ARTICLES
https://doi.org/10.1038/s41559-019-0878-2
nature ecology & evolution

The genetic history of admixture across inner Eurasia

Choongwon Jeong[1,2,33,34*], Oleg Balanovsky[3,4,34], Elena Lukianova[3], Nurzhibek Kahbatkyzy[5,6], Pavel Flegontov[7,8], Valery Zaporozhchenko[3,4], Alexander Immel[1], Chuan-Chao Wang[1,9], Olzhas Ixan[5], Elmira Khussainova[5], Bakhytzhan Bekmanov[5,6], Victor Zaibert[10], Maria Lavryashina[11], Elvira Pocheshkhova[12], Yuldash Yusupov[13], Anastasiya Agdzhoyan[3,4], Sergey Koshel[14], Andrei Bukin[15], Pagbajabyn Nymadawa[16], Shahlo Turdikulova[17], Dilbar Dalimova[17], Mikhail Churnosov[18], Roza Skhalyakho[4], Denis Daragan[4], Yuri Bogunov[3,4], Anna Bogunova[4], Alexandr Shtrunov[4], Nadezhda Dubova[19], Maxat Zhabagin[20,21], Levon Yepiskoposyan[22], Vladimir Churakov[23], Nikolay Pislegin[23], Larissa Damba[24], Ludmila Saroyants[25], Khadizhat Dibirova[3,4], Lubov Atramentova[26], Olga Utevska[26], Eldar Idrisov[27], Evgeniya Kamenshchikova[4], Irina Evseeva[28], Mait Metspalu[29], Alan K. Outram[30], Martine Robbeets[2], Leyla Djansugurova[5,6], Elena Balanovska[4], Stephan Schiffels[1], Wolfgang Haak[1], David Reich[31,32] and Johannes Krause[1*]

The indigenous populations of inner Eurasia—a huge geographic region covering the central Eurasian steppe and the northern Eurasian taiga and tundra—harbour tremendous diversity in their genes, cultures and languages. In this study, we report novel genome-wide data for 763 individuals from Armenia, Georgia, Kazakhstan, Moldova, Mongolia, Russia, Tajikistan, Ukraine and Uzbekistan. We furthermore report additional damage-reduced genome-wide data of two previously published individuals from the Eneolithic Botai culture in Kazakhstan (~5,400 BP). We find that present-day inner Eurasian populations are structured into three distinct admixture clines stretching between various western and eastern Eurasian ancestries, mirroring geography. The Botai and more recent ancient genomes from Siberia show a decrease in contributions from so-called 'ancient North Eurasian' ancestry over time, which is detectable only in the northern-most 'forest-tundra' cline. The intermediate 'steppe-forest' cline descends from the Late Bronze Age steppe ancestries, while the 'southern steppe' cline further to the south shows a strong West/South Asian influence. Ancient genomes suggest a northward spread of the southern steppe cline in Central Asia during the first millennium BC. Finally, the genetic structure of Caucasus populations highlights a role of the Caucasus Mountains as a barrier to gene flow and suggests a post-Neolithic gene flow into North Caucasus populations from the steppe.

Present-day human population structure is often marked by a correlation between geographic and genetic distances[1,2], reflecting continuous gene flow among neighbouring groups—a process known as 'isolation by distance'. However, there are also striking failures of this model, whereby geographically proximate populations can be quite distantly related. Such barriers to gene flow often correspond to major geographic features, such as the Himalayas[3] or the Caucasus Mountains[4]. Many cases also suggest the presence of social barriers to gene flow. For example, early Neolithic farming populations in Central Europe show a remarkable genetic

[1]Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany. [2]Eurasia3angle Research Group, Max Planck Institute for the Science of Human History, Jena, Germany. [3]Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia. [4]Federal State Budgetary Institution 'Research Centre for Medical Genetics', Moscow, Russia. [5]Department of Population Genetics, Institute of General Genetics and Cytology, Science Committee, Ministry of Education and Science of the Republic of Kazakhstan, Almaty, Kazakhstan. [6]Department of Molecular Biology and Genetics, Faculty of Biology and Biotechnology, Al-Farabi Kazakh National University, Almaty, Kazakhstan. [7]Department of Biology and Ecology, Faculty of Science, University of Ostrava, Ostrava, Czech Republic. [8]Faculty of Science, University of South Bohemia and Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic. [9]Department of Anthropology and Ethnology, Xiamen University, Xiamen, China. [10]Institute of Archeology and Steppe Civilization, Al-Farabi Kazakh National University, Almaty, Kazakhstan. [11]Kemerovo State Medical University, Kemerovo, Russia. [12]Kuban State Medical University, Krasnodar, Russia. [13]Institute of Strategic Research of the Republic of Bashkortostan, Ufa, Russia. [14]Faculty of Geography, Lomonosov Moscow State University, Moscow, Russia. [15]Transbaikal State University, Chita, Russia. [16]Mongolian Academy of Medical Sciences, Ulaanbaatar, Mongolia. [17]Center for Advanced Technologies, Ministry of Innovational Development, Tashkent, Uzbekistan. [18]Belgorod State University, Belgorod, Russia. [19]Institute of Ethnology and Anthropology, Russian Academy of Sciences, Moscow, Russia. [20]National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan. [21]National Center for Biotechnology, Astana, Kazakhstan. [22]Laboratory of Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences, Yerevan, Armenia. [23]Udmurt Institute of History, Language and Literature, Udmurt Federal Research Center, Ural Branch, Russian Academy of Sciences, Izhevsk, Russia. [24]Research Institute of Medical and Social Problems and Control, Healthcare Department of Tuva Republic, Kyzyl, Russia. [25]Leprosy Research Institute, Astrakhan, Russia. [26]V. N. Karazin Kharkiv National University, Kharkiv, Ukraine. [27]Astrakhan Branch, Russian Presidential Academy of National Economy and Public Administration under the President of the Russian Federation, Astrakhan, Russia. [28]Northern State Medical University, Arkhangelsk, Russia. [29]Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu, Estonia. [30]Department of Archaeology, University of Exeter, Exeter, UK. [31]Department of Genetics, Harvard Medical School, Boston, MA, USA. [32]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, USA. [33]Present address: School of Biological Sciences, Seoul National University, Seoul, Republic of Korea. [34]These authors contributed equally: Choongwon Jeong, Oleg Balanovsky. *e-mail: jeong@shh.mpg.de; krause@shh.mpg.de

homogeneity, suggesting minimal genetic exchange with local hunter-gatherer populations through the initial expansion; mixing of these two gene pools became evident only after thousands of years in the middle Neolithic[5]. Present-day Lebanese populations provide another example by showing a population stratification reflecting their religious community[6]. There are also examples of geographically very distant populations that are closely related; for example, people buried in association with artefacts of the Yamnaya horizon in the Pontic–Caspian steppe and the contemporaneous Afanasievo culture 3,000 km east in the Altai–Sayan Mountains[7,8].

The vast region of the Eurasian inland ('inner Eurasia' herein) is split into distinct ecoregions, such as the Eurasian steppe in Central Eurasia, boreal forests (taiga) in Northern Eurasia, and the Arctic tundra at the periphery of the Arctic Ocean (Fig. 1). These ecoregions stretch in an east–west direction within relatively narrow north–south bands. Various cultural features show a distribution that broadly mirrors the ecogeographic distinction in inner Eurasia. For example, indigenous peoples of the Eurasian steppe traditionally practise nomadic pastoralism[9,10], while Northern Eurasian peoples in the taiga mainly rely on reindeer herding and hunting[11]. The subsistence strategies in each of these ecoregions are often considered to be adaptations to the local environments[12].

At present, there is limited information about how environmental and cultural influences are mirrored in the genetic structure of inner Eurasians. Recent genome-wide studies of inner Eurasians mostly focused on detecting and dating genetic admixture in individual populations[13–16]. So far, only three studies have reported recent genetic sharing between geographically distant populations based on the analysis of 'identity-by-descent' segments[13,17,18]. One study reports long-distance sharing of large chromosomal pieces between Turkic populations based on a detailed comparison between Turkic-speaking groups and their non-Turkic neighbours[13]. The other two studies extend this approach to some Uralic and Yeniseian-speaking populations[17,18]. However, a comprehensive spatial genetic analysis of inner Eurasian populations is still lacking.

Ancient DNA studies have already shown that human populations of this region have dramatically transformed over time. For example, the Upper Palaeolithic genomes from the Mal'ta and Afontova Gora sites in Southern Siberia revealed a genetic profile, often called Ancient North Eurasians (ANE), which is deeply related to Palaeolithic/Mesolithic hunter-gatherers in Europe and also substantially contributed to the gene pools of present-day Native Americans, Siberians, Europeans and South Asians[19,20]. Studies of Bronze Age steppe populations found the appearance of additional Western Eurasian-related ancestries across the steppe from the Pontic–Caspian to the Altai–Sayan regions. Here, we collectively refer to them as Western Steppe herders (WSHs): the earlier populations associated with the Yamnaya and Afanasievo cultures (often called 'steppe Early and Middle Bronze Age') and the later ones associated with many cultures, such as Potapovka, Sintashta, Srubnaya and Andronovo, to name a few (often called 'steppe Middle and Late Bronze Age')[8]. The steppe Middle and Late Bronze Age gene pool was largely descended from the preceding steppe Early and Middle Bronze Age gene pool, with a substantial contribution from Late Neolithic Europeans[21]. Also, recent archaeogenetic studies trace multiple large-scale trans-Eurasian migrations over the past several millennia using ancient inner Eurasian genomes[22,23], including individuals from the Eneolithic Botai culture in Northern Kazakhstan in the fourth millennium BC[24]. These studies now provide a rich context for interpretation of the present-day population structure of inner Eurasians and characterization of ancient admixtures in fine resolution.

In this study, we analysed newly produced genome-wide data for 763 individuals belonging to 60 self-reported ethnic groups to provide a dense portrait of the genetic structure of inner Eurasians. We also produced damage-reduced genome-wide data of two ancient Botai individuals whose genome-wide data were recently published[23], to explore the genetic structure of pre-Bronze Age populations in inner Eurasia (Table 1). We aimed to characterize the genetic composition of inner Eurasians in fine resolution by applying both allele frequency- and haplotype-based methods. Based on the fine-scale genetic profile, we further explored whether and where barriers and conduits of gene flow exist in inner Eurasia.

Results

Present-day inner Eurasians form distinct east–west genetic clines mirroring geography. We generated genome-wide genotype data of 763 participants who represent a majority of large ethnic groups in Armenia, Georgia, Kazakhstan, Moldova, Mongolia, Russia, Tajikistan, Ukraine and Uzbekistan (Fig. 1 and Supplementary Table 1). We merged new data with published data of present-day[20,25,26] and ancient individuals[3,8,19–23,27–42] (Supplementary Table 2). The final dataset covers 581,230 autosomal single nucleotide polymorphisms (SNPs) in the Affymetrix Axiom Genome-wide Human Origins 1 (HumanOrigins) array platform[43].

In a principal component analysis (PCA) of Eurasian individuals, we find that PC1 separates Eastern and Western Eurasian populations, PC2 splits Eastern Eurasians along a north–south cline, and PC3 captures variation in Western Eurasians with Caucasus and Northeastern European populations at opposite ends (Fig. 2a and Supplementary Figs. 1 and 2). Inner Eurasians are scattered across PC1 in between, mirroring their geographic locations. Strikingly, they seem to be structured into three distinct west–east genetic clines running between different Western and Eastern Eurasian groups, instead of being evenly spaced in principal component space. The uppermost cline, composed of individuals from Northern Eurasia, mostly speaking Uralic or Yeniseian languages, connects Northeast Europeans and the Uralic (Samoyedic)-speaking Nganasans from Northern Siberia. The other two lower clines are occupied by individuals from the Eurasian steppe, mostly speaking Turkic and Mongolic languages. Both clines run into Turkic/Mongolic-speaking populations in Southern Siberia and Mongolia, and further into Tungusic-speaking populations in Manchuria and the Russian Far East in the East; however, they diverge in the west, with one heading to the Caucasus and the other heading to populations of the Volga–Ural area (Fig. 2 and Supplementary Fig. 2). Four groups (Daur, Mongola, Tu and Dungans) are located alongside other East Asian populations and displaced from the three inner Eurasian clines.

A model-based clustering analysis using ADMIXTURE shows a similar pattern (Fig. 2b and Supplementary Fig. 3). Overall, the proportions of ancestry components associated with Eastern or Western Eurasians are well correlated with longitude in inner Eurasians (Fig. 3). Notable outliers include known historical migrants such as Kalmyks, Nogais and Dungans. The Uralic- and Yeniseian-speaking populations, as well as Russians from multiple locations, derive most of their Eastern Eurasian ancestry from a component most enriched in Nganasans, while Turkic/Mongolic speakers have this component together with another component most enriched in populations from the Russian Far East, such as Ulchi and Nivkh (Supplementary Fig. 3). Turkic/Mongolic speakers comprising the bottom-most cline have a distinct Western Eurasian ancestry profile: they have a high proportion of a component most enriched in Mesolithic Caucasus hunter-gatherers[30] and Neolithic Iranians[20] and frequently harbour another component enriched in present-day South Asians (Supplementary Fig. 4). Based on the PCA and ADMIXTURE results, we heuristically assigned inner Eurasians to three clines: the 'forest-tundra' cline includes Russians and all Uralic and Yeniseian speakers; the 'steppe-forest' cline includes Turkic- and Mongolic-speaking populations from the Volga and Altai–Sayan regions and Southern Siberia; and the 'southern steppe' cline includes the rest of the populations.

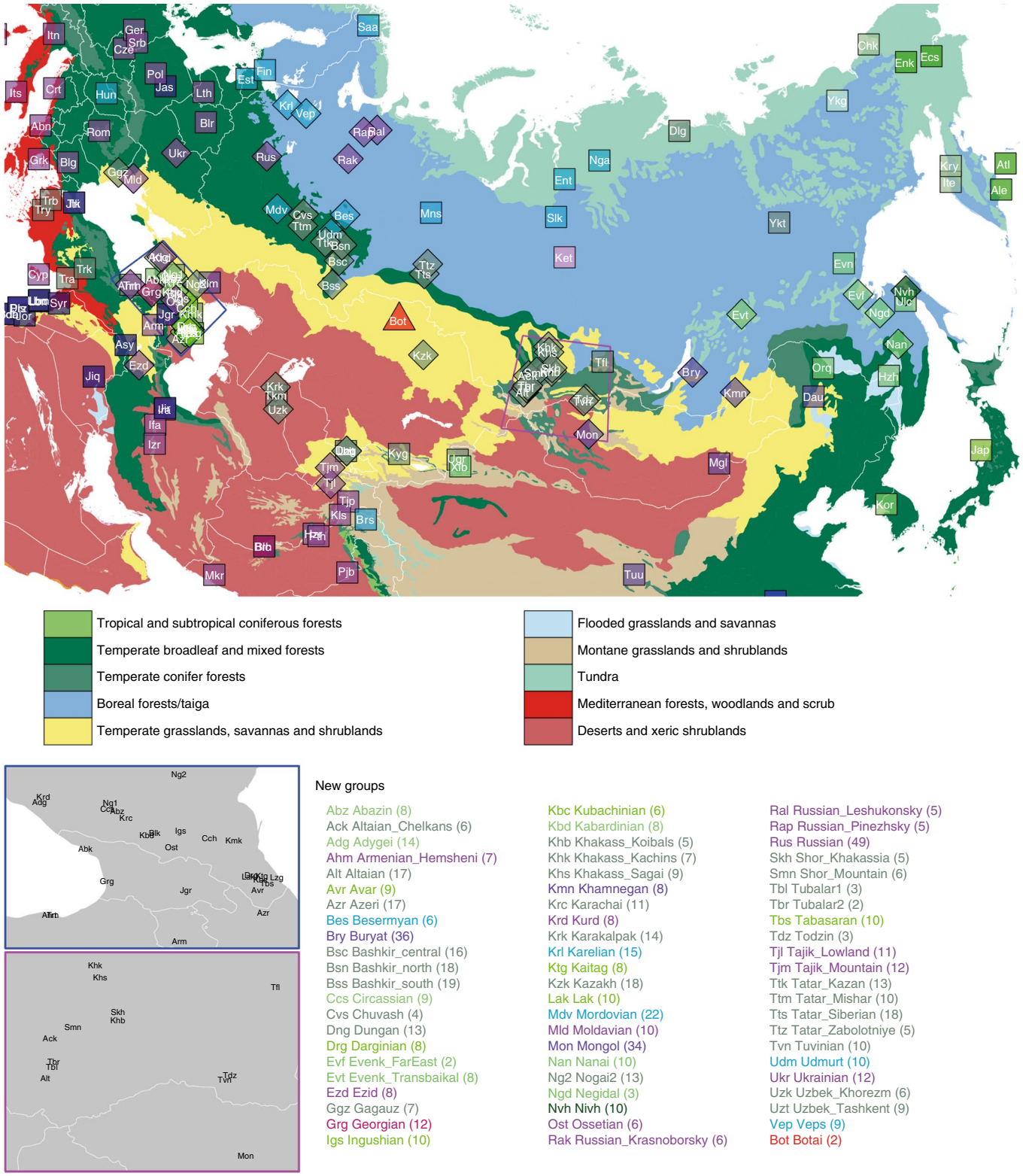

Fig. 1 | Geographic locations of the Eneolithic Botai, groups including newly sampled individuals, and nearby groups with published data. The Eneolithic Botai site is represented by a red triangle. The locations of the groups including newly sampled individuals (diamonds; $n = 65$) and nearby groups with published data (squares) are also shown. Mean latitude and longitude values for all individuals in each group were used. Two magnified plots for the Caucasus (blue) and Altai–Sayan regions (magenta) are included (bottom left). A list of the new groups, their three-letter codes and the number of new individuals (in parentheses) is shown in the bottom right. Present-day populations are colour-coded based on the language family in Figs. 1–3, following the key codes listed in Fig. 2. Corresponding information for the previously published groups, including definitions of the abbreviated codes, is provided in Supplementary Table 2. The map is overlayed with ecoregional information, divided into 14 biomes downloaded from https://ecoregions2017.appspot.com/ (credited to Ecoregions, Resolve). The main inner Eurasia map is on the Albers equal-area projection and was produced using the spTransform function in the R package rgdal version 1.2-5.

Table 1 | Sequencing statistics and radiocarbon dates of two Eneolithic Botai individuals analysed in this study

ID	Published ID	Genetic sex	Uncal. ^{14}C date[a]	Cal. ^{14}C date (2-σ)[b]	Number of reads sequenced	Mean autosomal coverage	Number of SNPs covered[c]	MT/Y haplogroup[d]	MT cont.[e]	X cont.[f]
TU45	BOT14	M	4620 ± 80 BP	3632–3100 BC	84,170,835	0.827×	169,053 (77,363)	K1b2/R1b1a1	0.02 (0.01–0.03)	0.0122 (0.0050)
BKZ001	BOT2016	F	4660 ± 25 BP	3517–3367 BC	69,678,735	2.420×	825,332 (432,078)	Z1/NA	0.01 (0–0.02)	NA

For these Botai individuals, we produced additional data. We provide corresponding individual IDs from a previous publication[23] ('Published ID'), radiocarbon dates, the number of total reads sequenced, mean autosomal coverage for the 1,240 K target sites, the number of SNPs covered at least once for the 1,240 K and HumanOrigins panels, uniparental haplogroups and contamination estimates. [a]The uncalibrated date of TU45 was published by Levine[70] under the ID OxA-4316. [b]Calibrated ^{14}C dates were calculated based on uncalibrated dates, using the OxCal version 4.3.2 program[71] with the INTCAL13 atmospheric curve[72]. [c]Number of SNPs in the 1,240 K panel (out of 1,233,013) or autosomal SNPs in the HumanOrigins array (out of 581,230; within parentheses) covered by at least one read. Only transversion SNPs were considered for the non-UDG libraries (both of the TU45 libraries, and one of two BKZ001 libraries). [d]Mitochondrial and Y chromosome haplogroups. [e]Contamination rate of mitochondrial reads, estimated using the Schmutzi program (95% confidence intervals in parentheses). [f]Nuclear contamination rate for the male (TU45) estimated based on X chromosome data using ANGSD software (s.e. in parentheses).

We separated four groups (Daur, Mongola, Tu and Dungans) as 'others' (Supplementary Table 2).

The genetic barriers splitting the inner Eurasians are also found in the estimated effective migration surface (EEMS) analysis[44] (Supplementary Fig. 5). Inferred barriers to gene flow are often co-localized with geographic features or genetic gaps. We observe a barrier overlapping with the Urals, one separating Beringian populations from the rest, one separating southern Siberians from Central and Northern Siberians, and one separating Caucasus populations from those further to the north. The Southern Siberian barrier matches with our distinction between the steppe-forest and forest-tundra populations, with the exception of the two northern-most Turkic-speaking populations—Yakuts and Dolgans. The Caucasus barrier also matches with our distinction between the southern steppe and steppe-forest populations. A local EEMS analysis on the Caucasus shows fine-scale barriers and conduits of gene flow, matching with the fine-scale structure within Caucasus populations (Supplementary Note 1).

High-resolution tests of admixture distinguish the genetic profile of source populations in the inner Eurasian clines. We performed both allele frequency-based three-population (f_3) tests and a haplotype-sharing-based GLOBETROTTER analysis to characterize the admixed gene pools of inner Eurasian groups. For these group-based analyses, we manually removed 87 outliers based on PCA results (Supplementary Table 1). We also split a few inner Eurasian groups showing genetic heterogeneity into subgroups based on PCA results and their sampling locations (Supplementary Table 1). This was done to minimize false positive admixture signals. Including two Aleut populations as positive control targets, we chose a total of 73 groups as the targets of admixture tests and another 26 groups (167 present-day and 93 ancient groups) as the 'sources' to represent worldwide genetic diversity (Supplementary Table 2).

Testing all possible pairs of 167 present-day 'source' groups as references, we detected highly significant f_3 statistics for 66 of 73 targets ($f_3 \leq -3$ standard error (s.e.); Supplementary Table 3). Negative f_3 values mean that allele frequencies of the target group are, on average, intermediate between those of the references, providing unambiguous evidence that the target population is a mixture of groups related, perhaps deeply, to the source populations[43]. Extending the references to include 93 ancient groups, the remaining 7 groups also have small f_3 statistics around 0 ($f_3 = -5.1$ to $+2.7$ s.e.). Reference pairs with the most negative f_3 statistics for the most part involve one Eastern and one Western Eurasian group, supporting the qualitative impression of east–west admixture from PCA and ADMIXTURE analyses. To highlight the difference between the distinct inner Eurasian clines, we looked into f_3 results with representative reference pairs comprising two ancient Western (Srubnaya to represent the steppe Middle and Late Bronze Age ancestry[21] and Chalcolithic Iranians to represent West/South Asian-related ancestry[20]; Supplementary Table 1) and three Eastern Eurasian groups (Mixe, Nganasan and Ulchi). In the southern steppe cline populations, reference pairs with Chalcolithic Iranians tend to produce more negative f_3 statistics than those with Srubnaya, while the opposite pattern is uniformly observed for the steppe-forest and forest-tundra populations (Fig. 4a). Reference pairs with Nganasans mostly result in more negative f_3 statistic than those with Ulchi in the forest-tundra populations, but the opposite pattern is dominant in the southern steppe populations. The steppe-forest cline populations show an intermediate pattern: seven northern groups (Chuvash, Bashkir_north, Tatar_Zabolotniye, Todzin, Tofalar, Dolgan and Yakut) have more negative f_3 values with Nganasans, while the others have more negative f_3 values with Ulchi. Most of these seven groups are also upward-shifted in PCA towards the forest-tundra cline, suggesting cross-talk between two clines.

To perform a higher-resolution characterization of the admixture landscape, we performed a haplotype-based GLOBETROTTER analysis. We took a regional approach, meaning that all 73 target groups were modelled as a patchwork of haplotypes from the 167 reference groups, but not those from any target. The goal of this approach was to minimize false negative results due to the sharing of admixture history between targets. All 73 targets show a robust signal of admixture (that is, a correlation of ancestry status shows a distinct pattern of decay over genetic distance in all bootstrap replicates (bootstrap $P < 0.01$ for all 73 targets; Supplementary Table 4)). When we consider the relative contribution of references (categorized into 12 groups (Supplementary Table 2)) to the 2 main sources of the admixture signal (date 1 and PC1), we observe a pattern comparable to the PCA, ADMIXTURE and f_3 results (Fig. 4b). The European references provide a major contribution for the Western Eurasian-related source in the forest-tundra and steppe-forest populations, while the Caucasus/Iranian references do so in the southern steppe populations. Similarly, Siberian references make the highest contribution to the Eastern Eurasian-related source in the forest-tundra populations, followed by the steppe-forest and southern steppe populations. Admixture date estimates from GLOBETROTTER range from 7–55 generations (200–1600 BP, using 29 years per generation[45]; Supplementary Fig. 6 and Supplementary Note 2). These match with previous reports using similar methodologies[13], but much younger observed admixtures in the Late Bronze and Iron Ages[8,39].

Admixture modelling of inner Eurasians shows multiple different temporal layers for present-day admixture clines. Using *F*-statistics-based approaches, we show that the Eneolithic Botai gene pool was closely related to the ANE ancestry and substantially contributed to the later Okunevo individuals (Supplementary Note 3). To test whether this ancient layer left a genetic legacy in later populations of inner Eurasia, we systematically explored diverse qpAdm-based admixture models of inner Eurasian populations.

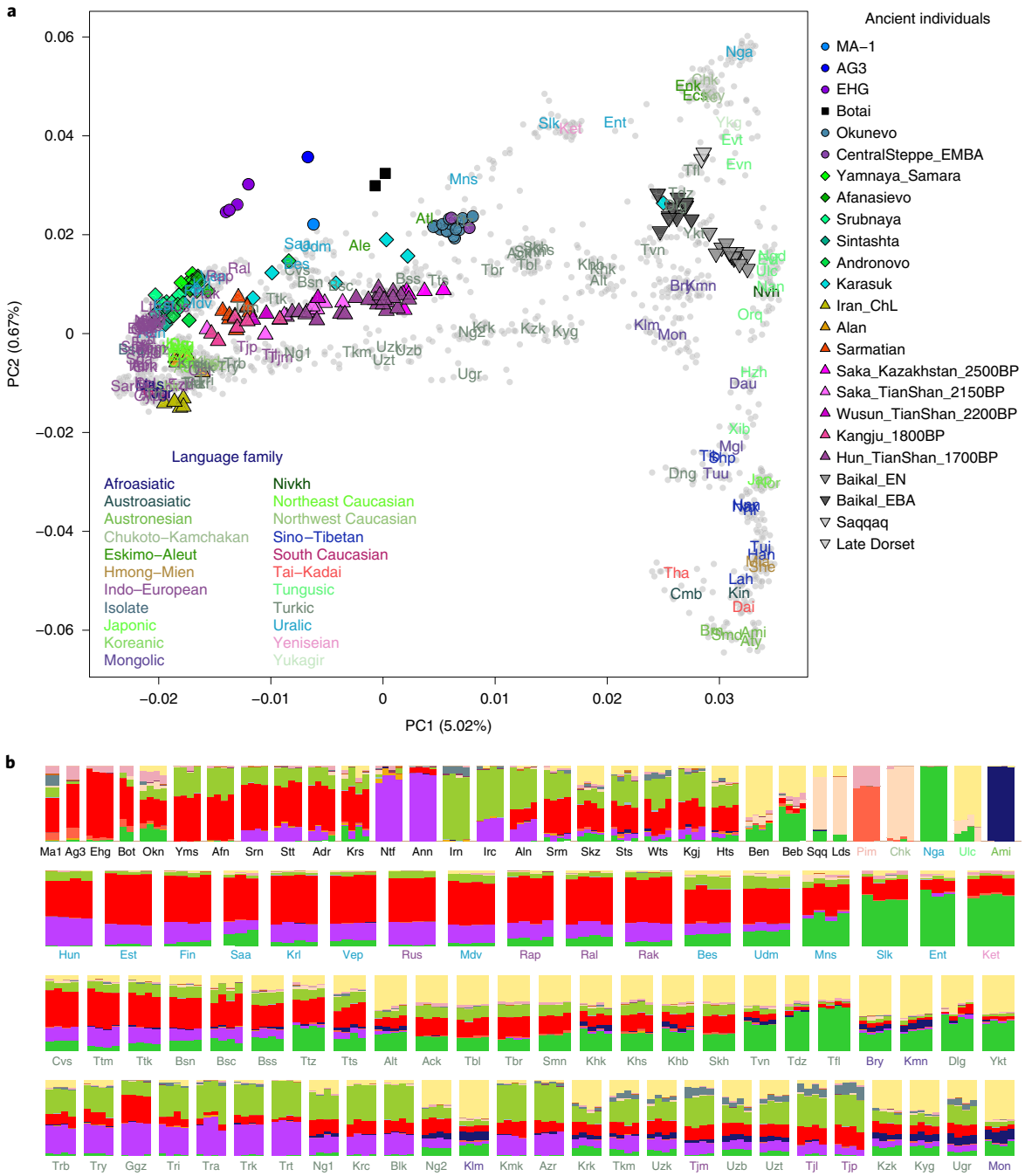

Fig. 2 | Genetic structure of inner Eurasian populations. a, The first two principal components of 2,077 Eurasian individuals separate Western and Eastern Eurasians (PC1) and Northeast and Southeast Asians (PC2). Most inner Eurasians are located between Western and Eastern Eurasians on PC1. Ancient individuals (colour-filled shapes) are projected onto principal components calculated based on contemporary individuals. Present-day individuals are marked by grey dots, with their per-group mean coordinates marked by three-letter codes (which are defined in Supplementary Table 2). Individuals are coloured by their language family. **b**, ADMIXTURE results for a chosen set of ancient and present-day groups ($K=14$). The top row shows ancient inner Eurasians and representative present-day Eastern Eurasians. The following three rows show forest-tundra, steppe-forest and southern steppe cline populations, respectively. Most inner Eurasians are modelled as a mixture of components primarily found in Eastern or Western Eurasians. The results for the full set of individuals are provided in Supplementary Fig. 3.

Two-way mixture of Ulchi/Nganasan and Srubnaya approximates the steppe-forest populations surprisingly well (χ^2 $P \geq 0.05$ and ≥ 0.01 for 12/24 and 18/24 populations, respectively; Supplementary Table 5). A more complex three-way model of Ulchi + Srubnaya + AG3 fits all steppe-forest populations (χ^2 $P \geq 0.05$ for 24/24 populations; Fig. 5 and Supplementary Table 5). Similarly, Nganasan + Srubnaya + AG3 provides a good fit to most populations, but with a negative contribution from AG3 (χ^2 $P \geq 0.05$ for 19/24 populations). We interpret this as reflecting minor heterogeneity in the Eastern Eurasian source, with average affinity to the ANE ancestry intermediate between Ulchi and Nganasan. Based on this admixture modelling, we suggest that the steppe-forest cline does not keep a detectable level of contribution from the older clines, the sources of which have higher ANE ancestry in both Western and Eastern Eurasian parts.

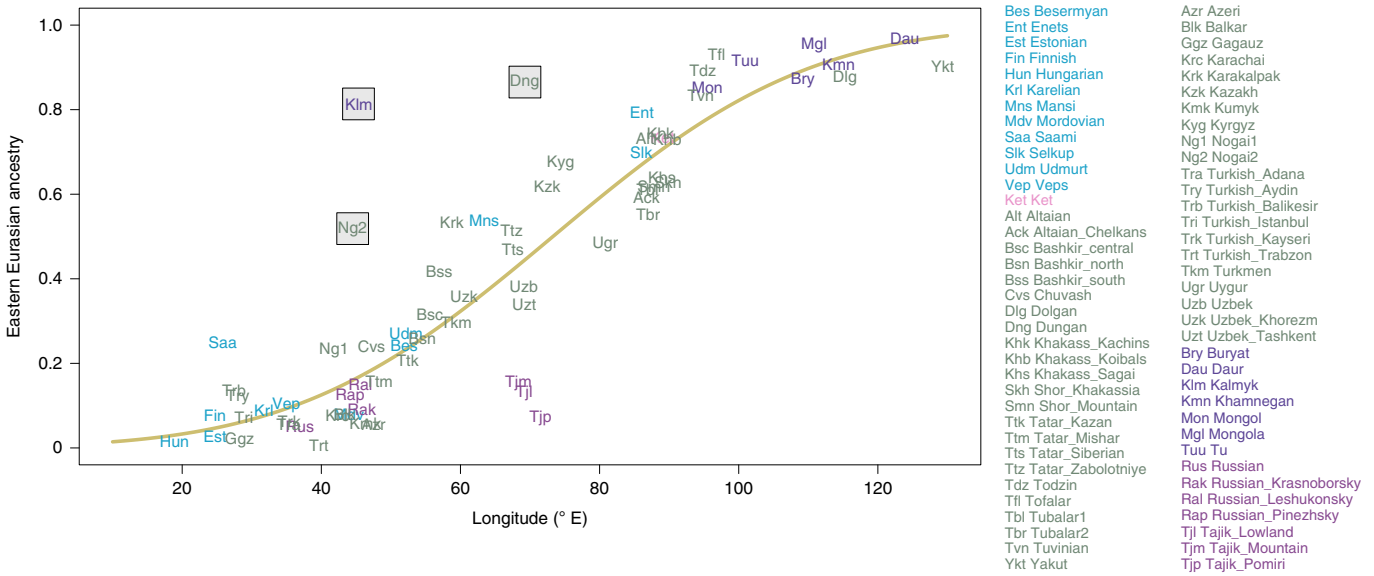

Fig. 3 | Correlation of longitude and ancestry proportion across inner Eurasian populations. Across inner Eurasian populations, mean longitudinal coordinates (*x* axis) and mean Eastern Eurasian ancestry proportions (*y* axis) are strongly correlated. Eastern Eurasian ancestry proportions were estimated from ADMIXTURE results with $K=14$ by summing six components maximized in Surui, Chipewyan, Itelmen, Nganasan, Atayal and Early Neolithic Russian Far East individuals ('Devil's Gate'), respectively (Supplementary Fig. 3). The yellow curve shows a probit regression fit following the model reported by Sedghifar et al.[69]. Three groups (Kalmyks, Dungans and Nogai2) are marked with a grey square due to their substantial deviation from the curve, as well as their historically known migration history.

In contrast, the southern steppe populations do not match with the Ulchi + Srubnaya model (χ^2 $P \leq 1.34 \times 10^{-7}$; Supplementary Table 6). Adding Chalcolithic Iranians as the third ancestry significantly improves the model fit with substantial contribution from them (χ^2 $P \leq 5.10 \times 10^{-5}$ with 7.0–64.6% contribution; Fig. 5 and Supplementary Table 6), although the three-way model still does not adequately explain the data. Ancient individuals from the Tian Shan region[22], dated to 2,200–1,100 BP, show a similar pattern (Supplementary Table 7). However, older individuals from Central Kazakhstan dated to 2,500 BP (Saka_Kazakhstan_2500BP in Fig. 2)[22] are adequately modelled as Nganasan + Srubnaya or Ulchi + Srubnaya + AG3 (χ^2 $P = 0.057$ and 0.824, respectively; Supplementary Table 7).

For the forest-tundra populations, the Nganasan + Srubnaya model is adequate only for the two Volga region populations, Udmurts and Besermyans (Fig. 5 and Supplementary Table 8). For the other populations west of the Urals, six from the northeastern corner of Europe are modelled with additional Mesolithic Western European hunter-gatherer (WHG) contribution (8.2–11.4%; Supplementary Table 8), while the rest need both WHG and early Neolithic European farmers (LBK_EN; Supplementary Table 2)[5,21]. Nganasan-related ancestry substantially contributes to their gene pools and cannot be removed from the model without a significant decrease in the model fit (4.1–29.0% contribution; χ^2 $P \leq 1.68 \times 10^{-5}$; Supplementary Table 8). For the 4 populations east of the Urals (Enets, Selkups, Kets and Mansi), for which the above models are not adequate, Nganasan + Srubnaya + AG3 provides a good fit (χ^2 $P \geq 0.018$; Fig. 5 and Supplementary Table 8). Using early Bronze Age populations from the Baikal Lake region ('Baikal_EBA'; Supplementary Table 2)[23] as a reference instead of Nganasan, the two-way model of Baikal_EBA + Srubnaya provides a reasonable fit (χ^2 $P \geq 0.016$; Supplementary Table 8) and the three-way model of Baikal_EBA + Srubnaya + AG3 is adequate but with negative AG3 contribution for Enets and Mansi (χ^2 $P \geq 0.460$; Supplementary Table 8). Bronze/Iron Age populations from Southern Siberia also show a similar ancestry composition with high ANE affinity (Supplementary Table 9). The additional ANE contribution beyond the Nganasan + Srubnaya model suggests a legacy from ANE-ancestry-rich clines before the Late Bronze Age.

Discussion

In this study, we analysed new genome-wide data of indigenous peoples from inner Eurasia, providing a dense representation for human genetic diversity in this vast region. Our finding of inner Eurasian populations being structured into three largely distinct clines shows a striking correlation between genes, geography and language (Figs. 1 and 2). Ecoregion-wide, the three clines match boreal forests and tundra, the forest-steppe zone and steppe/shrubland further to the south, respectively. Language-wide, they match the distribution of the Uralic-, and northern and southern Turkic-speaking languages. We acknowledge that the distinction of three clines is far from complete and that there are cases of intermediate patterns. For example, Turkic and Uralic speakers from the Volga region are genetically quite similar, but the Uralic speakers still have extra affinity with the Uralic speakers further to the east (for example, Nganasans; Supplementary Fig. 4b). Likewise, a number of Turkic-speaking populations (for example, Dolgans, Todzins, Tofalars and Tatar_Zabolotniye), living at the periphery or even inside the taiga belt, show a genetic influence from the forest-tundra cline (Fig. 4).

It may be viewed that our sampling scheme is not uniform geographically, although it gathers the vast majority of ethnic groups and is quite dense geographically. Indeed, the gaps between distinct genetic clines (with only a few groups located in between) tend to correspond to the gaps in sampling locations (Figs. 1 and 2). Although this non-uniformity of sampling largely results from the non-uniformity in the density of (language-defined) ethnic groups, it is important to organize a future study for further sampling of sparsely populated regions between the clines (for example, Central Kazakhstan or East Siberia).

The steppe cline populations derive their Eastern Eurasian ancestry from a gene pool similar to contemporary Tungusic speakers from the Amur river basin (Figs. 2 and 4), thus suggesting a genetic connection among the speakers of languages belonging to

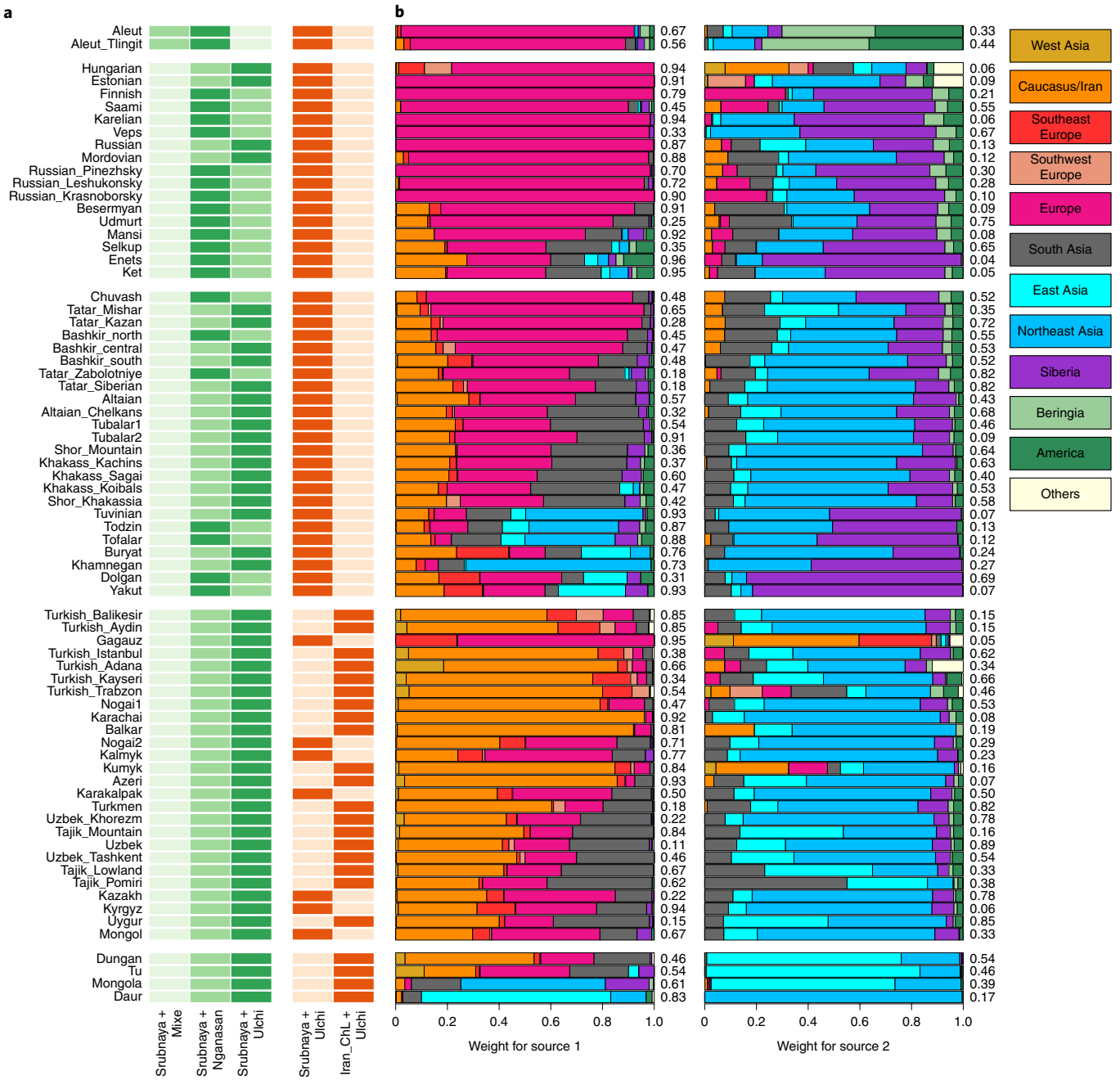

Fig. 4 | Characterization of the Western and Eastern Eurasian source ancestries in inner Eurasian populations. a, Admixture f_3 values are compared for different Eastern Eurasian (Mixe, Nganasan and Ulchi; green) and Western Eurasian references (Srubnaya and Chalcolithic Iranians (Iran_ChL); red). For each target group, darker shades mark more negative f_3 values. **b**, Weights of donor populations in two sources characterizing the main admixture signal (date 1 and PC1) in the GLOBETROTTER analysis. We merged 167 donor populations into 12 groups (top right). Target populations were split into five groups (from top to bottom): Aleuts; the forest-tundra cline populations; the steppe-forest cline populations; the southern steppe cline populations; and 'others'.

the Altaic macrofamily (Turkic, Mongolic and Tungusic families). Based on our results, as well as Early Neolithic genomes from the Russian Far East[38], we speculate that such a gene pool may represent the genetic profile of prehistoric hunter-gatherers in the Amur river basin. In contrast, a distinct Nganasan-related Eastern Eurasian ancestry in the forest-tundra cline suggests substantial separation between these two eastern ancestries. Nganasans have high genetic affinity with prehistoric individuals with the ANE ancestry in North Eurasia, such as the Upper Palaeolithic Siberians or the Mesolithic eastern European hunter-gatherers (EHG), which is exceeded only by Native Americans and by Beringians among Eastern Eurasians (Supplementary Fig. 7). Also, Northeast Asians are closer to Nganasans than they are to either Beringians, Native Americans or ancient Baikal populations, and the ANE affinity in East Asians is correlated well with their affinity with Nganasans (Supplementary Fig. 8). We hypothesize that Nganasans may be relatively isolated descendants of a prehistoric Siberian metapopulation with high ANE affinity, which formed present-day Northeast Asians by mixing with populations related to the Neolithic Northeast Asians[38].

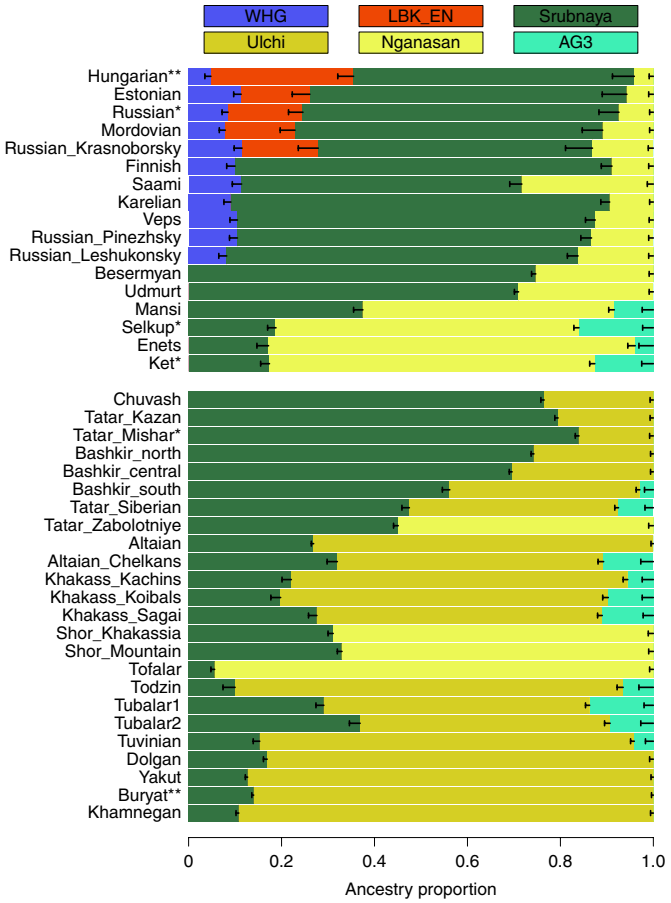

Fig. 5 | qpAdm-based admixture models for the forest-tundra and steppe-forest cline populations. For the forest-tundra population to the west of the Urals, Nganasan + Srubnaya + WHG + LBK_EN and its submodel provide a good fit, while additional ANE-related contribution (AG3) is required for those to the east of the Urals (Enets, Selkups, Kets and Mansi). For the steppe-forest populations, Srubnaya + Ulchi, Srubnaya + Ulchi + AG3 and Srubnaya + Nganasan provide a good fit. For each target population, we present the simplest best fitting model, that is, the one with the smallest number of references and with the biggest *P* value. The 5 cM jackknifing standard errors are marked by horizontal bars. Details of the model information are presented in Supplementary Tables 5 and 8.
*$P = 0.01$–0.05; **$P < 0.01$.

Forest-tundra populations to the east of the Urals, such as Selkups and Kets, show excess ANE affinity, suggesting a legacy from the ANE-ancestry-rich pre-Bronze Age gene pools (Supplementary Table 8). In contrast, admixture modelling finds that no contemporary steppe-forest cline population is required to have additional ANE ancestry beyond that which a mixture model of Bronze Age steppe plus present-day Eastern Eurasians can explain (Supplementary Table 5). This suggests that both Western and Eastern Eurasian ancestries of the steppe-forest populations were largely inherited from later gene flows since the Late Bronze Age (that is, Srubnaya-like WSH ancestry for the Western Eurasian part and present-day Tungusic speaker-related ancestry for the Eastern Eurasian part). Additional ancient genomes from Siberia will be critical to reconstruct changes in the ANE-related ancestries in Siberia over time and to understand the formation of the Nganasan gene pool.

The southern steppe populations differentiate from the steppe-forest populations to the north by having a strong genetic affinity broadly to West/South Asian ancestries (Supplementary Fig. 4 and Supplementary Table 6). Ancient Tian Shan populations dating back to 2,200 BP show the same property (Supplementary Table 7), while Sintashta culture-related WSH ancestry was widely reported in this region during the Late Bronze Age[46]. Together with the lack of West/South Asian affinity in the Saka culture individuals in Kazakhstan around 2,500 BP (Supplementary Table 7), we suggest a northward influx of West/South Asian-related ancestry into the Tian Shan region during the first half of the first millennium BC and into Kazakhstan further to the north slightly later.

It will be extremely important to expand the set of available ancient genomes across inner Eurasia. Inner Eurasia has functioned as a conduit for human migration and cultural transfer since the first appearance of modern humans in this region. As a result, we observe deep sharing of genes between Western and Eastern Eurasian populations in multiple layers: the Pleistocene ANE ancestry in Mesolithic EHG and contemporary Native Americans; Bronze Age steppe ancestry from Europe to Mongolia; and Nganasan-related ancestry extending from Western Siberia into Eastern Europe. More recent historical migrations, such as the westward expansions of Turkic and Mongolic groups, further complicate genomic signatures of admixture and have overwritten those from older events. Ancient genomes of Iron Age steppe individuals, already showing signatures of west–east admixture in the fifth to second century BC[39], provide further direct evidence for the hidden old layers of admixture, which is often difficult to appreciate from present-day populations, as shown by our finding of a discrepancy between the estimates of admixture dates from contemporary individuals and those from ancient genomes.

Methods

Study participants and genotyping. We collected samples from 763 participants from 9 countries (Armenia, Georgia, Kazakhstan, Moldova, Mongolia, Russia, Tajikistan, Ukraine and Uzbekistan). The sampling strategy included sampling a majority of large ethnic groups in the studied countries. Within groups, we sampled subgroups if they were known to speak different dialects. For ethnic groups with large area, we sampled within several districts across the area. We sampled individuals whose grandparents were all self-identified members of the given ethnic groups and were born within the studied district(s). Most of the ethnic Russian samples were collected from indigenous Russian areas (present-day Central Russia) and had been stored for years in the Estonian Biocentre. Samples from Mongolia, Tajikistan, Uzbekistan and Ukraine were collected partially in the framework of the Genographic Project. Most DNA samples were extracted from venous blood using the phenol-chloroform method. For this study, we identified 112 subgroups (belonging to 60 ethnic group labels) that were not previously genotyped on the HumanOrigins array platform[43], and selected an average of 7 individuals per subgroup (Fig. 1 and Supplementary Table 1). Genome-wide genotyping experiments were performed on the HumanOrigins array platform. We removed 18 individuals from further analysis either due to high genotype missing rates (>0.05; $n = 2$) or because they were outliers in PCA relative to other individuals from the same group ($n = 16$). The remaining 745 individuals assigned to 60 group labels were merged to published HumanOrigins datasets of worldwide contemporary populations[20] and of 4 Siberian ethnic groups (Enets, Kets, Nganasans and Selkups)[25]. Diploid genotype data of six contemporary individuals (two Saami, two Sherpa and two Tibetans) were obtained from the Simons Genome Diversity Project dataset[26]. We also added ancient individuals from published studies[3,8,19–23,27–42], by randomly sampling a single allele for 581,230 autosomal SNPs in the HumanOrigins array (Supplementary Table 2).

Sequencing of the ancient Botai genomes. We extracted genomic DNA from four skeletal remains belonging to two individuals, and built sequencing libraries either with no uracil-DNA glycosylase (UDG) treatment or with partial treatment following published protocols[47,48] (Table 1). Radiocarbon dating of BKZ001 was conducted by the Curt-Engelhorn-Centre for Archaeometry (Mannheim, Germany) for one of two bone samples used for DNA extraction. All libraries were barcoded with two library-specific 8 base pair indices[49]. The samples were manipulated in dedicated clean room facilities at the University of Tübingen or at the Max Planck Institute for the Science of Human History. Indexed libraries were enriched for about 1.24 million informative nuclear SNPs using the in-solution capture method ('1,240 K capture')[5,21].

Libraries were sequenced on an Illumina HiSeq 4000 platform with either single-end 75 bp or paired-end 50 bp cycles following the manufacturer's protocols. Output reads were demultiplexed by allowing up to one mismatch in each of two 8

base pair indices. FASTQ files were processed using EAGER version 1.92 (ref. [50]). Specifically, Illumina adapter sequences were trimmed using AdapterRemoval version 2.2.0 (ref. [51]) and the reads (30 base pairs or longer) were aligned onto the human reference genome (hg19) using BWA aln/samse version 0.7.12 (ref. [52]) with the relaxed edit distance parameter (-n 0.01). Seeding was disabled for reads from non-UDG libraries by adding an additional parameter (-l 9999). PCR duplicates were then removed using DeDup version 0.12.2 (ref. [50]), and reads with a Phred-scaled mapping quality score of <30 were filtered out using SAMtools version 1.3 (ref. [53]). We took several measurements to check the data authenticity. First, patterns of chemical damages typical of ancient DNA were tabulated using mapDamage version 2.0.6 (ref. [54]). Second, mitochondrial contamination for all of the libraries was estimated using Schmutzi[55]. Third, nuclear contamination for libraries derived from males was estimated by the contamination module in ANGSD version 0.910 (ref. [56]). Before genotyping, the first and last three bases of each read were masked for libraries with partial UDG treatment using the trimBam module in bamUtil version 1.0.13 (ref. [57]). To obtain haploid genotypes, we randomly chose one high-quality base (Phred-scaled base quality score ≥ 30) for each of the 1.24 million target sites using pileupCaller (https://github.com/stschiff/sequenceTools). We used masked reads from libraries with partial UDG treatment for transition SNPs and unmasked reads from all libraries for transversions. Mitochondrial consensus sequences were obtained using the log2fasta program in Schmutzi with a quality cut-off of 10, and subsequently assigned to haplogroups using HaploGrep2 (ref. [58]). The Y haplogroup R1b was assigned using the yHaplo program[59]. To estimate the phylogenetic position of the Botai Y haplogroup more precisely, Y chromosomal SNPs were called with SAMtools mpileup using bases with a quality score of ≥30: a total of 2,481 SNPs out of ~30,000 markers included in the 1,240 K capture panel were called with a mean read depth of 1.2×. Twenty-two SNP positions relevant to the up-to-date haplogroup R1b tree (www.isogg.org and www.yfull.com) confirmed that the sample was positive for the markers of the R1b-P297 branch but negative for its R1b-M269 sub-branch.

The frequency distribution map of this Y chromosomal clade was created with GeneGeo software[60,61] using the average weighted interpolation procedure with a radius of 1,200 km and a weight function inversely proportional to the cube of the distance. The initial frequencies were calculated as the proportion of samples positive for 'root' R1b marker M343 but negative for M269; these proportions were calculated for the 577 populations from the in-home Y-base database, which was compiled mainly from the published datasets.

Analysis of population structure. We performed a PCA of various groups using smartpca version 13050 in the EIGENSOFT version 6.0.1 package[62]. We used the 'lsqproject: YES' option to project individuals not used for calculating principal components (this procedure avoids bias due to missing genotypes). We performed unsupervised model-based genetic clustering as implemented in ADMIXTURE version 1.3.0 (ref. [63]). For this purpose, we used 118,387 SNPs with a minor allele frequency (MAF) of 1% or higher in 3,507 individuals, after pruning out linked SNPs by randomly removing one SNP from each pair with the coefficient of determination of their genotype values greater than 0.2 ($r^2 > 0.2$) using the '–indep-pairwise 200 25 0.2' command in PLINK version 1.90 (ref. [64]). For each value of K (that is, the number of ancestral populations) ranging from 2–20, we ran 5 replicates with different random seeds and took the one with the highest log likelihood value.

***F*-statistics analysis.** We computed various f_3 and f_4 statistics using the qp3Pop (version 400) and qpDstat (version 711) programs in the ADMIXTOOLS package[43]. We computed f_4 statistics with the 'f4mode: YES' option. For these analyses, we studied a total of 301 groups, including 73 inner Eurasian target groups and 167 contemporary and 93 ancient reference groups (Supplementary Table 2). We included two groups from the Aleutian Islands ('Aleut' and 'Aleut_Tlingit'; Supplementary Table 2) as positive control targets with known recent admixture. Aleut_Tlingits are Aleut individuals whose mitochondrial haplogroup lineages are related to Tlingits[31]. For each target, we calculated outgroup f_3 statistics of the form f_3(target, X; Mbuti) against all targets and references to quantify the overall allele sharing, and performed admixture f_3 tests of the form f_3(ref$_1$, ref$_2$; target) for all pairs of references to explore the admixture signal in the targets. We estimated standard error using a block jackknife with a 5 cM block[62].

We performed f_4 statistic-based admixture modelling using the qpAdm (version 632) program[20] in the ADMIXTOOLS package. We used a basic set of 7 outgroups, unless specified otherwise, to provide high enough resolution to distinguish various Western and Eastern Eurasian ancestries: Mbuti ($n = 10$; central African); Natufian ($n = 6$; early Holocene Levantine)[20]; Onge ($n = 11$; from the Andaman Islands); Neolithic Iranian ($n = 5$)[20]; Villabruna ($n = 1$; Palaeolithic European)[28]; Ami ($n = 10$; Taiwanese aborigine); and Mixe ($n = 10$; Central American). Before qpAdm modelling, we checked whether the reference groups were well distinguished by their relationship with the outgroups using the qpWave (version 400) program[65].

We used the qpGraph (version 6065) program in the ADMIXTOOLS package for graph-based admixture modelling. Starting with a graph of Mbuti, Ami and WHG, we iteratively added AG3 ($n = 1$; Palaeolithic Siberian)[28], EHG ($n = 4$; Mesolithic hunter-gatherers from Karelia or Samara)[5,23,28] and Botai by testing all of the possible topologies allowing up to one additional gene flow. After obtaining the best two-way admixture model for Botai, we tested additional three-way admixture models.

GLOBETROTTER analysis. We performed a GLOBETROTTER analysis of admixture for 73 inner Eurasian target populations to obtain haplotype-sharing-based evidence of admixture, independent of the allele frequency-based F statistics, as well as estimates of admixture dates and a fine-scale profile of their admixture sources[14]. We followed the regional approach described by Hellenthal et al.[14], in which target haplotypes can only be copied from the haplotypes of 167 contemporary reference groups, but not from those of the other target groups. This approach is recommended when multiple target groups share a similar admixture history[14], which is likely to be the case for our inner Eurasian populations.

We jointly phased the contemporary genome data without a prephased set of reference haplotypes, using SHAPEIT2 version 2.837 in its default setting[66]. We used a genetic map for the 1000 Genomes Project phase 3 data, downloaded from https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html. We used haplotypes from a total of 2,615 individuals belonging to 240 groups (73 recipients and 167 donors; Supplementary Table 2) for the GLOBETROTTER analysis. To reduce the computational burden and provide a more balanced set of donor populations, we randomly sampled 20 individuals if a group contained more than 20 individuals. Using these haplotypes, we performed GLOBETROTTER analysis following the recommended workflow[14]. We first ran 10 rounds of the expectation–maximization algorithm for chromosomes 4, 10, 15 and 22 in ChromoPainter version 2 with '-in' and '-iM' switches to estimate the chunk size and switch error rate parameters[67]. Both recipient and donor haplotypes were modelled as a patchwork of donor haplotypes. The 'chunk length' output was obtained by running ChromoPainter version 2 across all chromosomes, with the estimated parameters averaged over both recipient and donor individuals (-n 238.05 -M 0.000617341). We also generated ten painting samples for each recipient group by running ChromoPainter with the parameters averaged over all recipient individuals (-n 248.455 -M 0.000535236). Using the chunk length output and painting samples, we ran GLOBETROTTER with the 'prop.ind: 1' and 'null.ind: 1' options. We estimated the significance of the estimated admixture date by running 100 bootstrap replicates using the 'prop.ind: 0' and 'bootstrap.date.ind: 1' options; we considered date estimates between 1 and 400 generations as evidence of admixture[14]. For populations that gave evidence of admixture by this procedure, we repeated GLOBETROTTER analysis with the 'null:ind: 0' option[14]. We also compared admixture dates from GLOBETROTTER analysis with those based on weighted admixture linkage disequilibrium decay, as implemented in ALDER version 1.3 (ref. [68]). As the reference pair, we used (French, Eskimo_Naukan), (French, Nganasan), (Georgian, Ulchi), (French, Ulchi) and (Georgian, Ulchi) for target group categories 1–5, respectively, based on their genetic profiles (Supplementary Table 2). We used a minimum intermarker distance of 1.0 cM to account for linkage disequilibrium in the references.

EEMS analysis. To visualize the heterogeneity in the rate of gene flow across inner Eurasia, we performed the EEMS analysis[44]. We included a total of 1,214 individuals from 98 groups in the analysis (Supplementary Table 2). In this dataset, we kept 101,370 SNPs with a MAF ≥ 0.01 after linkage disequilibrium pruning ($r^2 \leq 0.2$). We computed the mean squared genetic difference matrix between all pairs of individuals using the 'bed2diffs_v1' program in the EEMS package. To reduce distortion in northern latitudes due to map projection, we used geographic coordinates in the Albers equal-area conic projection (+proj = aea + lat_1 = 50 + lat_2 = 70 + lat_0 = 56 + lon_0 = 100 + x_0 = 0 + y_0 = 0 + ellps = WGS84 + datum = WGS84 + units = m + no_defs). We converted the geographic coordinates of each sample and the boundary using the spTransform function in the R package rgdal version 1.2–5. We ran five initial Markov chain Monte Carlo (MCMC) runs of two million burn-ins and four million iterations with different random seeds and took a run with the highest likelihood. Starting from the best initial run, we set up another five MCMC runs of two million burn-ins and four million iterations as our final analysis. We used the following proposal variance parameters to keep the acceptance rate around 30–40%, as recommended by the developers[44]: qSeedsProposalS2 = 5000; mSeedsProposalS2 = 1000; qEffctProposalS2 = 0.0001; and mrateMuProposalS2 = 0.00005. We set up a total of 532 demes automatically with the 'nDemes = 600' parameter. We visualized the merged output from all five runs using the 'eems.plots' function in the R package rEEMSplots[44].

We performed the EEMS analysis for Caucasus populations in a similar manner, including a total of 237 individuals from 21 groups (Supplementary Table 2). In this dataset, we kept 95,442 SNPs with a MAF ≥ 0.01 after linkage disequilibrium pruning ($r^2 \leq 0.2$). We applied the Mercator projection of geographic coordinates to the map of Eurasia (+proj = merc + datum = WGS84). We ran five initial MCMC runs of 2 million burn-ins and 4 million iterations with different random seeds and took a run with the highest likelihood. Starting from the best initial run, we set up another five MCMC runs of one million burn-in and four million iterations as our final analysis. We used the following default following proposal variance parameters: qSeedsProposalS2 = 0.1; mSeedsProposalS2 = 0.01;

qEffctProposalS2 = 0.001; and mrateMuProposalS2 = 0.01. A total of 171 demes were automatically set up with the 'nDemes = 200' parameter.

Ethics. The study protocol was approved by the Ethics Committee of the Research Centre for Medical Genetics, Moscow, Russia. All 763 participants who contributed their genetic materials provided signed written informed consent.

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

Data availability

Genome-wide sequence data of two Botai individuals (BAM format) are available at the European Nucleotide Archive under the accession number PRJEB31152 (ERP113669). Eigenstrat-format array genotype data of 763 present-day individuals and 1,240 K pulldown genotype data of two ancient Botai individuals are available at the Edmond data repository of the Max Planck Society (https://edmond.mpdl.mpg.de/imeji/collection/Aoh9c69DscnxSNjm?q=).

Received: 1 June 2018; Accepted: 18 March 2019;
Published online: 29 April 2019

References

1. Li, J. Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
2. Wang, C., Zöllner, S. & Rosenberg, N. A. A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genet.* **8**, e1002886 (2012).
3. 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).
4. Yunusbayev, B. et al The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol. Biol. Evol.* **29**, 359–365 (2012).
5. Haak, W. et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
6. Haber, M. et al. Genome-wide diversity in the Levant reveals recent structuring by culture. *PLoS Genet.* **9**, e1003316 (2013).
7. Martiniano, R. et al. The population genomics of archaeological transition in west Iberia: investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genet.* **13**, e1006852 (2017).
8. Allentoft, M. E. et al. Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
9. Barfield, T. J. *The Nomadic Alternative* (Prentice Hall, 1993).
10. Frachetti, M. D. *Pastoralist Landscapes and Social Interaction in Bronze Age Eurasia* (Univ. California Press, 2009).
11. Burch, E. S. The caribou/wild reindeer as a human resource. *Am. Antiq.* **37**, 339–368 (1972).
12. Sherratt, A. The secondary exploitation of animals in the Old World. *World Archaeol.* **15**, 90–104 (1983).
13. Yunusbayev, B. et al. The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet.* **11**, e1005068 (2015).
14. Hellenthal, G. et al. A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
15. Flegontov, P. et al. Genomic study of the Ket: a Paleo-Eskimo-related ethnic group with significant ancient North Eurasian ancestry. *Sci. Rep.* **6**, 20768 (2016).
16. Pugach, I. et al. The complex admixture history and recent southern origins of Siberian populations. *Mol. Biol. Evol.* **33**, 1777–1795 (2016).
17. Triska, P. et al. Between Lake Baikal and the Baltic Sea: genomic history of the gateway to Europe. *BMC Genet.* **18**, 110 (2017).
18. Tambets, K. et al. Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biol.* **19**, 139 (2018).
19. Raghavan, M. et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014).
20. Lazaridis, I. et al. Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
21. Mathieson, I. et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
22. Damgaard, Pd. B. et al. 137 ancient human genomes from across the Eurasian steppes. *Nature* **557**, 369–374 (2018).
23. Damgaard, Pd. B. et al. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* **360**, eaar7711 (2018).
24. Levine, M. & Kislenko, A. New Eneolithic and early Bronze Age radiocarbon dates for north Kazakhstan and south Siberia. *Camb. Archaeol.* **7**, 297–300 (1997).
25. Flegontov, P. et al. Paleo-Eskimo genetic legacy across North America. Preprint at https://www.biorxiv.org/content/10.1101/203018v1 (2017).
26. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
27. Fu, Q. et al. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).
28. Fu, Q. et al. The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016).
29. Haber, M. et al. Continuity and admixture in the last five millennia of Levantine history from ancient Canaanite and present-day Lebanese genome sequences. *Am. J. Hum. Genet.* **101**, 274–282 (2017).
30. Jones, E. R. et al. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* **6**, 8912 (2015).
31. Lazaridis, I. et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
32. Lazaridis, I. et al. Genetic origins of the Minoans and Mycenaeans. *Nature* **548**, 214–218 (2017).
33. Raghavan, M. et al. The genetic prehistory of the New World Arctic. *Science* **345**, 1255832 (2014).
34. Rasmussen, M. et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014).
35. Rasmussen, M. et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010).
36. Rasmussen, M. et al. The ancestry and affiliations of Kennewick Man. *Nature* **523**, 455–458 (2015).
37. Saag, L. et al. Extensive farming in Estonia started through a sex-biased migration from the Steppe. *Curr. Biol.* **27**, 2185–2193 (2017).
38. Siska, V. et al. Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. *Sci. Adv.* **3**, e1601877 (2017).
39. Unterländer, M. et al. Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. *Nat. Commun.* **8**, 14615 (2017).
40. 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).
41. Kılınç, G. M. et al. The demographic development of the first farmers in Anatolia. *Curr. Biol.* **26**, 2659–2666 (2016).
42. McColl, H. et al. The prehistoric peopling of Southeast Asia. *Science* **361**, 88–92 (2018).
43. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
44. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).
45. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005).
46. Narasimhan, V. M. et al. The genomic formation of South and Central Asia. Preprint at https://www.biorxiv.org/content/10.1101/292581v1 (2018).
47. Dabney, J. et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl Acad. Sci. USA* **110**, 15758–15763 (2013).
48. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. *Phil. Trans. R. Soc. B* **370**, 20130624 (2015).
49. Kircher, M. in *Ancient DNA: Methods and Protocols* (eds Shapiro, B. & Hofreiter, M.) 197–228 (Humana Press, 2012).
50. Peltzer, A. et al. EAGER: efficient ancient genome reconstruction. *Genome Biol.* **17**, 60 (2016).
51. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
53. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
54. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).
55. Renaud, G., Slon, V., Duggan, A. T. & Kelso, J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16**, 224 (2015).
56. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
57. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* **25**, 918–925 (2015).
58. Weissensteiner, H. et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–W63 (2016).
59. Poznik, G. D. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. Preprint at https://www.biorxiv.org/content/10.1101/088716v1 (2016).
60. Balanovsky, O. et al. Parallel evolution of genes and languages in the Caucasus region. *Mol. Biol. Evol.* **28**, 2905–2920 (2011).

61. Koshel, S. in *Sovremennaya Geograficheskaya Kartografiya (Modern Geographic Cartography)* (eds Lourie, I. & Kravtsova, V.) 158–166 (Data+, 2012).
62. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
63. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
64. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
65. Reich, D. et al. Reconstructing native American population history. *Nature* **488**, 370–374 (2012).
66. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
67. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
68. Loh, P.-R. et al. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233–1254 (2013).
69. Sedghifar, A., Brandvain, Y., Ralph, P. & Coop, G. The spatial mixing of genomes in secondary contact zones. *Genetics* **201**, 243–261 (2015).
70. Levine, M. Botai and the origins of horse domestication. *J. Anthropol. Archaeol.* **18**, 29–78 (1999).
71. Bronk Ramsey, C. Bayesian analysis of radiocarbon dates. *Radiocarbon* **51**, 337–360 (2009).
72. Reimer, P. J. et al. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55**, 1869–1887 (2016).

Acknowledgements

We thank I. Mathieson and I. Lazaridis for helpful comments. The research leading to these results has received funding from the Max Planck Society, Max Planck Society Donation Award and European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement number 646612 to M.R.). Analysis of the Caucasus dataset was supported by RFBR grant 16-06-00364, and analysis of the Far East dataset was supported by Russian Scientific Fund project 17-14-01345. D.R. was supported by the US National Science Foundation HOMINID grant BCS-1032255, the US National Institutes of Health grant GM100233 and an Allen Discovery Center grant, and is an investigator of the Howard Hughes Medical Institute. P.F. was supported by IRP projects of the University of Ostrava, and by the Czech Ministry of Education, Youth and Sports (project OPVVV 16_019/0000759). C.-C.W. was funded by the Nanqiang Outstanding Young Talents Program of Xiamen University and the Fundamental Research Funds for the Central Universities. M.Z. has been funded by research grants from the Ministry of Education and Science of the Republic of Kazakhstan (numbers AP05134955 and 0114RK00492).

Author contributions

C.J., O.B., E.B., S.S., W.H., D.R. and J.K. conceived and coordinated the study. O.B., M.L., E.P., Y.Y., A.A., S.K., A.Bu., P.N., S.T., D.Dal., M.C., R.S., D.Dar., Y.B., A.Bo., A.S., N.D., M.Z., L.Y., V.C., N.P., L.Da., L.S., K.D., L.A., O.U., E.I., E.Ka., I.E., M.M. and E.B. contributed the present-day samples. N.K., O.I., E.Kh., B.B., V.Zai., L.Dj. and A.K.O. contributed the ancient Botai samples. N.K. and A.I. performed the ancient DNA laboratory works. C.J., O.B., E.L., V.Zap. and C.-C.W. conducted the population genetic analyses. C.J., O.B., S.S., W.H., P.F., M.R., L.Dj., D.R. and J.K. wrote the paper with input from all co-authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-019-0878-2.

Correspondence and requests for materials should be addressed to C.J. or J.K.

nature research

Corresponding author(s): Choongwon Jeong, Johannes Krause

Last updated by author(s): Feb 9, 2019

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	For sequence data processing, we used EAGER v1.92, AdapterRemoval v2.2.0, BWA v0.7.12, DeDup v0.12.2, samtools v1.3, mapDamage v2.0.6, ANGSD v0.910, bamUtil v1.0.3, pileupCaller, Schmutzi, HaploGrep2 and yHaplo. For data analysis, we used smartpca v13050, ADMIXTURE v1.3.0, PLINK v1.90, ADMIXTOOLS v3.0, GLOBETROTTER, SHAPEIT2 v2.837, EEMS and R v3.5.0.

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Genome-wide sequence data of two Botai individuals (BAM format) are available at the European Nucleotide Archive under the accession number PRJEB31152 (ERP113669). Eigenstrat-format array genotype data of 763 present-day individuals and 1240K pulldown genotype data of two ancient Botai individuals are available at the Edmond data repository of the Max Planck Society (https://edmond.mpdl.mpg.de/imeji/collection/Aoh9c69DscnxSNjm?q=).

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	We generated genome-wide genotype data of 763 individuals from 112 subgroups (belonging to 60 ethnic group labels) which were not previously genotyped on the Affymetrix Axiom® Genome-wide Human Origins 1 ("HumanOrigins") array platform and selected on average 7 individuals per subgroup. The number of individuals per subgroup is comparable to or bigger than previous panels of world-wide genetic diversity, such as the Human Genome Diversity Panel or the Simons Genome Diversity Panel. Ancient genomes are produced based on the skeletal sample availability.
Data exclusions	We removed 18 individuals from further analysis either due to high genotype missing rate (> 0.05; n=2) or due to being outliers in principal component analysis (PCA) relative to other individuals from the same group (n=16).
Replication	Multiple independent statistical methods (i.e. allele-frequency based F-statistics and haplotype-sharing-based methods GLOBETROTTER) were applied to test and characterize admixture in the target inner Eurasian populations.
Randomization	Analysis group labels are pre-defined based on the self-reported ancestry and linguistic affiliations. Therefore, randomization is not applicable.
Blinding	The goal of th study was to characterize the genetic profile of pre-defined analysis groups. Therefore, blinding is not applicable.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	We collected samples from 763 participants from nine countries (Armenia, Georgia, Kazakhstan, Moldova, Mongolia, Russia, Tajikistan, Ukraine, and Uzbekistan). The sampling strategy included sampling a majority of large ethnic groups in the studied countries. Within groups, we sampled subgroups if they were known to speak different dialects; for ethnic groups with large area, we sampled within several districts across the area.
Recruitment	We sampled individuals whose grandparents were all self-identified members of the given ethnic groups and were born within the studied district(s).
Ethics oversight	The study protocol was approved by the Ethics Committee of the Research Centre for Medical Genetics, Moscow, Russia. All 763 participants who contributed their genetic materials provided a signed written informed consent.

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