The formation of human populations in South and Central Asia

Narasimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., Lazaridis, I., Nakatsuka, N., Olalde, I., Lipson, M., Kim, A. M., Olivieri, L. M., Coppa, A., Vidale, M., Mallory, J., Moiseyev, V., Kitov, E., Monge, J., Adamski, N., ... Reich, D. (2019). The formation of human populations in South and Central Asia. *Science*, *365*(6457), [aat7487]. https://doi.org/10.1126/science.aat7487

Published in:
Science

Document Version:
Peer reviewed version

The Genomic Formation of South and Central Asia

Vagheesh Narasimhan[3], Nick Patterson[1,2,*], Swapan Mallick[1,3,4], Nadin Rohland[1,3], Priya Moorjani[40], Nathan Nakatsuka[3], Luca M. Olivieri[16], James Mallory[24], Vyacheslav Moiseyev[15], Janet Monge[36], Niraj Rai[42,43], Alexander Kim[3], Iñigo Olalde[3], Iosif Lazaridis[3], Rebecca Bernardos[3], Nicole Adamski[3,4], Nasreen Broomandkhoshbacht[3,4], Francesca Candilio[6], Olivia Cheronet[6,21], Matthew Ferry[3,4], Daniel Fernandes[6], Beatriz Gamarra[6], Daniel Gaudio[6], Denise Keating[6], Ann-Marie Lawson[3,4], Mark Lipson[3], Jonas Oppenheimer[3,4], Megan Michel[3,4], Mario Novak[6], Kendra Sirak[6,7], Viviane Slon[32], Kristin Stewardson[3,4], Zhao Zhang[3], Gaziz Akhatov[27], David Anthony[33], Anatoly N. Bagashev[49], Baurzhan Baitenaev[27], Gian Luca Bonora[45], Tatiana Chikisheva[12], Alfredo Coppa[39], Anatoly Derevianko[12], Katerina Douka[25,34], Nadezhda Dubova[10], Andrey Epimakhov[18], Antonina Ermolaev[27], Suzanne Freilich[21], Dorian Fuller[50], Alexander Goryachev[27], Andrey Gromov[15], Bryan Hanks[19], Eadaoin Harney[3,4,5], Margaret Judd[19], Erlan Kazizov[27], Egor Kitov[27,30], Aleksander Khokhlov[41], Donata Luiselli[44], Farhad Maksudov[28], Chris Meiklejohn[22], Deborah C. Merrett[23], Roberto Micheli[16], Zahir Muhammed[26], Samridin Mustafokulov[28,29], Ayushi Nanak[25], David Pettner[44], Dmitry Razhev[20], Stefania Sarno[44], Kulyan Sikhymbaeva[30], Sergey M. Slepchenko[37], Nadezhda Stepanova[12], Svetlana Svyatko[14,24], Sergey Vasilyev[38], Massimo Vidale[16,17], Dima Voyakin[27,48], Alisa Zubova[12,15], Matthias Meyer[32], Carles Lalueza-Fox[35], Nicole Boivin[25,+], Kumarasamy Thangaraj[42,+], Douglas Kennett[8,+], Michael Frachetti[7,8,+], Ron Pinhasi[6,21,+], David Reich[1,3,4,5,+]

* Contributed equally
+ Co-directed this work

To whom correspondence should be addressed: V.N. (vagheesh@mail.harvard.edu), N.P. (nickp@broadinstitute.org), or D.R. (reich@genetics.med.harvard.edu)

[note: this author list is just an initial proposal; please write to David Reich with suggestions for changes and let D.R. know if there are missing authors.]

[1] Broad Institute of Harvard and MIT, Cambridge, MA, 02142, USA
[2] Radcliffe Institute for Advanced Study, Harvard University, Cambridge MA 02138
[3] Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA
[4] Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, 02115, USA
[5] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138, USA
[6] Earth Institute, University College Dublin, Dublin 4, Ireland
[7] Department of Anthropology, Washington University in St.Louis, Saint Louis MO, 63112
[8] Spatial Analysis, Interpretation, and Exploration Laboratory, Washington University in St.Louis, Saint Louis MO, 63112
[10] Ethnic Ecology of the Institute of Ethnology and Anthropology Russian Academy of Sciences, Moscow, Russia
[11] Institute for Anthropological Research, 10000 Zagreb, Croatia.

[12] Institute of Archaeology and Ethnography, Siberian Branch, Russian Academy of Sciences, 630090 Novosibirsk, Russia
[14] CHRONO Centre for Climate, the Environment, and Chronology, Queen's University of Belfast, Belfast BT7 1NN, Northern Ireland, UK
[15] Peter the Great Museum of Anthropology and Ethnography (Kunstkamera), Russian Academy of Science, 199034 St. Petersburg, Russia
[16] ISMEO Italian Archaeological Mission in Pakistan, missing.city missing.postal.code, Pakistan
[17] Department of Cultural Heritage: Archaeology and History of Art, Cinema and Music, University of Padua, missing.postal.code Padua, Italy
[18] Institute of History and Archaeology (Ural Branch RAS) and South Ural State University, 454080 Chelyabinsk, Russia
[19] University of Pittsburgh, Department of Anthropology, Pittsburgh, PA, 15260, USA
[20] Institute of Problems of Development of the North of the Siberian Branch of the Russian Academy of Sciences, 625026 Tyumen, Russia
[21] Department of Anthropology, University of Vienna, 1090 Vienna, Austria
[22] Department of Anthropology, University of Winnipeg, Winnipeg, MB, R3B 2E9, Canada
[23] Department of Archaeology, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada
[24] School of Natural and Built Environment, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland, UK
[25] Max Planck Institute for the Science of Human History, 07745 Jena, Germany
[26] Department of Archaeology, Hazara University, Mansehra, 21300, Pakistan
[27] Institute of Archaeology A.Kh. Margulan, Almaty 050010, Kazakhstan
[28] Institute of Archaeology, Uzbek Akademy of Sciences, Samarkand 140151, Uzbekistan
[29] Afrosiab Museum, Samarkand 140151, Uzbekistan
[30] Center of Physical Anthropology, Institute of Ethnology and Anthropology RAS Lenina, Moscow 119334, Russia
[31] Central State Museum Republic of Kazakhstan, Samal-1 microdistrict, Almaty 050010, Kazakhstan
[32] Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany
[33] Anthropology Department, Hartwick College, Oneonta, New York 13820, USA
[34] Oxford Radiocarbon Accelerator Unit, Research Laboratory for Archaeology and the History of Art, University of Oxford, Oxford OX1 3QY, UK
[35] Institute of Evolutionary Biology, CSIC-Universitat Pompeu Fabra, Barcelona 08003, Spain
[36] University of Pennsylvania Museum of Archaeology and Anthropology, Philadelphia, PA 19104, USA
[37] Institute for Problems of the Development of the North, Siberian Branch of the Russian Academy of Sciences, 625026 Tyumen, Russia
[38] Institute of Ethnology and Anthropology, Russian Academy of Sciences, Moscow 119991, Russia
[39] Dipartimento di Biologia Ambientale, Sapienza Università di Roma, 00185 Roma, Italy
[40] Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley 94720, USA
[41] Samara State University of Social Sciences and Education, Samara 443099, Russia
[42] CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India
[43] Birbal Sahni Institute of Palaeosciences, Lucknow, India.
[44] Laboratory of Molecular Anthropology, Department of Biological, Geological and Environmental Sciences, University of Bologna, 40126 Bologna, Italy
[45] Archaeology of Asia Department, International Association of Mediterranean and Oriental Studies (ISMEO), RM00186 Roma, Italy
[46] Department of Anthropology and Institutes for Energy and the Environment, Pennsylvania State University, University Park, PA 16802, USA
[47] Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA
[48] Archaeological Expertise LLP, 050060 Almaty Kazakhstan
[49] Institute for Problems of the Development of the North, Siberian Branch of the Russian Academy of Sciences, Tyumen 625026, Russia
[50] Institute of Archaeology, University College London, London WC1H 0PY, UK

Abstract

The spread of farmers from the Near East in the 7th millennium BCE and pastoralists from the Eurasian steppe in the 3rd millennium BCE transformed the genetic makeup of Europe and India, but the process by which these ancestry types expanded East remains mysterious. We generated genome-wide ancient DNA from 327 individuals including never-before-sampled material cultures from the southeastern Steppe, the Bactria Margiana Archaeological Complex (BMAC), and the first data from South Asia—and show how almost all were formed from mixtures of seven deeply divergent populations. We document two West-to-East ancestry gradients in the Bronze Age—in the North overlaid onto steppe pastoralist ancestry, and in the South onto Iranian farmer related ancestry—and both with more ancestry related to Anatolian farmers in the West and Siberian hunter-gatherers in the East. We show how agro-pastoralists of the southeastern Steppe spread further South in the 2nd millennium BCE, bypassing the BMAC, to mix with peoples in the Indus Valley at the extreme of the southern gradient, thereby creating one of the two main source populations of South Asia today, the Ancestral North Indians (ANI). By co-analyzing with modern data we show that the other main source population, the Ancestral South Indians (ASI), arose in the last five thousand years as a mix of ~25% Iranian farmer ancestry and ~75% indigenous South Asian hunter-gatherer ancestry, showing that both extremes of present-day Indian variation were strongly affected by Iranian admixture. While the ANI-ASI mixture model works for most Indians, it fails for a subset of Brahmins and Bhumihar groups that are among the traditional custodians of texts written in early Sanskrit, and have a ratio of Steppe-to-Iranian-farmer-related ancestry significantly higher than in other groups. This provides a second genetic line of evidence—beyond the large-scale Middle Bronze Age spread of pastoralists from the Steppe we document here—for an origin of Indian Indo-European culture in the Steppe.

Ancient DNA Data

We generated whole-genome ancient DNA from 327 never-before-reported ancient individuals and higher quality data on 18 previously reported individuals. Almost all derive from three broad regions: 115 from Iran and the southern part of Central Asia sometimes called Turan ("Iran/Turan"), 164 from the western and central Steppe and northern forest zone encompassing present day Kazakhstan and Russia ("Steppe"), and 48 from northern Pakistan, the first data from South Asia ("South Asia"). Our dataset includes the first ancient DNA from Copper and Bronze Age eastern Iran and Turan (3700-1000 BCE from 12 sites); the first Neolithic hunter-gatherers from the Siberian forest zone (6400-4000 BCE from 2

sites); Copper and Bronze Age pastoralists from the steppe East of the Ural mountains including the first data from Kazakhstan (3200-1000 BCE from 35 sites); and Iron Age settlements in the Swat Valley of Pakistan (1200-0 BCE from 7 sites) (Figure 1; Supplementary Text; Extended Data Table 1; Online Table 1). To generate these data, we prepared samples in dedicated clean rooms, extracted DNA, built it into libraries for Illumina sequencing,(*1*) and screened it using previously described procedures (Methods).(*2, 3*) We enriched samples for DNA fragments overlapping about 1.24 million SNPs, sequenced the products, and performed quality control as described previously (Data S1, Data S2).(*3, 4*),(*5*) We also generated 178 new direct radiocarbon dates (Data S3). After grouping based on archaeological and chronological information and merging with previously reported data, our dataset included 731 ancient individuals which we co-analyzed with genome-wide data from 3,066 present-day individuals assessed at about 0.6 million single nucleotide polymorphisms (SNPs), 1,789 of which were from 246 ethnographically-distinct groups in South Asia (Data S4; Supplementary Text). We restricted most analyses to ancient samples having at least 15,000 SNPs when merged with the present-day individuals.

Analysis Strategy

We carried out Principal Component Analysis (PCA) projecting the ancient samples onto patterns of genetic variation in present-day Eurasians (Figure 1).(*6, 7*) This analysis revealed three major groupings, closely corresponding to the geographic regions of Steppe, Iran/Turan and South Asia, a pattern we replicate in ADMIXTURE clustering.(*8*) To formally test whether populations differ significantly in their ancestry within region, we used symmetry-f_4-statistics that measure whether pairs of populations differ in their degree of allele sharing to a third population, and admixture-f_3 statistics to test formally for mixture (Text S1).(*9*) We tested the fit of mixture models using *qpAdm*,(*10*) which evaluates whether the set of f_4-statistics relating a set of tested populations to outgroup populations is consistent with mixtures of a pre-specified number of sources and if so infers mixture proportions. We can successfully model almost every population as a mixture of seven "distal" ancestry sources: (1) Anatolian farmer-related ancestry, (2) Western Hunter Gatherer (WHG) ancestry, (3) Iranian farmer-related ancestry, (4) Eastern European Hunter-Gatherer (EHG) related ancestry, (5) West Siberian hunter-gatherer related ancestry, (6) East Asian related ancestry, and (7) South Asian hunter-gatherer ancestry. We also used *qpAdm* to identify "proximal" models for each group as mixtures of temporally preceding groups. We identified numerous alternative proximal models that fit, but the qualitative findings tended to be consistent.

Formation of the Ancestry Gradient in the South

We analyzed our new data together with previously published data to examine the genetic transformations that accompanied the spread of agriculture eastward from Iran beginning in the 8th millennium BCE. Our analysis confirms that early Iranian farmers occupy an extreme position in the genetic variation of West Eurasians (Figure 1, Figure S1),(*11, 12*) while later groups in Iran were admixed between this type of ancestry and that related to early Anatolian farmers.(*11*) Using our new data from Copper and Bronze Age eastern Iran and Turan (present-day Uzbekistan, Tajikistan and Turkmenistan), we show that the Anatolian admixture so far only documented in a single place in Iran was part of a West-to-East Copper Age cline of decreasing Anatolian farmer-related admixture ranging from ~70% in Copper Age Anatolia to ~33% in Eastern Iran to ~3% in far Eastern Turan (Figure 1; Supplementary Text). In the eastern part of this cline (Turan and eastern Iran) we detect admixture related to west Siberian hunter-gatherers, proving that ancestry related to North Eurasian hunter-gatherers impacted Turan well before the spread of steppe pastoralists (*Steppe_EMBA*).

Moving to the Middle Bronze Age in Turan, we examined 69 ancient individuals from four urban sites of the BMAC and post-BMAC spread over 2500-1500 BCE. The great majority fall in a cluster that is similar to the preceding groups in having a large Iranian farmer related ancestry component (~68%) with smaller components of Anatolian farmer related ancestry (~18%) and west Siberian hunter-gatherer related ancestry (~14%), suggesting that the BMAC population coalesced from preceding pre-urban populations and ruling out the hypothesis(*13*) that the BMAC were a link in the chain-of-transmission that brought the steppe pastoralist ancestry ubiquitous in South Asia today (Supplementary Text). The data instead suggest gene flow in the opposite direction: the main BMAC cluster has a proportion of ancestry from an indigenous Asian source, which we can model as derived from ancestors of the SPGT, some of whom may have expanded North in the preceding millennium.

The individuals buried in the BMAC necropolises also included striking outliers that provide critical insight about population transformations in the region. First, around ~2000 BCE, we observe 3 outliers with west Siberian hunter-gatherer related ancestry of a type present in Kazakhstan over the preceding and succeeding millennia but no ancestry from Yamnaya-related steppe pastoralists (*Steppe_EMBA*). However, *Steppe_EMBA* ancestry in the admixed form carried by groups associated with the Andronovo material culture was very common in

Turan between 1800-1500 BCE as all outliers from three sites in this period had high proportions of it consistent with the archaeological evidence of intensified contact in this period.(*14*) The outliers in the BMAC thus provide evidence for a southward movement of Steppe ancestry through this region that only began to have a major impact after the turn of the 2nd millennium BCE. Second, at ~2000 BCE we observe a BMAC outlier with an ancestry profile extremely similar to ancient DNA samples from the Swat Valley of northern Pakistan from the Swat Protohistoric Grave culture (*SPGT*) who lived approximately a millennium later (1200-800 BCE). Both the BMAC outlier and the Swat individuals are distinctive in having ~18% ancestry from South Asian hunter-gatherers. Based on evidence of trade between BMAC and the contemporaneous Indus Valley Culture (IVC) towns,(*15-17*) and the similarity to the Swat samples, we hypothesize that this outlier was from a family that migrated from the IVC, which if true would make it the first DNA sample from this culture. Rigorous testing using *qpAdm* reveals that the SPGT differ subtly in ancestry from the migrant to the BMAC, and are consistent with harboring 10% *Andonovo_SE* ancestry. A parsimonious explanation is that the *SPGT* were isolated descendants of the same population that produced the southern migrant to the BMAC but with a small amount of *Andronovo_SE* admixture of the same type that as we show admixed to a larger extent with one of the two major ancestral populations of South Asia, the ANI. We also report data from two later sites in the Swat Valley from 450-0 BCE, showing that after the *SPGT,* the region was impacted by movements of people with increasing fractions of South Asian ancestry more similar to what is observed across the region today.

Formation of the Ancestry Gradient in the North

Three samples from the West Siberian forest zone dated to 6400-4000 BCE are critical to this study as they are of a never-before-reported ancestry that we call *West_Siberian_HG* (West Siberian Hunter-Gatherer) that can be modeled as 30% derived from Eastern European Hunter-Gatherers (EHG), 50% from Ancestral North Eurasians like an ~24,000 year old Siberian,(*18*) and about 20% related to East Asians. This ancestry existed not just in the forest zone but also in the southern steppe and in Turan as it contributed about 80% of the ancestry of an early 3rd millenium BCE from Kazakhstan, and also contributed to multiple outlier individuals from 2nd millennium sites in Kazakhstan and Turan. Using these samples and ones from previously reported ancestry types, we document the formation of a West-to-East gradient of ancestry by 2500-1400 BCE, characterized by a declining proportion of Anatolian farmer-related ancestry of ~40% in *Corded_Ware*, and ~26% in the *Srubnaya*,

Sintashta and *Andronovo_NW* cultures—all superimposed on a substrate of early Bronze Age Steppe Pastoralist-related ancestry (*Steppe_EMBA*). This reflects a previously reported phenomenon in which following westward movement of *Steppe_EMBA* groups into central Europe and admixture with local European farmers, there was eastward genetic flow beyond the Urals, with the European (Anatolian) farmer-related ancestry being diluted by admixture with previously established *Steppe_EMBA* groups (Figure 2).(*19, 20*) Our new data also extend this gradient to the far southeast – to present-day Kazakhstan – and show that the *Andronovo_SE* grouping there had distinctively higher *West_Siberian_Neolithic* derived ancestry and less Anatolian Neolithic-related admixture than previously reported groups from the Northwest ("*Andronovo_NW"*).(*19*) This is important since as shown below, this signature makes *Andronovo_SE* a plausible source for present-day groups in South Asia who have too little Anatolian Neolithic related ancestry to be fit by any previously reported Late Bronze Age steppe groups (Supplementary Text).(*11*) The significant ancestry difference between *Andronovo_NW* between *Andronovo_SE* accords with the archaeological evidence for a difference in economic subsistence between these two groups with extensive reliance agriculture in the southeast,(*21*) which is intriguing as it could be related to the differences in mobility patterns and expansion of *Andronovo_SE* further South that we document here.

Our large sample sizes also allow us to document ancestry heterogeneity which is informative about the history. Our analysis of 51 newly reported samples from the Kamennyi Ambar V cemetery from the Sintashta culture, the largest ancient DNA study of a single site to date in the ancient DNA literature, reveals in addition to the main cluster of xx individuals, three groups of outliers with direct radiocarbon dates contemporaneous with the other samples and elevated proportions of *Steppe_EMBA* or *West_Siberian_Neolithic* related ancestry, providing evidence that these fortified steppe sites harbored people of diverse ancestries living side-by-side (Supplementary Text). Second, samples from three sites from the southern and eastern end of the Steppe, dated to xxxx-xxxx BCE and contemporaneous with the late BMAC, show evidence of significant admixture from Iranian farmer related populations demonstrating northward gene flow from Turan into the steppe in this period just as there was southward gene flow through Turan and into South Asia. Third, from xxxx-xxxx BCE we observe multiple sites that derive up to ~25% of their ancestry from a source related to present day East Asians and the rest from populations from *Steppe_EMBA* cluster. This type of ancestry became widespread in the region by the early Iron Age (Scytho-Sarmatian period)(*22*) and these samples provide a minimum date for when it arrived.

Iranian Farmer Related Ancestry at Both Extremes of the Indian Cline

We co-analyzed our newly reported ancient DNA data with data from diverse present-day Indians(*23*) to gain insight into the deep ancestry sources of the "Indian Cline,"(*24-26*) the primary driver of genetic variation across South Asia, represented here by 140 groups that fall on this cline in a Principal Component Analysis (Supplementary Text). Previous work has shown that the Indian Cline can be well modeled as having arisen from a mixture of two statistically reconstructed ancestral populations (the ANI and the ASI) largely between 2000-0 BCE.(*24, 25*) Ancient DNA analysis has furthermore shown that Indian Cline groups descend more deeply from at least three ancestral populations: early Iranian farmers, *Steppe_EMBA*, and an unsampled South Asian hunter-gatherer group related to indigenous Andamanese (*Onge*)(*11*) (we refer to this deep Asian ancestry source as the "Ancient Ancestral South Indians" (AASI), to distinguish it from the more Asian extreme of Indian Cline, referred to in this study as the ASI). To shed light on the mixture events that transformed this minimum of three ancestral populations into two (the ANI and ASI), we used *qpAdm* to evaluate if each of the Indian Cline groups in turn, *Onge*, and diverse pairs of ancient West Eurasian groups from the Copper Age onward, were consistent with descending from just three ancestral populations relative to distantly related outgroups (Supplementary Text). We obtain fits models for the great majority of Indian Cline groups when one source was *Onge*, one was *Steppe_EMBA*, and third was a Turan Copper or Bronze Age group (Figure 3, Supplementary Text). (We also obtained fits for a variety of northern populations when one source was *SPGT*, a more proximal set of models we return to below.)

Commented [DR1]: Maybe add a table showing these fits

We found that the per-group *qpAdm* estimates for these three source populations are statistically noisy, and we therefore developed new methodology that allows us to jointly fit the data from all Indian Cline groups within a single hierarchical model. The analysis both confirms that almost all groups on the Indian Cline can be jointly modeled as a mixture of two populations (Supplementary Text), and produces estimates of the functional relationship between the ancestry components: (*Steppe_EMBA*) = 62% – 82%(*Onge*-related), with standard errors of ±3% (Supplementary Text). Setting *Onge*-related ancestry to 1 to determine what would be expected if the ASI had no West Eurasian related ancestry, we infer negative *Steppe_EMBA* ancestry, a nonsensical result. Setting *Steppe_EMBA*=0%, the smallest proportion it can possible be, we find that the ASI must have had no less than 25±3% Iranian farmer-related ancestry. In fact, our estimate in real Indian Cline groups

shows that groups with this ancestry type exist today contradicting previous suggestions by some of the primary authors of the present study that unmixed descendants of the ASI may no longer exist in India (Figure 3).(*24*) To further probe the finding of Iranian farmer-related ancestry in the ASI, we computed statistics of the form f_4*(Steppe_EMBA, Tepe_Hissar; Onge, Test)*, where *Test* is an Indian Cline group with α ANI ancestry. This is expected to be α f_4*(Steppe_EMBA, Tepe_Hissar; Onge, ANI)* under the null hypothesis that the ASI had no West Eurasian related ancestry, and thus is expected to increase in magnitude with more West Eurasian-related ancestry. In fact, the magnitude becomes significantly smaller with increasing α (Z = -7.7 standard errors from zero), confirming our finding that the ASI were related to early Iranian farmers (Supplementary Text).

To further understand the deep relationship between these groups, we built an admixture graph using *qpGraph(9)* that co-models the *Palliyar* (one of the present-day Indian Cline groups consistent with being direct descendants of the ASI) and the *Juang* (an Austroasiatic speaking group in India with the least West Eurasian relatedness), and show that it fits when the ASI are ~27% Iranian farmer-related in ancestry and the *Juang* also harbor ancestry from an AASI population without Iranian admixture (Figure 3). This admixture graph along with related *qpAdm* analyses is also notable in showing that *Tepe_Hissar* fits without *AASI* admixture, and thus there is no evidence for AASI gene flow into ancient Iran implying that the patterns we are observing are driven in major part by gene flow into South Asia (Figure; Supplementary Text). The fitted admixture graph also reveals the deep ancestry of the indigenous hunter-gather population of India, showing that South Asia harbors an extraordinarily anciently divergent branch of Asian human variation that split off around the same time as *Onge* and *Australian* aboriginal ancestors separated from each other. Our results are consistent with a model in which essentially all the ancestry of present-day Asians derive from a West-to-East migration, which first budded off the AASI, the *Onge*, and finally, *Australians*.(*27*) These splits must have occurred in a ~5,000 year period prior to the Denisovan gene flow into the ancestors of *Papuans* and *Australians* 49,000-44,000 years ago(*28*) and after *Neanderthal* admixture into all non-Africans 54,000-49,000 years ago, known to have occurred ~12% earlier.(*29, 30*)

Using admixture linkage disequilibrium, we estimate a date of 134±22 generations ago for the Iranian farmer and AASI-related admixture in the *Palliyar*, corresponding to a 90%

confidence interval of 4,800-2,700 years ago assuming 28 years per generation.(*30*) Thus, the ASI must have been formed during Bronze Age population transformations in the Indian subcontinent not associated with groups of steppe origin. The finding that the ASI were not fully formed until after 5,000 years ago is also consistent with the fact that Austroasiatic cluster groups like the *Juang*—who descend from groups that likely arrived in South Asia only after around 5,000 years ago based on the expansion of East Asian rice farming technology likely associated with the spread of Austroasiatic languages(*31*)–have a higher ratio of AASI-to-Iranian farmer related ancestry than the ASI (Figure 3), implying that groups of variable Iranian- and AASI-related ancestry were present in peninsular India around that time and the ASI had not yet overspread India.

Constraints on the Origins of the Indian Cline in Light of Ancient DNA

The model of the Indian Cline as a three-way admixture of groups related to *Onge, Steppe_EMBA,* and *Tepe_Hissar,* while useful for documenting the presence of Iranian farmer-related admixture in both the ANI and ASI, is not in fact plausible. One set of reasons is archaeological—there is no evidence for a *Steppe_EMBA* presence in Kazakhstan or further south in the ancient genomes from Iran and Central Asia we analyze here, whereas we document direct evidence of *Andronovo_SE* ancestry moving southward between 1800-1500 BCE. The other sets of reasons are genetic. First, nearly all *Steppe_EMBA* males to date harbor Y chromosome haplogroup R1b that is nearly absent in India today whereas the *Andronovo_SE* cluster like all the other Middle and Late Bronze Age steppe groups harbored a high frequency (in the case of *Andronovo_SE,* 100%) of Y chromosome haplogroup R1a which is common in both eastern Europe and in South Asia (Figure 1).(*32, 33*) Second, we detect no *Steppe_EMBA* admixture in sites in Kazakhstan (Dali and Kairan) or Turan until the second millennium BCE at which point this ancestry (in admixed form in *Andronovo_SE*) suffuses Kazakhstan and is evident in Turan in outlier individuals at multiple BMAC sites. Third, in the Supplementary Text we show that Swat Valley Iron Age samples (*SPGT)* can be fit in *qpAdm* as an admixture of *Gonur2_Andronovo_SE*, but not *Gonur2-Steppe_EMBA*. Fourth, beginning in 1500 BCE as we show in this study and intensifying in the Iron Age,East Asian admixture becomes ubiquitous in Turan and Kazakhstan, suggesting that populations after this time—including Scythians, Kushans, and Huns sometimes suggested as sources for the steppe ancestry influences in India—are unlikely to have contributed to the great majority of groups in South Asia who have negligible East Asian related ancestry.(*22*)

Taken together, these four lines of evidence provide a compelling case that *Andronovo_SE* dispersed southward through Turan into the northern part of South Asia in the first half of the 2nd millennium BCE.

Motivated by the finding that *Steppe_EMBA* was implausible as a direct source of the steppe pastoralist related ancestry in South Asia, we turned to the other class of fitting models in Supplementary Text: the ones involving groups related to *Palliyar* (to represent the ASI), *Andronovo_SE* and either *SPGT* or the outlier *BMAC* individual *Gonur2*. Figure 3 shows that when we use *Gonur2* as the source of the Iranian-related ancestry, the ANI can be modeled as xx% *Andronovo_SE* and yy% from the population of which *Gonur2* was a part. It is tempting to think that this population derived from peoples who were part of the Indus Valley Complex (IVC), but without ancient DNA data we cannot rule out the possibility that many IVC had a much higher proportion of *AASI* ancestry and that admixing populations like *Gonur2* were limited to the northern fringe of the IVC. However, we can definitively rule out the possibility that the main cluster of the *BMAC* combined with *Andronovo_SE* to form the ANI as models of this type fail (Supplementary Text). These results suggest that the BMAC was an ancestry cul-de-sac: a population that was affected by the same demographic forces that later impacted South Asia (southward gene flow from *Andronovo_SE*), but was bypassed by other southward-oriented genetic pulses.

The history of the formation of ASI is less clear. Iranian farmer-related groups plausibly first spread into South Asia from the Iranian plateau some time after 7000 BCE when there is the first archaeologically attested spread of Iranian farming technology to the hills surrounding the Indus Valley, although there is not yet any ancient DNA from the Indus Valley to test this. However, the 2850-750 BCE date of admixture suggests that even if this ancestry came into South Asia around this time, it did not immediately homogenize with the AASI, and instead continued to admix with it well after the beginning of the Bronze Age. One plausible scenario is that the AASI-Iranian admixture occurred as a result of the spread of agriculture and pastoralism from the Indus Valley into peninsular South Asia 3000-2000 BCE(*17, 20, 34-37*) However, an alternative is that the southward movement of *Andronovo_SE* groups precipitated the formation of both the ANI and the ASI: the ANI through admixture with peoples with ancestry like *Gonur2* and *SPGT* who were plausibly similar in ancestry to the IVC, and the ASI as unfavorable circumstances pushed these people into peninsular India after ~1700 BCE to mix with the AASI.

Conclusions

These results shed light on the spread of Indo-European languages in two ways. First, our documentation of a large-scale southward spread of *Andronovo_SE* groups in the 2nd millenium provides a prime candidate for the spread of late proto-Indo-European languages southward into India from the Steppe where they were likely spread by Yamnaya pastoralists at the beginning of the 3rd millenium.(*13*) It has been argued—controversially—that material culture remains from *Andronovo_SE* groups have connections to early Vedic culture in India,(*13*) and our results of an ancestry link to later cultures to the South adds weight to this theory. Second, our genetic analysis also provides an entirely new line of evidence for a connection of steppe ancestry to Indo-European culture. When we used *qpAdm* to test if a mixture of ANI and ASI is a fit to the data for all 140 Indian Cline groups, we found five with poor fits and a significantly higher ratio of *Andronovo_SE* to *Gonur2*-like ancestry than other groups in the Indian Cline. These were all groups of priestly status (Brahmins or Bhumihars), with the strongest signal seen in *Brahmin_Tiwari* ($P<2\times10^{-7}$) (Supplementary Text; Online Table 2). Thus, at the time the ANI began mixing with the ASI, there was likely a substructured meta-population including subgroups with relatively higher proportions of steppe ancestry that may have had a disproportional role in spreading early Vedic culture, and this stratification has been preserved even to this day by the strong endogamy rules in India. A working model for the formation of Indian Cline groups is shown in Figure 4.

Taken together, our results reveal a remarkable parallel between the prehistory of two sub-continents of Eurasia: South Asia and Europe.(*19, 38*) In both regions, agriculture spread from an origin in the Near East after 7000 BCE. In South Asia this occurred via the Iranian plateau, and in Europe via western Anatolia, with the technological spreads mediated in both cases by movements of people. An admixed population formed by the incoming farmers and resident hunter-gatherers developed over thousands of years – in South Asia ASI, and in Europe Middle Neolithic, although the admixture proportions differ, for example in the ASI they were around 70% hunter-gatherer which is much higher than in the European Middle Neolithic where they were around 20% hunter-gatherer. (Speculatively, the higher hunter-gatherer proportion in South Asia than in Europe reflects the fact that the winter/summer rainfall barrier to the spread of farming was greater in South Asia than the latitude barrier in Europe, which allowed the hunter-gatherers more time to adapt to the new technologies and mix with the farmers.) A new wave of migrants then arrived into both subcontinents, who

were a mixture of ancestry ultimately related to *Yamnaya* Steppe pastoralists, and the farmers they encountered along their path. In South Asia this mixed population became the ANI, and in Europe these were the people buried with Corded Ware pottery, with their shared ancestry related to *Yamnaya* steppe pastoralists providing support for the hypothesis that expansions of the *Yamnaya* or people related to them were a primary driver for Indo-European language spread both in Europe and in India.(*38, 39*) These two mixed populations then mixed in turn, forming gradients of ancestry both in Europe via mixtures of Middle Neolithic and Corded Ware-like groups, and in India via mixtures of ANI and ASI (Figure 5).

Data availability

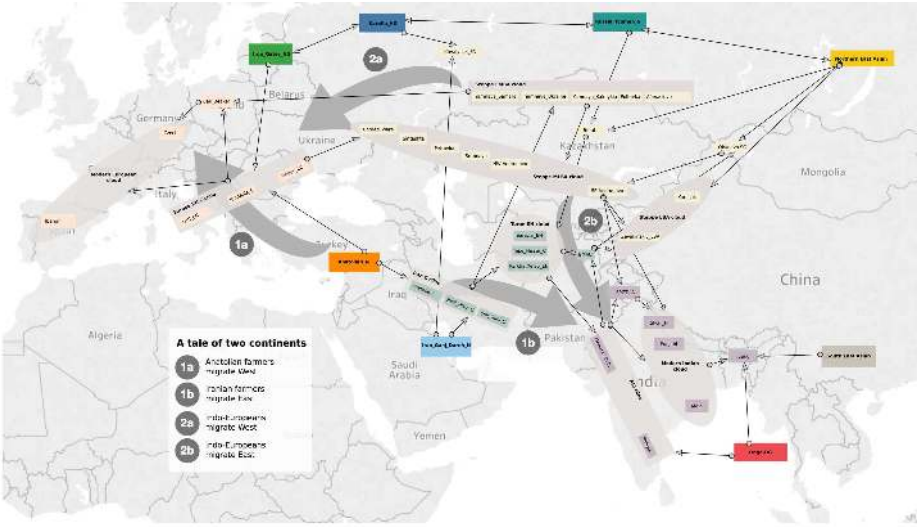
All sequencing data are available from the European Nucleotide Archive, accession number XXXXXXXX [to be made available on publication]. Genotype data obtained by random sampling of sequences at approximately 1.24 million analyzed positions are available from the Reich Lab website at [to be made available on publication].

Acknowledgments

We are grateful to the Minusinsk Regional Museum n.a. N. M. Martyanov for sharing some of the skeletal samples analyzed in this study. N.P. carried out this work while a fellow at the Radcliffe Institute for Advanced Study at Harvard University. T.C. and A.D. were supported by the Russian Science Foundation (project no. 14-50-00036). 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.

Figure 1: Overview of the newly reported ancient DNA data. (A) Number of newly samples passing our analysis thresholds and their age range is shown by site. (B) Locations, color-coded by analysis grouping. (C) Projections onto PCAs of all Eurasians. (D) ADMIXTURE analysis, components maximized in West Siberian Hunter Gatherer, Anatolian farmers, Iranian farmers, Indigenous South Asians and Western Hunter Gatherers in blue, orange, teal, red and green respectively (E) Y-chromosome calls.

Figure 2: Modeling results. (A) Approximate geographic positions of relevant samples and a parsimonious set of proximal models reflecting admixture between populations as modeled by *qpAdm*. Populations modeled by a mixture of two sources are shown in rectangles and populations modeled as a mixture of three or more sources are shown in ellipses. Pre-Copper Age or modern outgroup populations reflecting deeply divergent ancestry are shown in colors, and in (B) we break down the ancestry of the different sites into components from these sources.



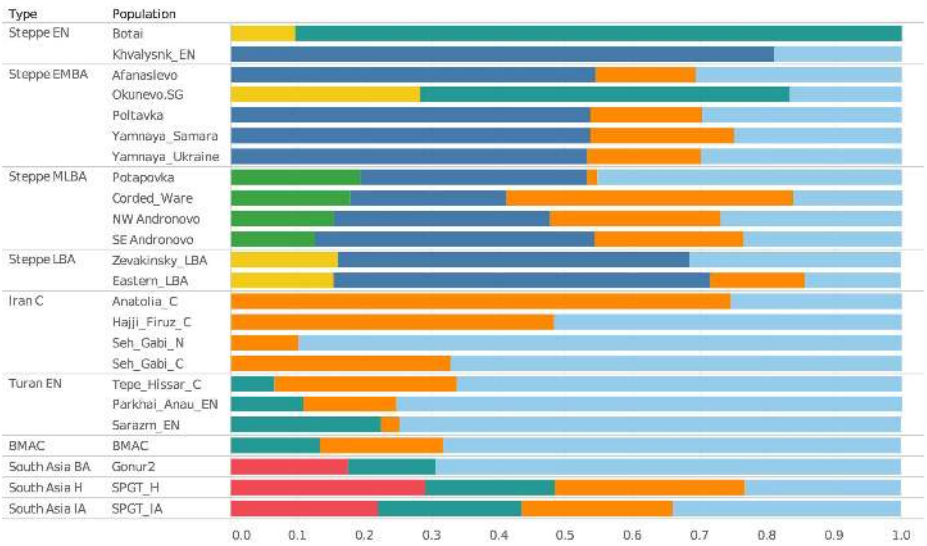
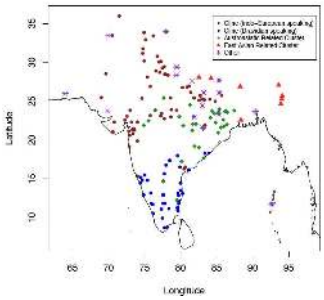


Figure 3: Origins of the Indian Cline in Light of Ancient DNA. (A) We analyzed 246 South Asians genotyped on the Affymetrix Human Origins array, colored here by language group and whether we analyze them as part of the Indian Cline (B) (B, C) Modeling using *qpAdm* shows two fits of the Indian Cline using three source populations. In B we show *Maximum A Posteriori* estimates of Steppe_EMBA (Yamnaya) - related, AASI (Onge) -related, and Iranian (Tepe_Hissar) farmer-related ancestry, showing that groups with all ASI (defined as minimal Steppe) ancestry had 25% ± 3% Iranian farmer-related ancestry and still exist today. In C we show *Maximum A Posteriori* estimates of Steppe (Southeastern Andronovo) - related, ASI (Palliyar) - related, and Iranian-related (Gonur2) ancestry. (D) Admixture graph fit to the data.

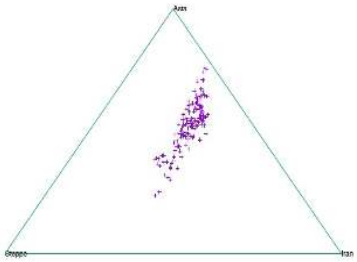
A (locations of analyzed groups)



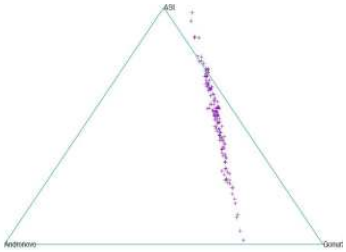
B (minimum P-value for qpAdm fit at two extremes of the Cline)

Steppe pastoralist-related source \ Iranian farmer-related source	*Caucasus Hunter-Gatherers (CHG)*	*Armenian Early Bronze Age*	*Hajji Firuz (Western Iran)*	*Seh_Gabi (Western Iran)*	*Tepe Hissar (Eastern Iran)*	*Bactria Margiana Archaeological Complex - BMAC*	*Swat Protohistoric Grave Complex - SPGT (Pakistan)*
Steppe_EMBA	-	-	-	-	0.05	0.10	0.48
Srubnaya	-	-	-	-	-	-	0.44
Sintashta	-	-	-	-	-	-	0.46
Andronovo_NW	-	-	-	-	-	-	0.37
Andronovo_SE	-	-	-	-	-	-	0.58

C (Distal model of the Cline)



D (Proximal model of the Cline)



E (admixture graph relating data from diverse groups)

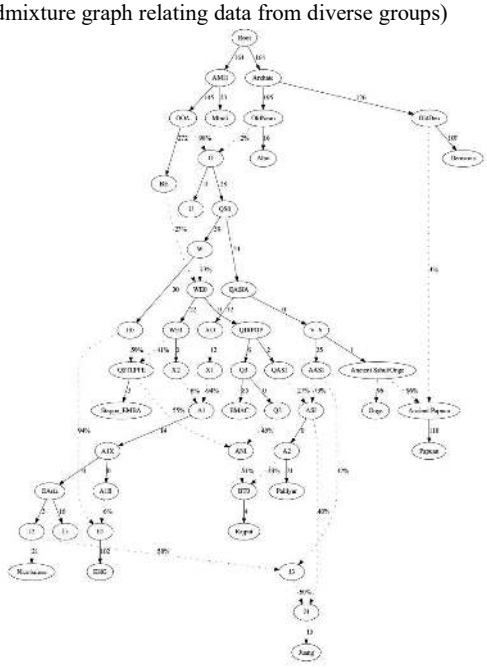


Figure 4. A Working Model Relating Iran, the Steppe, Central Asia, and South Asia in the Bronze Age

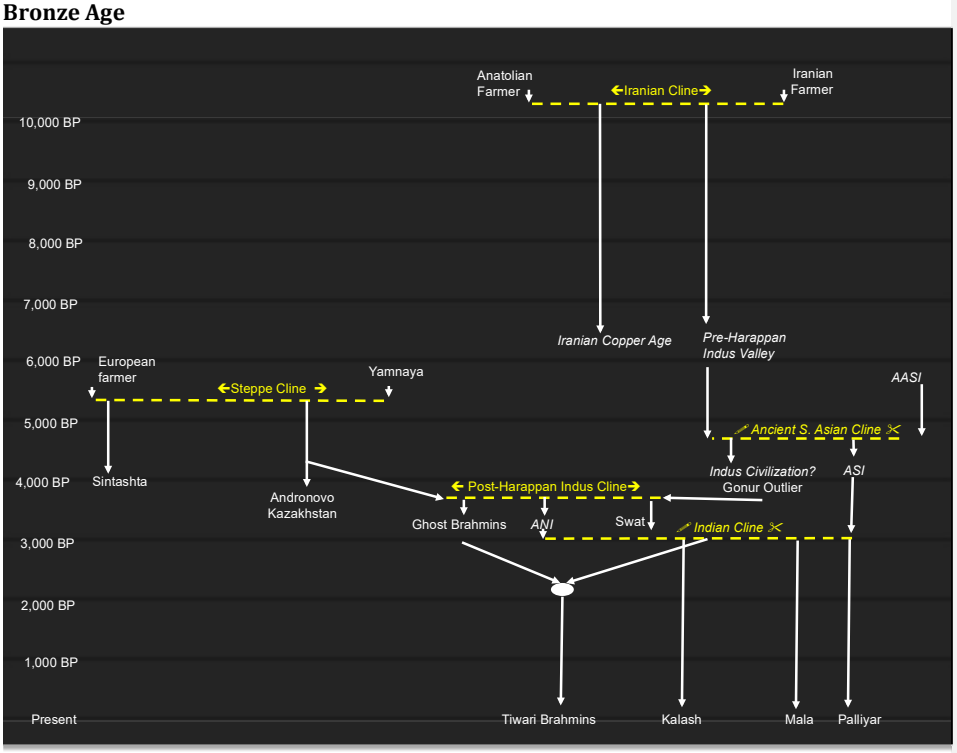
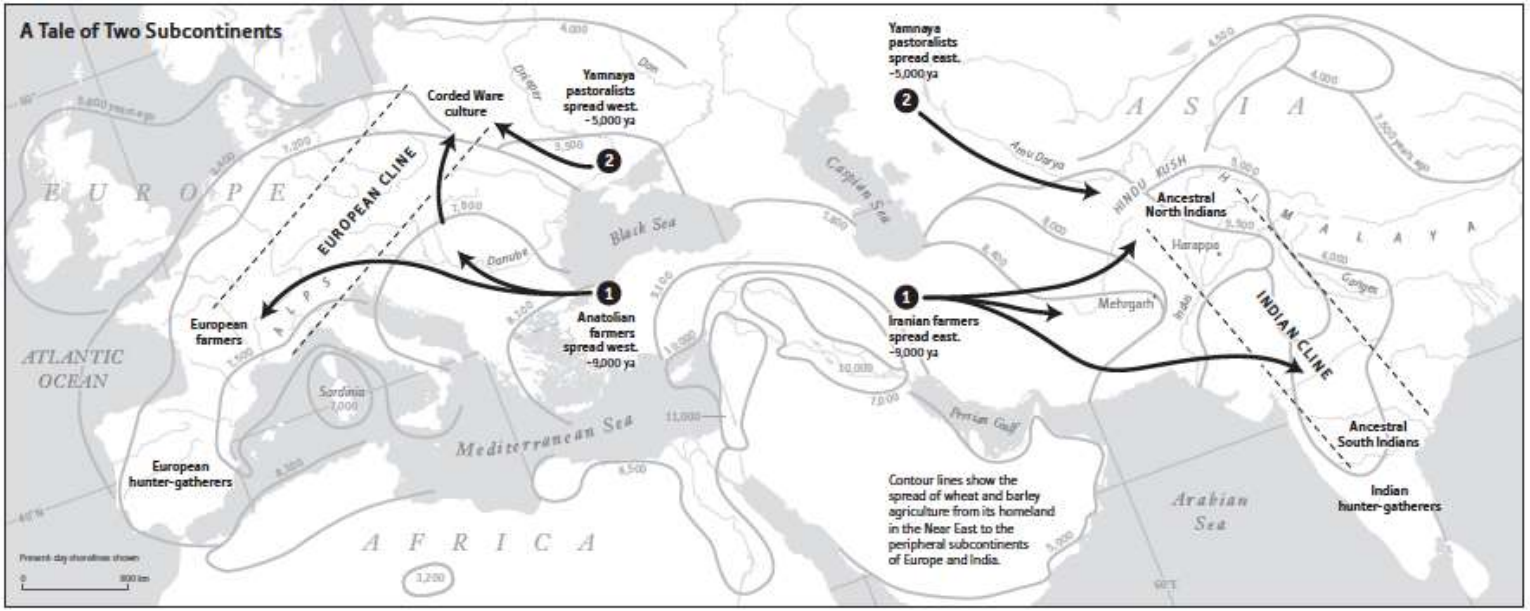


Figure 5: A Tale of Two Subcontinents. The prehistory of South Asia and Europe are parallel in both being impacted by two successive migrations, the first from the Near East after 7000 BCE bringing farmers who mixed with local hunter-gatherers, and the second from the Steppe after 3000 BCE bringing people who spoke Indo-European languages who mixed with those they encountered during their migratory movement. Mixtures of these mixed populations then produced the clines of ancestry present in both South Asia and in Europe today, which are (imperfectly) correlated to geography. The plot shows in contour lines the time of the expansion of Near Eastern agriculture. Human movements and mixtures, which also plausibly contributed to the spread of languages, are shown with arrows. [revised drawing agriculture contours with color and changing the Yamnaya eastward spread to go through the steppe rather than the desert]



Online Methods

Principal Components Analysis (PCA): We carried out principal components analysis using the *smartpca* package of *EIGENSOFT* 7.2.1.(*7*) We used default parameters and added two options, (lsqproject: YES and numoutlieriter: 0 options) in order to project our ancient samples to the PCA space here. We used two basis sets for the projection. The first based on 991 present day West Eurasians and the second based on x present day East Eurasians. For each population analyzed we show the results of our analysis for both the West and East Eurasian PCA space. As part of this analysis, we also computed the Fst between populations using the parameters inbreed: YES, and fstonly: YES.

ADMIXTURE clustering analysis: Using PLINK2 (Chang et al. 2015), we first pruned our dataset using the --geno 0.7 option to ensure that we only performed our analysis on sites which had at least 70% of samples with a called genotype. On this data, we ran ADMIXTURE(*8*) with 10 replicates and only report the version of the run with the highest likelihood. We show results for K=6 for the set of ancient samples we reported. This provides the most resolution at disambiguated the sources of Neolithic ancestry present in our newly reported samples from Central Asia

***f*-statistics:** We used the *qp3pop* and *qpDstat* packages in ADMIXTOOLS(*40*) to perform f_3 and f_4 statistics. We used the inbreed:YES parameter to compute f_3 statistics with an ancient population as a target, as a test for admixture with all published and newly reported ancient genomes as sources. With the f4Mode:YES parameter in *qpDstat,* we computed two sets of f_4 statistics. The first is what we call a test cross-comparison statistic, where we compare each newly reported test population against the other with respect to a set of ancient populations that we show encapsulates various streams of ancestry present in our reported data (where Test is one of Iran_Ganj_Dareh_Neolithic, Karelia_HG, Boisman_MN, Onge, LBK_EN, AfontovaGora3, Ukraine_Mesolithic) and an outgroup (Mbuti). This take the form f_4(Reported 1, Reported 2, Test, Outgroup). The second is a comparison of these very same neolithic populations against each other, with respect to one of our reported populations and takes the form of f_4(Test 1, Test 2, Reported, Outgroup), where the Test populations remain the same as that used in the test cross-comparison statistic. We call this the reported cross-comparison statistic.

Formally modelling admixture history: We used the *qpAdm* methodology(*10*) in the ADMIXTOOLS package(*40*) to estimate the proportions of ancestry in a *Test* population deriving

Commented [DR2]: Add sections here on laboratory methods

from a mixture of *N* 'reference' populations by exploiting (but not explicitly modeling) shared genetic drift with a set of 'Outgroup' populations. We set the details: YES parameter, which reports a normally distributed Z-score estimated with a block jackknife.

Heirarchical model of the Indian cline: We estimated ancestry proportions using *qpAdm* as described above to obtain mixture proportions for the proportion of Steppe-related Iranian-related and AASI-related ancestries and their relevant covariance matrices for each population on the Indian cline. We then jointly modelled these estimates using a bi-variate normal model (since the sum of the 3 proportions sum to 1) and estimate the mean and covariance of these proportions across all samples on the Indian cline using maximum likelihood estimation. Then, using this estimated matrix, we tested that the cline could be modelled by a mixture of two populations, the ANI and the ASI in two ways. First we examined that the covariance matrix estimated is singular, implying that knowledge of one estimated proportion of ancestry of one of the ancestry components revealed knowledge of the other two, as expected in a two way mix. Second, after showing that the first was true, we examined the difference between the expected and observed ratios of the ancestry proportions of individual populations on this generative model that was obtained from fitting all the populations simultaneously. This process resulted in several populations deviating from expectation and we discuss their relevance in the main text and Supplementary Text.

Abbreviations

We have used the following abbreviations in population labels: N, Neolithic; EN, Early Neolithic; C, Copper Age; EMBA, Early to middle Bronze Age; MLBA, Middle to late Bronze Age; IA, Iron Age.

References

1. J. Dabney *et al.*, Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 15758-15763 (2013).
2. N. Rohland, E. Harney, S. Mallick, S. Nordenfelt, D. Reich, Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **370**, 20130624 (2015).
3. Q. Fu *et al.*, DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 2223-2227 (2013).
4. Q. Fu *et al.*, An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, (2015).
5. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation Sequencing Data. *BMC bioinformatics* **15**, 356 (2014).
6. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS genetics* **2**, e190 (2006).
7. K. J. Galinsky *et al.*, Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *American journal of human genetics* **98**, 456-472 (2016).
8. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655-1664 (2009).
9. N. J. Patterson *et al.*, Ancient Admixture in Human History. *Genetics* **192**, 1065-1093 (2012).
10. W. Haak *et al.*, Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207-211 (2015).
11. I. Lazaridis *et al.*, Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419-424 (2016).
12. F. Broushaki *et al.*, Early Neolithic genomes from the eastern Fertile Crescent. *Science*, (2016).
13. D. W. Anthony, *The horse, the wheel, and language : how bronze-age riders from the Eurasian steppes shaped the modern world*. (Princeton University Press, Princeton, NJ, 2007), pp. xii, 553 p.
14. B. Cerasetti, *WALKING IN THE MURGHAB ALLUVIAL FAN (SOUTHERN TURKMENISTAN): AN INTEGRATED APPROACH BETWEEN OLD AND NEW PROVIDES NEW INTERPRETATIONS ABOUT THE INTERACTION BETWEEN SETTLED AND NOMADIC PEOPLE*. (2014), vol. BAR International Series 2690, pp. 105-114.
15. G. L. Possehl, *The indus civilization : a contemporary perspective*. (AltaMira Press, Walnut Creek, CA, 2002), pp. xi, 276 p.
16. N. A. Dubova, A. B. Saipov, S. M. Junusbayev, Interaction between Steppe and Agricultural Tribes during the Bronze Age: Mophological Aspects. *International Journal of Anthropology* **31**, 109-125 (2016).
17. C. J. Stevens *et al.*, Between China and South Asia: A Middle Asian corridor of crop dispersal and agricultural innovation in the Bronze Age. *The Holocene* **26**, 1541-1555 (2016).
18. M. Raghavan *et al.*, Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87-91 (2014).
19. M. E. Allentoft *et al.*, Population genomics of Bronze Age Eurasia. *Nature* **522**, 167-+ (2015).
20. I. Mathieson *et al.*, Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499-503 (2015).

21. M. D. Frachetti, C. E. Smith, C. M. Traub, T. Williams, Nomadic ecology shaped the highland geography of Asia's Silk Roads. *Nature* **543**, 193-198 (2017).
22. M. Unterlander *et al.*, Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. *Nature communications* **8**, 14615 (2017).
23. P. M. Nathan Joel Nakatsuka, Niraj Rai, Biswanath Sarkar, Arti Tandon, Nick Patterson, Gandham SriLakshmi Bhavani, Katta Mohan Girisha, Mohammed S Mustak, Sudha Srinivasan, Amit Kaushik, Saadi Abdul Vahab, Sujatha M Jagadeesh, Kapaettu Satyamoorthy, Lalji Singh, David Reich, Kumarasamy Thangaraj, The promise of disease gene discovery in South Asia. *biorxiv.org* https://doi.org/10.1101/047035.
24. D. Reich, K. Thangaraj, N. Patterson, A. L. Price, L. Singh, Reconstructing Indian population history. *Nature* **461**, 489-494 (2009).
25. P. Moorjani *et al.*, Genetic evidence for recent population mixture in India. *American journal of human genetics* **93**, 422-438 (2013).
26. A. Basu, N. Sarkar-Roy, P. P. Majumder, Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 1594-1599 (2016).
27. S. Mallick *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201-206 (2016).
28. S. Sankararaman, S. Mallick, N. Patterson, D. Reich, The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current biology : CB* **26**, 1241-1247 (2016).
29. Q. Fu *et al.*, Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449 (2014).
30. P. Moorjani *et al.*, A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 5652-5657 (2016).
31. P. S. Bellwood, *First Farmers: The origins of agricultural societies*. (Blackwell, Malden, MA, 2005), pp. xix, 360 p.
32. P. A. Underhill *et al.*, The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. *European journal of human genetics : EJHG* **23**, 124-131 (2015).
33. M. Silva *et al.*, A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals. *BMC evolutionary biology* **17**, 88 (2017).
34. D. Q. Fuller, in *Examining the Farming/Language Dispersal Hypothesis*. (McDonald Institute for Archaeological Research, 2003), pp. 191-213.
35. D. Q. Fuller, in *The evolution and history of human populations in South Asia,* M. D. Petraglia, B. Allchin, Eds. (Springer, Dordrecht, The Netherlands, 2007), pp. 393-443.
36. I. Mathieson *et al.*, The Genomic History of Southeastern Europe. *biorXiv.org*, doi.org/10.1101/135616 (2017).
37. M. Lipson *et al.*, Parallel ancient genomic transects reveal complex population history of early European farmers. *bioRxiv*, doi.org/10.1101/114488 (2017).
38. W. Haak *et al.*, Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, (2015).
39. M. E. Allentoft *et al.*, Population genomics of Bronze Age Eurasia. *Nature* **522**, 167-172 (2015).
40. N. Patterson *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).