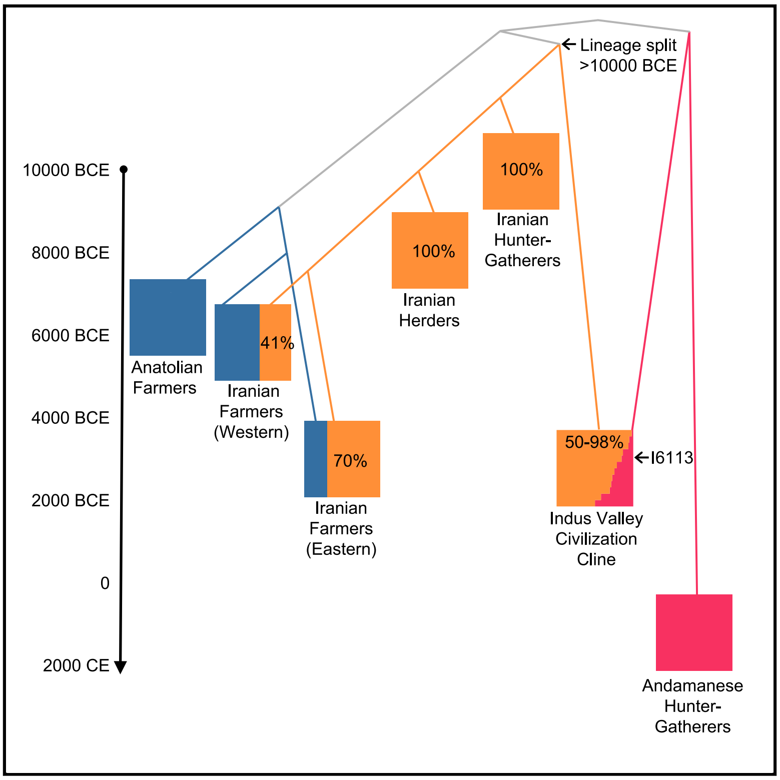
Cell

An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers

Graphical Abstract



Authors

Vasant Shinde, Vagheesh M. Narasimhan, Nadin Rohland, ..., Nick Patterson, Niraj Rai, David Reich

Correspondence

vasant.shinde@dcpune.ac.in (V.S.),
vagheesh@mail.harvard.edu (V.M.N.),
nirajrai@bsip.res.in (N.R.),
reich@genetics.med.harvard.edu (D.R.)

In Brief

A genome from the Indus Valley Civilization is from a population that is the largest source for South Asians. The population has no detectable ancestry from Steppe pastoralists or from Anatolian and Iranian farmers, suggesting farming in South Asia arose from local foragers rather than from large-scale migration from the West.

Highlights

- The individual was from a population that is the largest source of ancestry for South Asians
- Iranian-related ancestry in South Asia split from Iranian plateau lineages >12,000 years ago
- First farmers of the Fertile Crescent contributed little to no ancestry to later South Asians

Shinde et al., 2019, Cell *179*, 1–7
October 17, 2019 © 2019 Elsevier Inc.
https://doi.org/10.1016/j.cell.2019.08.048

Please cite this article in press as: Shinde et al., An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers, Cell (2019), https://doi.org/10.1016/j.cell.2019.08.048

An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers

Vasant Shinde,[1,14,15,*] Vagheesh M. Narasimhan,[2,14,*] Nadin Rohland,[2] Swapan Mallick,[2,3,4] Matthew Mah,[2,3,4] Mark Lipson,[2] Nathan Nakatsuka,[2] Nicole Adamski,[2,3] Nasreen Broomandkhoshbacht,[2,3,10] Matthew Ferry,[2,3] Ann Marie Lawson,[2,3] Megan Michel,[2,3,11,12] Jonas Oppenheimer,[2,3,13] Kristin Stewardson,[2,3] Nilesh Jadhav,[1] Yong Jun Kim,[1] Malavika Chatterjee,[1] Avradeep Munshi,[1] Amrithavalli Panyam,[1] Pranjali Waghmare,[1] Yogesh Yadav,[1] Himani Patel,[5] Amit Kaushik,[6] Kumarasamy Thangaraj,[7] Matthias Meyer,[8] Nick Patterson,[4,9] Niraj Rai,[5,7,15,*] and David Reich[2,3,4,15,16,*]

[1]Department of Archaeology, Deccan College Post-Graduate and Research Institute, Pune 411006, India
[2]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[3]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA
[4]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[5]Birbal Sahni Institute of Palaeosciences, Lucknow 226007, India
[6]Amity Institute of Biotechnology, Amity University, Noida 201313, India
[7]CSIR-Centre for Cellular and Molecular Biology, Hyderabad 500 007, India
[8]Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany
[9]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[10]Present address: Department of Anthropology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[11]Present address: Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[12]Present address: Max Planck-Harvard Research Center for the Archaeoscience of the Ancient Mediterranean, Cambridge, MA 02138, USA
[13]Present address: Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[14]These authors contributed equally
[15]Senior author
[16]Lead Contact
*Correspondence: vasant.shinde@dcpune.ac.in (V.S.), vagheesh@mail.harvard.edu (V.M.N.), nirajrai@bsip.res.in (N.R.), reich@genetics.med.harvard.edu (D.R.)

https://doi.org/10.1016/j.cell.2019.08.048

SUMMARY

We report an ancient genome from the Indus Valley Civilization (IVC). The individual we sequenced fits as a mixture of people related to ancient Iranians (the largest component) and Southeast Asian hunter-gatherers, a unique profile that matches ancient DNA from 11 genetic outliers from sites in Iran and Turkmenistan in cultural communication with the IVC. These individuals had little if any Steppe pastoralist-derived ancestry, showing that it was not ubiquitous in northwest South Asia during the IVC as it is today. The Iranian-related ancestry in the IVC derives from a lineage leading to early Iranian farmers, herders, and hunter-gatherers before their ancestors separated, contradicting the hypothesis that the shared ancestry between early Iranians and South Asians reflects a large-scale spread of western Iranian farmers east. Instead, sampled ancient genomes from the Iranian plateau and IVC descend from different groups of hunter-gatherers who began farming without being connected by substantial movement of people.

INTRODUCTION

The mature Indus Valley Civilization (IVC), also known as the Harappan Civilization, was spread over northwestern South Asia from 2600 to 1900 BCE and was one of the first large-scale urban societies of the ancient world, characterized by systematic town planning, elaborate drainage systems, granaries, and standardization of weights and measures. The inhabitants of the IVC were cosmopolitan, with multiple cultural groups living together in large regional urban centers like Harappa (Punjab), Mohenjodaro (Sindh), Rakhigarhi (Haryana), Dholavira (Kutch/Gujarat), and Ganweriwala (Cholistan) (Figure 1A) (Mughal, 1990; Possehl, 1982, 1990; Shaffer and Lichtenstein, 1989). Rakhigarhi is one of the largest known IVC sites (Figures 1B and 1C), and seven dates from charcoal at depths of 9–23 m have point estimates of 2800–2300 BCE, which largely fall within the mature phase of the IVC (Shinde et al., 2018; Vahia et al., 2016). As part of the archaeological effort, we attempted to generate ancient DNA data for a subset of the excavated burials.

RESULTS

In dedicated clean rooms, we obtained powder from 61 skeletal samples from the Rakhigarhi cemetery, which lies ∼1 km west of the ancient town (Table S1). We extracted DNA (Dabney et al., 2013; Korlević et al., 2015) and converted the extracts into libraries (Rohland et al., 2015), some of which we treated with uracil-DNA glycosylase (UDG) to greatly reduce the error rates associated with the characteristic cystosine-to-uracil lesions of ancient DNA. We enriched all libraries for sequences overlapping both the mitochondrial genome and ∼3,000 targeted nuclear positions (Olalde et al., 2018) and sequenced the enriched libraries either

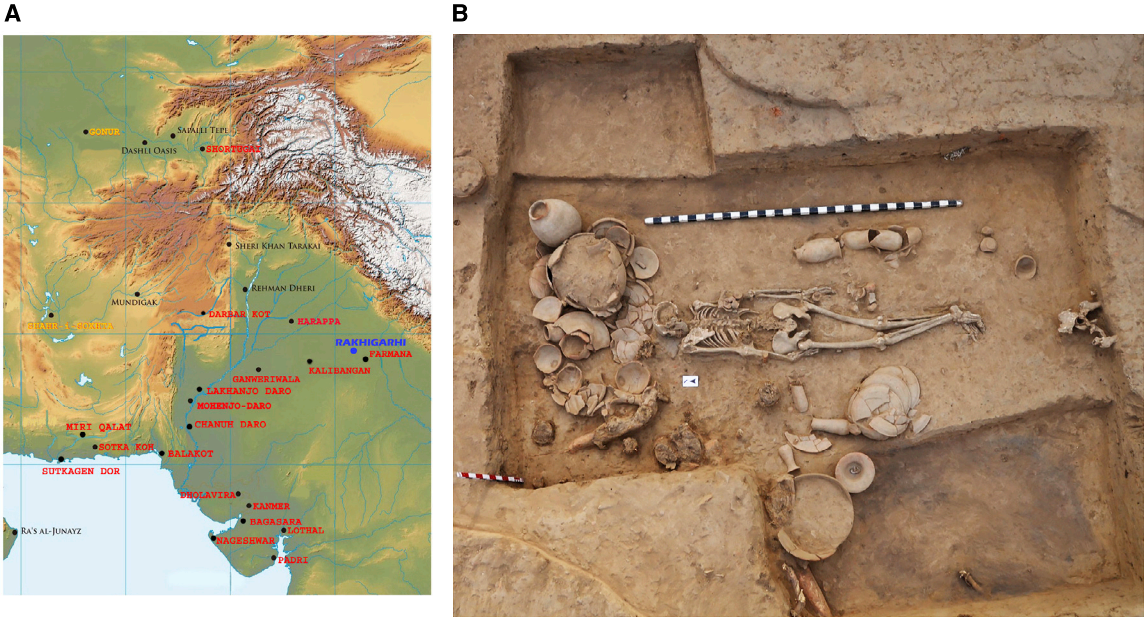


Figure 1. Archeological Context of the Individual Who Yielded Ancient DNA

(A) We label the geographic location of the archaeological site of Rakhigarhi (blue) and other significant Harappan sites (red) to define the geographic range of the IVC. We label in black sites in the north and west with which IVC people were in cultural contact and specifically highlight in yellow the sites of Gonur and Shahr-i-Sokhta, which are the source of the 11 outlier individuals who genetically form a cline of which the Rakhigarhi individual is a part.

(B) Photograph of the I6113 burial (skeletal code RGR7.3, BR-01, HS-02) and associated typical IVC grave goods illustrating typical North-South orientation of IVC burials. High-resolution images of IVC-style ceramics associated with the grave are shown in Figure S1.

on an Illumina NextSeq500 instrument using paired 2 × 76 base pair (bp) reads or on Illumina HiSeq X10 instruments using paired 2 × 150 bp reads. After trimming adapters and merging sequences overlapping by at least 15 bp (allowing up to one mismatch), we mapped to both the mitochondrial genome *rsrs* (Behar et al., 2012) and the human genome reference *hg19* (Li and Durbin, 2010) (Table S1). After inspecting the screening results, we enriched a subset of libraries for ~1.2 million SNPs and sequenced the enriched libraries and processed the data as described above mapping to *hg19* only (Fu et al., 2015; Haak et al., 2015; Mathieson et al., 2015). For the most promising sample, which had the genetic identification code I6113 and the archaeological skeletal code RGR7.3, BR-01, HS-02 (Figures 1B and S1), we created, enriched, and sequenced a total of 109 double- and single-stranded libraries from five extractions (Meyer et al., 2012; Glocke and Meyer, 2017; Rohland et al., 2018; Gansauge et al., 2017) (only the initial library was UDG treated). After removing 41 libraries (from one extraction) that had significantly lower coverage, and merging data from the remaining 68 libraries, we had 86,440 SNPs covered at least once. Almost all of these 68 libraries showed cytosine-to-thymine mismatch rates to the human reference genome in the final 5′ and 3′ nucleotides greater than 10%, consistent with the presence of authentic ancient DNA ("ancient DNA damage"). However, when we stratified the pooled data by sequence length, we found lower damage rates particularly for sequences of length >50 bp (Figure S2; STAR Methods). This was suggestive of the presence of contamination, and to increase confidence that our analyses were not biased by contamination, we restricted the data to molecules that showed cytosine-to-thymine mismatches characteristic of ancient DNA. This resulted in data at 31,760 SNPs. The ratio of damage-restricted sequences mapping to the Y chromosome to sequences mapping to both the Y and X chromosomes was in the range expected for a female. After building a mitochondrial DNA consensus using damage-restricted sequences, we determined that its haplogroup was U2b2, which is absent in whole mitochondrial genomes sequences available from about 400 ancient Central Asians; today, this specific haplogroup is nearly exclusive to South Asia (Narasimhan et al., 2019).

In principal-component analysis (PCA) (Figure 2A), I6113 projects onto a previously defined genetic gradient represented in 11 individuals from two sites in Central Asia in cultural contact with the IVC (3 from Gonur in present-day Turkmenistan and 8 from Shahr-i-Sokhta in far eastern Iran); these individuals were previously identified via a formal statistical procedure as significant outliers relative to the majority of samples at these two sites (they represent only 25% of the total) and were called the *Indus Periphery Cline* (Narasimhan et al., 2019). Despite having only modest SNP coverage, the error bars for the positioning of I6113 in the PCA are sufficiently small to show that this individual is not only significantly different in ancestry from the primary ancient populations of Bronze Age Gonur and Shahr-i-Sokhta but also does not fall within the variation of present-day South Asians. We obtained qualitatively consistent results when analyzing the data using ADMIXTURE (Alexander et al., 2009), with I6113 again similar to the 11 outlier individuals in harboring a mixture of ancestry related to ancient Iranians and tribal southern Indians. None of these individuals had evidence of "Anatolian

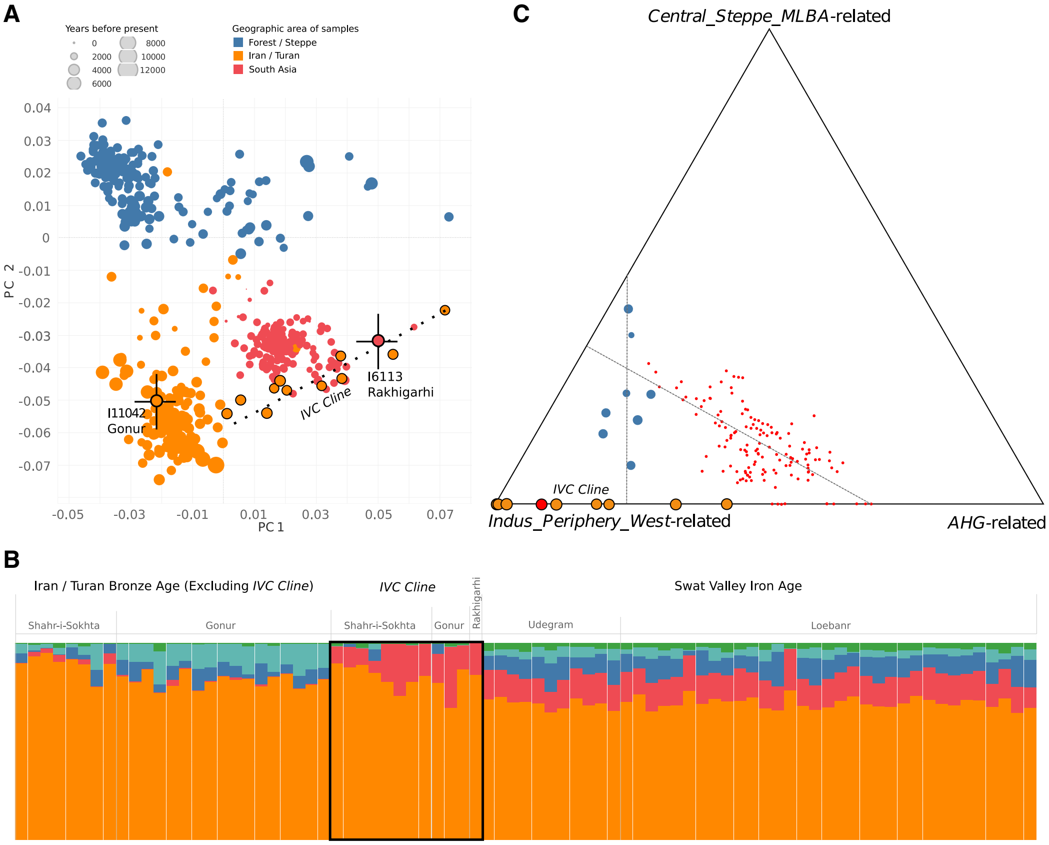


Figure 2. Population Genetic Analysis

(A) PCA of ancient DNA from South and Central Asia projected onto a basis of whole-genome sequencing data from present-day Eurasians. I6113 and I11042 (a non-outlier individual from the site of Gonur of similar data quality) are shown along with black error bars indicating 1 SE as estimated using a chromosomal block jackknife. I6113's position in the PCA is inconsistent with that of present-day South Asians and with Iranian groups but is consistent with a set of 11 outliers who represent 25% of analyzed individuals at the sites of Gonur and Shahr-i-Sokhta and who with I6113 form the *IVC Cline* (the points representing all 12 individuals are outlined in black).

(B) ADMIXTURE analysis of individuals from South and Central Asia shown with components in orange, teal, blue, green, and red maximized in Iranian farmers, Anatolian farmers, Eastern European hunter-gatherers, Western European hunter-gatherers, and Andamanese hunter-gatherers, respectively.

(C) Estimated proportions of three ancestral lineages in ancient and present-day individuals. The three components are maximized in Middle-to-Late Bronze Age Steppe pastoralists (*Central_Steppe_MLBA*), the reconstructed hunter-gatherer population of South Asia (represented by *Andamanese Hunter-Gatherers* [*AHG*] who we use as a proxy that we hypothesize is descended deeply in time from the same ancestral population), and *Indus_Periphery_West* (an individual on the *IVC Cline* who has one of the lowest proportions of *AHG*-related ancestry along with the highest-quality data). Individuals that fit a two-way model of mixture between these three sources are shown on the triangle edges, whereas individuals that could only be fit with a three-way model are shown in the interior. I6113 is shown on the *IVC Cline* as a red dot and the other *IVC Cline* individuals are shown as orange dots (all outlined in black), later individuals who formed as mixtures between people on the *IVC Cline* and people with Steppe ancestry are shown as green dots, and diverse modern South Asian groups who formed as a mixture of two later mixed groups are shown as blue dots

farmer-related" ancestry, a term we use to refer to the lineage found in ancient genomes from 7th millennium BCE farmers from Anatolia (Mathieson et al., 2015). This Anatolian farmer-related ancestry was absent in all sampled ancient genomes from Iranian herders or hunter-gatherers dating from the 12th through the 8th millennia BCE, who instead carried a very different ancestry profile also present in mixed form in South Asia that we call "Iranian related" (Broushaki et al., 2016; Lazaridis et al., 2016).

We used *qpAdm* to test highly divergent populations that have been shown to be effective for modeling diverse West and South Eurasian groups as potential sources for I6113 (Narasimhan et al., 2019). If one of these population fits, it does not mean it is the true source; instead, it means that it and the true source population are consistent with descending without mixture from the same homogeneous ancestral population that potentially lived thousands of years before. The only fitting two-way models were mixtures of a group related to herders from the western Zagros mountains of Iran and also to either Andamanese hunter-gatherers (73% ± 6% Iranian-related ancestry; p = 0.103 for overall model fit) or East Siberian hunter-gatherers (63% ± 6% Iranian-related ancestry; p = 0.24) (the fact that the latter two populations both fit reflects that they have the same phylogenetic relationship to the non-West Eurasian-related component of I6113 likely due to shared ancestry deeply in time). This is the same class of models previously shown to fit the 11 outliers that form the *Indus Periphery Cline* (Narasimhan et al., 2019), and indeed, I6113 fits as a genetic clade with the pool of *Indus Periphery Cline* individuals in *qpAdm* (p = 0.42). Multiple lines of evidence suggest that the genetic similarity of I6113 to the *Indus Periphery Cline* individuals is due to gene flow from South Asia rather than in the reverse direction. First, of the 44 individuals with good-quality data we have from Gonur and Shahr-i-Sokhta, only 11 (25%) have this ancestry profile; it would be surprising to see this ancestry profile in the one individual we analyzed from Rakhigarhi if it was a migrant from regions where this ancestry profile was rare. Second, of the three individuals at Shahr-i-Sokhta who have material culture linkages to Baluchistan in South Asia, all are *IVC Cline* outliers, specifically pointing to movement out of South Asia (Narasimhan et al., 2019). Third, both the *IVC Cline* individuals and the Rakhigarhi individual have admixture from people related to present-day South Asians (ancestry deeply related to Andamanese hunter-gatherers) that is absent in the non-outlier Shahr-i-Sokhta samples and is also absent in Copper Age Turkmenistan and Uzbekistan (Narasimhan et al., 2019), implying gene flow from South Asia into Shahr-i-Sokhta and Gonur, whereas our modeling does not necessitate reverse gene flow. Based on these multiple lines of evidence, it is reasonable to conclude that individual I6113's ancestry profile was widespread among people of the IVC at sites like Rakhigarhi, and it supports the conjecture (Narasimhan et al., 2019) that the 11 outlier individuals in the *Indus Periphery Cline* are migrants from the IVC living in non-IVC towns. We rename the genetic gradient represented in the combined set of 12 individuals the "*IVC Cline*" and then use higher-coverage individuals from this cline in lieu of I6113 to carry out fine-scale modeling of this ancestry profile.

Modeling the individuals on the *IVC Cline* using the two-way models previously fit for diverse present-day South Asians (Narasimhan et al., 2019), we find that, as expected from the PCA, it does not fit the two-way mixture that drives variation in modern South Asians as it is significantly depleted in Steppe pastoralist-related ancestry adjusting for its proportion of Iranian-related ancestry (p = 0.018 from a two-sided Z test). Modeling the *IVC Cline* using the simpler two-way admixture model without Steppe pastoralist-derived ancestry previously shown to fit the 11 outliers (Narasimhan et al., 2019), I6113 falls on the more Iranian-related end of the gradient, revealing that Iranian-related ancestry extended to the eastern geographic extreme of the IVC and was not restricted to individuals at its Iranian and Central Asian periphery. The estimated proportion of ancestry related to tribal groups in southern India in I6113 is smaller than in present-day groups, suggesting that since the time of the IVC there has been gene flow into the part of South Asia where Rakhigarhi lies from both the northwest (bringing more Steppe ancestry) and southeast (bringing more ancestry related to tribal groups in southern India). The genetic profile that we document in this individual, with large proportions of Iranian-related ancestry but no evidence of Steppe pastoralist-related ancestry, is no longer found in modern populations of South Asia or Iran, providing further validation that the data we obtained from this individual reflects authentic ancient DNA.

To obtain insight into the origin of the Iranian-related ancestry in the *IVC Cline*, we co-modeled the highest-coverage individual from the *IVC Cline*, *Indus_Periphery_West* (who also happens to have one of the highest proportions of Iranian-related ancestry) with other ancient individuals from across the Iranian plateau representing early hunter-gatherer and food-producing groups: a ∼10,000 BCE individual from Belt Cave in the Alborsz Mountains, a pool of ∼8000 BCE early goat herders from Ganj Dareh in the Zagros Mountains, a pool of ∼6000 BCE farmers from Hajji Firuz in the Zagros Mountains, and a pool of ∼4000 BCE farmers from Tepe Hissar in Central Iran. Using *qpGraph* (Patterson et al., 2012), we tested all possible simple trees relating the Iranian-related ancestry component of these groups, accounting for known admixtures (Anatolian farmer-related admixture into Hajji Firuz and Tepe Hissar and Andamanese hunter-gatherer-related admixture in the *IVC Cline*) (Figure S3), using an acceptance criterion for the model fitting that the maximum $|Z|$ scores between observed and expected *f*-statistics was <3 or that the Akaike Information Criterion (AIC) was within 4 of the best-fit (Burnham and Anderson, 2004). The only consistently fitting models specified that the Iranian-related lineage contributing to the *IVC Cline* split from the Iranian-related lineages sampled from ancient genomes of the Iranian plateau before the latter separated from each other (Figure 3 represents one such model consistent with our data). We confirmed this result by using symmetry tests that we applied first to stimulated data (Figure S4) and then evaluated the relationships among the Iranian-related lineages, correcting for the effects of Anatolian farmer-related, Andamanese hunter-gatherer-related, and West Siberian hunter-gatherer-related admixture (STAR Methods). We find that 94% of the resulting trees supported the Iranian-related lineage in the *IVC Cline* being the first to separate from the other lineages, consistent with our modeling results.

Our evidence that the Iranian-related ancestry in the *IVC Cline* diverged from lineages leading to ancient Iranian hunter-gatherers, herders, and farmers prior to their ancestors' separation places constraints on the spread of Iranian-related ancestry across the combined region of the Iranian plateau and South Asia, where it is represented in all ancient and modern genomic data sampled to date. The Belt Cave individual dates to ∼10,000 BCE, definitively before the advent of farming anywhere in Iran, which implies that the split leading to the Iranian-related component in the *IVC Cline* predates the advent of farming there as well (Figure 3). Even if we do not consider the results from the low-coverage Belt Cave individual, our analysis shows that the Iranian-related lineage present in the *IVC Cline* individuals split before the date of the ∼8000 BCE Ganj Dareh individuals, who lived in the Zagros mountains of the Iranian plateau before crop farming began there around ∼7000–6000 BCE. Thus, the Iranian-related ancestry in the *IVC Cline* descends from a different group of hunter-gatherers from the ancestors of the earliest known farmers or herders in the western Iranian plateau. We also highlight a second line of evidence against the hypothesis that eastward migrations of descendants of western Iranian farmers or herders were the source of the Iranian-related ancestry in the *IVC Cline.* An independent study has shown that all ancient genomes from Neolithic and Copper Age crop farmers of the Iranian plateau harbored

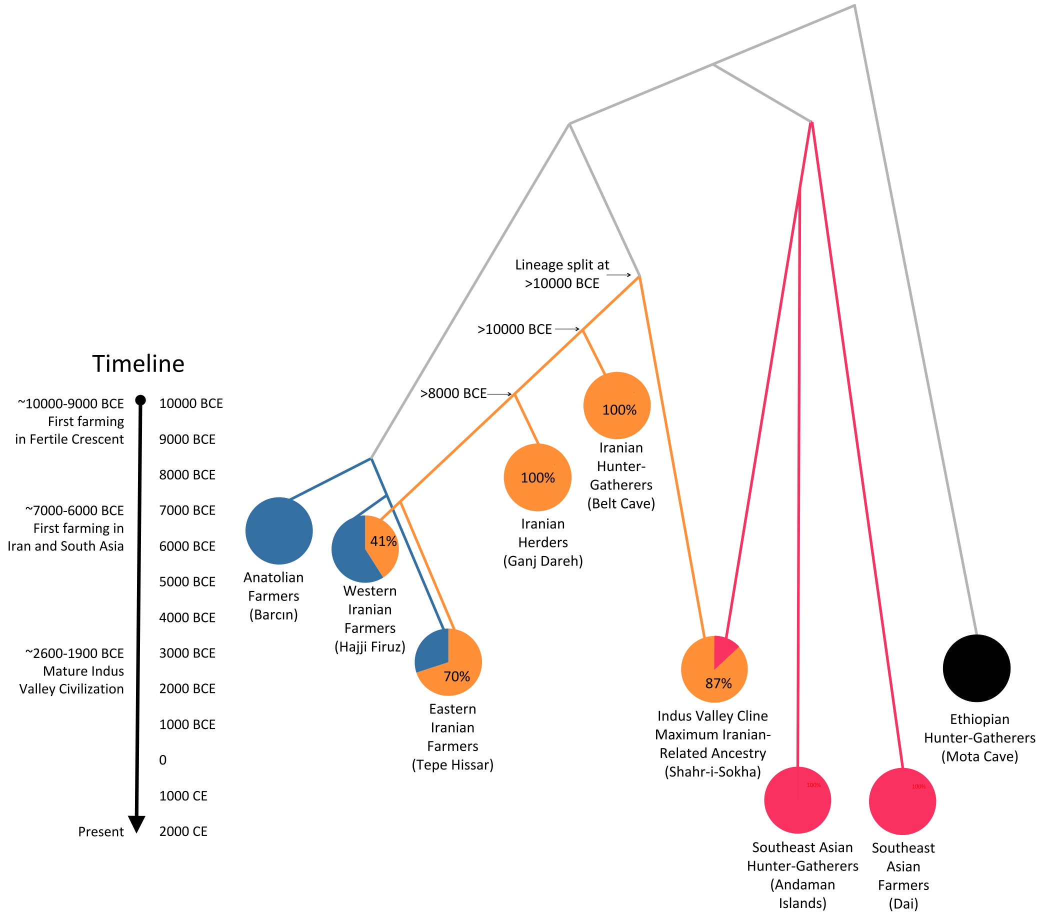


Figure 3. Best-Fitting Admixture Graph Relating Populations with Iranian-Related Ancestry

The Iranian-related subtree is shown in green, the Anatolian farmer-related subtree in blue, the southeast Asian-related subtree in red, and mixed populations as pie charts, with the proportion of Iranian-related ancestry labeled. The dates of the analyzed populations are shown on the vertical axis and provide minima on the population split dates. The observation that the Iranian-related lineage contributing to the *IVC Cline* split earlier than Belt Cave at ~10,000 BCE and Ganj Dareh at ~8000 BCE—an inference that is replicated in the other fitting Admixture Graphs—is incompatible with the hypothesis that the advent of farming in South Asia after ~7000–6000 BCE was associated with a large-scale eastward migration bringing ancestry from people related to western Zagros mountain farmers or herders across the Iranian plateau to South Asia. A schematic version of this figure that adds in the *IVC Cline* individuals not included in the Admixture Graph fitting is shown in the Graphical Abstract.

Anatolian farmer-related ancestry not present in the earlier herders of the western Zagros (Narasimhan et al., 2019). This includes western Zagros farmers (~59% Anatolian farmer-related ancestry at ~6000 BCE at Hajji Firuz) and eastern Alborsz farmers (~30% Anatolian farmer-related ancestry at ~4000 BCE at Tepe Hissar). That the 12 sampled individuals from the *IVC Cline* harbored negligible Anatolian farmer-related ancestry thus provides an independent line of evidence (in addition to their deep-splitting Iranian-related lineage that has not been found in any sampled ancient Iranian genomes to date) that they did not descend from groups with ancestry profiles characteristic of all sampled Iranian crop-farmers (Narasimhan et al., 2019). While there is a small proportion of Anatolian farmer-related ancestry in South Asians today, it is consistent with being entirely derived from Steppe pastoralists who carried it in mixed form and who spread into South Asia from ~2000–1500 BCE (Narasimhan et al., 2019).

DISCUSSION

These findings suggest that in South Asia as in Europe, the advent of farming was not mediated directly by descendants of the world's first farmers who lived in the fertile crescent. Instead, populations of hunter-gatherers—in Eastern Anatolia in the case of Europe (Feldman et al., 2019) and in a yet-unsampled location in the case of South Asia—began farming without large-scale movement of people into these regions. This does not mean that movements of people were unimportant in the introduction of farming economies at a later date; for example, ancient DNA studies have documented that the introduction of farming to Europe after ~6500 BCE was mediated by a large-scale expansion of Western Anatolian farmers who descended largely from early hunter-gatherers of Western Anatolia (Feldman et al., 2019). It is possible that in an analogous way, an early farming population expanded dramatically within South Asia, causing large-scale population turnovers that helped to spread this economy within the region. Whether this occurred is still unverified and could be determined through ancient DNA studies from just before and after the farming transitions in South Asia.

Our results also have linguistic implications. One theory for the origins of the now-widespread Indo-European languages in South Asia is the “Anatolian hypothesis,” which posits that the spread of these languages was propelled by movements of people from Anatolia across the Iranian plateau and into South Asia associated with the spread of farming. However, we have shown that the ancient South Asian farmers represented in the *IVC Cline* had negligible ancestry related to ancient Anatolian farmers as well as an Iranian-related ancestry component distinct from sampled ancient farmers and herders in Iran. Since language proxy spreads in pre-state societies are often accompanied by large-scale movements of people (Bellwood, 2013), these results argue against the model (Heggarty, 2019) of a trans-Iranian-plateau route for Indo-European language spread into South Asia. However, a natural route for Indo-European languages to have spread into South Asia is from Eastern Europe via Central Asia in the first half of the 2nd millennium BCE, a chain of transmission that did occur as has been documented in detail with ancient DNA. The fact that the Steppe pastoralist ancestry in South Asia matches that in Bronze Age Eastern Europe (but not Western Europe [de Barros Damgaard et al., 2018; Narasimhan et al., 2019]) provides additional evidence for this theory, as it elegantly explains the shared distinctive features of Balto-Slavic and Indo-Iranian languages (Ringe et al., 2002).

Our analysis of data from one individual from the IVC, in conjunction with 11 previously reported individuals from sites in cultural contact with the IVC, demonstrates the existence of an ancestry gradient that was widespread in farmers to the northwest of peninsular India at the height of the IVC, that had little if any genetic contribution from Steppe pastoralists or western Iranian farmers or herders, and that had a primary impact on the ancestry of later South Asians. While our study is sufficient to demonstrate that this ancestry profile was a common feature of the IVC, a single sample—or even the gradient of 12 likely IVC samples we have identified—cannot fully characterize a cosmopolitan ancient civilization. An important direction for future work will be to carry out ancient DNA analysis of additional individuals across the IVC range to obtain a quantitative understanding of how the ancestry of IVC people was distributed and to characterize other features of its population structure.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - ○ Contextual date for individual I6113
  - ○ Ancient DNA Data Generation
  - ○ Assessing samples for authenticity of ancient DNA
  - ○ Autosomal Contamination Estimates
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ ADMIXTURE clustering analysis
  - ○ Hierarchical modeling
  - ○ Determination of the phylogeny of Iranian-related ancestry
  - ○ Building scaffolds of all possible topologies of Iranian-related ancestry
  - ○ Methodology for model selection of admixture graphs
  - ○ Results from the model selection of tested admixture graphs
  - ○ Robustness of the model selection procedure
  - ○ Alternative approaches to determining phylogeny
- DATA AND CODE AVAILABILITY
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2019.08.048.

ACKNOWLEDGMENTS

We thank Richard Meadow for detailed comments on the archaeological context and several anonymous reviewers for critiques of the manuscript and suggestions for improvements. We thank Anna Koutoulas and the Genomics Platform at the Broad Institute for sequencing support. We thank Elena Essel and Sarah Nagel for help in the laboratory. K.T. was supported by the FTT project (IndiGen) fund (MLP1809) of the Council of Scientific and Industrial Research (CSIR), Government of India. N.N. was supported by National Institutes of Health grant GM007753. D.R. is an Investigator of the Howard Hughes Medical Institute, and his ancient DNA laboratory work was supported by National Science Foundation HOMINID grant BCS-1032255, by National Institutes of Health grant GM100233, by an Allen Discovery Center grant, and by grant 61220 from the John Templeton Foundation.

AUTHOR CONTRIBUTIONS

Conceptualization, V.S., V.M.N., N.Ro., N.P., N.Ra., and D.R.; Formal Analysis, V.M.N., N.Ro., S.M., M.Ma., M.L., N.N., M.Me., N.P., and D.R.; Investigation, V.M.N., N.Ro, S.M., N.A., N.B., M.F., A.M.L., M.Mi., J.O., K.S., M.Me., N.P., N.Ra., and D.R.; Resources, V.S., N.J., Y.J.K., M.C., A.M., A.P., P.W., Y.Y., H.P., A.K., K.T., M.Me., N.Ra., and D.R.; Data Curation, V.B., S.M., M.Ma., N.N., M.Me., and D.R.; Writing, V.M.N. and D.R.; Supervision, V.S., N.Ro., K.T., M.Me., N.P., N.Ra., and D.R.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 9, 2019
Revised: August 26, 2019
Accepted: August 27, 2019
Published: September 5, 2019

REFERENCES

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Behar, D.M., van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N.M., Kivisild, T., Torroni, A., and Villems, R. (2012). A “Copernican” reassessment of the human mitochondrial DNA tree from its root. Am. J. Hum. Genet. *90*, 675–684.

Bellwood, P.S. (2013). 11 Human migrations and the histories of major language families. In The Encyclopedia of Global Human Migration, I. Ness, ed. (Chichester, UK: Wiley-Blackwell), pp. 87–95.

Broushaki, F., Thomas, M.G., Link, V., López, S., van Dorp, L., Kirsanow, K., Hofmanová, Z., Diekmann, Y., Cassidy, L.M., Díez-Del-Molino, D., et al. (2016). Early Neolithic genomes from the eastern Fertile Crescent. Science *353*, 499–503.

Burnham, K.P., and Anderson, D.R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociol. Methods Res. *33*, 261–304.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., and Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. USA *110*, 15758–15763.

Dales, G.F. (1991). Some specialized ceramic studies at Harappa. In Harappa Excavations 1986-1990, Monographs in World Archaeology No. 3, R.H. Meadow, ed. (Prehistory Press), pp. 61–69.

de Barros Damgaard, P., Martiniano, R., Kamm, J., Moreno-Mayar, J.V., Kroonen, G., Peyrot, M., Barjamovic, G., Rasmussen, S., Zacho, C., Baimukhanov, N., et al. (2018). The first horse herders and the impact of early Bronze Age steppe expansions into Asia. Science *360*, eaar7711.

Feldman, M., Fernández-Domínguez, E., Reynolds, L., Baird, D., Pearson, J., Hershkovitz, I., May, H., Goring-Morris, N., Benz, M., Gresky, J., et al. (2019). Late Pleistocene human genome suggests a local origin for the first farmers of central Anatolia. Nat. Commun. *10*, 1218.

Fu, Q., Hajdinjak, M., Moldovan, O.T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., et al. (2015). An early modern human from Romania with a recent Neanderthal ancestor. Nature *524*, 216–219.

Gansauge, M.T., Gerber, T., Glocke, I., Korlevic, P., Lippik, L., Nagel, S., Riehl, L.M., Schmidt, A., and Meyer, M. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. Nucleic Acids Res. *45*, e79.

Glocke, I., and Meyer, M. (2017). Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. Genome Res. *27*, 1230–1237.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature *522*, 207–211.

Heggarty, P. (2019). Prehistory through language and archaeology (Routledge), pp. 616–644.

Kelleher, J., Etheridge, A.M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Comput. Biol. *12*, e1004842.

Korlević, P., Gerber, T., Gansauge, M.-T., Hajdinjak, M., Nagel, S., Aximu-Petri, A., and Meyer, M. (2015). Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. Biotechniques *59*, 87–93.

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., et al. (2016). Genomic insights into the origin of farming in the ancient Near East. Nature *536*, 419–424.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lipson, M., and Reich, D. (2017). A Working Model of the Deep Relationships of Diverse Modern Human Genetic Lineages Outside of Africa. Mol. Biol. Evol. *34*, 889–902.

Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietrusewsky, M., Pryce, T.O., Willis, A., Matsumura, H., Buckley, H., et al. (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science *361*, 92–95.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature *528*, 499–503.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226.

Meyer, M., Fu, Q., Aximu-Petri, A., Glocke, I., Nickel, B., Arsuaga, J.-L., Martínez, I., Gracia, A., de Castro, J.M.B., Carbonell, E., and Pääbo, S. (2014). A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature *505*, 403–406.

Mughal, M.R. (1990). Further evidence of the Early Harappan Culture in the Greater Indus Valley: 1971–90. South Asian Stud. *6*, 175–199.

Narasimhan, V.M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., Lazaridis, I., Nakatsuka, N., Olalde, I., Lipson, M., et al. (2019). The formation of human populations in South and Central Asia. Science *365*, eaat7487.

Olalde, I., Brace, S., Allentoft, M.E., Armit, I., Kristiansen, K., Booth, T., Rohland, N., Mallick, S., Szécsényi-Nagy, A., Mittnik, A., et al. (2018). The Beaker phenomenon and the genomic transformation of northwest Europe. Nature *555*, 190–196.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093.

Possehl, G.L. (1982). The Harappan Civilization: A contemporary perspective. Harappan Civilization (Oxford University Press), pp. 16–28.

Possehl, G.L. (1990). Revolution in the Urban Revolution: The Emergence of Indus Urbanization. Annu Rev Anthropol. *19*, 261–282.

Posth, C., Nakatsuka, N., Lazaridis, I., Skoglund, P., Mallick, S., Lamnidis, T.C., Rohland, N., Nägele, K., Adamski, N., Bertolini, E., et al. (2018). Reconstructing the Deep Population History of Central and South America. Cell *175*, 1185–1197.e22.

Ringe, D., Warnow, T., and Taylor, A. (2002). Indo-European and Computational Cladistics. Trans. Philol. Soc. *100*, 59–129.

Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20130624.

Rohland, N., Glocke, I., Aximu-Petri, A., and Meyer, M. (2018). Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. Nat. Protoc. *13*, 2447–2461.

Shaffer, J.G., and Lichtenstein, D.A. (1989). Ethnicity and change in the Indus Valley cultural tradition. In Old Problems and New Perspectives in the Archaeology of South Asia, Wisconsin Archaeological Reports, *Vol. 2*, J.M. Kenoyer, ed. (Madison: University of Wisconsin,), pp. 117–126.

Shinde, V.S., Kim, Y.J., Woo, E.J., Jadhav, N., Waghmare, P., Yadav, Y., Munshi, A., Chatterjee, M., Panyam, A., Hong, J.H., et al. (2018). Archaeological and anthropological studies on the Harappan cemetery of Rakhigarhi, India. PLoS ONE *13*, e0192299.

Skoglund, P., Northoff, B.H., Shunkov, M.V., Derevianko, A.P., Pääbo, S., Krause, J., and Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc. Natl. Acad. Sci. USA *111*, 2229–2234.

Vahia, M.N., Kumar, P., Bhogale, A., Kothari, D.C., Chopra, S., Shinde, V.S., Jadhav, N., and Shastri, R. (2016). Radiocarbon dating of charcoal samples from Rakhigarhi using AMS. Curr. Sci. *111*, 27–28.

Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.J., Kronenberg, F., Salas, A., and Schönherr, S. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. *44*, W58–W63.

Please cite this article in press as: Shinde et al., An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers, Cell (2019), https://doi.org/10.1016/j.cell.2019.08.048

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Ancient skeletal element	This study	Sample ID: I6113
Chemicals, Peptides, and Recombinant Proteins		
Pfu Turbo Cx Hotstart DNA Polymerase	Agilent Technologies	Cat#: 600412
Herculase II Fusion DNA Polymerase	Agilent Technologies	Cat#: 600679
2x HI-RPM hybridization buffer	Agilent Technologies	Cat#: 5190-0403
50% (w/v) PEG 8000	Anatrace	Cat#: OPTIMIZE-82 100 ML
0.5 M EDTA pH 8.0	BioExpress	Cat#: E177
Silica magnetic beads	G-Biosciences	Cat#: 786-915
Sera-Mag Magnetic Speed-beads Carboxylate-Modified (1 µm, 3EDAC/PA5)	GE LifeScience	Cat#: 6.51521E+13
USER enzyme	New England Biolabs	Cat#: M5505
Uracil Glycosylase Inhibitor (UGI)	New England Biolabs	Cat#: M0281
Bst DNA Polymerase2.0, large frag.	New England Biolabs	Cat#: M0537
T4 RNA Ligase Reaction Buffer	New England Biolabs	Cat#: B0216L
PE buffer concentrate	QIAGEN	Cat#: 19065
Buffer PB	QIAGEN	Cat#: 19066
Proteinase K	Sigma Aldrich	Cat#: P6556
Guanidine hydrochloride	Sigma Aldrich	Cat#: G3272
3M Sodium Acetate (pH 5.2)	Sigma Aldrich	Cat#: S7899
Water	Sigma Aldrich	Cat#: W4502
Tween-20	Sigma Aldrich	Cat#: P9416
Isopropanol	Sigma Aldrich	Cat#: 650447
Ethanol	Sigma Aldrich	Cat#: E7023
5M NaCl	Sigma Aldrich	Cat#: S5150
1M NaOH	Sigma Aldrich	Cat#: 71463
20% SDS	Sigma Aldrich	Cat#: 5030
PEG-8000	Sigma Aldrich	Cat#: 89510
1 M Tris-HCl pH 8.0	Sigma Aldrich	Cat#: AM9856
dNTP Mix	Thermo Fisher Scientific	Cat#: R1121
AccuPrime Pfx DNA Polymerase	Thermo Fisher Scientific	Cat#: 12344032
ATP	Thermo Fisher Scientific	Cat#: R0441
FastAP Thermosensitive Alkaline Phosphatase	Thermo Fisher Scientific	Cat#: EF0651
Klenow Fragment	Thermo Fisher Scientific	Cat#: EP0052
10x Buffer Tango	Thermo Fisher Scientific	Cat#: BY5
T4 Polynucleotide Kinase	Thermo Fisher Scientific	Cat#: EK0032
T4 DNA Polymerase	Thermo Fisher Scientific	Cat#: EP0062
T4 DNA Ligase	Thermo Fisher Scientific	Cat#: EL0011
T4 DNA Ligase, HC	Thermo Fisher Scientific	Cat#: EL0013
SDS, 20% Solution	Thermo Fisher Scientific	Cat#: AM9820
Maxima SYBR Green kit	Thermo Fisher Scientific	Cat#: K0251
Maxima Probe/ROX qPCR Master Mix	Thermo Fisher Scientific	Cat#: K0231
50x Denhardt's solution	Thermo Fisher Scientific	Cat#: 750018
SSC Buffer (20x)	Thermo Fisher Scientific	Cat#: AM9770
GeneAmp 10x PCR Gold Buffer	Thermo Fisher Scientific	Cat#: 4379874

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Dynabeads MyOne Streptavidin T1	Thermo Fisher Scientific	Cat#: 65602
Dynabeads MyOne Streptavidin C1	Thermo Fisher Scientific	Cat#: 65002
Salmon sperm DNA	Thermo Fisher Scientific	Cat#: 15632-011
Human Cot-I DNA	Thermo Fisher Scientific	Cat#: 15279011
DyNAmo HS SYBR Green qPCR Kit	Thermo Fisher Scientific	Cat#: F410L
Critical Commercial Assays		
High Pure Extender from Viral Nucleic Acid Large Volume Kit	Roche	Cat#: 5114403001
MinElute PCR Purification Kit	QIAGEN	Cat#: 28006
NextSeq 500/550 High Output Kit v2 (150 cycles)	Illumina	Cat#: FC-404-2002
HiSeq X Ten Reagent kit v2.5	Illumina	Cat#: FC-501-2501
Deposited Data		
Raw and analyzed data	This paper	ENA: PRJEB34154
Software and Algorithms		
Samtools	Li et al., 2009	http://samtools.sourceforge.net/
BWA	Li and Durbin, 2010	http://bio-bwa.sourceforge.net/
ADMIXTOOLS	Patterson et al., 2012	https://github.com/DReichLab/AdmixTools
R	https://www.r-project.org/	https://www.r-project.org/
SeqPrep	https://github.com/jstjohn/SeqPrep	https://github.com/jstjohn/SeqPrep
smartpca	Patterson et al., 2006	https://www.hsph.harvard.edu/alkes-price/software/
ADMIXTURE	Alexander et al., 2009	http://software.genetics.ucla.edu/admixture/download.html
PMDtools	Skoglund et al., 2014	https://github.com/pontussk/PMDtools
Haplogrep 2	Weissensteiner et al., 2016	http://haplogrep.uibk.ac.at/

LEAD CONTACT AND MATERIALS AVAILABILITY

This study did not generate new unique reagents. Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, David Reich (reich@genetics.med.harvard.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We attempted to generate genome-wide data from skeletal remains of 61 ancient individuals from the IVC site of Rakhigarhi. Only a single sample yielded enough authentic ancient DNA for analysis: I6113, Rakhigarhi, Haryana, India (n = 1). We report the archeological context dates for this burial in Method Details. The skeletal samples from Rakhigarhi were excavated by the archaeological team led by V.S. at the Deccan College Post-Graduate and Research Institute in Pune India and sampled by the ancient DNA group led by N.Ra. at the Birbal Sahni Institute of Palaeosciences in Lucknow India. Analysis using the methods implemented by the ancient DNA group led by D.R. at Harvard Medical School in Boston USA was approved by a Memorandum of Understanding between Deccan College and Harvard Medical School executed in February 2016.

METHOD DETAILS

Contextual date for individual I6113

The collagen preservation for the human bones at Rakhigarhi cemetery is so poor that Accelerator Mass Spectrometry radiocarbon dating on skeletal remains is very difficult; multiple attempts showed a carbon-to-nitrogen ratio outside the range appropriate for dating, including five attempts we made specifically on skeletal elements from I6113. However, the cemetery can be securely dated based on archaeological context. First, the only evidence of human occupation of the site is in the Harappan period and hence all the excavated remains are likely to belong to that period. Second, all the characteristic features of the Harappan burial customs and features are present in the cemetery, including a separation from the main habitation area (about 1 km), and typical Harappan artifacts including pottery (Figure S1), beads made of semi-precious stones, and bangles of copper, shell or terracotta, all of which are indistinguishable from artifacts found in the main habitation area. As discussed in the text, the main habitation area has 7 radiocarbon dates based on charcoal spanning 2800-2300 BCE, largely falling within the mature IVC (Shinde et al., 2018; Vahia et al., 2016).

Ancient DNA Data Generation

Table S1 presents details of genetic results on the 252 libraries we generated on 61 distinct samples. To represent I6113, we generated data from 109 libraries (27 double-stranded (Rohland et al., 2015) and 82 single-stranded (Meyer et al., 2014; Gansauge et al., 2017), and then filtered out 41 single-stranded libraries (all the libraries from a single extraction) that tended to have much lower coverage. For the remaining 68 libraries, we restricted the data to sequences with evidence of characteristic ancient DNA damage in the final nucleotide using PMDtools (Skoglund et al., 2014).

Assessing samples for authenticity of ancient DNA

We prioritized individual I6113, who yielded appreciable amounts of DNA as well as a relatively high rate of cytosine-to-thymine mismatches to the reference sequence in the final nucleotide of the libraries, for additional library preparation and sequencing (Table S1).

For I6113 we generated a total of 27 double-stranded libraries (Rohland et al., 2015) (1 UDG-treated and 26 not UDG-treated) using powder from both the otic capsule and semicircular canals of the petrous bone, and also generated an additional 82 single-stranded libraries (all non-UDG-treated) using powder from the semicircular canal and one of two different extraction procedures (Glocke and Meyer, 2017; Rohland et al., 2018). Out of these 109 libraries, nearly all of the 41 made from single-stranded libraries prepared using the extract made with Buffer G (Glocke and Meyer, 2017) had low coverage (< 100 targeted SNPs covered) and low damage in the final nucleotide (Table S1), consistent with the extreme sensitivity of extracts made using this buffer to inhibition especially for single-stranded libraries (Glocke and Meyer, 2017; Rohland et al., 2018). We therefore removed all libraries prepared from this extract from analysis and proceeded with the remaining 68.

The number of DNA sequences obtained from each library of I6113 was insufficient for assessment of contamination on a per-library basis. We therefore pooled data across the 68 libraries for I6113, and found that 208,111 SNPs were covered by at least one sequence. Examining the number of sequences mapping to the Y chromosome as a fraction of that mapping to both the X and Y, we found a ratio of 0.047. On data for many other ancient individuals subject to ~1.2 million SNP enrichment, we have empirically found that this ratio is less than about 0.03 for uncontaminated libraries from females, and above 0.35 for uncontaminated males. Thus, I6113 has evidence of a mixture of human DNA from both males and females, and thus contamination.

To identify subsets of the molecules that are highly likely to be authentic, we analyzed the fraction of sequences that retained typical signatures of ancient DNA damage based on a characteristic cytosine-to-thymine mismatches to a reference sequence at their ends (Meyer et al., 2014; Skoglund et al., 2014), stratified by the lengths of the molecules (Figures S2A and S2B). We carried out this analysis not only for I6113, but also for a previously published ancient DNA sample from Southeast Asia from a similar time period (I4011) comprised of a merge of data from 21 UDG-treated double-stranded libraries (Lipson et al., 2018). The libraries from I6113 have high rates of damage (up to ~50%) indicative of a high proportion of genuine ancient DNA. The rate of damage for I6113 decreases dramatically for lengths greater that 50 bp, suggesting that longer molecules are more likely to be contaminated. We further found that sequences that were damaged on one end of the ancient DNA molecules (showing cytosine-to-thymine (C-to-T) mismatches relative to the reference sequence) also had an enhanced chance of damage on the other, as expected if damage restriction enriches for authentic DNA in single-stranded (Meyer et al., 2014) (Figure S2C).

The resulting dataset contains sequences covering 31,760 SNPs at least once. Its ratio of Y chromosome sequences to X+Y chromosome sequences is 0.026, consistent with being an uncontaminated female (and the anthropological determination).

Autosomal Contamination Estimates

We estimated contamination using an algorithm based on breakdown of linkage disequilibrium (Posth et al., 2018). This software measures contamination levels by comparing the haplotype distribution of a tested sample to the haplotype distribution of an external reference panel. We used Sri Lankan Tamils sampled from the United Kingdom (STU) from the 1000 Genomes Project (Auton et al., 2015) as the reference panel. The algorithm was run in the usually conservative "uncorrected" mode to attain maximal power.

QUANTIFICATION AND STATISTICAL ANALYSIS

ADMIXTURE clustering analysis

We pruned the data using PLINK2 to retain only sites for which at least 70% of individuals had a non-missing genotype (Chang et al., 2015). We then ran ADMIXTURE (Alexander et al., 2009) with 10 replicates and report the replicate with the highest likelihood. In Figure 2, we show the results for clustering using K = 6 components.

***f*-statistics** We computed *f*-statistics using the packages in ADMIXTOOLS (Patterson et al., 2012). To test for admixture we ran f_3-statistics using the inbreed:YES parameter with an ancient population as a target. To estimate the ancestry proportion for a test population given a set of source populations and a set of outgroups, we used the *qpAdm* methodology (*18*) in ADMIXTOOLS.

Hierarchical modeling

To model a given sample as part of an established genetic cline determined by a set of other populations, we used an approach described in (Narasimhan et al., 2019), where the Supplemental Materials give the full details. We begin by obtaining ancestry proportions for a set of samples on a genetic cline, and jointly model these in a single generative model taking advantage of the fact that the proportions for the three ancestral sources must sum to 1. We estimate the mean and covariance of these sources using a bivariate normal distribution via maximum likelihood. We evaluate if the test population can be fit as deriving from the same original three sources as those we just modeled on the genetic cline using *qpAdm*, and if there is a fit, evaluate if the observed ratios of the ancestry proportions of the test population fit with the expected values from the generative model established by the cline. We compute a p value based on the empirically determined covariance matrices.

Determination of the phylogeny of Iranian-related ancestry

We wished to examine the relationship of the Iranian-related ancestry present in the *IVC Cline* to that of ancient Iranian plateau groups reported in the ancient DNA literature.

We first focused on a set of populations chosen to represent a diverse group of early hunter-gatherers and farmers from across the geographic area of present-day Iran. Our approach was to identify a set of phylogenies consistent with the data and then, based on the known dates of the samples, to make inferences on minimum split times between lineages.

The individuals or groups of individuals we examined were:

1. *Belt_Cave_M (BC)* (n = 1) – A Mesolithic individual from the Alborz mountains of Central Iran. Due to the evidence of contamination in the data from this individual, we used a damage-restricted version of this dataset resulting in 30,722 SNPs.
2. *Ganj_Dareh_N (GD)* (n = 8) – Early goat herders from the Zagros Mountains of western Iran. The highest coverage individual has data from 938,523 SNPs.
3. *Hajji_Firuz_C (HF)* (n = 5) – Late Neolithic and early Copper Age individuals from the Zagros Mountains of Western Iran. The highest coverage individual has data from 916,581 SNPs.
4. *Tepe_Hissar_C (TH)* (n = 12) – Copper Age and Early Bronze Age individuals from the Central Iranian Plateau. The highest coverage individual has data from 745,066 SNPs.
5. *Indus_Periphery_West (IP)* (n = 1) – Member of the *IVC Cline* which includes the Rakhigarhi individual I6113. We represent the Iranian-related ancestry in this cline with I8728, an individual with one of the highest proportions of Iranian-related ancestry who also happens to have the highest coverage of all individuals on this cline.

As documented in ref. (Narasimhan et al., 2019), the Hajji Firuz and Tepe Hissar pools of individuals have evidence of admixture related to Anatolian farmers while the *Indus Periphery Cline* individuals = have significant proportions of ancestry related (deeply in time) to southeast Asian hunter-gatherers.

Building scaffolds of all possible topologies of Iranian-related ancestry

We were interested in understanding the relationships among the Iranian-related lineages contributing to these 5 populations, treating the non-Iranian-related components such as the Anatolian farmer-related ancestry as nuisances that we need to model out assuming a topology in which the lineages leading to Tepe Hissar (PTA) and Hajji Firuz (PHA) formed a separate clade from Anatolian farmers (in the next sub-section we show that our results are robust to the choice of the topology relating the Anatolian farmer-related lineages).

There are 3 distinct topologies according to which these 5 populations could be related, which we call "Serial Founder" (Figure S3A), "Single Outgroup" (Figure S3B), and "Two Clades" (Figure S3C). Within these topologies there are multiple permutations for how the 5 individual populations could relate, depending on how the 5 Iranian-related populations fit into "slots" on the topology.

We used *qpWave* (Patterson et al., 2012) to evaluate all 120 possible ways for the 5 populations to be grafted onto each of the open "Slots" or positions, taking care to account for the correct admixing source for the populations that were admixed. In some topologies the assignment of populations to the slots did not alter the graph topology (when two populations were a clade with respect to the others. Therefore, there were only 30, 15 and 60 different models that were unique for the Serial Founder, Single Outgroup and Two Clades phylogenies respectively, though in our results we show all 120 possibilities for the assignment of a "Slot" to a population.

Methodology for model selection of admixture graphs

In the previous section we described the assignment of populations to "Slots" and the creation of a large number of admixture graphs.

For each admixture graph produced there are two metrics that we used to evaluate fit. The first is a list of residuals above a particular Z-score and the second is a score for the weighted error for the fitted statistics based on the graph in comparison to the empirically observed statistics, $S(G) = -1/2(g-f)'Q^{-1}(g-f)$. Here f is the vector of observed f-statistics. g is the corresponding vector of statistics fit on the graph with the specified topology, and Q is an estimated covariance matrix determined empirically (Patterson et al., 2012).

As a first screen for a working model, we begin by selecting models that have their largest residual with an absolute Z-score below 3. This is a standard approach in the literature and while this may provide a practical threshold for rejecting models, this alone is not sufficient to adjudicate between two models whose worst fitting residuals are both close to the threshold. We wanted to compare how similar two models were with respect to their statistical likelihood. If our admixture graph models were nested within one another, we could do this using a Likelihood Ratio Test as described in (Lipson and Reich, 2017). In this approach, the log-likelihood of two models, one involving admixture from a certain population and one without were compared and their difference can then be compared using a chi-square test.

However this approach cannot be applied in the present setting as the models we are testing are not nested within one another and the number of parameters of all of the models are the same. To enable us to determine the level of support one particular model has when compared to another we use Akaike Information Criterion (AIC) as a measure of model fit. Since the number of parameters between two models remains the same, the actual computed score remains the same whether we use AIC or a Bayesian Information Criterion (BIC).

We used a set of guidelines outlined in (Burnham and Anderson, 2004) for performing model selection using AIC. Specifically we computed $\delta_\mathrm{i} = AIC_i - AIC_{\min}$ where AIC_i is the AIC of the *i*-th model, and $AIC_{\min}$ is the lowest AIC obtained among all the models we tested across all topologies (SO, SF and TC). The models can then be compared using the following guidelines from (Burnham and Anderson, 2004):

1. $\delta_i \leq 2$, the *i*-th model is nearly as plausible as the best fitting model;
2. $2 < \delta_i \leq 4$, the *i*-th model is consistent with the data but considerably less probable than the best fitting model;
3. $4 < \delta_i \leq 7$, the *i*-th model is much less likely than the best fitting model;
4. $7 < \delta_i$, the *i*-th model has essentially no support.

Based on this published set of criteria we chose to accept all models for which the difference in AIC was < 4.

Among the samples we analyzed was a Mesolithic individual from Central Iran, *Belt_Cave_M*, which after restricting to damaged sequences (to address the evidence of contamination in this individual) reflects data from just 30,722 autosomal SNPs, and thus co-modeling with this individual restricts the number of SNPs available for admixture graph fitting. To address this, we repeated the admixture graph fitting removing this particular individual which improved our SNP coverage by more than 10-fold, allowing us to remove models that were plausible simply because of a lack of data and ensuring that the fit of a particular admixture graph was not due to our inability to reject it at lower coverage. As a further criterion for model selection, we restricted to the intersection between the fitting models analyzed with and without the Belt Cave individual.

Results from the model selection of tested admixture graphs

In Table S3A, we observe that all working models exclude the "Two Clades" topology, regardless of how populations are assigned to "slots." We also observe that all fitting populations have *Indus_Periphery_West* or *Hajji_Firuz* as an outgroup with respect to all other groups at the AIC < 4 threshold, with only models with *Indus_Periphery_West* if we use AIC < 2.

Robustness of the model selection procedure

We explored if modifying the topology relating the South Asian hunter-gatherer-related components (Table S3B) and that relating the Anatolian farmer-related components (Table S3C) changes our inferences and found that they did not except in one notable way. Previously our model selection criterion had not been successful at distinguishing the earliest diverging of the Iranian-related populations. i.e., either *Indus_Periphery_West* or *Hajji_Firuz_C,* but under a different topology of the Anatolian farmer-related component in *Hajji_Firuz_C* we see that models with *Indus_Periphery_West* as the earliest diverging split are strongly supported over the other working models.

We also exploring allowing all the five populations to be admixed with all source populations thereby allowing a much freer model. We find that our results showing that the *Indus_Periphery_West* being the first to split are robust to this perturbation (Table S3D).

Taken together, this analysis shows a clear branching structure that involves the *Indus_Periphery_West* ancestry as the first to split, followed by the others which are not distinguishable. There are two marginally fitting models in which *Hajji_Firuz* is the first to split (Table S3A), but even if these models are correct they do not change the inference that the Iranian-related ancestry in *Indus_Periphery_West* split from the lineages leading to those in Belt Cave, Tepe Hissar, and Ganj Dareh before they separated from each other, which is the only inference we need for our main conclusions.

Alternative approaches to determining phylogeny

The major confounder when inferring trees and examining their topology as determined by shared drift amongst different populations is admixture. In analyses described above, we dealt with this by modeling known admixtures into populations, which showed that the changing of the topology of the admixing sources does not affect the inference we obtain about the internal phylogeny of the Iranian related component of the ancestry of the test populations. As an alternative approach to exploring these issues, we obtained unbiased estimates of the allele frequencies for the Iranian-related component in each of the samples by subtracting the expectation from the admixing sources, and then performed symmetry tests to reconstruct the phylogenetic relationships.

Prior to implementing this procedure on real data, we began by confirming that if the relevant admixing source populations or populations related to those source populations were available and the proportion of their admixture known, then it was possible to recover the internal phylogeny of the populations even though there is significant admixture present in the data.

To verify this we simulated the phylogeny described in Figure S4 using the *msprime* coalescent simulator (Kelleher et al., 2016). We used standard mutation and recombination rates and sampled 1 million positions in 10 individuals from each population. We converted these to haploid genotypes by random sampling. We were interested in whether we could recover the internal phylogeny of the *pp5* node. The choice of this particular topology and set of admixing populations mirrors the structure of the admixture graph that we think may be a reasonable match to our real data.

In the first step of the process, we computed the allele frequencies per SNP for the populations for which we were interested in obtaining a phylogeny, namely 3, 4, 5 and 6. We then subtracted the relevant allele frequencies of the admixing populations which were known in this setting. For example, we subtracted the allele frequency of population 8 from population 6, weighted by the admixture proportion 50%. We then computed all statistics of the form $f_4(0,A,B,Test)$, where A, B and Test could be any of the populations 3, 4, 5 and 6.

In Table S3E, we show that for samples without admixture correction there are no simple trees that are compatible with the data (at the $|Z|>3$ level). However, after subtacting the allele frequencies in the appropriate manner, we obtain passing symmetry statistics for exactly the pairs of populations we expected based on the simulated topology (Table S3F). This suggests that if we account for the correct admixing population as well as the proportion of admixture, it is possible to recover the phylogeny of a set of populations even though they might be admixed even to levels of 50% as was the case in simulations. In the next section, we apply this procedure to the admixture graph that we constructed using the real data.

Applying this approach to real data, we estimated admixture-corrected allele frequencies fo the Iranian-related ancestry in four populations carrying it: *Ganj_Dareh_N*, *Hajji_Firuz_C*, *Tepe_Hissar_C* and *Indus_Periphery_West* (we dropped *Belt_Cave_M* as it was too low in coverage to produce meaningful results). As part of our inference we needed to infer proportions of non-Iranian-related ancestry in these populations, and to do this we utilized the *qpAdm* framework developed in (Narasimhan et al., 2019), to estimate the proportions of *AHG*- and *Anatolia_N*-related ancestry in each. This produced admixture point estimates that were in line with our admixture graph fits as well as a covariance matrix measuring uncertainty.

To account for uncertainty, we carried out this procedure sampled 1000 times from the point estimates and covariance matrix of admixture proportions and produced 1000 samples. For each of these samples we subtracted the allele frequencies in *AHG* and *Anatolia_N* (weighting by the admixture proportion), related ancestries and computed all possible triplets of f_4-statistics as we had done for the simulated data. We computed f_4-statistics using a $|Z|>3$ threshold to determine whether there continued to remain significant evidence of admixture relating the populations. Unlike with the simulated data where we knew *a priori* the exact mixing proportions and admixing sources, uncertainty about the true admixture proportion in this context reduced power. We observed 465 cases where the inferred tree was (IP,(GD,(HF,TH))), 29 cases where the inferred tree was (IP,(HF,(GD,TH))), and 506 cases where no tree fit. Thus, using this procedure we only observe two viable tree topologies, both of which involve *Indus_Periphery_West* as the population splitting first, mirroring the topology produced using the admixture graph methodology.

DATA AND CODE AVAILABILITY

All newly reported sequencing data are available from the European Nucleotide Archive under accession number ENA: PRJEB34154.

ADDITIONAL RESOURCES

The Supplemental Data includes an Excel spreadsheet listing results on all 252 libraries generated for this study, an Excel spreadsheet listing all 61 samples for which attempts were made to extract ancient DNA, and an Excel spreadsheet reporting other statistics described in the paper.

Supplemental Figures

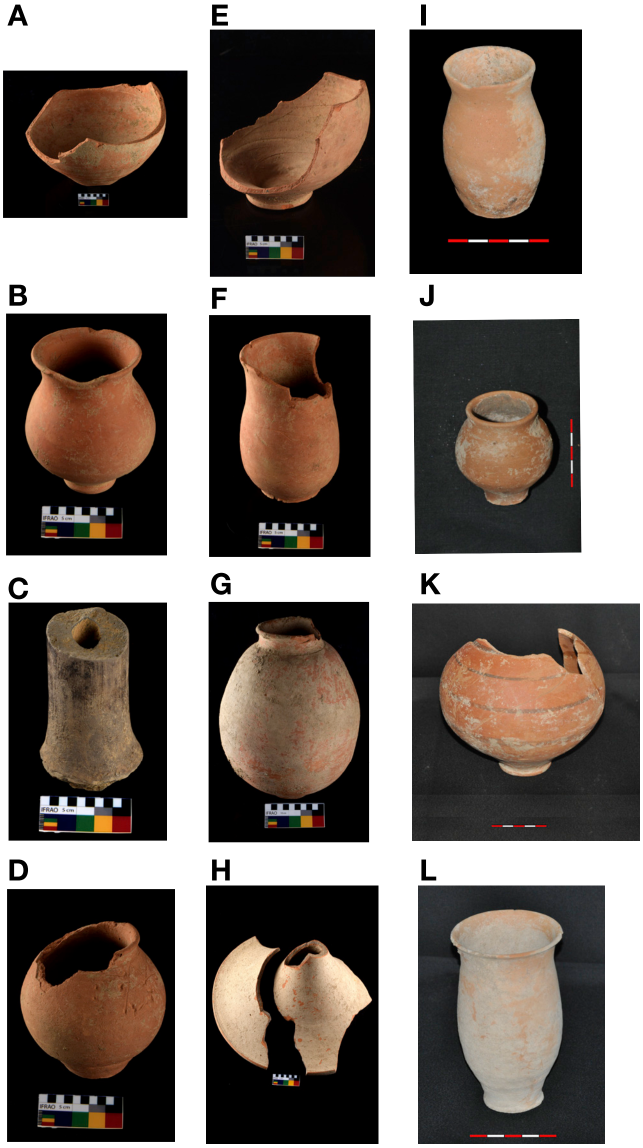


Figure S1. Ceramics in the Grave of Individual I6113 / RGR7.3, BR-01, HS-02, Related to Figure 1 and STAR Methods

All objects are from the Harappan level and were recovered from north of the head of the buried individual unless otherwise noted (see Figure 1B). In the notations that follow we specifically indicate aspects of the ceramics that are characteristically Harappan. (A) Lower portion and base of red wheel-made vessel of Harappan style. (B) Complete medium-size red globular pot of Harappan style. (C) Central column of a ceramic stand broken at both ends and burned. See vessel (H) for the base of such a stand. Dishes-, plates-, and bowls-on-stands are characteristic Harappan ceramic vessels often found intact in graves. (D) A red-slipped ware medium-size globular pot. There are lines as well as indentations on the upper right side just below the neck of the vessel. These could be examples of ancient graffiti or possibly even of the ‘‘Indus script’’. (E) Broken specimen of a medium-size red globular pot. There are a series of linear marks both near the base and on the inside that suggest that such a pot was wheel-thrown, which was common for Harappan vessels. (F) Broken specimen of a red ceramic beaker. (G) Nearly complete specimen of a large red vessel. The distinct rim suggests that it was made separately and then attached to the upper body of the vessel. The light brown clay that now only partially covers the vessel may have been intentionally applied before the vessel was put into the grave. (H) Broken stand of a dish- or plate- or bowl-on-stand. The red color can be seen peeking through the light brown clay. The photograph of the grave (Figure 1B) reveals that almost all ceramic vessels are covered with such clay. This is a phenomenon that is known from the site of Harappa itself where even elaborately decorated vessels were covered with clay before being lowered into the tomb (Dales, 1991). (I) Complete beaker placed near the head of the skeleton. (J) Complete goblet of an unusual dark red-brown color. (K) Broken red vessel with black-painted horizontal lines. (L) Complete beaker placed near the right leg of the skeleton.

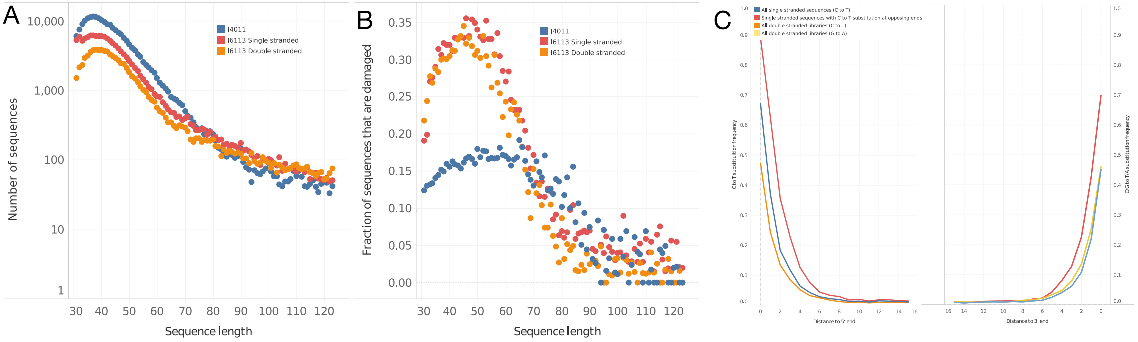


Figure S2. Quality Control to Identify a Subset of Authentic Sequences, Related to STAR Methods

(A) The distribution of the number of sequences of a given length for two different samples - I6113 from Rakhigarhi, and I4011 which is previously reported data of a similar chronological age from Myanmar. For I6113, sequence length distributions are shown separately for pools of the double-stranded (DS) and single-stranded (SS) libraries. (B) The fraction of sequences showing characteristic ancient DNA damage as a function of sequence length for I6113 and I4011. For I6113, sequence length distributions are shown separately for double and single-stranded libraries. (C) Frequency of C-to-T substitutions for both ends of the sequences as a function of distance from the end for I6113, after merging all the non-UDG-treated libraries (dropping the UDG-treated one for this analysis). The profile shows damage patterns characteristic of authentic ancient DNA in the manner expected for each library preparation protocol. Restricting to sequences with C-to-T substitutions on one end of a fragment results in an increase in the damage rate on the other end.

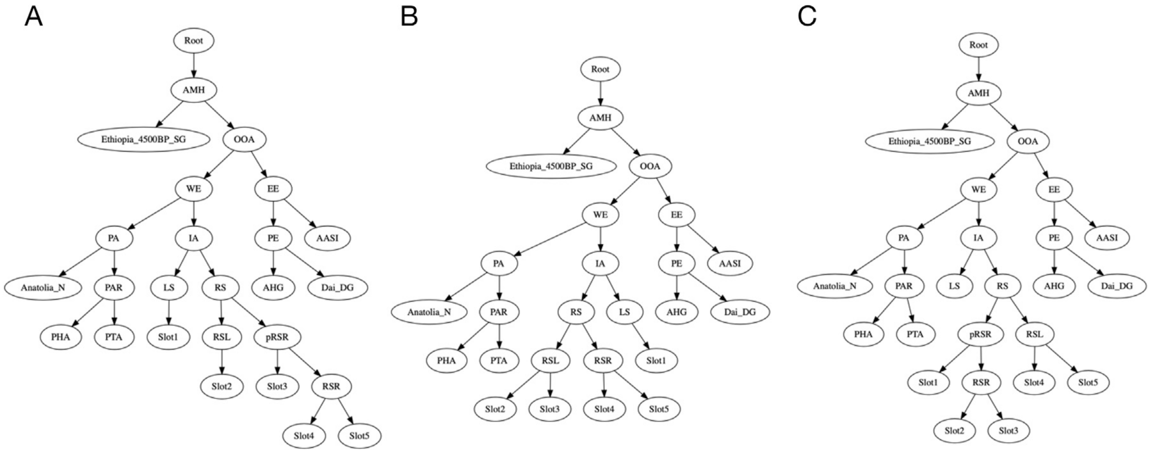


Figure S3. Phylogenies of Iranian-Related Populations Tested for Fits to the Data, Related to STAR Methods

(A) “Serial Founder” model: A first population splits, then a second, then a third, then a fourth and fifth. (B) “Single Outgroup” model: A single population splits. Thereafter two pairs of populations diverge from a common source. (C) “Two Clades” model: The 5 populations split into two groups, one with 3 populations in a clade and the second with 2 populations in a clade.

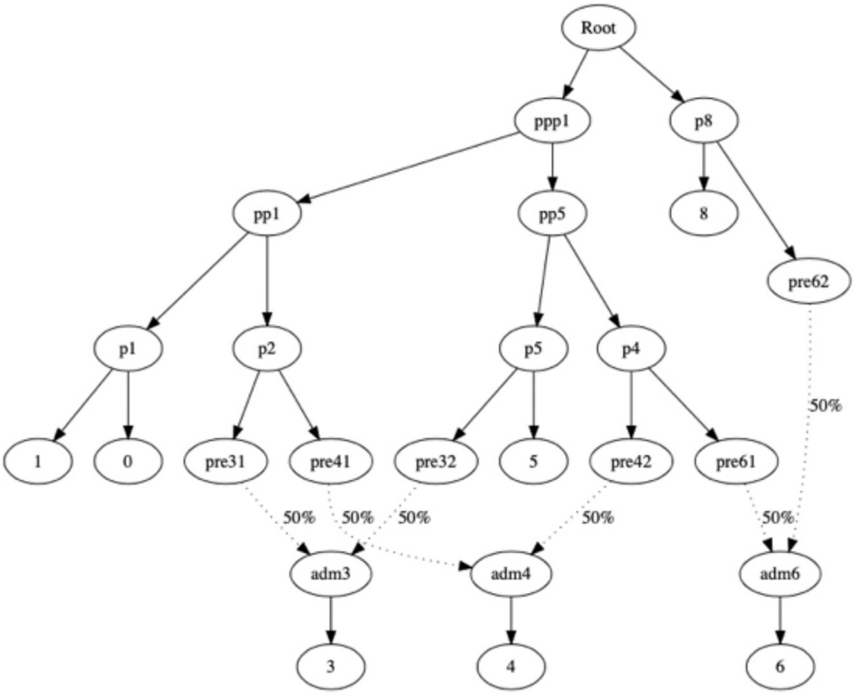


Figure S4. Simulated Phylogeny, Related to STAR Methods

This was the phylogeny used to test our inference procedure.