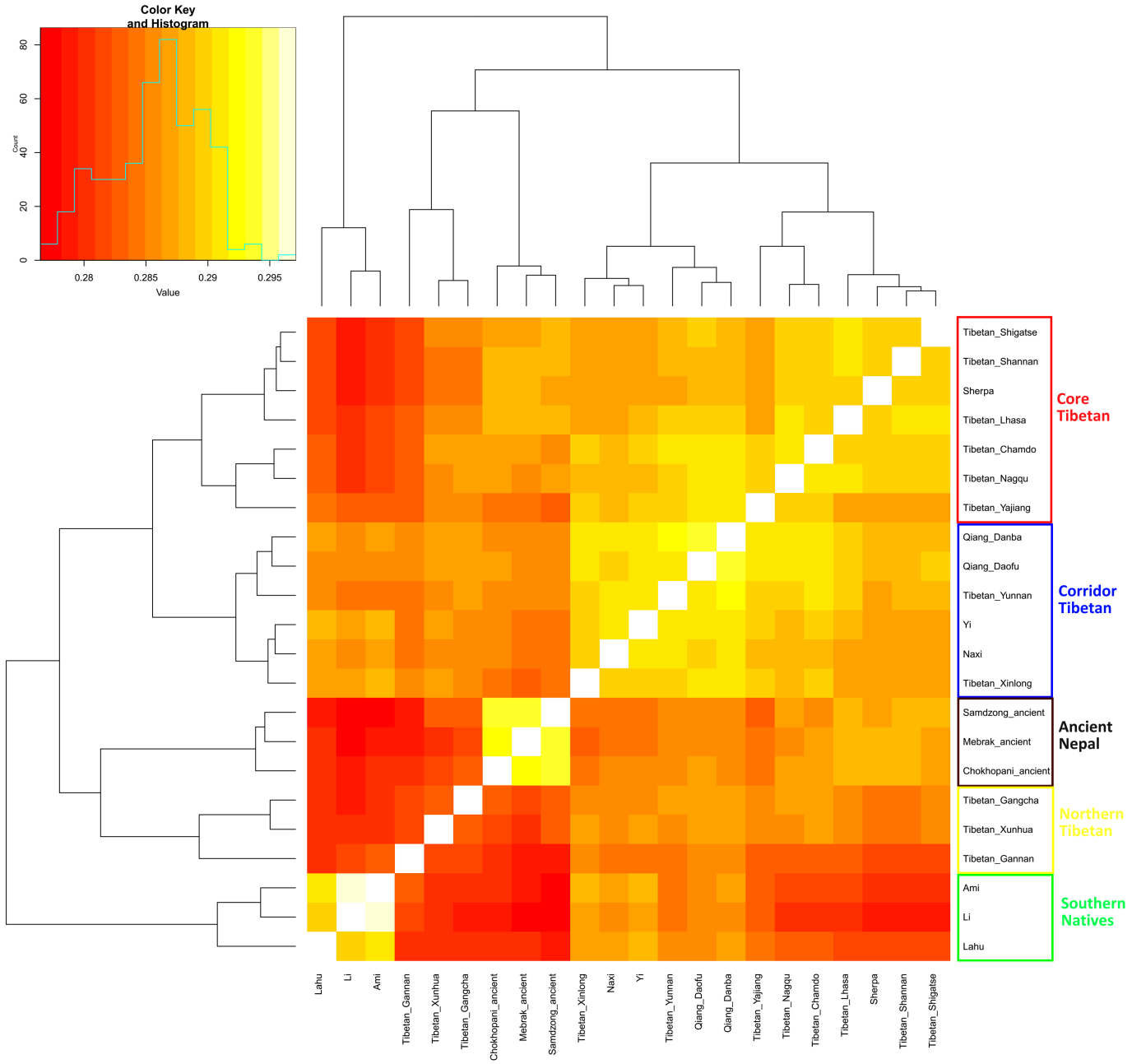
Extended Data Fig. 4 | Admixture plot at $K = 15$ using the human origin dataset. a–**f**, We grouped the populations roughly into six groups based on geographical and genetic affinity. **a**, Populations mainly from Africa (yellow), America (magenta), West Eurasia (dark green and light brown) and Oceania (light magenta). **b**, Populations mainly from Mongolia (blue) and Siberia (purple). **c**, Populations mainly from southern China and Southeast Asia (light blue). **d**, Populations mainly from the Tibetan Plateau (olive) and Neolithic Yellow River Basin (red). **e**, Mainly Han Chinese groups from China (light blue and red). **f**, Populations mainly from the Amur River Basin (blue and red) and northeast Asia.



Extended Data Fig. 5 | Estimates of population split times. a, Cross-coalescence rates for selected population pairs. We ran MSMC for four pairs of populations: Tibetan–Ami, Tibetan–Atayal, Tibetan–Ulchi and Tibetan–Mixe. We used one individual from each population in this analysis. The modern genomic data for those individuals are from the Simons Genome Diversity Project. The times are calculated based on the mutation rate and generation time specified on the *x* axis. **b**, Cross-coalescence rates for selected population pairs. The same analysis as shown in **a** but using MSMC2 instead of MSMC, and using two individuals per population except for the Tibetan–Atayal pair, for which we used only one.



Extended Data Fig. 6 | Admixture graph model. This figure is the same as Fig. 2 except we show the fitted genetic drifts on each lineage. We used all available sites in the dataset comprising 1,237,207 SNPs, restricting to transversions only to confirm that the same model fit (Supplementary Information section 3). We started with a skeleton tree that fits the data for Denisovan, Mbuti, Onge, Tianyuan and Luxembourg Loschbour and one admixture event. We grafted on Mongolia East Neolithic, Late Neolithic farmers from the Upper Yellow River, Liangdao 2, Japan Jomon, Nepal Chokhopani, Taiwan Hanben and Late Neolithic farmers from the West Liao River in turn, adding them consecutively to all possible edges in the tree and retaining only graph solutions that provided no differences of $|Z| < 3$ between fitted and estimated statistics (maximum $|Z| = 2.95$ here). We used the MSMC and MSMC2 relative population split time estimates to constrain models. Deep splits are not well constrained because of the minimal availability of data on East Asian populations from the Upper Paleolithic. **a**, Locations and dates of the East Asian individuals used in model fitting, with colours indicating whether the majority ancestry is from the hypothesized coastal expansion (green), interior expansion south (red) and interior expansion north (blue). The map is based on the 'Google Map Layer' from ArcGIS Online Basemaps (map data ©2020 Google). The grey circles represent sampled populations and white circles represent unsampled hypothesized nodes. **b**, In the model visualization, we colour lineages modelled as deriving entirely from one of these expansions, and also colour populations according to ancestry proportions. Dashed lines represent admixture (proportions are marked), and we show the amount of genetic drift on each lineage in units of $F_{ST} \times 1{,}000$.



Extended Data Fig. 7 | Shared genetic drift among Tibetan groups, measured by $f_3(X, Y;$ Mbuti). Lighter colours indicate more shared drift. Lahu groups with the Southeast Asian cluster probably due to substantial admixture. The Tibetan_Yajiang are geographically in the Tibeto-Burman Corridor but group with Core Tibetan individuals, presumably reflecting less genetic admixture from people of the Southeast Asian cluster.

Extended Data Table 1 | Population information for newly genotyped present-day individuals

Population	Language	Location	Latitude	Longitude	N
Tibetan	Tibetic, Sino-Tibetan	Chamdo, Tibet, China	31.1	97.2	12
Tibetan	Tibetic, Sino-Tibetan	Gangcha, Qinghai, China	37.3	100.2	20
Tibetan	Tibetic, Sino-Tibetan	Gannan, Gansu, China	35	102.9	5
Tibetan	Tibetic, Sino-Tibetan	Lhasa, Tibet, China	30	91.1	9
Tibetan	Tibetic, Sino-Tibetan	Nagqu, Tibet, China	31.5	92.1	7
Tibetan	Tibetic, Sino-Tibetan	Shannan, Tibet, China	29.2	91.8	10
Tibetan	Tibetic, Sino-Tibetan	Shigatse, Tibet, China	29.3	88.9	10
Tibetan	Tibetic, Sino-Tibetan	Xinlong, Sichuan, China	31	100.3	10
Tibetan	Tibetic, Sino-Tibetan	Xunhua, Qinghai, China	35.8	102.5	4
Tibetan	Tibetic, Sino-Tibetan	Yajiang, Sichuan, China	30	101	10
Tibetan	Tibetic, Sino-Tibetan	Yunnan, China	27.8	99.7	4
Qiang	Qiangic, Sino-Tibetan	Daofu, Sichuan, China	30.9	101.1	11
Qiang	Qiangic, Sino-Tibetan	Danba, Sichuan, China	30.8	101.9	9
Han	Chinese, Sino-Tibetan	Chongqing, China	29.3	106.3	3
Han	Chinese, Sino-Tibetan	Fujian, China	26.1	119.3	5
Han	Chinese, Sino-Tibetan	Guangdong, China	23.2	113.2	7
Han	Chinese, Sino-Tibetan	Henan, China	34.8	113.6	5
Han	Chinese, Sino-Tibetan	Hubei, China	30.5	114.3	5
Han	Chinese, Sino-Tibetan	Jiangsu, China	32.1	118.8	7
Han	Chinese, Sino-Tibetan	Shandong, China	36.6	117	10
Han	Chinese, Sino-Tibetan	Shanghai, China	31.2	121.5	2
Han	Chinese, Sino-Tibetan	Shanxi, China	37.9	112.5	8
Han	Chinese, Sino-Tibetan	Sichuan, China	30.7	104.1	7
Han	Chinese, Sino-Tibetan	Zhejiang, China	30.3	120.2	5
Zhuang	Tai, Tai-Kadai	Guangxi, China	22.8	108.4	22
Li	Hlai, Tai–Kadai	Hainan, China	18.5	110	4
Dong	Kam-Sui, Tai–Kadai	Guizhou, China	26.7	106.6	13
Dong	Kam-Sui, Tai–Kadai	Hunan, China	27.4	109.2	7
Mulam	Kam-Sui, Tai–Kadai	Luocheng, Guangxi, China	24.8	108.9	17
Maonan	Kam-Sui, Tai–Kadai	Huanjiang, Guangxi, China	24.8	108.3	17
Gelao	Kra, Tai-Kadai	Longlin, Baise, Guangxi, China	24.8	105.3	10
Bonan	Mongolic	Jishishan, Gansu, China	35.7	102.8	10
Dongxiang	Mongolic	Linxia, Gansu, China	35.6	103.2	7
Yugur-Eastern	Mongolic	Sunan, Gansu, China	38.9	99.6	16
Kazakh	Kipchak, Turkic	Kazak Autonomous County of Aksay, Gansu, China	38.5	94.3	8
Kyrgyz	Kipchak, Turkic	Urumqi, Xinjiang, China	43.8	87.7	13
Yugur-Western	Turkic	Sunan, Gansu, China	38.9	99.6	1
Salar	Oghuz, Turkic	Xunhua, Qinghai, China	35.8	102.5	8
Bahun	Nepali, Indo-European	Nepal	27.4	85.3	5
Gurung	Tamangic, Sino-Tibetan	Nepal	27.4	86.2	5
Magar	Magaric, Sino-Tibetan	Nepal	27.4	86.2	6
Newar	Sino-Tibetan	Nepal	27.4	85.3	8
Rai	Kiranti/Nepali	Nepal	27.4	85.3	5
Sherpa	Tibetic, Bodish, Sino-Tibetan	Nepal	27.4	85.3	4
Tamang	Tamangic, Sino-Tibetan	Nepal	27.4	86.2	8
Tharu	Indo-Aryan, Indo-European	Nepal	27.4	86.2	5

Extended Data Table 2 | Kinship detected between pairs of individuals

Region	Site	Family ID	N	Individuals	Relationship	Date
Japan	Rokutsu	Rokutsu.Family	2	I13886-I13887	Brothers	2136-1982 calBCE [intersection]
China	Wuzhuangguoliang	Wuzh.Family1	2	S95-S97	1st degree relatives	3400-2800 BCE
Taiwan	Hanben	Hanben.Family1	2	I3611-I3612	2nd or 3rd degree relatives	133-324 calCE [based on I3611]
Taiwan	Hanben	Hanben.Family2	2	I15156-I8072	1st degree relatives	1-800 CE
Taiwan	Hanben	Hanben.Family3	3	I8078-I3735-I3734	I8078-I1375 1st degree relatives; I3734 is a 2-3rd relative of I8078	376-532 calCE [based on I3735]
Russia	Boisman-2	Boisman.Family1	6	I3356-I14819-I14771-I14772-I14773-I14774	father-mother-son-daughter-son2-daughter2	3705-3633 calBCE [based on I3356]
Russia	Boisman-2	Boisman.Family2	2	I1206-I1192	1st degree relatives	4935-4803 calBCE [intersection]
Russia	Boisman-2	Boisman.Family3	2	I14307-I14308	1st degree relatives	4841-4706 calBCE [based on I14308]
Mongolia	Marzyn	Marzyn.Family	3	I11696-I11697-I11698	2nd or 3rd degree relatives	5620-5484 calBCE [intersection]
Mongolia	Ulaangom	Ulaangom.Family1	2	I7029-I6230	father-son	346-172 calBCE [intersection]
Mongolia	Ulaangom	Ulaangom.Family2	2	I6231-I6232	2nd or 3rd degree relatives	357-208 calBCE [intersection]
Mongolia	Ulaangom	Ulaangom.Family3	2	I12970-I7028	1st or 2nd degree relatives	382-231 calBCE [intersection]
Mongolia	Ulaangom	Ulaangom.Family4	2	I6224-I6225	siblings	370-197 calBCE [based on I6224]