Published in final edited form as:
Nat Ecol Evol. 2019 June ; 3(6): 966–976. doi:10.1038/s41559-019-0878-2.

The genetic history of admixture across inner Eurasia

Choongwon Jeong[#1,2,*], **Oleg Balanovsky**[#3,4], **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], **Johannes Krause**[1,*]

[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, SC MES RK, Almaty, Kazakhstan

[6]Department of Molecular Biology and Genetics, Kazakh National University by al-Farabi, Almaty, Kazakhstan

[*]Correspondence: jeong@shh.mpg.de (C.J.), krause@shh.mpg.de (J.K.).

Life Science Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Ethics Statement. 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.

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=).

Author Contributions

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

Competing Interests

The authors declare no competing interests.

Europe PMC Funders Author Manuscripts

[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 361005, China

[10]Institute of Archeology and Steppe Civilization, Kazakh National University by al-Farabi, Almaty, Kazakhstan

[11]Kemerovo State Medical University, Krasnaya 3, 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 under the Minisrty of Innovational Development, Tashkent, Uzbekistan

[18]Belgorod State University, Belgorod, Russia

[19]The Institute of Ethnology and Anthropology of the 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 of National Academy of Sciences, Yerevan, Armenia

[23]Udmurt Institute of History, Language and Literature of Udmurt Federal Research Center of the Ural Branch of the Russian Academy of Sciences, Russia

[24]Research Institute of Medical and Social Problems and Control of the 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 of the Russian 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 51010, Estonia

[30]Department of Archaeology, University of Exeter, Exeter EX4 4QE, UK

[31]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA

[32]Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA

[#] These authors contributed equally to this work.

Abstract

The indigenous populations of inner Eurasia, a huge geographic region covering the central Eurasian steppe and the northern Eurasian taiga and tundra, harbor 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 contribution from so-called “ancient North Eurasian” ancestry over time, 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 neighboring 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 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 artifacts 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 eco-geographic distinction in inner Eurasia. For example, indigenous peoples of the Eurasian steppe traditionally practice nomadic pastoralism[9,10], while northern Eurasian peoples in the taiga

mainly rely on reindeer herding and hunting11. The subsistence strategies in each of these ecoregions are often considered to be adaptations to the local environments12.

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 populations13–16. So far only three studies have reported recent genetic sharing between geographically distant populations based on the analysis of “identity-by-descent” segments13,17,18. One study reports a long-distance extra genetic sharing between Turkic populations based on a detailed comparison between Turkic-speaking groups and their non-Turkic neighbors13. The other two studies extend this approach to some Uralic and Yeniseian-speaking populations17,18. 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 Paleolithic 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 Paleolithic/Mesolithic hunter-gatherers in Europe and also substantially contributed to the gene pools of present-day Native Americans, Siberians, Europeans and South Asians19,20. 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 as “Western Steppe Herders (WSH)”: the earlier populations associated with the Yamnaya and Afanasievo cultures (often called “steppe Early and Middle Bronze Age”; “steppe_EMBA”) 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”; “steppe_MLBA”)8. The steppe_MLBA gene pool was largely descended from the preceding steppe_EMBA gene pool, with a substantial contribution from Late Neolithic Europeans.21 Also, recent archaeogenetic studies trace multiple large-scale trans-Eurasian migrations over the last several millennia using ancient inner Eurasian genomes22,23, including individuals from the Eneolithic Botai culture in northern Kazakhstan in the 4th millennium BC24. These studies now provide a rich context to interpret present-day population structure of inner Eurasians and to characterize ancient admixtures in fine resolution.

In this study, we analyzed 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 published23, to explore the genetic structure of pre-Bronze Age populations in inner Eurasia (Table 1). We aimed at characterizing 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 if and where the 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 Table S1). We merged new data with published data of present-day20,25,26 and ancient individuals3,8,19–23,27–42 (Table S2). The final data set covers 581,230 autosomal single nucleotide polymorphisms (SNPs) in the Affymetrix Axiom® Genome-wide Human Origins 1 (“HumanOrigins”) array platform43.

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. 1-2). 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 PC 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, 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 (“CHG”)30 and Neolithic Iranians (“Iran_N”)20 and frequently harbor another component enriched in present-day South Asians (Supplementary Fig. 4). Based on the PCA and ADMIXTURE results, we heuristically assign inner Eurasians into 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 the Altai-Sayan regions and southern Siberia, and the “southern steppe” cline includes the rest of populations. We

separate four groups (Daur, Mongola, Tu and Dungans) as “others” (Supplementary Table 2).

The genetic barriers splitting the inner Eurasians are also found in the EEMS (“estimated effective migration surface”) analysis44 (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 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 260 groups (167 present-day and 93 ancient groups) as the “sources” to represent world-wide genetic diversity (Supplementary Table 2).

Testing all possible pairs of 167 present-day “source” groups as references, we detect highly significant f_3 statistics for 66 of 73 targets (< -3 SE; standard error; 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 seven groups also have small f_3 statistics around zero (-5.1 SE to +2.7 SE). Reference pairs with the most negative f_3 statistics for the most part involve one eastern and one western Eurasian groups supporting the qualitative impression of east-west admixture from PCA and ADMIXTURE analysis. 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 MLBA_steppe ancestry21 and Chalcolithic Iranians (“Iran_ChL”) to represent West/South Asian-related ancestry20; 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 with Nganasans while the others have more negative f_3 with Ulchi. Most of these seven groups are also upward-shifted in PCA toward the forest-tundra cline, suggesting a 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 modeled 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 sharing of admixture history between targets. All 73 targets show a robust signal of admixture: i.e. 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 the relative contribution of references, categorized to 12 groups (Supplementary Table 2), into the two main sources of the admixture signal ("date 1 PC 1") is considered, we observe a pattern comparable to 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 ones. Admixture date estimates from GLOBETROTTER range 7-55 generations (200-1600 BP; years before present; using 29 years per generation45; Supplementary Fig. 6 and Supplementary Note 2). These match with previous reports using similar methodologies13, but much younger observed admixtures in the Late Bronze and Iron Ages8,39.

Admixture modeling of inner Eurasians shows multiple different temporal layers for present-day admixture clines

Using *F*-statistic-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 if this ancient layer left a genetic legacy in later populations of inner Eurasia, we systematically explored diverse qpAdm-based admixture models to inner Eurasian populations.

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 negative contribution from AG3 (χ^2 $p \geq 0.05$ for 19/24 populations). We interpret this as reflecting a minor heterogeneity in the eastern Eurasian source, with average affinity to the ANE ancestry is intermediate between Ulchi and Nganasan. Based on this admixture modeling, 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.

In contrast, the southern steppe populations do not match with the Ulchi+Srubnaya model (χ^2 $p \leq 1.34\times10^{-7}$; Supplementary Table 6). Adding Chalcolithic Iranians as the third ancestry significantly improves model fit with substantial contribution from them (χ^2 $p \leq 5.10\times10^{-5}$ with 7.0-64.6% contribution; Fig. 5 and Supplementary Table 6), although the three-way model still does not adequately explain data. Ancient individuals from the Tian Shan region22, 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")22 are adequately modeled 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 modeled with additional Mesolithic western European hunter-gatherers ("WHG") contribution (8.2-11.4%; Supplementary Table 8), while the rest need both WHG and early Neolithic European farmers (EEF; represented by "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 model fit (4.1% to 29.0% contribution; χ^2 $p \leq 1.68\times10^{-5}$; Supplementary Table 8). For the four populations east of the Urals (Enets, Selkups, Kets and Mansi), for which the above models are not adequate, Nganasan+Srubnaya+AG3 provide a good fit (χ^2 $p \geq 0.018$; Fig. 5 and Supplementary Table 8). Substituting Nganasan to early Bronze Age populations from the Baikal Lake region ("Baikal_EBA"; Supplementary Table 2)23, the two-way model of Baikal_EBA+Srubnaya provides a reasonable fit (χ^2 $p \geq 0.016$; Supplementary Table 8) and three-way model of Baikal_EBA+Srubnaya+AG3 are 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 prior to Late Bronze Age.

Discussion

In this study, we analyzed 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. 1-2). Ecoregion-wide, the three clines match boreal forests and tundra, forest-steppe zone and steppe/shrub-land further to the south, respectively. Language-wide, they match the distribution of the Uralic, 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 (e.g. Nganasans; Supplementary Fig. 4b). Likewise, a number of Turkic-speaking populations

(e.g. Dolgans, Todzins, Tofalars and Tatar_Zabolotniye), living at the periphery or even inside of the taiga belt, do show a genetic influence from the forest-tundra cline (Fig. 4).

It may be viewed that our sampling scheme is not uniform geographically, although gathering the vast majority of ethnic groups and 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 (Fig. 1-2). 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 on sparsely populated regions between the clines (e.g. 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 the Altaic macrofamily (Turkic, Mongolic and Tungusic families). Based on our results as well as early Neolithic genomes from the Russian Far East38, we speculate that such a gene pool may represent the genetic profile of prehistoric hunter-gatherers in the Amur river basin. On the other hand, a distinct Nganasan-related eastern Eurasian ancestry in the forest-tundra cline suggests a 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 Paleolithic Siberians or the Mesolithic 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 meta-population with high ANE affinity, which formed present-day Northeast Asians by mixing with populations related to the Neolithic Northeast Asians38.

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 modeling finds that no contemporary steppe-forest cline population is required to have additional ANE ancestry beyond what 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 are largely inherited from later gene flows since Late Bronze Age: 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 Nganasan gene pool.

The southern steppe populations differentiate from the steppe-forest ones 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 up 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 Age46. 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 5^{th} to 2^{nd} century BC39, provide further direct evidence for the hidden old layers of admixture, which is often difficult to appreciate from present-day populations as shown in 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 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. 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 Biocenter; 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 via the phenol-chloroform method. For this study we identified 112 subgroups (belonging to 60 ethnic group labels) which were not previously genotyped on the Affymetrix Axiom® Genome-wide Human Origins 1 (“HumanOrigins”) array platform43 and selected on average 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 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). The remaining 745 individuals assigned to 60 group labels were merged to published HumanOrigins data sets of world-wide contemporary populations20 and of four 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

Europe PMC Funders Author Manuscripts

the Simons Genome Diversity Panel data set26. We also added ancient individuals from published studies3,8,19–23,27–42, by randomly sampling a single allele for 581,230 autosomal single nucleotide polymorphisms (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 protocols47,48 (Table 1). Radiocarbon dating of BKZ001 was conducted by the CEZ Archaeometry gGmbH (Mannheim, Germany) for one of two bone samples used for DNA extraction. All libraries were barcoded with two library-specific 8-mer indices49. 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 (MPI-SHH). Indexed libraries were enriched for about 1.24 million informative nuclear SNPs using the in-solution capture method (“1240K capture”)5,21.

Libraries were sequenced on the Illumina HiSeq 4000 platform with either single-end 75 bp (SE75) or paired-end 50 bp (PE50) cycles following manufacturer’s protocols. Output reads were demultiplexed by allowing up to 1 mismatch in each of two 8-mer indices. FASTQ files were processed using EAGER v1.9250. Specifically, Illumina adapter sequences were trimmed using AdapterRemoval v2.2.051, aligned reads (30 base pairs or longer) onto the human reference genome (hg19) using BWA aln/samse v0.7.1252 with 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 v0.12.250 and reads with Phred-scaled mapping quality score < 30 were filtered out using Samtools v1.353. We did several measurements to check data authenticity. First, patterns of chemical damages typical to ancient DNA were tabulated using mapDamage v2.0.654. Second, mitochondrial contamination for all libraries was estimated by Schmutzi55. Third, nuclear contamination for libraries derived from males was estimated by the contamination module in ANGSD v0.91056. Prior to genotyping, the first and last 3 bases of each read were masked for libraries with partial UDG treatment using the trimBam module in bamUtil v1.0.1357. 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 (Ts) SNPs and used unmasked reads from all libraries for transversions (Tv). Mitochondrial consensus sequences were obtained by the log2fasta program in Schmutzi with the quality cutoff 10 and subsequently assigned to haplogroups using HaploGrep258. Y haplogroup R1b was assigned using the yHaplo program59. To estimate the phylogenetic position of the Botai Y haplogroup more precisely, Y chromosomal SNPs were called with Samtools mpileup using bases with quality score ≥ 30: a total of 2,481 SNPs out of ~30,000 markers included in the 1240K capture panel were called with mean read depth of 1.2. Twenty-two SNP positions relevant to the up-to-date haplogroup R1b tree (www.isogg.org; www.yfull.com) confirmed that the sample was positive for the markers of R1b-P297 branch but negative for its R1b-M269 sub-branch.

The frequency distribution map of this Y chromosomal clade was created by the GeneGeo software60,61 using the average weighed interpolation procedure with the weight function of degree 3 and radius 1,200 km. The initial frequencies were calculated as 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 principal component analysis (PCA) of various groups using smartpca v13050 in the EIGENSOFT v6.0.1 package62. We used the "*lsqproject: YES*" option to project individuals not used for calculating PCs (this procedure avoids bias due to missing genotypes). We performed unsupervised model-based genetic clustering as implemented in ADMIXTURE v1.3.063. For that purpose, we used 118,387 SNPs with minor allele frequency (maf) 1% or higher in 3,507 individuals after pruning out linked SNPs ($r^2 > 0.2$) using the "--indep-pairwise 200 25 0.2" command in PLINK v1.9064. For each value of K ranging from 2 to 20, we ran 5 replicates with different random seeds and took one with the highest log likelihood value.

F-statistics analysis

We computed various f_3 and f_4 statistics using the qp3Pop (v400) and qpDstat (v711) programs in the ADMIXTOOLS package43. 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 Tlingits31. For each target, we calculated outgroup f_3 statistic of the form f_3(Target, X; Mbuti) against all targets and references to quantify overall allele sharing and performed admixture f_3 test of the form f_3(Ref_1, Ref_2; Target) for all pairs of references to explore the admixture signal in targets. We estimated standard error (SE) using a block jackknife with 5 centiMorgan (cM) block62.

We performed f_4 statistic-based admixture modeling using the qpAdm (v632) program20 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), Iran_N (n=5; Neolithic Iranian)20, Villabruna (n=1; Paleolithic European)28, Ami (n=10; Taiwanese aborigine) and Mixe (n=10; Central American). Prior to qpAdm modeling, we checked if the reference groups are well distinguished by their relationship with the outgroups using the qpWave (v400) program65.

We used the qpGraph (v6065) program in the ADMIXTOOLS package for graph-based admixture modeling. Starting with a graph of (Mbuti, Ami, WHG), we iteratively added AG3 (n=1; Paleolithic Siberian)28, EHG (n=4; Mesolithic hunter-gatherers from Karelia or Samara)5,23,28, and Botai onto the graph by testing all 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 sources14. We followed the "regional" approach described in 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 history14, which is likely to be the case for our inner Eurasian populations.

We jointly phased the contemporary genome data without a pre-phased set of reference haplotypes, using SHAPEIT2 v2.837 in its default setting66. 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 computational burden and to provide 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 workflow14. We first ran 10 rounds of the expectation-maximization (EM) algorithm for chromosomes 4, 10, 15 and 22 in ChromoPainter v2 with "-in" and "-iM" switches to estimate chunk size and switch error rate parameters67. Both recipient and donor haplotypes were modeled as a patchwork of donor haplotypes. The "chunk length" output was obtained by running ChromoPainter v2 across all chromosomes with the estimated parameters averaged over both recipient and donor individuals ("-n 238.05 -M 0.000617341"). We also generated 10 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 chunklength output and painting samples, we ran GLOBETROTTER with the "prop.ind: 1" and "null.ind: 1" options. We estimated significance of 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 admixture14. For populations that gave evidence of admixture by this procedure, we repeated GLOBETROTTER analysis with the "null:ind: 0" option14. We also compared admixture dates from GLOBETROTTER analysis with those based on weighted admixture linkage disequilibrium (LD) decay, as implemented in ALDER v1.368. As the reference pair, we used (French, Eskimo_Naukan), (French, Nganasan), (Georgian, Ulchi), (French, Ulchi) and (Georgian, Ulchi) for the target group categories 1 to 5, respectively, based on their genetic profile (Supplementary Table 2). We used a minimum inter-marker distance of 1.0 cM to account for LD in the references.

EEMS analysis

To visualize the heterogeneity in the rate of gene flow across inner Eurasia, we performed the EEMS ("estimated effective migration surface") analysis44. We included a total of 1,214

Europe PMC Funders Author Manuscripts

individuals from 98 groups in the analysis (Supplementary Table 2). In this dataset, we kept 101,370 SNPs with maf ≥ 0.01 after LD 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 geographic coordinates of each sample and the boundary using the "spTransform" function in the R package rgdal v1.2-5. 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 2 million burn-ins and 4 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 developers44: qSeedsProposalS2 = 5000, mSeedsProposalS2 = 1000, qEffctProposalS2 = 0.0001, 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 rEEMSplots44.

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 maf ≥ 0.01 after LD 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 1 million burn-in and 4 million iterations as our final analysis. We used the default following proposal variance parameters: qSeedsProposalS2 = 0.1, mSeedsProposalS2 = 0.01, qEffctProposalS2 = 0.001, mrateMuProposalS2 = 0.01. A total of 171 demes were automatically set up with the "nDemes = 200" parameter.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Iain Mathieson and Iosif Lazaridis for their helpful comments. The research leading to these results has received funding from the Max Planck Society, the Max Planck Society Donation Award and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 646612 granted 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 U.S. National Science Foundation HOMINID grant BCS-1032255, the U.S. National Institutes of Health grant GM100233, by 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 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 MES RK No. AP05134955 and No. 0114RK00492.

Europe PMC Funders Author Manuscripts

References

1. Li JZ, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008; 319:1100–1104. [PubMed: 18292342]
2. Wang C, Zöllner S, Rosenberg NA. A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet. 2012; 8:e1002886. [PubMed: 22927824]
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. 2016; 113:7485–7490. [PubMed: 27325755]
4. Yunusbayev B, et al. The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Mol Biol Evol. 2012; 29:359–365. [PubMed: 21917723]
5. Haak W, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015; 522:207–211. [PubMed: 25731166]
6. Haber M, et al. Genome-wide diversity in the Levant reveals recent structuring by culture. PLoS Genet. 2013; 9:e1003316. [PubMed: 23468648]
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. 2017; 13:e1006852. [PubMed: 28749934]
8. Allentoft ME, et al. Population genomics of Bronze Age Eurasia. Nature. 2015; 522:167–172. [PubMed: 26062507]
9. Barfield, TJ. The nomadic alternative. Prentice Hall; Englewood Cliffs, NJ: 1993.
10. Frachetti, MD. Pastoralist landscapes and social interaction in Bronze Age Eurasia. Univ of California Press; Berkeley, CA: 2009.
11. Burch ES. The caribou/wild reindeer as a human resource. Am Antiquity. 1972; 37:339–368.
12. Sherratt A. The secondary exploitation of animals in the Old World. World Archaeol. 1983; 15:90–104.
13. Yunusbayev B, et al. The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. PLoS Genet. 2015; 11:e1005068. [PubMed: 25898006]
14. Hellenthal G, et al. A genetic atlas of human admixture history. Science. 2014; 343:747–751. [PubMed: 24531965]
15. Flegontov P, et al. Genomic study of the Ket a Paleo-Eskimo-related ethnic group with significant ancient North Eurasian ancestry. Sci Rep. 2016; 6
16. Pugach I, et al. The complex admixture history and recent southern origins of Siberian populations. Mol Biol Evol. 2016; 33:1777–1795. [PubMed: 26993256]
17. Triska P, et al. Between Lake Baikal and the Baltic Sea genomic history of the gateway to Europe. BMC Genet. 2017; 18:110. [PubMed: 29297395]
18. Tambets K, et al. Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. Genome Biology. 2018; 19:139. [PubMed: 30241495]
19. Raghavan M, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2014; 505:87–91. [PubMed: 24256729]
20. Lazaridis I, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016; 536:419–424. [PubMed: 27459054]
21. Mathieson I, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015; 528:499–503. [PubMed: 26595274]
22. Damgaard, PdB; , et al. 137 ancient human genomes from across the Eurasian steppes. Nature. 2018; 557:369–374. [PubMed: 29743675]
23. Damgaard, PdB; , et al. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. Science. 2018; doi: 10.1126/science.aar7711
24. Levine M, Kislenko A. New Eneolithic and early Bronze Age radiocarbon dates for north Kazakhstan and south Siberia. Camb Archaeol. 1997; 7:297–300.
25. Flegontov P, et al. Paleo-Eskimo genetic legacy across North America. bioRxiv. 2017
26. Mallick S, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016; 538:201–206. [PubMed: 27654912]

Europe PMC Funders Author Manuscripts

27. Fu Q, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014; 514:445–449. [PubMed: 25341783]
28. Fu Q, et al. The genetic history of Ice Age Europe. Nature. 2016; 534:200–205. [PubMed: 27135931]
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. 2017; 101:274–282. [PubMed: 28757201]
30. Jones ER, et al. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. Nat Commun. 2015; 6:8912. [PubMed: 26567969]
31. Lazaridis I, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014; 513:409–413. [PubMed: 25230663]
32. Lazaridis I, et al. Genetic origins of the Minoans and Mycenaeans. Nature. 2017; 548:214–218. [PubMed: 28783727]
33. Raghavan M, et al. The genetic prehistory of the New World Arctic. Science. 2014; 345
34. Rasmussen M, et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature. 2014; 506:225–229. [PubMed: 24522598]
35. Rasmussen M, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010; 463:757–762. [PubMed: 20148029]
36. Rasmussen M, et al. The ancestry and affiliations of Kennewick Man. Nature. 2015; 523:455–458. [PubMed: 26087396]
37. Saag L, et al. Extensive farming in Estonia started through a sex-biased migration from the Steppe. Curr Biol. 2017; 27:2185–2193. e2186. [PubMed: 28712569]
38. Siska V, et al. Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. Sci Adv. 2017; 3:e1601877. [PubMed: 28164156]
39. Unterländer M, et al. Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. Nat Commun. 2017; 8:14615. [PubMed: 28256537]
40. Yang MA, et al. 40,000-year-old individual from Asia provides insight into early population structure in Eurasia. Curr Biol. 2017; 27:3202–3208.e3209. [PubMed: 29033327]
41. Kılınç GM, et al. The demographic development of the first farmers in Anatolia. Curr Biol. 2016; 26:2659–2666. [PubMed: 27498567]
42. McColl H, et al. The prehistoric peopling of Southeast Asia. Science. 2018; 361:88–92. [PubMed: 29976827]
43. Patterson N, et al. Ancient admixture in human history. Genetics. 2012; 192:1065–1093. [PubMed: 22960212]
44. Petkova D, Novembre J, Stephens M. Visualizing spatial population structure with estimated effective migration surfaces. Nat Genet. 2016; 48:94–100. [PubMed: 26642242]
45. Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol. 2005; 128:415–423. [PubMed: 15795887]
46. Narasimhan VM, et al. The genomic formation of South and Central Asia. bioRxiv. 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. 2013; 110:15758–15763. [PubMed: 24019490]
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. 2015; 370:20130624. [PubMed: 25487342]
49. Kircher, M. Ancient DNA: methods and protocols. Shapiro, Beth; Hofreiter, Michael, editors. Humana Press; 2012. 197–228.
50. Peltzer A, et al. EAGER: efficient ancient genome reconstruction. Genome Biol. 2016; 17:60. [PubMed: 27036623]
51. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016; 9:88. [PubMed: 26868221]
52. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

53. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
54. Jónsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013; 29:1682–1684. [PubMed: 23613487]
55. Renaud G, Slon V, Duggan AT, Kelso J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biol. 2015; 16:224. [PubMed: 26458810]
56. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014; 15:356. [PubMed: 25420514]
57. Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. 2015; 25:918–925. [PubMed: 25883319]
58. Weissensteiner H, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016; 44:W58–W63. [PubMed: 27084951]
59. Poznik GD. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. bioRxiv. 2016
60. Balanovsky O, et al. Parallel evolution of genes and languages in the Caucasus region. Mol Biol Evol. 2011; 28:2905–2920. [PubMed: 21571925]
61. Koshel, S. Sovremennaya geograficheskaya kartografiya (Modern Geographic Cartography). Lourie, I, Kravtsova, V, editors. Data+; 2012. 158–166.
62. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2:e190. [PubMed: 17194218]
63. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19:1655–1664. [PubMed: 19648217]
64. Chang CC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4:7. [PubMed: 25722852]
65. Reich D, et al. Reconstructing native American population history. Nature. 2012; 488:370–374. [PubMed: 22801491]
66. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013; 10:5–6. [PubMed: 23269371]
67. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012; 8:e1002453. [PubMed: 22291602]
68. Loh P-R, et al. Inferring admixture histories of human populations using linkage disequilibrium. Genetics. 2013; 193:1233–1254. [PubMed: 23410830]
69. Sedghifar A, Brandvain Y, Ralph P, Coop G. The spatial mixing of genomes in secondary contact zones. Genetics. 2015; 201:243–261. [PubMed: 26205988]
70. Levine M. Botai and the origins of horse domestication. J Anthropol Archaeol. 1999; 18:29–78.
71. Bronk Ramsey C. Bayesian analysis of radiocarbon dates. Radiocarbon. 2009; 51:337–360.
72. Reimer PJ, et al. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. Radiocarbon. 2016; 55:1869–1887.

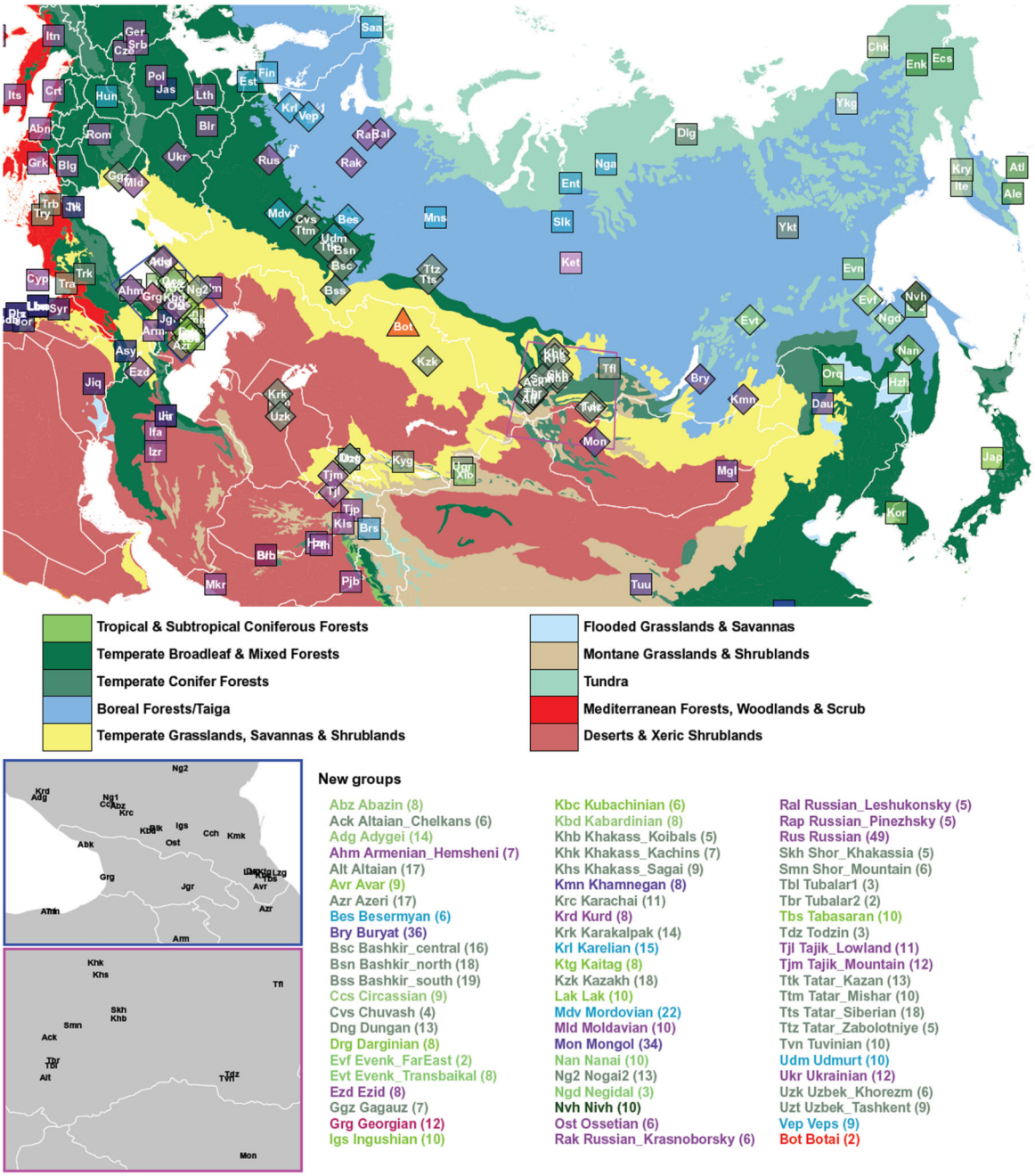

Fig. 1. Geographic locations of the Eneolithic Botai site (red triangle), 65 groups including newly sampled individuals (filled diamonds) and nearby groups with published data (filled squares).

Mean latitude and longitude values across all individuals under each group label were used. Two zoom-in plots for the Caucasus (blue) and the Altai-Sayan (magenta) regions are presented in the lower left corner. A list of new groups, their three-letter codes, and the number of new individuals (in parenthesis) are provided at the bottom. Present-day populations are color-coded based on the language family for Figs. 1-3, following key codes listed in Fig. 2. Corresponding information for the previously published groups is provided in Supplementary Table 2. The map is overlayed with ecoregional information, divided into

Europe PMC Funders Author Manuscripts

14 biomes, downloaded from https://ecoregions2017.appspot.com/ (credited to Ecoregions 2017 © Resolve). The main inner Eurasian map is on the Albers equal area projection and was produced using the spTransform function in the R package rgdal v1.2-5.

Europe PMC Funders Author Manuscripts

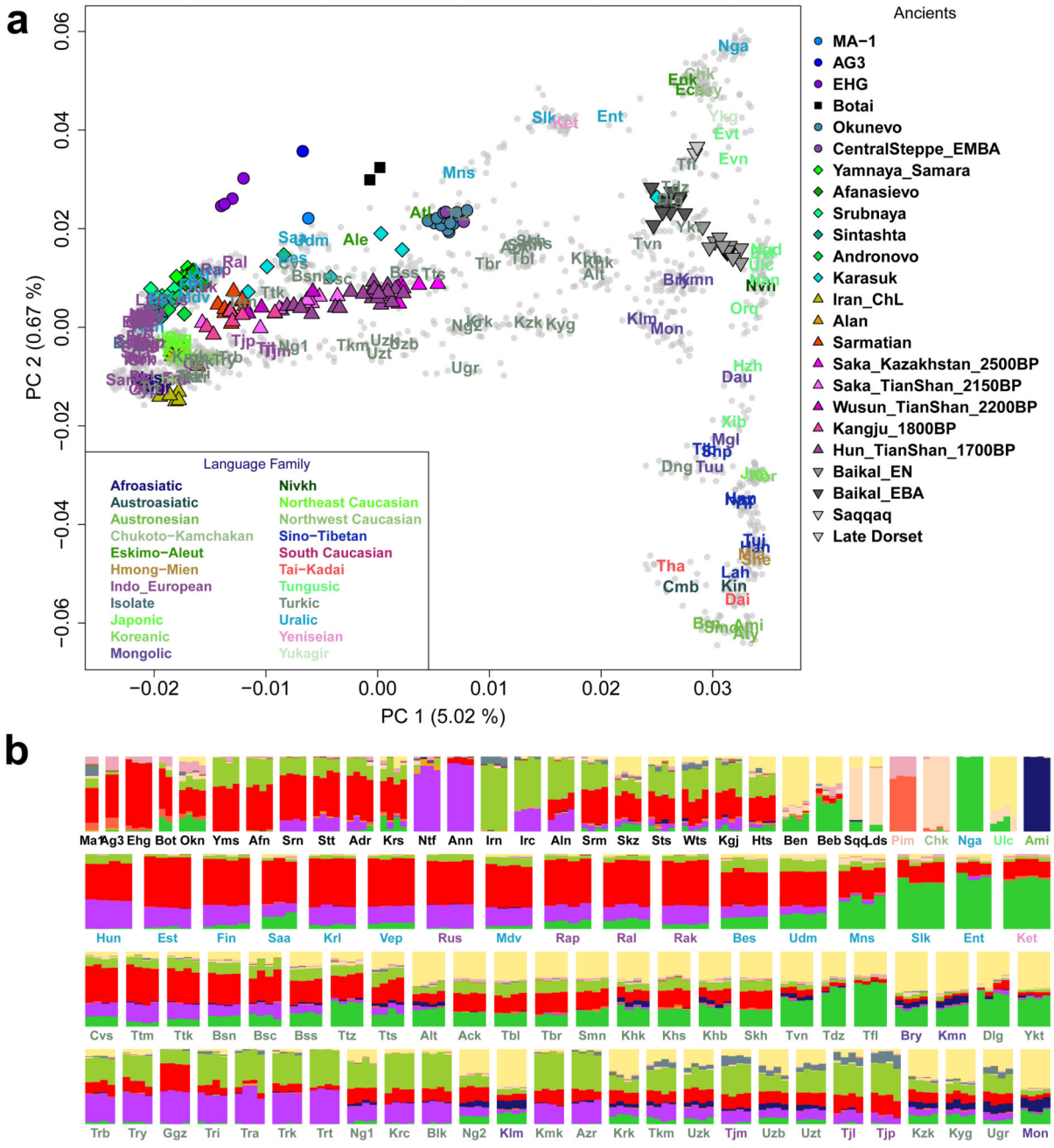

Fig. 2. The genetic structure of inner Eurasian populations.

(a) The first two PCs 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 (color-filled shapes) are projected onto PCs calculated based on contemporary individuals. Present-day individuals are marked by grey dots, with their per-group mean coordinates marked by three-letter codes listed in Supplementary Table 2. Individuals are colored by their language family. (b) ADMIXTURE results for a chosen set of ancient and present-day groups (K = 14). The top

Europe PMC Funders Author Manuscripts

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. Most inner Eurasians are modeled as a mixture of components primarily found in eastern or western Eurasians. Results for the full set of individuals are provided in Supplementary Fig. 3.

Europe PMC Funders Author Manuscripts

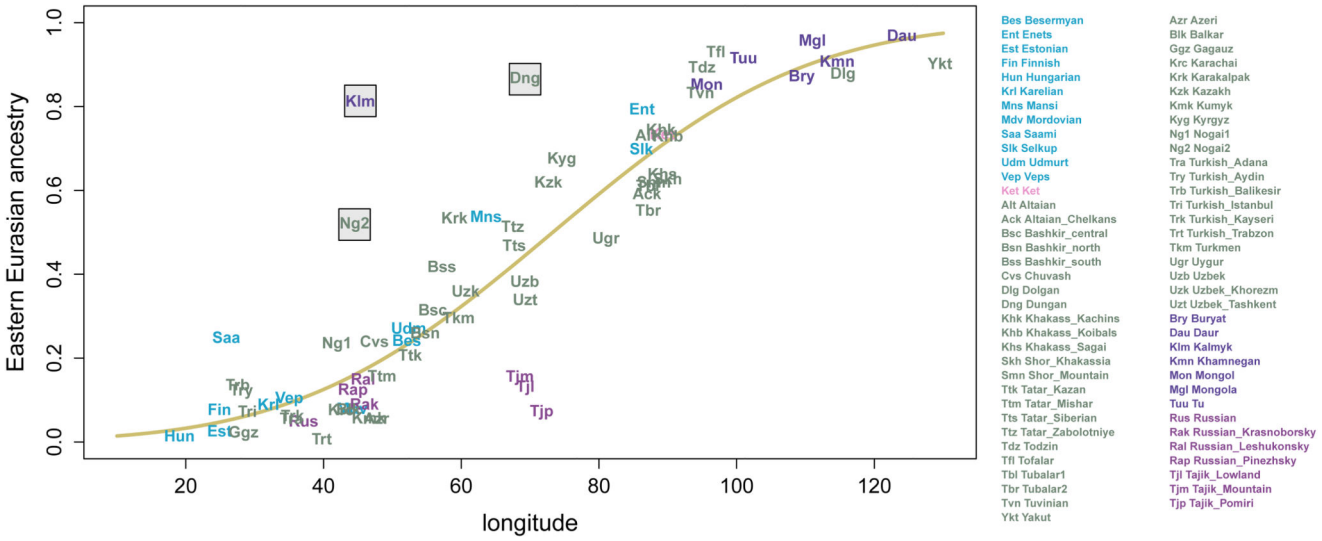

Fig. 3. Correlation of longitude and ancestry proportion across inner Eurasian populations.

Across inner Eurasian populations, mean longitudinal coordinates (x-axis) and mean eastern Eurasian ancestry proportions (y-axis) are strongly correlated. Eastern Eurasian ancestry proportions are estimated from ADMIXTURE results with K=14 by summing up 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 in Sedghifar et al.69. Three groups (Kalmyks, Dungans, Nogai2) are marked with grey square due to their substantial deviation from the curve as well as their historically known migration history.

Europe PMC Funders Author Manuscripts

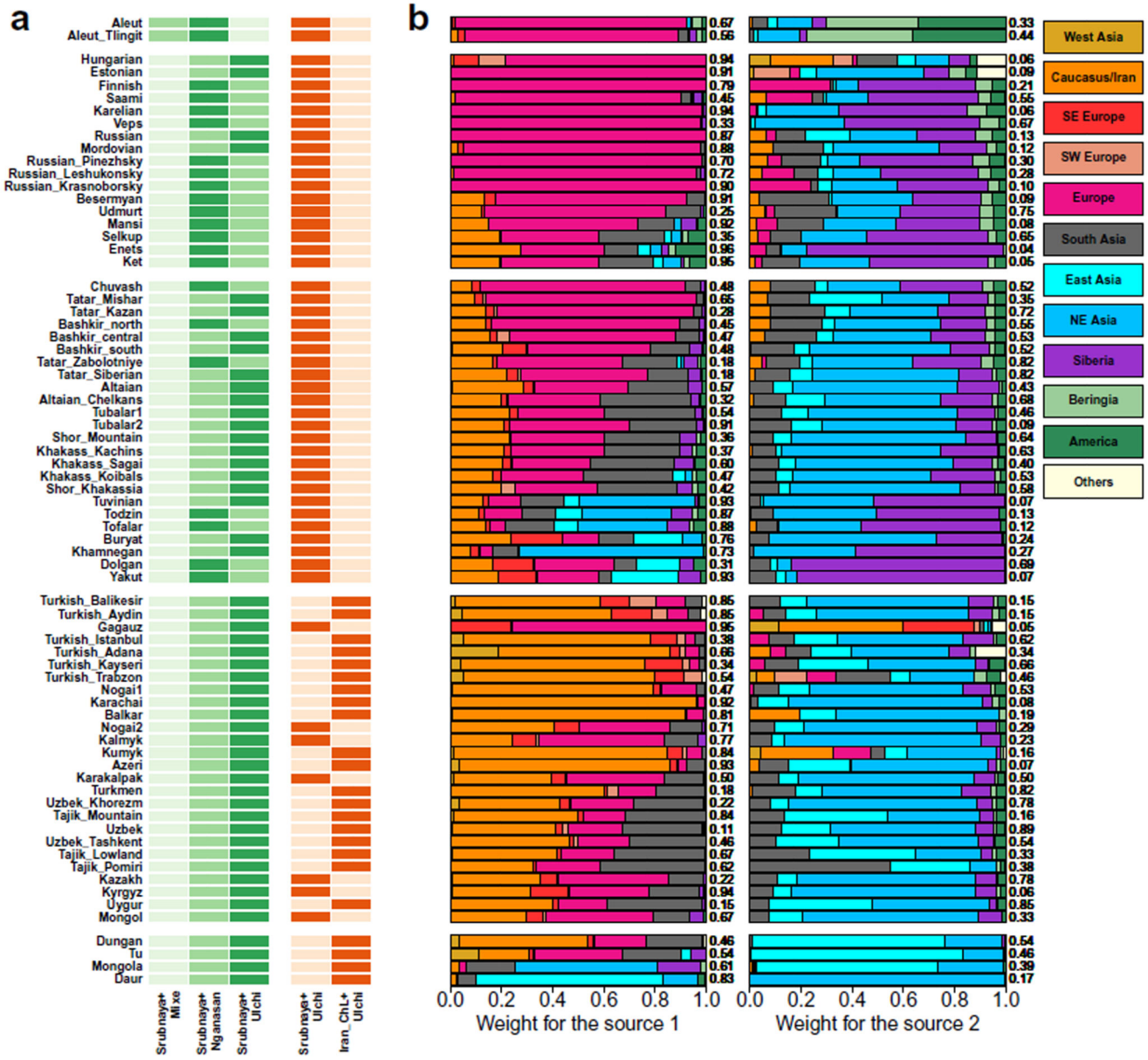

Fig. 4. Characterization of the western and eastern Eurasian source ancestries in inner Eurasian populations.

(a) Admixture f_3 values are compared for different eastern Eurasian references (Mixe, Nganasan, Ulchi; left) or western Eurasian ones (Srubnaya, Iran_ChL; right). 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 PC 1") in the GLOBETROTTER analysis. We merged 167 donor populations into 12 groups, as listed on the top right side. Target populations are split into five groups: Aleuts, the forest-tundra cline populations, the steppe-forest cline populations, the southern steppe cline populations and the rest of four populations ("others"), from the top to bottom.

Europe PMC Funders Author Manuscripts

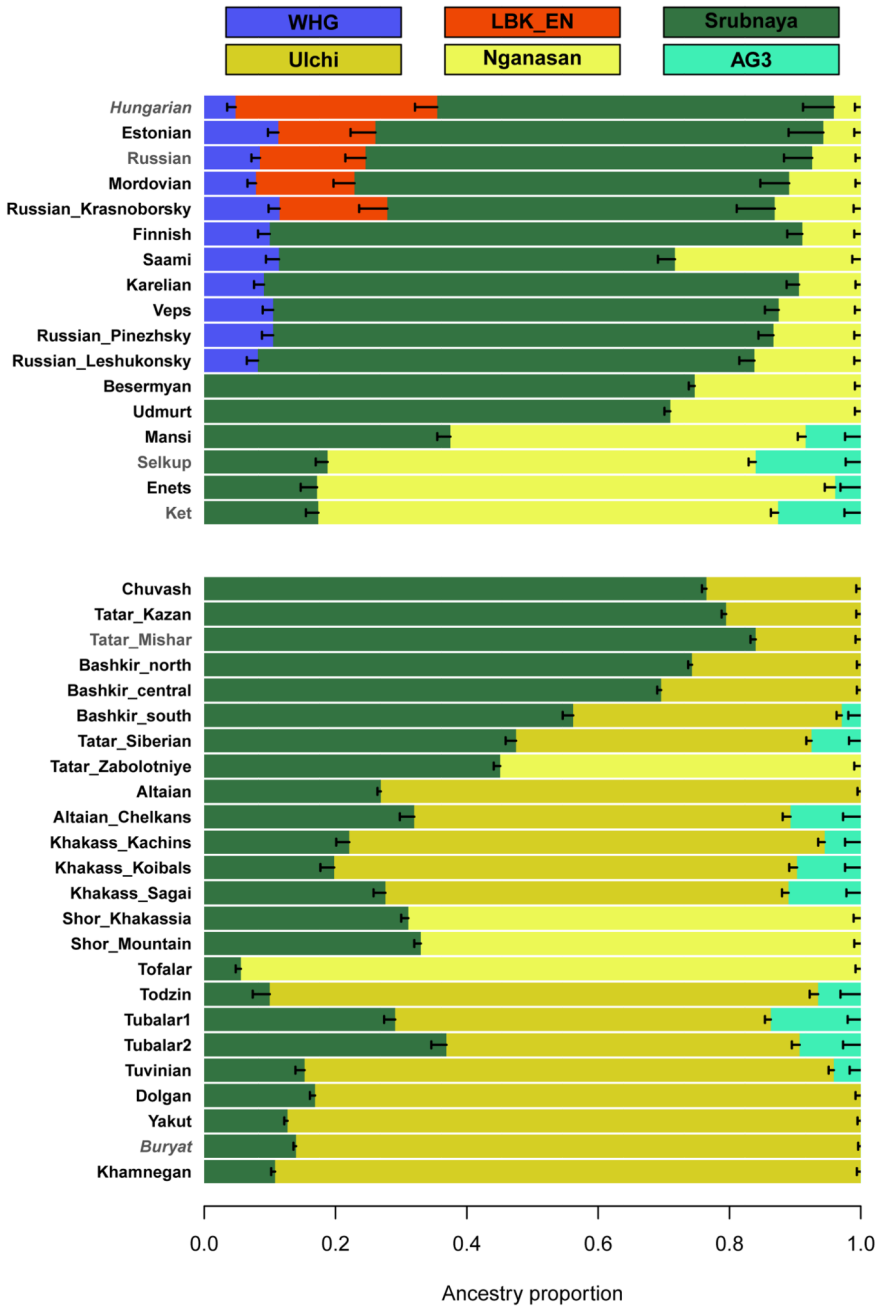

Fig. 5. qpAdm-based admixture models for the forest-tundra and steppe-forest cline populations.

For the forest-tundra population to the west of the Urals, Nganasan+Srubnaya+WHG +LBK_EN or its submodel provides 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, or Srubnaya +Nganasan provides a good fit. 5 cM jackknifing standard errors are marked by the horizontal bar. Models with p-value between 0.01 and 0.05 are marked by grey color and

those with p-value < 0.01 are marked by grey color and italic font. Details of the model information are presented in Supplementary Tables 5 and 8.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Sequencing statistics and radiocarbon dates of two Eneolithic Botai individuals analyzed in this study.

For Botai individuals we produced additional data, we provide corresponding individual ID from a previous publication23 (“Published ID”), radiocarbon date, the number of total reads sequenced, mean autosomal coverage for the 1240K target sites, the number of SNPs covered at least once for the 1240K and HumanOrigins panels, uniparental haplogroup and contamination estimates.

ID	Published ID	Genetic Sex	Uncal. ^{14}C Date	Cal. ^{14}C Date (2-sigma)[b]	# of reads sequenced	Mean autosomal coverage	# of SNPs covered[c]	MT / Y haplogroup	MT.cont[d]	X.cont[e]
TU45	BOT14	M	4620 ± 80[a]	3632-3100 cal. BCE	84,170,835	0.827x	169,053 (77,363)	K1b2 / R1b1a1	0.02 (0.01-0.03)	0.0122 (0.0050)
BKZ001	BOT2016	F	4660 ± 25	3517-3367 cal. BCE	69,678,735	2.420x	825,332 (432,078)	Z1 / NA	0.01 (0.00-0.02)	NA

[a] The uncalibrated date of TU45 was published in Levine (1999) under the ID OxA-431670.

[b] The calibrated ^{14}C dates are calculated based on uncalibrated dates, by the OxCal v4.3.2 program71 using the INTCAL13 atmospheric curve72.

[c] The number of SNPs in the 1240K panel (out of 1,233,013) or autosomal SNPs in the HumanOrigins array (out of 581,230; within the parenthesis) covered at least by one read. Only transversion SNPs are considered for the non-UDG libraries (both of the TU45 libraries, one of two BKZ001 libraries).

[d] The contamination rate of mitochondrial reads estimated by the Schmutzi program (95% confidence interval in parentheses)

[e] The nuclear contamination rate for the male (TU45) estimated based on X chromosome data by ANGSD software (standard error in parentheses)