

The Genomic Formation of South and Central Asia

Authors: Vagheesh M. Narasimhan^{1,*}, Nick Patterson^{2,3,*}, Priya Moorjani^{4,5,+}, Iosif Lazaridis¹, Mark Lipson¹, Swapan Mallick^{1,2,6}, Nadin Rohland^{1,2}, Rebecca Bernardos¹, Alexander M. Kim^{1,7}, Nathan Nakatsuka^{1,8}, Iñigo Olalde¹, Alfredo Coppa⁹, James Mallory¹⁰, Vyacheslav Moiseyev¹¹, Janet Monge¹², Luca M. Olivieri¹³, Nicole Adamski^{1,6}, Nasreen Broomandkhoshbacht^{1,6}, Francesca Candilio^{12,14,15}, Olivia Cheronet^{14,16,17}, Brendan J. Culleton^{18,19}, Matthew Ferry^{1,6}, Daniel Fernandes^{14,16,17,20}, Beatriz Gamarra^{14,16}, Daniel Gaudio¹⁴, Mateja Hajdinjak²¹, Éadaoin Harney^{1,6,22}, Thomas K. Harper^{18,19}, Denise Keating¹⁴, Ann Marie Lawson^{1,6}, Megan Michel^{1,6,23}, Mario Novak^{14,24}, Jonas Oppenheimer^{1,6}, Niraj Rai^{25,26}, Kendra Sirak^{14,27}, Viviane Slon²¹, Kristin Stewardson^{1,6}, Zhao Zhang¹, Gaziz Akhatov²⁸, Anatoly N. Bagashev²⁹, Bauryzhan Baitanayev²⁸, Gian Luca Bonora³⁰, Tatiana Chikisheva³¹, Anatoly Derevianko³¹, Enshin Dmitry²⁹, Katerina Douka^{32,33}, Nadezhda Dubova³⁴, Andrey Epimakhov^{35,36}, Suzanne Freilich¹⁷, Dorian Fuller³⁷, Alexander Goryachev²⁹, Andrey Gromov¹¹, Bryan Hanks³⁸, Margaret Judd³⁸, Erlan Kazizov²⁸, Aleksander Khokhlov³⁹, Egor Kitov³⁴, Elena Kupriyanova⁴¹, Pavel Kuznetsov³⁹, Donata Luiselli⁴², Farhod Maksudov⁴³, Christopher Meiklejohn⁴⁴, Deborah Merrett⁴⁵, Roberto Micheli^{13,46}, Oleg Mochalov³⁹, Zahir Muhammed^{32,47}, Samariddin Mustafokulov^{43,48}, Ayushi Nayak³², Rykun M. Petrovna⁴⁹, Davide Pettener⁴², Richard Potts⁵⁰, Dmitry Razhev²⁹, Stefania Sarno⁴², Kulyan Sikhymbaeva⁴⁰, Sergey M. Slepchenko²⁹, Nadezhda Stepanova³¹, Svetlana Svyatko^{10,51}, Sergey Vasilyev³⁴, Massimo Vidale^{13,52}, Dmitriy Voyakin^{28,53}, Antonina Yermolayeva²⁸, Alisa Zubova^{11,31}, Vasant S. Shinde⁵⁴, Carles Lalueza-Fox⁵⁵, Matthias Meyer²¹, David Anthony⁵⁶, Nicole Boivin^{32,+}, Kumarasamy Thangaraj^{25,+}, Douglas J. Kennett^{18,19,+}, Michael Frachetti^{57,58,+}, Ron Pinhasi^{14,17,+}, David Reich^{1,2,6,59,+}

* 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)

Affiliations

¹ Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

² Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

³ Radcliffe Institute for Advanced Study, Harvard University, Cambridge, MA 02138, USA

⁴ Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA

⁵ Center for Computational Biology, University of California, Berkeley, CA 94720, USA

⁶ Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA

⁷ Department of Anthropology, Harvard University, Cambridge, MA 02138, USA

⁸ Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA

⁹ Dipartimento di Biologia Ambientale, Sapienza Università di Roma, Rome 00185, Italy

¹⁰ School of Natural and Built Environment, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland, UK

- 46 ¹¹ Peter the Great Museum of Anthropology and Ethnography (Kunstkamera), Russian Academy
47 of Science, St. Petersburg 199034, Russia
- 48 ¹² University of Pennsylvania Museum of Archaeology and Anthropology, Philadelphia, PA
49 19104, USA
- 50 ¹³ ISMEO Italian Archaeological Mission in Pakistan, 19200 Saidu Sharif (Swat), Pakistan
- 51 ¹⁴ Earth Institute, University College Dublin, Dublin 4, Ireland
- 52 ¹⁵ Soprintendenza Archeologia, Belle Arti e Paesaggio per la Città Metropolitana di Cagliari e le
53 Province di Oristano e Sud Sardegna, Cagliari 09124, Italy
- 54 ¹⁶ School of Archaeology, University College Dublin, Dublin 4, Ireland
- 55 ¹⁷ Department of Anthropology, University of Vienna, 1090 Vienna, Austria
- 56 ¹⁸ Department of Anthropology, Pennsylvania State University, University Park, PA 16802, USA
- 57 ¹⁹ Institutes for Energy and the Environment, Pennsylvania State University, University Park, PA
58 16802, USA
- 59 ²⁰ CIAS, Department of Life Sciences, University of Coimbra, Coimbra 3000-456, Portugal
- 60 ²¹ Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany
- 61 ²² Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA
62 02138, USA
- 63 ²³ Department of Human Evolutionary Biology, Harvard University, Cambridge MA, 02138,
64 USA
- 65 ²⁴ Institute for Anthropological Research, Zagreb 10000, Croatia
- 66 ²⁵ CSIR-Centre for Cellular and Molecular Biology, Hyderabad 500 007, India
- 67 ²⁶ Birbal Sahni Institute of Palaeosciences, Lucknow 226007, India
- 68 ²⁷ Department of Anthropology, Emory University, Atlanta, GA 30322, USA
- 69 ²⁸ Institute of Archaeology A.Kh. Margulan, Almaty 050010, Kazakhstan
- 70 ²⁹ Tyumen Scientific Centre SB RAS, Institute of the Problems of Northern Development,
71 Tyumen 625003, Russia
- 72 ³⁰ Archaeology of Asia Department, ISMEO - International Association of Mediterranean and
73 Oriental Studies, Rome RM00186, Italy
- 74 ³¹ Institute of Archaeology and Ethnography, Siberian Branch, Russian Academy of Sciences,
75 Novosibirsk 630090, Russia
- 76 ³² Department of Archaeology, Max Planck Institute for the Science of Human History, Jena
77 07745, Germany
- 78 ³³ Oxford Radiocarbon Accelerator Unit, Research Laboratory for Archaeology and the History
79 of Art, University of Oxford, Oxford OX1 3QY, UK
- 80 ³⁴ Institute of Ethnology and Anthropology, Russian Academy of Sciences, Moscow 119991,
81 Russia
- 82 ³⁵ Institute of History and Archaeology, Ural Branch RAS, Yekaterinburg 620990, Russia
- 83 ³⁶ South Ural State University, Chelyabinsk 454080, Russia
- 84 ³⁷ Institute of Archaeology, University College London, London WC1H 0PY, UK
- 85 ³⁸ University of Pittsburgh, Department of Anthropology, Pittsburgh, PA 15260, USA
- 86 ³⁹ Samara State University of Social Sciences and Education, Samara 443099, Russia
- 87 ⁴⁰ Central State Museum Republic of Kazakhstan, Samal-1 Microdistrict, Almaty 050010,
88 Kazakhstan
- 89 ⁴¹ Scientific and Educational Center of Study on the Problem of Nature and Man, Chelyabinsk
90 State University, Chelyabinsk 454021, Russia

- 91 ⁴² Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum –
92 University of Bologna, Bologna 40126, Italy
- 93 ⁴³ Institute for Archaeological Research, Uzbekistan Academy of Sciences, Samarkand 140151,
94 Uzbekistan
- 95 ⁴⁴ Department of Anthropology, University of Winnipeg, Winnipeg, MB, R3B 2E9, Canada
- 96 ⁴⁵ Department of Archaeology, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada
- 97 ⁴⁶ MiBACT - Soprintendenza Archeologia, Belle Arti e Paesaggio del Friuli Venezia Giulia,
98 34135 Trieste, Italy
- 99 ⁴⁷ Department of Archaeology, Hazara University, Mansehra 21300, Pakistan
- 100 ⁴⁸ Afrosiab Museum, Samarkand 140151, Uzbekistan
- 101 ⁴⁹ Tomsk State National Research University, Tomsk 634050, Russia
- 102 ⁵⁰ Human Origins Program, National Museum of Natural History, Smithsonian Institution,
103 Washington, DC 20013, USA
- 104 ⁵¹ CHRONO Centre for Climate, the Environment, and Chronology, Queen's University of
105 Belfast, Belfast BT7 1NN, Northern Ireland, UK
- 106 ⁵² Department of Cultural Heritage: Archaeology and History of Art, Cinema and Music,
107 University of Padua, Padua 35139, Italy
- 108 ⁵³ Archaeological Expertise LLP, Almaty 050060, Kazakhstan
- 109 ⁵⁴ Department of Archaeology, Deccan College Post-Graduate and Research Institute, Pune
110 411006, India
- 111 ⁵⁵ Institute of Evolutionary Biology, CSIC-Universitat Pompeu Fabra, Barcelona 08003, Spain
- 112 ⁵⁶ Anthropology Department, Hartwick College, Oneonta, New York 13820, USA
- 113 ⁵⁷ Department of Anthropology, Washington University in St. Louis, St. Louis, MO 63112, USA
- 114 ⁵⁸ Spatial Analysis, Interpretation, and Exploration Laboratory, Washington University in St.
115 Louis, St. Louis, MO 63112, USA
- 116 ⁵⁹ Max Planck-Harvard Research Center for the Archaeoscience of the Ancient Mediterranean,
117 Cambridge, MA 02138, USA

118 **Abstract**

119 The genetic formation of Central and South Asian populations has been unclear because of an
120 absence of ancient DNA. To address this gap, we generated genome-wide data from 362 ancient
121 individuals, including the first from eastern Iran, Turan (Uzbekistan, Turkmenistan, and
122 Tajikistan), Bronze Age Kazakhstan, and South Asia. Our data reveal a complex set of genetic
123 sources that ultimately combined to form the ancestry of South Asians today. We document a
124 southward spread of genetic ancestry from the Eurasian Steppe, correlating with the
125 archaeologically known expansion of pastoralist sites from the Steppe to Turan in the Middle
126 Bronze Age (2300-1500 BCE). These Steppe communities mixed genetically with peoples of the
127 Bactria Margiana Archaeological Complex (BMAC) whom they encountered in Turan (primarily
128 descendants of earlier agriculturalists of Iran), but there is no evidence that the main
129 BMAC population contributed genetically to later South Asians. Instead, Steppe communities
130 integrated farther south throughout the 2nd millennium BCE, and we show that they mixed with
131 a more southern population that we document at multiple sites as outlier individuals exhibiting a
132 distinctive mixture of ancestry related to Iranian agriculturalists and South Asian hunter-gathers.
133 We call this group *Indus Periphery* because they were found at sites in cultural contact with the
134 Indus Valley Civilization (IVC) and along its northern fringe, and also because they were
135 genetically similar to post-IVC groups in the Swat Valley of Pakistan. By co-analyzing ancient
136 DNA and genomic data from diverse present-day South Asians, we show that *Indus Periphery*-
137 related people are the single most important source of ancestry in South Asia—consistent with
138 the idea that the *Indus Periphery* individuals are providing us with the first direct look at the
139 ancestry of peoples of the IVC—and we develop a model for the formation of present-day South
140 Asians in terms of the temporally and geographically proximate sources of *Indus Periphery*-
141 related, Steppe, and local South Asian hunter-gatherer-related ancestry. Our results show how
142 ancestry from the Steppe genetically linked Europe and South Asia in the Bronze Age, and
143 identifies the populations that almost certainly were responsible for spreading Indo-European
144 languages across much of Eurasia.

145

146 **One Sentence Summary:** Genome wide ancient DNA from 357 individuals from Central and
147 South Asia sheds new light on the spread of Indo-European languages and parallels between the
148 genetic history of two sub-continent, Europe and South Asia.

149 **Main text**

150

151 **Ancient DNA Data and Analysis Strategy**

152

153 We generated whole-genome ancient DNA data from 362 previously unreported ancient
154 individuals and higher quality data from 17 previously reported individuals. Almost all derive
155 from three broad regions: 132 from Iran and the southern part of Central Asia (present-day
156 Turkmenistan, Uzbekistan, and Tajikistan, which we call Turan; “Iran/Turan”), 165 from the
157 western and central Steppe and northern forest zone encompassing present day Kazakhstan and
158 Russia (“Forest Zone/Steppe”), and 65 from northern Pakistan (“South Asia”). Our dataset
159 includes the first published ancient DNA data from 1) Chalcolithic and Bronze Age eastern Iran
160 and Turan (5600-1200 BCE from 12 sites); 2) early ceramic-using hunter-gatherers from the
161 western Siberian forest zone (6200-4000 BCE from 2 sites); 3) Chalcolithic and Bronze Age
162 pastoralists from the Steppe east of the Ural mountains, including the first ancient data from
163 Bronze Age Kazakhstan (4700-1000 BCE from 20 sites); and 4) the first ever ancient DNA from
164 South Asia from Iron Age and historical settlements in the Swat Valley of Pakistan (1200 BCE –
165 1 CE from 7 sites) (**Fig. 1, Supplementary Materials, Data S1**). To generate these data, we
166 prepared samples in dedicated clean rooms, extracted DNA (*1, 2*), constructed libraries for
167 Illumina sequencing (*3, 4*), and screened them using previously described procedures (*5-7*). We
168 enriched the libraries for DNA overlapping around 1.24 million single nucleotide polymorphisms
169 (SNPs), sequenced the products on Illumina instruments, and performed quality control (**Data**
170 **S1**) (*5, 6, 8*). We also report 186 new direct radiocarbon dates on human bone (**Data S2**). After
171 grouping individuals based on archaeological and chronological information and merging with
172 previously reported data, our dataset included 612 ancient individuals that we then co-analyzed
173 with genome-wide data from present-day individuals genotyped at around 600,000 SNPs, 1,789
174 of which were from 246 ethnographically-distinct groups in South Asia (**Data S3;**
175 **Supplementary Materials**) (*9-11*). We restricted analyses to ancient samples covered by at least
176 15,000 SNPs. We use *Italic* font to refer to genetic groupings and normal font to indicate
177 archaeological cultures or sites.

178

179 We carried out principal component analysis (PCA) by projecting the ancient individuals onto
180 the patterns of genetic variation in present-day Eurasians (**Fig. 1**) (*12, 13*). This revealed three

181 major groupings, closely corresponding to the geographic regions of the Forest Zone/Steppe,
182 Iran/Turan and South Asia, a pattern we replicate in ADMIXTURE clustering (14). To test
183 formally whether populations differ significantly in their ancestry within regions, we used
184 symmetry- f_4 -statistics measuring whether pairs of populations differ in their degree of allele
185 sharing to a third population, and admixture- f_3 -statistics to test formally for mixture
186 (**Supplementary Materials**). We tested the fit of mixture models using *qpAdm*, which evaluates
187 whether all possible f_4 -statistics relating a set of tested populations to outgroup populations is
188 consistent with mixtures of a pre-specified number of sources and if so estimates proportions of
189 ancestry (5). We can model almost every population as a mixture of seven deeply divergent
190 “distal” ancestry sources (usually closely related to populations for which we have data, but in
191 some cases deeply related):

- 192
- 193 • “Anatolian agriculturalist-related”: represented by 7th millennium BCE western Anatolian
194 agriculturalists (6)
 - 195 • “Western European Hunter-Gatherer (*WHG*)-related”: represented by Mesolithic western
196 Europeans (5, 10, 15, 16)
 - 197 • “Iranian agriculturalist-related”: represented by 8th millennium BCE pastoralists from the
198 Zagros Mountains of Iran (17, 18)
 - 199 • “Eastern European Hunter-Gatherer (*EHG*)-related”: represented by hunter-gatherers from
200 diverse sites in Eastern Europe (5, 6)
 - 201 • “West Siberian Hunter-Gatherer (*West_Siberian_HG*)-related”: a newly documented deep
202 source of Eurasian ancestry represented here by three samples
 - 203 • “East Asian-related”: represented in this study by Han Chinese
 - 204 • “Ancient Ancestral South Indian (*AASI*)-related”: a hypothesized South Asian Hunter-Gatherer
205 lineage related deeply to present-day indigenous Andaman Islanders (19)

206

207 We also used *qpAdm* to identify “proximal” models for each group as mixtures of temporally
208 preceding groups. This often identified multiple alternative models that were equally good fits to
209 the data. These analyses were nevertheless useful because we could identify patterns that were
210 qualitatively consistent across models. The discussion that follows presents an overview of these

211 analyses, while the **Supplementary Materials** presents the full details. **Table 1** summarizes the
212 key findings that emerge from our analysis.

213

214 **Iran/Turan**

215 We analyzed our newly generated data together with previously published data to examine the
216 genetic transformations that accompanied the spread of agriculture eastward from Iran beginning
217 in the 7th millennium BCE (20, 21). Our analysis confirms that early Iranian agriculturalists from
218 the Zagros Mountains harbor a distinctive type of West Eurasian ancestry (17, 18) (**Fig. 1**), while
219 later groups across a broad geographic region were admixed between this type of ancestry and
220 that related to early Anatolian agriculturalists. (In this paper we use the term “agriculturalists” to
221 refer both to crop cultivation and/or herding, and accordingly refer to the people of the Zagros
222 Mountains who kept domesticated goats as agriculturalists (17, 22, 23).) We show that there was
223 a west-to-east cline of decreasing Anatolian agriculturalist-related admixture ranging from ~70%
224 in Chalcolithic Anatolia to ~33% in eastern Iran, to ~3% in far eastern Turan (**Fig. 1**;
225 **Supplementary Materials**). The timing of the establishment of this cline is consistent with the
226 dates of spread of wheat and barley agriculture from west to east (in the 7th to 6th millennia
227 BCE), suggesting the possibility that individuals of Anatolian ancestry may have contributed to
228 spreading agriculturalist economies not only westward to Europe, but also eastward to Iran (21,
229 24, 25). An increase of Anatolian agriculturalist-related ancestry was also proposed for the Pre-
230 Pottery agriculturalists from the Levant in comparison to the earlier Natufian hunter-gatherers
231 (17), further supporting this hypothesis. However, without data on the distribution of
232 Anatolian/Iranian-agriculturalist ancestry in early agriculturalists in Mesopotamia, it is difficult
233 to determine when the cline was established. In the far eastern part of this cline (eastern Iran and
234 Turan) we also detect admixture related to *West_Siberian_HG*, proving that North Eurasian
235 admixture impacted Turan well before the spread of Yamnaya-related Steppe pastoralists
236 (*Steppe_EMBA*).

237

238 From Bronze Age Turan, we report 69 ancient individuals (2300-1400 BCE) from four urban
239 sites of the Bactria Margiana Archaeological Complex (BMAC) and its immediate successors.
240 The great majority of individuals fall in a genetic cluster that is similar, albeit not identical, to the
241 preceding groups in Turan in harboring a large proportion of early Iranian agriculturalist-related

242 ancestry (~60% in the *BMAC*) with smaller components of Anatolian agriculturalist-related
243 ancestry (~21%) and *West_Siberian_HG*-related ancestry (~13%) suggesting that the main
244 *BMAC* cluster coalesced from preceding pre-urban populations in Turan (which in turn likely
245 derived from earlier eastward spreads from Iran). The absence in the *BMAC* cluster of the
246 *Steppe_EMBA* ancestry that is ubiquitous in South Asia today—along with *qpAdm* analyses that
247 rule out *BMAC* as a substantial source of ancestry in South Asia (**Fig. 3A**)—suggests that while
248 the *BMAC* was affected by the same demographic forces that later impacted South Asia (the
249 southward movement of Middle to Late Bronze Age Steppe pastoralists described in the next
250 section), it was also bypassed by members of these groups who hardly mixed with *BMAC* people
251 and instead mixed with peoples further south. In fact, the data suggest that instead of the main
252 *BMAC* population having a demographic impact on South Asia, there was a larger effect of gene
253 flow in the reverse direction, as the main *BMAC* genetic cluster is slightly different from the
254 preceding Turan populations in harboring ~5% of their ancestry from the *AASI*.

255

256 We also observe outlier individuals at multiple sites, revealing interactions among populations
257 that would be difficult to appreciate without the large sample sizes reported here.

258

259 First, around ~2300 BCE in Turan, we observe two outliers at the *BMAC* site of Gonur with
260 *West_Siberian_HG*-related ancestry of a type that we observe at multiple sites in Kazakhstan
261 over the preceding and succeeding millennia. The most plausible explanation is that this ancestry
262 is that of indigenous populations associated with the Kelteminar culture, the native hunter-
263 gatherers of the region who covered a vast area of Central Asia before the *BMAC* (26). Future
264 ancient DNA data from Kelteminar contexts will make it possible to determine whether it is
265 indeed the case that the genetic ancestry of Kelteminar people was similar to that of
266 *West_Siberia_HG*. Importantly, in the 3rd millennium BCE we do not find any individuals with
267 ancestry derived from Yamnaya-related Steppe pastoralists in Turan. Thus, *Steppe_EMBA*
268 ancestry was not yet widespread across the region.

269

270 Second, between 2100-1700 BCE, we observe *BMAC* outliers from three sites with
271 *Steppe_EMBA* ancestry in the admixed form typically carried by the later Middle to Late Bronze
272 Age Steppe groups (*Steppe_MLBA*). This documents a southward movement of Steppe ancestry

273 through this region that only began to have a major impact around the turn of the 2nd millennium
274 BCE.

275
276 Third, between 3100-2200 BCE we observe an outlier at the BMAC site of Gonur, as well as two
277 outliers from the eastern Iranian site of Shahr-i-Sokhta, all with an ancestry profile similar to 41
278 ancient individuals from northern Pakistan who lived approximately a millennium later in the
279 isolated Swat region of the northern Indus Valley (1200-800 BCE). These individuals had
280 between 14-42% of their ancestry related to the AASI and the rest related to early Iranian
281 agriculturalists and *West_Siberian_HG*. Like contemporary and earlier samples from Iran/Turan
282 we find no evidence of Steppe-pastoralist-related ancestry in these samples. In contrast to all
283 other Iran/Turan samples, we find that these individuals also had negligible Anatolian
284 agriculturalist-related admixture, suggesting that they might be migrants from a population
285 further east along the cline of decreasing Anatolian agriculturalist ancestry. While we do not
286 have access to any DNA directly sampled from the Indus Valley Civilization (IVC), based on (a)
287 archaeological evidence of material culture exchange between the IVC and both BMAC to its
288 north and Shahr-i-Sokhta to its east (27), (b) the similarity of these outlier individuals to post-
289 IVC Swat Valley individuals described in the next section (27), (c) the presence of substantial
290 *AASI* admixture in these samples suggesting that they are migrants from South Asia, and (d) the
291 fact that these individuals fit as ancestral populations for present-day Indian groups in *qpAdm*
292 modeling, we hypothesize that these outliers were recent migrants from the IVC. Without ancient
293 DNA from individuals buried in IVC cultural contexts, we cannot rule out the possibility that the
294 group represented by these outlier individuals, which we call *Indus_Periphery*, was limited to the
295 northern fringe and not representative of the ancestry of the entire Indus Valley Civilization
296 population. In fact, it was certainly the case that the peoples of the Indus Valley were genetically
297 heterogeneous as we observe one of the *Indus_Periphery* individuals having ~42% *AASI*
298 ancestry and the other two individuals having ~14-18% *AASI* ancestry (but always mixes of the
299 same two proximal sources of *AASI* and Iranian agriculturalist-related ancestry). Nevertheless,
300 these results show that *Indus_Periphery* were part of an important ancestry cline in the wider
301 Indus region in the 3rd millennium and early 2nd millennium BCE. As we show in what follows,
302 peoples related to this group had a pivotal role in the formation of subsequent populations in
303 South Asia.

304
305 Using a newly developed approach for estimating dates of admixture in ancient genomes (an
306 adaptation of a previous method to measure ancestry covariance among pairs of neighboring
307 positions in the genome; **Supplementary Materials**), we estimate that the time of admixture
308 between Iranian agriculturalist-related ancestry and AASI ancestry in the three *Indus_Periphery*
309 samples was 53 ± 15 generations ago on average, corresponding to a 95% confidence interval of
310 about 4700-3000 BCE assuming 28 years per generation (28). This places a minimum date on the
311 first contact between these two types of ancestries.

312 313 **The Steppe**

314 Three individuals from the West Siberian forest zone with direct dates ranging from 6200 BCE
315 to 4000 BCE play an important role in this study as they are representatives of a never-before-
316 reported mixture of ancestry that we call *West_Siberian_HG*: ~30% derived from *EHG*, ~50%
317 from Ancestral North Eurasians (defined as being related deeply to 22000-15000 BCE Siberians
318 (29, 30)), and ~20% related to present-day East Asians. This ancestry type also existed in the
319 southern Steppe and in Turan, as it formed about 80% of the ancestry of an early 3rd millennium
320 BCE agro-pastoralist from Dali, Kazakhstan, and also contributed to multiple outlier individuals
321 from 2nd millennium sites in Kazakhstan and Turan (**Fig. 2**).

322
323 Using the *West_Siberian_HG* individuals as a reference population along with other pre-
324 Chalcolithic groups that have been previously reported in the ancient DNA literature, we
325 document the presence of a genetically relatively homogeneous population spread across a vast
326 region of the eastern European and trans-Ural Steppe between 2000-1400 BCE (*Steppe_MLBA*)
327 (17). Many of the samples from this group are individuals buried in association with artifacts of
328 the Corded Ware, Srubnaya, Petrovka, Sintashta and Andronovo complexes, all of which
329 harbored a mixture of *Steppe_EMBA* ancestry and ancestry from European Middle Neolithic
330 agriculturalists (*Europe_MN*). This is consistent with previous findings showing that following
331 westward movement of eastern European populations and mixture with local European
332 agriculturalists, there was an eastward reflux back beyond the Urals (6, 16, 31). Our new dataset
333 enhances our understanding of the *Steppe_MLBA* cluster by including many sites in present-day
334 Kazakhstan and as far east as the Minusinsk Basin of Russia—and in doing so allows us to

335 appreciate previously undetected substructure. All previously reported samples fall into a
336 subcluster we call *Steppe_MLBA_West* that harbors ~26% *Europe_MN* ancestry and ~74%
337 *Steppe_EMBA* ancestry. With our newly reported data we now also detect a previously
338 unappreciated subcluster, *Steppe_MLBA_East*, which is significantly differentiated ($p=7\times 10^{-6}$
339 from *qpAdm*), with ~8% *West_Siberian_HG*-related ancestry and proportionally less of the other
340 ancestry components, suggesting that people carrying *Steppe_MLBA_West* ancestry admixed
341 with *West_Siberian_HG*-related peoples as they spread further east.

342

343 As in Iran/Turan, the outlier individuals provide key additional information.

344

345 First, our analysis of 50 newly reported individuals from the Kamennyi Ambar V cemetery from
346 the Sintashta culture reveals three groups of outliers, in addition to the main cluster of 40
347 individuals. These outliers have elevated proportions of *Steppe_EMBA*, *West_Siberian_HG* or
348 *East Asian*-related ancestry (and direct dates that are contemporaneous with the other
349 individuals), thereby showing that this fortified site harbored people of diverse ancestries living
350 side-by-side.

351

352 Second, samples from three sites from the southern and eastern end of the Steppe dated to 1600-
353 1500 BCE (Dashti-kozy, Taldysay and Kyzylbulak) show evidence of significant admixture from
354 Iranian agriculturalist-related populations, demonstrating northward gene flow from Turan into
355 the Steppe at the same time as there was southward movement of *Steppe_MLBA* ancestry
356 through Turan and into South Asia. These findings are consistent with evidence of a high degree
357 of human mobility both to the north and south along the Inner Asian Mountain Corridor (32, 33).

358

359 Third, we observe samples from multiple sites dated to 1700-1500 BCE (Maitan, Kairan,
360 Oy_Dzhaylau and Zevakinskiy) that derive up to ~25% of their ancestry from a source related to
361 present-day East Asians and the remainder from *Steppe_MLBA*. A similar ancestry profile
362 became widespread in the region by the Late Bronze Age, as documented by our time transect
363 from Zevakinskiy and samples from many sites dating to 1500-1000 BCE, and was ubiquitous
364 by the Scytho-Sarmatian period in the Iron Age (34). This observation decreases the probability
365 that populations in the 1st millennium BCE and 1st millennium CE—including Scythians,

366 Kushans, and Huns, sometimes suggested as sources for the Steppe ancestry influences in South
367 Asia today (17)—contributed to the majority of South Asians, which have negligible East Asian
368 ancestry in our analysis. It is possible that there were unsampled groups in Central Asia with
369 negligible East Asian admixture that could have migrated later to South Asia. However, at least
370 some (possibly all) of the Steppe pastoralist ancestry in South Asia owes its origins to southward
371 pulses in the 2nd millennium BCE, as indeed we prove directly through our observation of this
372 ancestry in the Swat Iron Age individuals dating to ~1000 BCE (discussed further below).

373

374 **South Asia**

375 Previous work has shown that the Indian Cline—a gradient of different proportions of West
376 Eurasian related ancestry in South Asia—can be well modeled as having arisen from a mixture
377 of two statistically reconstructed ancestral populations (the *ANI* and the *ASI*), which mixed
378 mostly after 2000 BCE (35, 36). Ancient DNA analysis has furthermore revealed that the
379 populations along the Indian Cline actually descend more deeply in time from at least three
380 ancestral populations (17), with ancestry from groups related to early Iranian agriculturalists,
381 *Steppe_EMBA*, and *Onge*.

382

383 To shed light on the mixture events that transformed this minimum of three ancestral populations
384 into two (the *ANI* and *ASI*), we used *qpAdm* to search for triples of source populations—the
385 *AASI*, all sampled ancient Iran/Turan-related groups, and all sampled ancient Steppe groups—
386 that could fit as sources for South Asians. As South Asian test populations we used an Indian
387 Cline group with high ANI ancestry (*Punjabi.DG*), one with high ASI ancestry (*Mala.DG*), early
388 Iron Age Swat Valley samples (*Swat Protohistoric Grave Type - SPGT*), and Early Historic Swat
389 Valley samples (*Butkara_IA*). **Fig. 3A** shows that the only models that fit all four test South
390 Asians groups are combinations that involve the *AASI*, *Indus_Periphery* and *Steppe_MLBA* (in
391 the analyses that follow, we therefore pooled the *Steppe_MLBA*). The evidence that the
392 *Steppe_MLBA* cluster is a plausible source for the Steppe ancestry in South Asia is also
393 supported by Y chromosome evidence, as haplogroup R1a which is of the Z93 subtype common
394 in South Asia today (37, 38) was of high frequency in *Steppe_MLBA* (68%) (16), but rare in
395 *Steppe_EMBA* (absent in our data).

396

397 To obtain a richer understanding of the ancestry of the entire Indian Cline, we took advantage of
398 previously published genome-wide data from 246 ethnographically diverse groups from South
399 Asia (*II*), from which we sub-selected 140 groups that fall on a clear gradient in PCA to
400 represent the Indian Cline (the other groups either fall off the cline due to additional African or
401 East Asian-related ancestry or had small sample size or heterogeneous ancestry). The per-group
402 *qpAdm* estimates for the proportions of ancestry from these three sources are statistically noisy.
403 We therefore developed new methodology that allows us to jointly fit the data from all Indian
404 Cline groups within a hierarchical model. The analysis confirms that the great majority of all
405 groups on the Indian Cline can be jointly modeled as a mixture of two populations, and the
406 analysis also produces an estimate of the functional relationship between the ancestry
407 components. Setting *Steppe_MLBA* to its smallest possible proportion of zero to estimate the
408 minimum fraction of *Indus_Periphery* ancestry that could have existed in the *ASI*, we obtain
409 ~39%. Setting *AASI* to its smallest possible proportion of zero to estimate the maximal fraction
410 of *Indus_Periphery* ancestry that could have existed in the *ANI*, we obtain ~72%. In fact, we find
411 four tribal groups from southern India (*Palliyar*, *Ulladan*, *Malayan*, and *Adiyan*) with close to
412 the maximum mathematically allowed proportion of *Indus_Periphery*-related ancestry, and we
413 find a population in northern Pakistan (*Kalash*) with close to the minimum. Thus, nearly
414 unmixed descendants of the *ASI* and *ANI* exist as isolated groups in South Asia today.

415
416 We built an admixture graph using *qpGraph* co-modeling *Palliyar* (as a representative of the
417 *ASI*) and *Juang* (an Austroasiatic speaking group in India with low West Eurasian-relatedness),
418 and show that it fits when the *ASI* have ~27% Iranian agriculturalist-related ancestry and the
419 *Juang* also harbor ancestry from an *AASI* population without Iranian admixture (**Fig. 3**). This
420 model is also notable in showing that early Iranian agriculturalists fit without *AASI* admixture,
421 and thus the patterns we observe are driven by gene flow into South Asia and not the reverse
422 (**Fig. 3; Supplementary Materials**). The fitted admixture graph also reveals that the deep
423 ancestry of the indigenous hunter-gather population of India represents an anciently divergent
424 branch of Asian human variation that split off around the same time that East Asian, *Onge* and
425 *Australian* aboriginal ancestors separated from each other. This finding is consistent with a
426 model in which essentially all the ancestry of present-day eastern and southern Asians (prior to

427 West Eurasian-related admixture) derives from a single eastward spread, which gave rise in a
428 short span of time to the lineages leading to *AASI*, East Asians, *Onge*, and *Australians* (19).

429
430 Using admixture linkage disequilibrium, we estimate a date of 107 ± 11 generations ago for the
431 Iranian agriculturalist and AASI-related admixture in the *Palliyar*, corresponding to a 95%
432 confidence interval of 1700-400 BCE assuming 28 years per generation (28). This date is
433 consistent with a previous estimate of 110 ± 12 generations ago for the *Kalash* (39). These
434 results suggest that the *ASI* and *ANI* were both largely unformed at the beginning of the 2nd
435 millennium BCE, and imply that the *ASI* may have formed in the course of the spread of West
436 Asian domesticates into peninsular India beginning around 3000 BCE (where they were
437 combined with local domesticates to form the basis of the early agriculturalist economy of South
438 India (40)), or alternatively in association with eastward spread of material culture from the
439 Indus Valley after the IVC declined (41). Further evidence for a Bronze Age formation of the
440 *ASI* comes from our analysis of Austroasiatic-speaking groups in India such as *Juang*, who have
441 a higher ratio of *AASI*-to-Iranian agriculturalist-related ancestry than the *ASI* (**Fig. 3,**
442 **Supplementary Materials**). Austroasiatic speakers likely descend from populations that arrived
443 in South Asia in the 3rd millennium BCE (based on hill cultivation systems associated with the
444 spread of Austroasiatic languages (20)), and our genetic results show that when Austroasiatic
445 speakers arrived they mixed with groups with elevated ratios of *AASI*- to Iranian-agriculturalist-
446 related ancestry than are found in the *ASI*, showing that the *ASI* had not yet overspread
447 peninsular India.

448
449 Finally, we examined our Swat Valley time transect from 1200 BCE to 1 CE. While the earliest
450 group of samples (*SPGT*) is genetically very similar to the *Indus_Periphery* samples from the
451 sites of *Gonur* and *Shahr-i-Sokhta*, they also differ significantly in harboring *Steppe_MLBA*
452 ancestry (~22%). This provides direct evidence for *Steppe_MLBA* ancestry being integrated into
453 South Asian groups in the 2nd millennium BCE, and is also consistent with the evidence of
454 southward expansions of *Steppe_MLBA* groups through Turan at this time via outliers from the
455 main *BMAC* cluster from 2000-1500 BCE. Later samples from the Swat time transect from the
456 1st millennium BCE had higher proportions of Steppe and AASI derived ancestry more similar to

457 that found on the Indian Cline, showing that there was an increasing percolation of Steppe
458 derived ancestry into the region and additional admixture with the ASI through time.

459

460 **Implications for Archaeology and Linguistics**

461 Our evidence that a population with both Iranian agriculturalist and South Asian hunter gatherer
462 ancestry (*Indus_Periphery*) was established in the 3rd millennium BCE—and that its Iranian
463 agriculturalist-related and *AASI* ancestry sources mixed at an average time of around 4700-3000
464 BCE—shows that this type of Iranian agriculturalist-related ancestry must have reached the
465 Indus Valley by the 4th millennium BCE. However, it is very possible that Iranian agriculturalist-
466 related ancestry was widespread in South Asia even earlier, as wheat and barley agriculture as
467 well as goat and sheep herding spread into South Asia after the 7th millennium BCE, as attested
468 at sites such as Mehrgarh in the hills surrounding the Indus Valley (20, 21), and these
469 domesticates could have been carried by movements of people. Regardless of when these
470 agricultural species arrived, the genetic data show that *Indus_Periphery*-related ancestry
471 contributed in large proportions to both the *ANI* and *ASI*, and that these two groups both formed
472 in the 2nd millennium BCE, overlapping the decline of the IVC and major changes in settlement
473 patterns in the northern part of the Indian subcontinent (41). A parsimonious hypothesis is that as
474 *Steppe_MLBA* groups moved south and mixed with *Indus_Periphery*-related groups at the end of
475 the IVC to form the *ANI*, other *Indus_Periphery*-related groups moved further south and east to
476 mix with *AASI* groups in peninsular India to form the *ASI*. This is consistent with suggestions
477 that the spread of the IVC was responsible for dispersing Dravidian languages (42-44), although
478 scenarios in which Dravidian languages derive from pre-Indus languages of peninsular India are
479 also entirely plausible as *ASI* ancestry is mostly derived from the *AASI*.

480

481 Our results also shed light on the question of the origins of the subset of Indo-European
482 languages spoken in India and Europe (45). It is striking that the great majority of Indo-European
483 speakers today living in both Europe and South Asia harbor large fractions of ancestry related to
484 Yamnaya Steppe pastoralists (corresponding genetically to the *Steppe_EMBA* cluster),
485 suggesting that “Late Proto-Indo-European”—the language ancestral to all modern Indo-
486 European languages—was the language of the Yamnaya (46). While ancient DNA studies have
487 documented westward movements of peoples from the Steppe that plausibly spread this ancestry

488 to Europe (5, 31), there has not been ancient DNA evidence of the chain of transmission to South
489 Asia. Our documentation of a large-scale genetic pressure from *Steppe_MLBA* groups in the 2nd
490 millennium BCE provides a prime candidate, a finding that is consistent with archaeological
491 evidence of connections between material culture in the Kazakh middle-to-late Bronze Age
492 Steppe and early Vedic culture in India (46).

493
494 Our analysis also provides an entirely new line of evidence for a linkage between Steppe
495 ancestry and Indo-European culture. When we used *qpAdm* to test if a mixture of ANI and ASI is
496 a fit to the data for all 140 Indian Cline groups, we found 10 groups with poor fits and a
497 significantly elevated ratio of *Steppe_MLBA*- to *Indus_Periphery*-related ancestry compared to
498 the expectation for the model ($Z \geq 3$). We found the strongest two signals in *Brahmin_Tiwari*
499 ($p=2 \times 10^{-5}$) and *Brahmin_UP* ($p=4 \times 10^{-5}$), and more generally there was a striking enrichment of
500 a $Z \geq 3$ signals in groups of traditionally priestly status in northern India (57% of groups with $Z \geq 3$
501 were *Brahmins* or *Bhumihars* even though these groups comprised only 11% of the 74 groups we
502 analyzed in northern India). Although the enrichment for Steppe ancestry is not found in the
503 southern Indian groups, the Steppe enrichment in the northern groups is striking as *Brahmins* and
504 *Bhumihars* are among the traditional custodians of texts written in early Sanskrit. A possible
505 explanation is that the influx of *Steppe_MLBA* ancestry into South Asia in the mid-2nd
506 millennium BCE created a meta-population of groups with different proportions of Steppe
507 ancestry, with ones having relatively more Steppe ancestry having a central role in spreading
508 early Vedic culture. Due to strong endogamy in South Asia—which has kept some groups
509 isolated from their neighbors for thousands of years (35)—some of this substructure within
510 Indian population still persists.

511
512 We finally highlight a remarkable parallel between the prehistory of two sub-continent of
513 Eurasia: South Asia and Europe. In both regions, West Asian agricultural technology spread
514 from an origin in the Near East in the 7th and 6th millennia BCE (**Fig. 4**). In South Asia this
515 occurred via the Iranian plateau, and in Europe via western Anatolia, with the technological
516 spreads mediated in both cases by movements of people. An admixed population was then
517 formed by the mixing of incoming agriculturalists and resident hunter-gatherers—in South Asia
518 eventually giving rise to the *Indus_Periphery* and *ASI* and in Europe the Middle Neolithic

519 genetic cluster *Europe_MN*. In both Europe and South Asia, populations related to the *Yamnaya*
520 Steppe pastoralists arrived after this agriculturalist and hunter-gatherer admixture took place,
521 interacting with local populations to produce mixed groups, which then mixed further with
522 already resident agriculturalist populations to produce genetic groupings such as those found
523 associated with Corded Ware and central European Bell Beaker artifacts in much of Europe, and
524 the *ANI* genetic cluster in South Asia. These mixed groups then mixed further to produce the
525 major gradients of ancestry in both regions. Future studies of populations from South Asia and
526 the linguistically related Iranian world will extend and add nuance to the model presented here.

527 **Figure Legends**

528

529 **Fig. 1 Overview of ancient DNA data.** (A) Number of newly reported samples passing our
530 analysis thresholds and their date range is shown by site. (B) Locations, color-coded by analysis
531 grouping. (C) Projections of ancient samples onto PCA axes computed using present-day
532 Eurasians. (D) ADMIXTURE analysis, with components maximized in *West_Siberian_HG*,
533 Anatolian agriculturalists, Iranian agriculturalists, indigenous South Asians and *WHG* in blue,
534 orange, teal, red and green, respectively. (E) Y-chromosome haplogroups. N, Neolithic; C,
535 Chalcolithic; BA, Bronze Age; IA, Iron Age; H, Historic; E/M/L, Early/Middle/Late; o, outlier.
536

537 **Fig. 2 Modeling results.** (A) Admixture events originating from 7 “Distal” populations leading
538 to the formation of the modern Indian cloud shown geographically. Clines or 2-way mixtures of
539 ancestry are shown in rectangles, and clouds (3-way mixtures) are shown in ellipses. (B) A
540 schematic model of events originating from 7 “Distal” populations leading to the formation of
541 the modern Indian cline, shown chronologically. (C) Admixture proportions as estimated
542 using *qpAdm* for populations reflected in A and B.
543

544 **Fig. 3 The Genomic Origins of Indians.** (A) We used *qpAdm* to model four groups that are
545 representative of major sources of South Asian ancestry over the last few thousand years
546 (*Punjabi.DG*, *Mala.DG*, *SPGT*, and *Butkara_IA*) as mixtures of *Onge*, an Iran/Turan-related
547 population, and a Steppe-related group, and report the minimum p-value (highlighting cases at
548 $p > 0.01$). The only working models involve a combination of *Indus_Periphery* and a
549 *Steppe_MLBA* group (note that the *Steppe_MLBA_West* group includes a subset
550 *Sintashta_MLBA* and *Srubnaya*). (B) For all 140 Indian Cline groups, we give *Maximum A*
551 *Posteriori* fits for this model. Significant outliers ($|Z| > 2$) are shown, and include a cluster of
552 *Brahmins* (filled circles) and *Bhumihars* (filled squares) with excess Steppe pastoralist-related
553 ancestry compared to others with similar West Eurasian ancestry proportion. (C) Admixture
554 graph fit supports Iranian agriculturalist-related admixture into South Asia but no gives evidence
555 of AASI-related admixture into ancient Iran; dotted lines show admixture events.
556

557 **Fig. 4 A Tale of Two Subcontinents.** The prehistory of South Asia and Europe are parallel in
558 both being impacted by two successive spreads, the first from the Near East after 7000 BCE
559 bringing agriculturalists who mixed with local hunter-gatherers, and the second from the Steppe
560 after 3000 BCE bringing people who spoke Indo-European languages and who mixed with those
561 they encountered during their migratory movement. Mixtures of these mixed populations then
562 produced the rough clines of ancestry present in both South Asia and in Europe today (albeit
563 with more variable proportions of local hunter-gatherer-related ancestry in Europe than in India),
564 which are (imperfectly) correlated to geography. The plot shows in contour lines the time of the
565 expansion of Near Eastern agriculture. Human movements and mixtures, which also plausibly
566 contributed to the spread of languages, are shown with arrows.

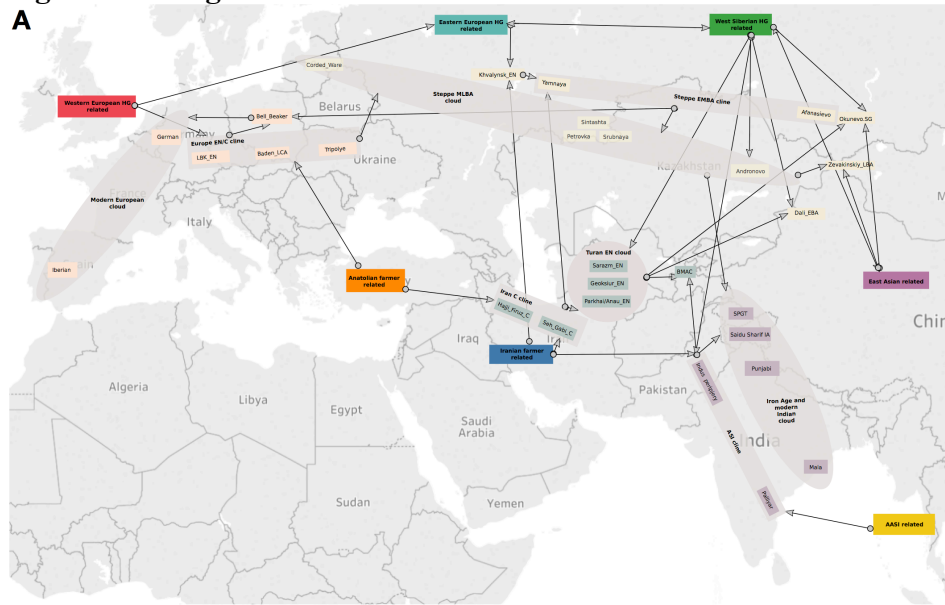
567 **Table 1 Summary of Key Findings**

<p>Iran/Turan</p> <ol style="list-style-type: none">1. There was a west-to-east gradient of ancestry across Eurasia in the Chalcolithic and Bronze Ages, with more Anatolian agriculturalist-related ancestry in the west and more <i>West_Siberian_HG</i> or <i>AASI</i>-related ancestry in the east, all superimposed on primary ancestry related to early Zagros agriculturalists. The establishment of the Anatolian ancestry gradient corresponds to the spread of crop-based agriculture across this region, raising the possibility that people of Anatolian ancestry spread this technology east just as they helped spread it west into Europe. However, Anatolian agriculturalist-related ancestry is absent in the <i>Indus_Periphery</i> samples, showing that if such people were instrumental in bringing crop farming eastward to Iran, diffusion of ideas brought it further east to South Asia.2. The primary population of the BMAC was largely derived from preceding local Chalcolithic peoples and had little if any Steppe pastoralist ancestry of the type that is ubiquitous in South Asia today. Instead of being a source for South Asia, the <i>BMAC</i> received admixture from South Asia.3. Outlier analysis shows no evidence of Steppe pastoralist ancestry in groups surrounding BMAC sites prior to 2100 BCE, but suggests that between 2100-1700 BCE, the BMAC communities were surrounded by peoples carrying such ancestry.4. We document a distinctive ancestry type—58%-86% Iranian agriculturalist-related ancestry with little Anatolian agriculturalist-related admixture, and 14%-42% <i>AASI</i> ancestry—that was present at two sites known to be in close cultural contact with the Indus Valley Culture (IVC). Combined with similar ancestry about a millennium later in the post-IVC Swat Valley, this documents an <i>Indus_Periphery</i> population during the flourishing of the IVC, which we show formed by admixture 4700-3000 BCE.
<p>The Steppe</p> <ol style="list-style-type: none">1. In the Kazakh Steppe and Minusinsk Basin during the Middle to Late Bronze Age, ancestry typical of pastoralists in the western Steppe (<i>Steppe_MLBA_West</i>) admixed with ancestry related to earlier <i>West_Siberian_HG</i>-related groups to form a distinctive <i>Steppe_MLBA_East</i> cluster.2. Outlier analysis shows that by 1600 BCE in the Middle to Late Bronze Age of the Kazakh Steppe, there were numerous individuals with admixture from Turan, providing genetic evidence of northward movement into the Steppe in this period.3. By 1500 BCE, there were numerous individuals in the Kazakh Steppe with East Asian-related admixture, the same type of ancestry that was widespread by the Scythian period (34). This ancestry is hardly present in the two primary ancestral populations of South Asia—<i>ANI</i> and <i>ASI</i>—suggesting that Steppe ancestry widespread in South Asia derived from earlier southward movements.
<p>South Asia</p> <ol style="list-style-type: none">1. After exploring a wide range of models of present-day and ancient South Asia, we identify a unique class of models that fits geographically and temporally South Asians: a mixture of <i>AASI</i>, <i>Indus_Periphery</i>, and <i>Steppe_MLBA</i>. We reject <i>BMAC</i> as a primary source of ancestry in South Asians.2. A population of which the <i>Indus_Periphery</i> samples were a part played a pivotal role in the formation of the two proximal sources of ancestry in South Asia, the <i>ANI</i> and <i>ASI</i>. Both ends of the Indian Cline had major components of <i>Indus_Periphery</i> admixture: ~39% for the <i>ASI</i> and ~72% for the <i>ANI</i>. Today there are groups in South Asia with very similar ancestry to the <i>ASI</i> and <i>ANI</i>.3. Much of the formation of both the <i>ASI</i> and <i>ANI</i> occurred in the 2nd millennium BCE. Thus, the events that formed both the <i>ASI</i> and <i>ANI</i> overlapped the decline of the IVC.4. The <i>ASI</i> were not a clade with the earlier hunter-gatherer populations of South Asia (<i>AASI</i>), but harbored significant amounts of ancestry related to early Iranian agriculturalists, likely transmitted though the IVC.

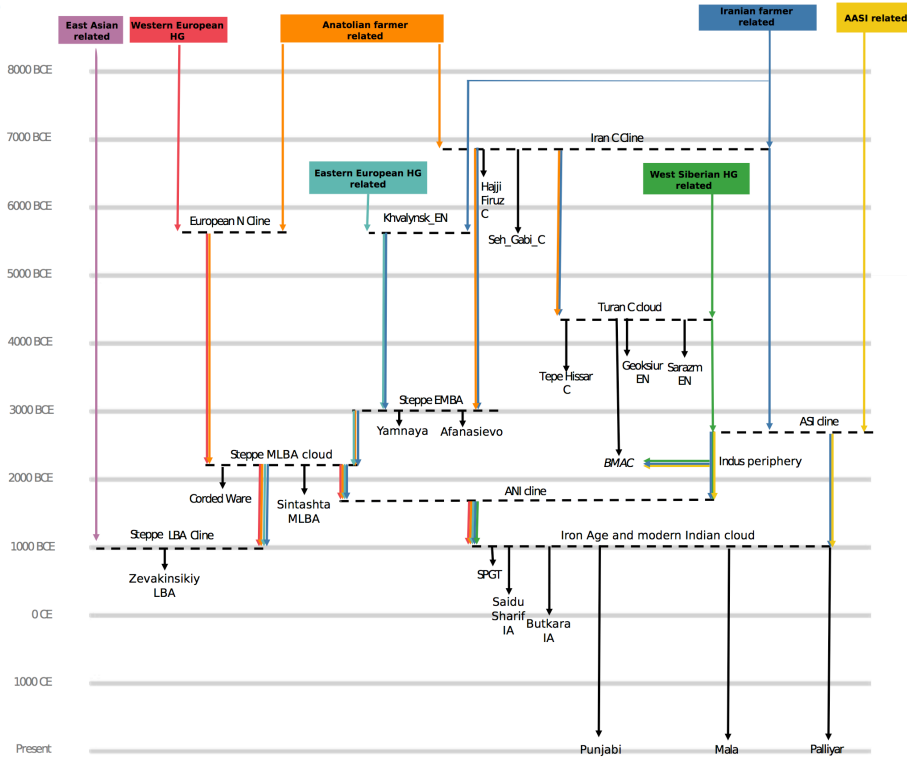
568

Fig. 2 Modeling results

A



B



C

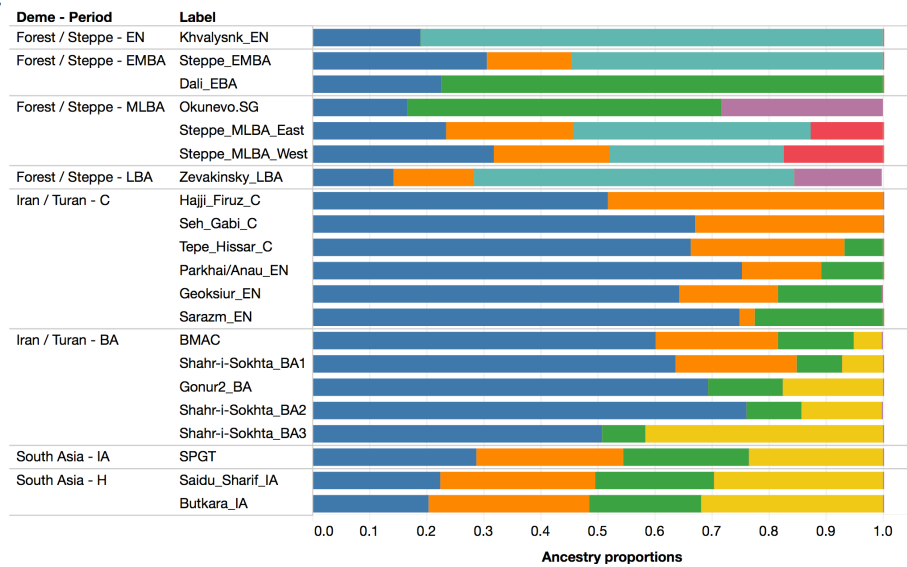
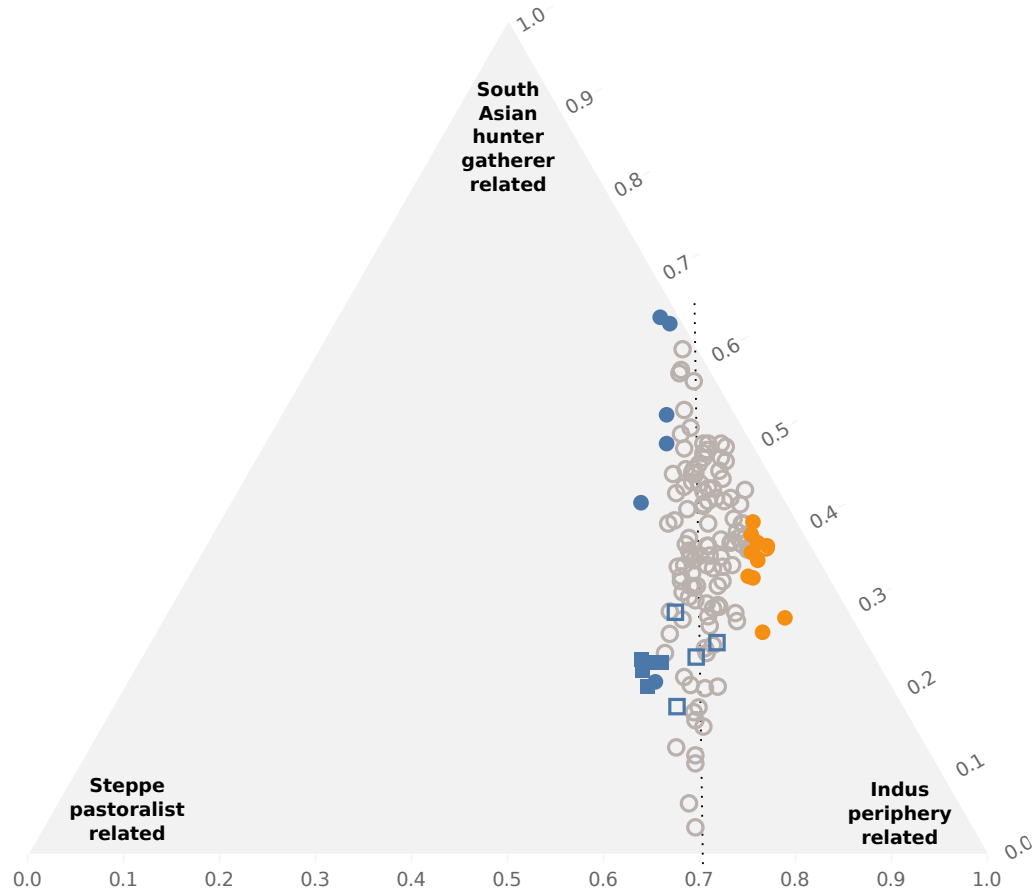


Fig. 3 The genomic origin of Indians

A

Tested combinations	CHG	Armenia_EBA	Hajji_Firuz_C	Seh_Gabl_C	Tepe_Hissar_C	Geokslur_EN	BMAC	Indus_Periphery
<i>Khvalynsk_EN</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>Steppe_EMBA</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>Okunevo_SG</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>Sintashta_MLBA</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1877
<i>Srubnaya</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2024
<i>Steppe_MLBA_West</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1599
<i>Steppe_MLBA_East</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0834
<i>Steppe_LBA</i>	0.0000	0.0000	0.0000	0.0000	0.0001	0.0045	0.0000	0.0000

B



C

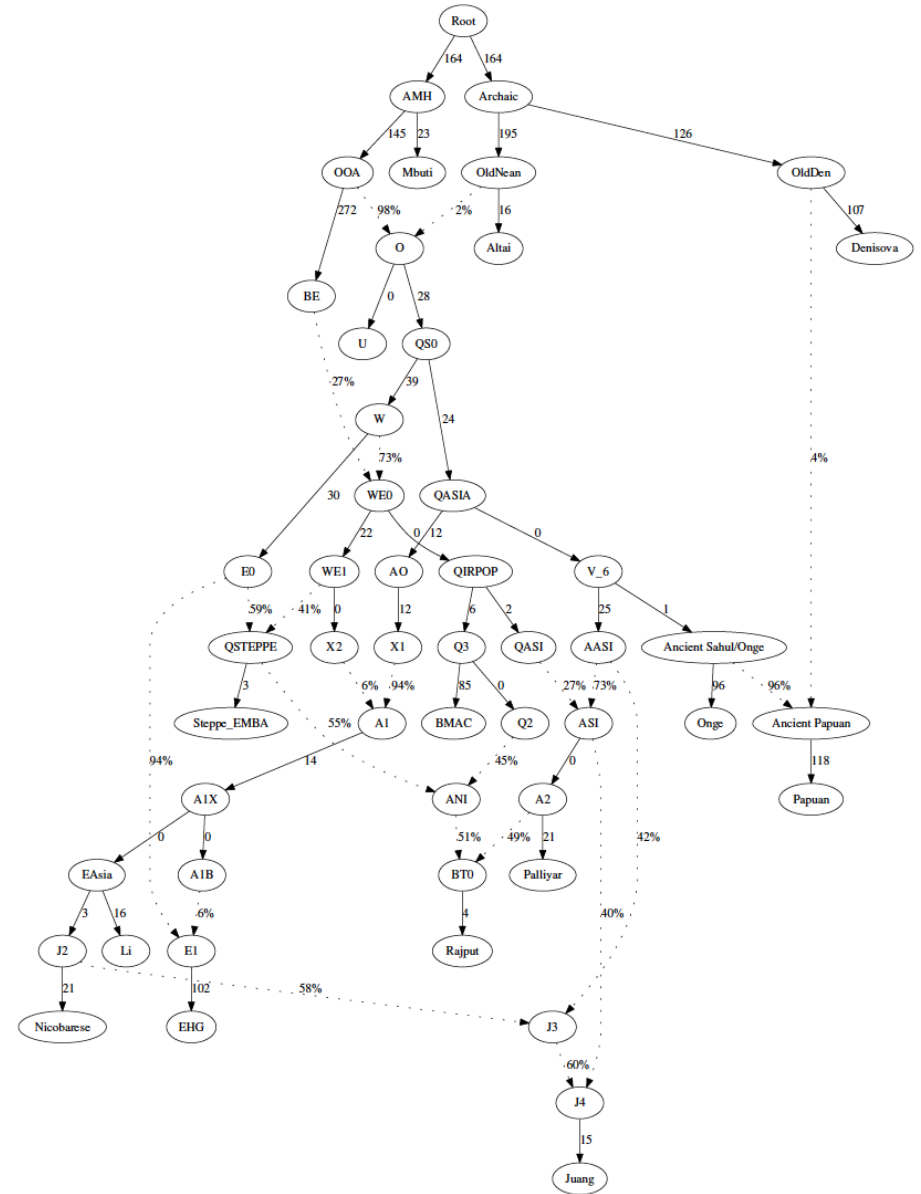
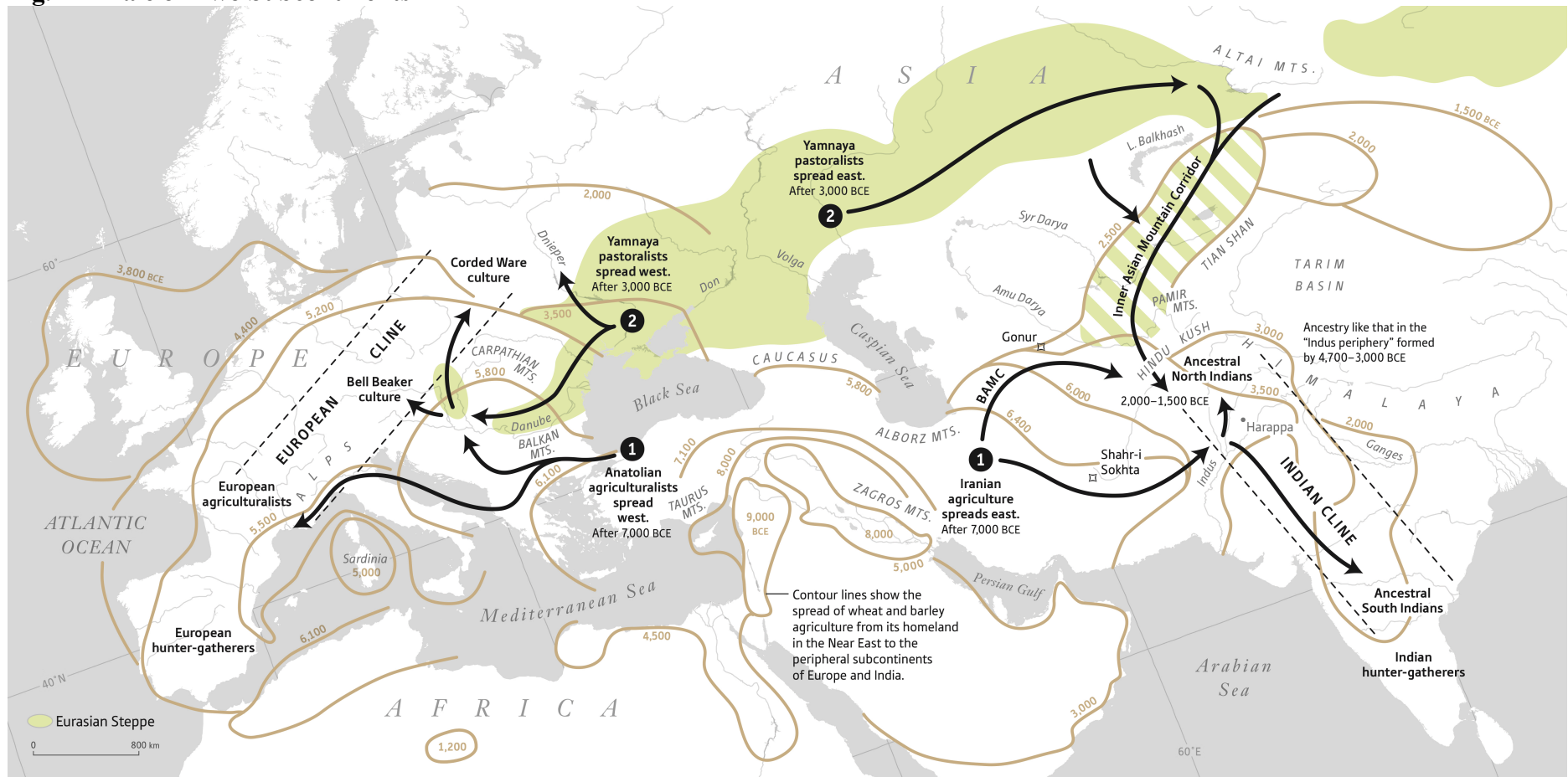


Fig. 4 A Tale of Two Subcontinents



1 **Materials and Methods**

2
3 **Ancient DNA Laboratory Work** We prepared powder from human skeletal remains either at field
4 sites using a method for extracting petrous bone powder by drilling directly from the cranial base
5 (47), or in dedicated clean rooms at Harvard Medical School, University College Dublin, or the
6 Max Planck Institute for Evolutionary Anthropology in Leipzig Germany.

7
8 All the molecular work except for that of a single sample (Darra-i-Kur) was carried out at Harvard
9 Medical School (HMS). At HMS, we extracted DNA using a method that is optimized to retain
10 small DNA fragments (1, 2). We converted the DNA into a form that could be sequenced using a
11 double-stranded library preparation protocol, usually pre-treating with the enzyme Uracil-DNA
12 Glycosylase (UDG) to reduce the characteristic cytosine-to-thymine errors in ancient DNA (4). For
13 some libraries, we substituted the MinElute columns used for cleaning up reactions with magnetic
14 beads, and the MinElute column-based PCR cleanup at the end of library preparation with SPRI
15 beads (48, 49). We enriched the libraries both for sequences overlapping mitochondrial DNA (50),
16 and for sequences overlapping about 1.24 million nuclear targets after two rounds of enrichment (5,
17 6, 8). We sequenced the enriched products on an Illumina NextSeq500 using v.2 150 cycle kits for
18 2×76 cycles and 2×7 cycles, and sequenced up to the point that the expected number of new SNPs
19 covered per 100 additional read pairs sequenced was approximately less than 1.

20
21 To analyze the data computationally, we separated read pairs into individuals based on searching
22 for the expected two indices and two barcodes, allowing up to one mismatch from the expected
23 sequence in each case. We removed adapters and merged together sequences requiring a 15 base
24 pair overlap (allowing up to one mismatch), using a modified version of Seqprep
25 (<https://github.com/jstjohn/SeqPrep>), which takes the highest quality base in the merged regions.
26 We mapped the resulting single-ended sequences were mapped to the GRCh37 human reference
27 (from the 1000 Genomes project) using the *samse* command of the Burrows-Wheeler Aligner tool
28 (*BWA*) (version 0.6.1) (51). We trimmed two nucleotides from the end of each sequence, and then
29 randomly selected a single sequence at each site covered by at least one sequence in each
30 individual to represent their genotype at that position (“pseudo-haploid” genotyping). For each
31 sample we generated “pseudo-haploid” calls at the 1.24 million target sites, selecting sequences
32 that have a minimum mapping quality of $MAPQ \geq 10$, restricting to nucleotides with a minimum
33 base quality of 20, and trimming 2 base pairs from each end of the reads.

34

35 For Darra-i-Kur, we prepared a single-stranded DNA library (L5082) at the Max-Planck-Institute
36 for Evolutionary Anthropology (MPI-EVA) in Leipzig, Germany, as part of a previous project (52).
37 The previous study only analyzed mitochondrial DNA, and for the current study, the library was
38 enriched for molecules overlapping target the same panel of 1.24 million nuclear targets using two
39 rounds of hybridization capture (5, 6, 8). We sequenced the enriched libraries on 2 lanes of an
40 Illumina HiSeq2500 platform in a double index configuration (2x76 cycles) (53), and we called
41 sites using *FreeIbis* (54). We merged overlapping paired-end and trimmed using *leeHom* (55). We
42 used *BWA* to align the captured data to the human reference genome (GRCh37 from the 1000
43 Genomes project) (51). Only sequences showing a perfect match to the expected index combination
44 were retained for downstream analyses.

45
46 We assessed evidence for ancient DNA authenticity by measuring the rate of damage in the first
47 nucleotide (flagging individuals as potentially contaminated if they had a less than 3% cytosine to
48 thymine substitution rate in the first nucleotide for a UDG-treated library and less than 10%
49 substitution rate for a non-UDG-treated library). We used *contamix* to determine evidence of
50 contamination based on polymorphism in mitochondrial DNA (56), and ANGSD to determine
51 evidence of contamination based on polymorphism on the X chromosome in males (57).

52
53 **Principal component analysis (PCA)** We carried out PCA using the *smartpca* package of
54 *EIGENSOFT* 7.2.1 (13). We used default parameters and added two options (`lsqproject:YES` and
55 `numoutlieriter:0` options) in order to project our ancient samples onto the PCA space. We used two
56 basis sets for the projection: the first based on 1,340 present-day Eurasians genotyped on the
57 Affymetrix Human Origins array, and the second based on a subset of 991 present-day West
58 Eurasians (5, 10, 58). These projections are shown repeatedly in the **Supplementary Materials**,
59 and the whole-Eurasian projection is shown in **Fig. 1**. As part of this analysis, we also computed
60 the F_{ST} between groups using the parameters `inbreed:YES` and `fstonly:YES`.

61
62 **ADMIXTURE clustering analysis** Using PLINK2 (59), we first pruned our dataset using the `--`
63 `geno 0.7` option to ensure that we only performed our analysis on sites that had at least 70% of
64 samples with a called genotype. We then ran ADMIXTURE (14) with 10 replicates, reporting the
65 replicate with the highest likelihood. We show results for $K=6$ in **Fig. 1**, as we found in practice
66 that this provides the most resolution for disambiguating the sources of pre-Chalcolithic ancestry in
67 our newly reported samples.

68

69 ***f*-statistics** We used the *qp3pop* and *qpDstat* packages in ADMIXTOOLS to compute f_3 -statistics
70 and f_4 -statistics. We used the `inbred:YES` parameter to compute f_3 -statistics as a test for admixture
71 with an ancient population as a target, with all published and newly reported ancient genomes as
72 sources. Using the `f4Mode:YES` parameter in *qpDstat*, we also computed two sets of f_4 -symmetry
73 statistics to evaluate if pairs of populations are consistent with forming a clade relative to a
74 comparison population. The first is a statistic where we compare all possible pairs of newly
75 reported ancient groups (*Reported1* and *Reported2*) to a panel of *Test* populations that encompass
76 diverse pre-Chalcolithic and more widespread genetic variation (*Test* is one of
77 *Iran_Ganj_Dareh_Neolithic*, *Karelia_HG*, *Han*, *Onge*, *LBK_EN*, *AfontovaGora3*,
78 *Ukraine_Mesolithic*). Thus, we compute a statistic of the form $f_4(\textit{Reported1}, \textit{Reported2}; \textit{Test},$
79 *Mbuti* African outgroup). The second is a comparison of each newly reported group in turn against
80 all possible pairs of *Test* populations, using statistics of the form $f_4(\textit{Test1}, \textit{Test2}; \textit{Reported}, \textit{Mbuti})$.

81

82 **Formally modeling admixture history** We used the *qpAdm* methodology (5) in the
83 ADMIXTOOLS package to estimate the proportions of ancestry in a *Test* population deriving from
84 a mixture of N ‘reference’ populations by exploiting (but not explicitly modeling) shared genetic
85 drift with a set of ‘Outgroup’ populations. We set the `details:YES` parameter, which reports a
86 normally distributed Z -score for the fit (estimated with a block jackknife).

87

88 **Hierarchical model of the Indian Cline** We used *qpAdm* as described above to obtain estimates
89 for the proportion of Steppe-related, Iranian agriculturalist-related and *AASI*-related ancestries and
90 their relevant covariance matrices for each population on the Indian cline. We then jointly modeled
91 these estimates using a bivariate normal model (since the three proportions sum to 100%) and
92 inferred the mean and covariance of the two parameters across all samples on the Indian cline using
93 maximum likelihood estimation. Then, using this inferred matrix, we tested whether the cline could
94 be modeled by a mixture of two populations, the *ANI* and the *ASI*, in two ways. First, we examined
95 whether the covariance matrix is singular, implying that knowledge of one estimated proportion of
96 ancestry of one of the ancestry components revealed knowledge of the other two, as expected in a
97 two-way mixture. Second, if we were able to establish that this was the case, we examined the
98 difference between the expected and observed ratios of the ancestry proportions of individual
99 populations in this generative model obtained from fitting all the populations simultaneously. This
100 process resulted in a handful of populations deviating from expectation, as discussed in the main
101 text and **Supplementary Materials**.

102 **Supplementary Materials:**

103 Materials and Methods

104 Online Tableau Server for visualizing data.

105 Data S1-S3

106

107 **References and notes**

- 108 1. J. Dabney *et al.*, Complete mitochondrial genome sequence of a Middle Pleistocene cave
109 bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy*
110 *of Sciences of the United States of America* **110**, 15758-15763 (2013).
- 111 2. P. Korlevic *et al.*, Reducing microbial and human contamination in DNA extractions from
112 ancient bones and teeth. *BioTechniques* **59**, 87-93 (2015).
- 113 3. M. Meyer *et al.*, A high-coverage genome sequence from an archaic Denisovan individual.
114 *Science* **338**, 222-226 (2012).
- 115 4. N. Rohland, E. Harney, S. Mallick, S. Nordenfelt, D. Reich, Partial uracil-DNA-glycosylase
116 treatment for screening of ancient DNA. *Philosophical transactions of the Royal Society of*
117 *London. Series B, Biological sciences* **370**, 20130624 (2015).
- 118 5. W. Haak *et al.*, Massive migration from the steppe was a source for Indo-European
119 languages in Europe. *Nature* **522**, 207-211 (2015).
- 120 6. I. Mathieson *et al.*, Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*
121 **528**, 499-503 (2015).
- 122 7. I. Olalde *et al.*, The Beaker phenomenon and the genomic transformation of northwest
123 Europe. *Nature* **555**, 190-196 (2018).
- 124 8. Q. Fu *et al.*, An early modern human from Romania with a recent Neanderthal ancestor.
125 *Nature* **524**, 216-219 (2015).
- 126 9. N. J. Patterson *et al.*, Ancient Admixture in Human History. *Genetics* **192**, 1065-1093
127 (2012).
- 128 10. I. Lazaridis *et al.*, Ancient human genomes suggest three ancestral populations for present-
129 day Europeans. *Nature* **513**, 409-413 (2014).
- 130 11. N. Nakatsuka *et al.*, The promise of discovering population-specific disease-associated
131 genes in South Asia. *Nat. Genet.* **49**, 1403-1407 (2017).
- 132 12. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS genetics*
133 **2**, e190 (2006).
- 134 13. K. J. Galinsky *et al.*, Fast Principal-Component Analysis Reveals Convergent Evolution of
135 ADH1B in Europe and East Asia. *American journal of human genetics* **98**, 456-472 (2016).
- 136 14. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in
137 unrelated individuals. *Genome research* **19**, 1655-1664 (2009).
- 138 15. I. Olalde *et al.*, Derived immune and ancestral pigmentation alleles in a 7,000-year-old
139 Mesolithic European. *Nature* **507**, 225-228 (2014).
- 140 16. I. Mathieson *et al.*, The genomic history of southeastern Europe. *Nature* **555**, 197-203
141 (2018).
- 142 17. I. Lazaridis *et al.*, Genomic insights into the origin of farming in the ancient Near East.
143 *Nature* **536**, 419-424 (2016).
- 144 18. F. Broushaki *et al.*, Early Neolithic genomes from the eastern Fertile Crescent. *Science*,
145 (2016).
- 146 19. S. Mallick *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse
147 populations. *Nature* **538**, 201-206 (2016).
- 148 20. D. Q. Fuller, in *Examining the Farming/Language Dispersal Hypothesis*. (McDonald
149 Institute for Archaeological Research, 2003), pp. 191-213.
- 150 21. D. Q. Fuller, in *The evolution and history of human populations in South Asia*, M. D.
151 Petraglia, B. Allchin, Eds. (Springer, Dordrecht, The Netherlands, 2007), pp. 393-443.
- 152 22. M. Gallego-Llorente *et al.*, The genetics of an early Neolithic pastoralist from the Zagros,
153 Iran. *Sci Rep* **6**, 31326 (2016).

- 154 23. G. Barker, Goucher, C., in *The Cambridge World History: Volume 2: A World with*
155 *Agriculture, 12,000 BCE–500 CE*, G. Barker, Goucher, C., Ed. (Cambridge University
156 Press, Cambridge, 2015), pp. 1-25.
- 157 24. A. J. Ammerman, L. L. Cavalli-Sforza, *The neolithic transition and the genetics of*
158 *populations in Europe*. (Princeton University Press, Princeton, N.J., 1984), pp. xv, 176 p.
- 159 25. C. J. Stevens *et al.*, Between China and South Asia: A Middle Asian corridor of crop
160 dispersal and agricultural innovation in the Bronze Age. *The Holocene* **26**, 1541-1555
161 (2016).
- 162 26. L. Sverchkov, *Tokhary, drevnie indoevropeytsy v tsentral'noy Azii*. 2012, Ed., (SMI-ASIA,
163 Tashkent).
- 164 27. G. L. Possehl, The Middle Asian Interaction Sphere: Trade and contact in the 3rd
165 millennium BC. *Expedition* **49**, 40-42 (2004).
- 166 28. P. Moorjani *et al.*, A genetic method for dating ancient genomes provides a direct estimate
167 of human generation interval in the last 45,000 years. *Proceedings of the National Academy*
168 *of Sciences of the United States of America* **113**, 5652-5657 (2016).
- 169 29. M. Raghavan *et al.*, Upper Palaeolithic Siberian genome reveals dual ancestry of Native
170 Americans. *Nature* **505**, 87-91 (2014).
- 171 30. Q. Fu *et al.*, The genetic history of Ice Age Europe. *Nature advance online publication*,
172 (2016).
- 173 31. M. E. Allentoft *et al.*, Population genomics of Bronze Age Eurasia. *Nature* **522**, 167-+
174 (2015).
- 175 32. M. D. Frachetti, Multiregional Emergence of Mobile Pastoralism and Nonuniform
176 Institutional Complexity across Eurasia. *Current Anthropology* **53**, 2-38 (2012).
- 177 33. M. D. Frachetti, C. E. Smith, C. M. Traub, T. Williams, Nomadic ecology shaped the
178 highland geography of Asia's Silk Roads. *Nature* **543**, 193-198 (2017).
- 179 34. M. Unterlander *et al.*, Ancestry and demography and descendants of Iron Age nomads of
180 the Eurasian Steppe. *Nature communications* **8**, 14615 (2017).
- 181 35. D. Reich, K. Thangaraj, N. Patterson, A. L. Price, L. Singh, Reconstructing Indian
182 population history. *Nature* **461**, 489-494 (2009).
- 183 36. P. Moorjani *et al.*, Genetic evidence for recent population mixture in India. *American*
184 *journal of human genetics* **93**, 422-438 (2013).
- 185 37. P. A. Underhill *et al.*, The phylogenetic and geographic structure of Y-chromosome
186 haplogroup R1a. *European journal of human genetics : EJHG* **23**, 124-131 (2015).
- 187 38. M. Silva *et al.*, A genetic chronology for the Indian Subcontinent points to heavily sex-
188 biased dispersals. *BMC evolutionary biology* **17**, 88 (2017).
- 189 39. G. Hellenthal *et al.*, The Kalash Genetic Isolate? The Evidence for Recent Admixture.
190 *American journal of human genetics* **98**, 396-397 (2016).
- 191 40. C. A. Murphy, Fuller, D. Q., in *A Companion to South Asia in the Past*, G. R. G. R. Schug,
192 Walimbe, S.R., Ed. (2016).
- 193 41. L. Giosan *et al.*, Fluvial landscapes of the Harappan civilization. *Proceedings of the*
194 *National Academy of Sciences of the United States of America* **109**, E1688-1694 (2012).
- 195 42. I. Mahadevan, Dravidian Proof of the Indus Script via the Rig Veda: A Case Study. *Bulletin*
196 *of the Indus Research Centre* **4**, (2014).
- 197 43. A. Parpola, *The Roots of Hinduism : The Early Aryans and the Indus Civilization*. (Oxford
198 University Press, New York, 2015), pp. xvi, 363 pages.
- 199 44. V. Kolipakam *et al.*, A Bayesian phylogenetic study of the Dravidian language family.
200 *Royal Society Open Science* **5**, (2018).
- 201 45. C. Renfrew, *Archaeology and language : the puzzle of Indo-European origins*. (Cambridge
202 University Press, New York, 1988), pp. xiv, 346 p., 348 p. of plates.

- 203 46. D. W. Anthony, *The horse, the wheel, and language : how bronze-age riders from the*
204 *Eurasian steppes shaped the modern world*. (Princeton University Press, Princeton, NJ,
205 2007), pp. xii, 553 p.
- 206 47. K. A. Sirak *et al.*, A minimally-invasive method for sampling human petrous bones from
207 the cranial base for ancient DNA analysis. *BioTechniques* **62**, 283-289 (2017).
- 208 48. M. M. DeAngelis, D. G. Wang, T. L. Hawkins, Solid-phase reversible immobilization for
209 the isolation of PCR products. *Nucleic acids research* **23**, 4742-4743 (1995).
- 210 49. N. Rohland, D. Reich, Cost-effective, high-throughput DNA sequencing libraries for
211 multiplexed target capture. *Genome research*, (2012).
- 212 50. T. Maricic, M. Whitten, S. Paabo, Multiplexed DNA sequence capture of mitochondrial
213 genomes using PCR products. *PloS one* **5**, e14004 (2010).
- 214 51. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform.
215 *Bioinformatics* **26**, 589-595 (2010).
- 216 52. K. Douka *et al.*, Direct radiocarbon dating and DNA analysis of the Darra-i-Kur
217 (Afghanistan) human temporal bone. *Journal of human evolution* **107**, 86-93 (2017).
- 218 53. M. Kircher, S. Sawyer, M. Meyer, Double indexing overcomes inaccuracies in multiplex
219 sequencing on the Illumina platform. *Nucleic acids research* **40**, e3 (2012).
- 220 54. G. Renaud, M. Kircher, U. Stenzel, J. Kelso, freeIbis: an efficient basecaller with calibrated
221 quality scores for Illumina sequencers. *Bioinformatics* **29**, 1208-1209 (2013).
- 222 55. G. Renaud, U. Stenzel, J. Kelso, leeHom: adaptor trimming and merging for Illumina
223 sequencing reads. *Nucleic acids research* **42**, e141 (2014).
- 224 56. Q. Fu *et al.*, A revised timescale for human evolution based on ancient mitochondrial
225 genomes. *Current biology : CB* **23**, 553-559 (2013).
- 226 57. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation
227 Sequencing Data. *BMC bioinformatics* **15**, 356 (2014).
- 228 58. N. Patterson *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).
- 229 59. C. C. Chang *et al.*, Second-generation PLINK: rising to the challenge of larger and richer
230 datasets. *Gigascience* **4**, 7 (2015).
- 231
- 232

233 **Acknowledgments:** We are grateful to Richard Meadow and Ajita Patel for critical comments. We
234 are grateful to the Minusinsk Regional Museum of N. M. Martyanov for sharing some of the
235 skeletal samples analyzed in this study. We are grateful to Orazak Ismagulov and Ainagul
236 Ismagulova for facilitating access to some of the Kazakh material. **Funding:** N.P. carried out this
237 work while a fellow at the Radcliffe Institute for Advanced Study at Harvard University. P.M. was
238 supported by a Burroughs Wellcome Fund CASI award. N.N. is supported by an NIGMS
239 (GM007753) fellowship. T.C. and A.D. were supported by the Russian Science Foundation (project
240 no. 14-50-00036). D.P., S.S. and D.L. were supported by European Research Council ERC-2011-
241 AdG 295733 grant (Langelin). M.R. acknowledges support from RFBR grant № 18-09-00779.
242 Radiocarbon work supported by the NSF Archaeometry program BCS-1460369 to D.J.K. and
243 B.J.C. and by the NFS Archaeology program BCS-1725067 to D.J.K. and T.Ha. K.T. was
244 supported by the Council of Scientific and Industrial Research (CSIR), Government of India, New
245 Delhi. N.B., A.N., and Z.M. were supported by the Max Planck Society. D.R. was supported by the
246 U.S. National Science Foundation HOMINID grant BCS-1032255, the U.S. National Institutes of
247 Health grant GM100233, by an Allen Discovery Center grant, and is an investigator of the Howard
248 Hughes Medical Institute. **Competing interests:** The authors declare no competing interests. **Data
249 and materials availability:** All sequencing data are available from the European Nucleotide
250 Archive, accession number XXXXXXXX [to be made available on publication]. Genotype data
251 obtained by random sampling of sequences at approximately 1.24 million analyzed positions or at
252 approximately 600,000 positions (when merged with genotyping data from diverse present-day
253 individuals) are available to researchers who write David Reich (reich@genetics.med.harvard.edu)
254 a signed letter containing the following text: “For the data that is indicated as “signed letter only”
255 (a) I will not distribute the data outside my collaboration; (b) I will not post the data publicly; (c) I
256 will make no attempt to connect the genetic data to personal identifiers for the samples; (d) I will
257 use the data only for studies of population history; (e) I will not use the data for any selection
258 studies; (f) I will not use the data for medical or disease-related analyses; (g) I will not use the data
259 for commercial purposes.”
260