**QUEEN'S
UNIVERSITY
BELFAST**

The formation of human populations in South and Central Asia

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

Published in:
Science

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
[Link to publication record in Queen's University Belfast Research Portal](#)

Publisher rights

© 2020 American Association for the Advancement of Science.

This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

General rights

Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

The Genomic Formation of South and Central Asia

Vagheesh Narasimhan³, Nick Patterson^{1,2,*}, Swapan Mallick^{1,3,4}, Nadin Rohland^{1,3}, Priya Moorjani⁴⁰, Nathan Nakatsuka³, Luca M. Olivieri¹⁶, James Mallory²⁴, Vyacheslav Moiseyev¹⁵, Janet Monge³⁶, Niraj Rai^{42,43}, Alexander Kim³, Iñigo Olalde³, Iosif Lazaridis³, Rebecca Bernardos³, Nicole Adamski^{3,4}, Nasreen Broomandkhoshbacht^{3,4}, Francesca Candilio⁶, Olivia Cheronet^{6,21}, Matthew Ferry^{3,4}, Daniel Fernandes⁶, Beatriz Gamarra⁶, Daniel Gaudio⁶, Denise Keating⁶, Ann-Marie Lawson^{3,4}, Mark Lipson³, Jonas Oppenheimer^{3,4}, Megan Michel^{3,4}, Mario Novak⁶, Kendra Sirak^{6,7}, Viviane Slon³², Kristin Stewardson^{3,4}, Zhao Zhang³, Gaziz Akhatov²⁷, David Anthony³³, Anatoly N. Bagashev⁴⁹, Baurzhan Baitenaev²⁷, Gian Luca Bonora⁴⁵, Tatiana Chikisheva¹², Alfredo Coppa³⁹, Anatoly Derevianko¹², Katerina Douka^{25,34}, Nadezhda Dubova¹⁰, Andrey Epimakhov¹⁸, Antonina Ermolaev²⁷, Suzanne Freilich²¹, Dorian Fuller⁵⁰, Alexander Goryachev²⁷, Andrey Gromov¹⁵, Bryan Hanks¹⁹, Eadaoin Harney^{3,4,5}, Margaret Judd¹⁹, Erlan Kazizov²⁷, Egor Kitov^{27,30}, Aleksander Khokhlov⁴¹, Donata Luiselli⁴⁴, Farhad Maksudov²⁸, Chris Meiklejohn²², Deborah C. Merrett²³, Roberto Micheli¹⁶, Zahir Muhammed²⁶, Samridin Mustafokulov^{28,29}, Ayushi Nanak²⁵, David Pettner⁴⁴, Dmitry Razhev²⁰, Stefania Sarno⁴⁴, Kulyan Sikhymbaeva³⁰, Sergey M. Slepchenko³⁷, Nadezhda Stepanova¹², Svetlana Svyatko^{14,24}, Sergey Vasilyev³⁸, Massimo Vidale^{16,17}, Dima Voyakin^{27,48}, Alisa Zubova^{12,15}, Matthias Meyer³², Carles Lalueza-Fox³⁵, Nicole Boivin^{25,+}, Kumarasamy Thangaraj^{42,+}, Douglas Kennett^{8,+}, Michael Frachetti^{7,8,+}, Ron Pinhasi^{6,21,+}, David Reich^{1,3,4,5,+}

* Contributed equally

+ Co-directed this work

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

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

¹ Broad Institute of Harvard and MIT, Cambridge, MA, 02142, USA

² Radcliffe Institute for Advanced Study, Harvard University, Cambridge MA 02138

³ Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA

⁴ Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, 02115, USA

⁵ Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138, USA

⁶ Earth Institute, University College Dublin, Dublin 4, Ireland

⁷ Department of Anthropology, Washington University in St. Louis, Saint Louis MO, 63112

⁸ Spatial Analysis, Interpretation, and Exploration Laboratory, Washington University in St. Louis, Saint Louis MO, 63112

¹⁰ Ethnic Ecology of the Institute of Ethnology and Anthropology Russian Academy of Sciences, Moscow, Russia

¹¹ Institute for Anthropological Research, 10000 Zagreb, Croatia.

45 ¹² Institute of Archaeology and Ethnography, Siberian Branch, Russian Academy of Sciences, 630090
46 Novosibirsk, Russia
47 ¹⁴ CHRONO Centre for Climate, the Environment, and Chronology, Queen's University of Belfast,
48 Belfast BT7 1NN, Northern Ireland, UK
49 ¹⁵ Peter the Great Museum of Anthropology and Ethnography (Kunstkamera), Russian Academy of
50 Science, 199034 St. Petersburg, Russia
51 ¹⁶ ISMEO Italian Archaeological Mission in Pakistan, **missing.city missing.postal.code**, Pakistan
52 ¹⁷ Department of Cultural Heritage: Archaeology and History of Art, Cinema and Music, University of
53 Padua, **missing.postal.code** Padua, Italy
54 ¹⁸ Institute of History and Archaeology (Ural Branch RAS) and South Ural State University, 454080
55 Chelyabinsk, Russia
56 ¹⁹ University of Pittsburgh, Department of Anthropology, Pittsburgh, PA, 15260, USA
57 ²⁰ Institute of Problems of Development of the North of the Siberian Branch of the Russian Academy
58 of Sciences, 625026 Tyumen, Russia
59 ²¹ Department of Anthropology, University of Vienna, 1090 Vienna, Austria
60 ²² Department of Anthropology, University of Winnipeg, Winnipeg, MB, R3B 2E9, Canada
61 ²³ Department of Archaeology, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada
62 ²⁴ School of Natural and Built Environment, Queen's University Belfast, Belfast BT7 1NN, Northern
63 Ireland, UK
64 ²⁵ Max Planck Institute for the Science of Human History, 07745 Jena, Germany
65 ²⁶ Department of Archaeology, Hazara University, Mansehra, 21300, Pakistan
66 ²⁷ Institute of Archaeology A.Kh. Margulan, Almaty 050010, Kazakhstan
67 ²⁸ Institute of Archaeology, Uzbek Academy of Sciences, Samarkand 140151, Uzbekistan
68 ²⁹ Afrosiab Museum, Samarkand 140151, Uzbekistan
69 ³⁰ Center of Physical Anthropology, Institute of Ethnology and Anthropology RAS Lenina, Moscow
70 119334, Russia
71 ³¹ Central State Museum Republic of Kazakhstan, Samal-1 microdistrict, Almaty 050010, Kazakhstan
72 ³² Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany
73 ³³ Anthropology Department, Hartwick College, Oneonta, New York 13820, USA
74 ³⁴ Oxford Radiocarbon Accelerator Unit, Research Laboratory for Archaeology and the History of
75 Art, University of Oxford, Oxford OX1 3QY, UK
76 ³⁵ Institute of Evolutionary Biology, CSIC-Universitat Pompeu Fabra, Barcelona 08003, Spain
77 ³⁶ University of Pennsylvania Museum of Archaeology and Anthropology, Philadelphia, PA 19104,
78 USA
79 ³⁷ Institute for Problems of the Development of the North, Siberian Branch of the Russian Academy of
80 Sciences, 625026 Tyumen, Russia
81 ³⁸ Institute of Ethnology and Anthropology, Russian Academy of Sciences, Moscow 119991, Russia
82 ³⁹ Dipartimento di Biologia Ambientale, Sapienza Università di Roma, 00185 Roma, Italy
83 ⁴⁰ Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley 94720,
84 USA
85 ⁴¹ Samara State University of Social Sciences and Education, Samara 443099, Russia
86 ⁴² CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India
87 ⁴³ Birbal Sahni Institute of Palaeosciences, Lucknow, India.
88 ⁴⁴ Laboratory of Molecular Anthropology, Department of Biological, Geological and Environmental
89 Sciences, University of Bologna, 40126 Bologna, Italy
90 ⁴⁵ Archaeology of Asia Department, International Association of Mediterranean and Oriental Studies
91 (ISMEO), RM00186 Roma, Italy
92 ⁴⁶ Department of Anthropology and Institutes for Energy and the Environment, Pennsylvania State
93 University, University Park, PA 16802, USA
94 ⁴⁷ Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA
95 02115, USA
96 ⁴⁸ Archaeological Expertise LLP, 050060 Almaty Kazakhstan
97 ⁴⁹ Institute for Problems of the Development of the North, Siberian Branch of the Russian Academy of
98 Sciences, Tyumen 625026, Russia
99 ⁵⁰ Institute of Archaeology, University College London, London WC1H 0PY, UK

100 **Abstract**

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

124

125 **Ancient DNA Data**

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

134 sites); Copper and Bronze Age pastoralists from the steppe East of the Ural mountains
 135 including the first data from Kazakhstan (3200-1000 BCE from 35 sites); and Iron Age
 136 settlements in the Swat Valley of Pakistan (1200-0 BCE from 7 sites) (Figure 1;
 137 Supplementary Text; Extended Data Table 1; Online Table 1). To generate these data, we
 138 prepared samples in dedicated clean rooms, extracted DNA, built it into libraries for Illumina
 139 sequencing,⁽¹⁾ and screened it using previously described procedures (Methods).^(2, 3) We
 140 enriched samples for DNA fragments overlapping about 1.24 million SNPs, sequenced the
 141 products, and performed quality control as described previously (Data S1, Data S2).^{(3, 4),(5)}
 142 We also generated 178 new direct radiocarbon dates (Data S3). After grouping based on
 143 archaeological and chronological information and merging with previously reported data, our
 144 dataset included 731 ancient individuals which we co-analyzed with genome-wide data from
 145 3,066 present-day individuals assessed at about 0.6 million single nucleotide polymorphisms
 146 (SNPs), 1,789 of which were from 246 ethnographically-distinct groups in South Asia (Data
 147 S4; Supplementary Text). We restricted most analyses to ancient samples having at least
 148 15,000 SNPs when merged with the present-day individuals.

149

150 Analysis Strategy

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

168

169 **Formation of the Ancestry Gradient in the South**

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

183

184 Moving to the Middle Bronze Age in Turan, we examined 69 ancient individuals from four
185 urban sites of the BMAC and post-BMAC spread over 2500-1500 BCE. The great majority
186 fall in a cluster that is similar to the preceding groups in having a large Iranian farmer related
187 ancestry component (~68%) with smaller components of Anatolian farmer related ancestry
188 (~18%) and west Siberian hunter-gatherer related ancestry (~14%), suggesting that the
189 BMAC population coalesced from preceding pre-urban populations and ruling out the
190 hypothesis(13) that the BMAC were a link in the chain-of-transmission that brought the
191 steppe pastoralist ancestry ubiquitous in South Asia today (Supplementary Text). The data
192 instead suggest gene flow in the opposite direction: the main BMAC cluster has a proportion
193 of ancestry from an indigenous Asian source, which we can model as derived from ancestors
194 of the SPGT, some of whom may have expanded North in the preceding millennium.

195

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

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

224 **Formation of the Ancestry Gradient in the North**

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

236 *Sintashta* and *Andronovo_NW* cultures—all superimposed on a substrate of early Bronze Age
 237 Steppe Pastoralist-related ancestry (*Steppe_EMBA*). This reflects a previously reported
 238 phenomenon in which following westward movement of *Steppe_EMBA* groups into central
 239 Europe and admixture with local European farmers, there was eastward genetic flow beyond
 240 the Urals, with the European (Anatolian) farmer-related ancestry being diluted by admixture
 241 with previously established *Steppe_EMBA* groups (Figure 2).(19, 20) Our new data also
 242 extend this gradient to the far southeast – to present-day Kazakhstan – and show that the
 243 *Andronovo_SE* grouping there had distinctively higher *West_Siberian_Neolithic* derived
 244 ancestry and less Anatolian Neolithic-related admixture than previously reported groups from
 245 the Northwest (“*Andronovo_NW*”).(19) This is important since as shown below, this
 246 signature makes *Andronovo_SE* a plausible source for present-day groups in South Asia who
 247 have too little Anatolian Neolithic related ancestry to be fit by any previously reported Late
 248 Bronze Age steppe groups (Supplementary Text).(11) The significant ancestry difference
 249 between *Andronovo_NW* and *Andronovo_SE* accords with the archaeological evidence
 250 for a difference in economic subsistence between these two groups with extensive reliance
 251 agriculture in the southeast,(21) which is intriguing as it could be related to the differences in
 252 mobility patterns and expansion of *Andronovo_SE* further South that we document here.
 253
 254 Our large sample sizes also allow us to document ancestry heterogeneity which is informative
 255 about the history. Our analysis of 51 newly reported samples from the Kamennyi Ambar V
 256 cemetery from the Sintashta culture, the largest ancient DNA study of a single site to date in
 257 the ancient DNA literature, reveals in addition to the main cluster of xx individuals, three
 258 groups of outliers with direct radiocarbon dates contemporaneous with the other samples and
 259 elevated proportions of *Steppe_EMBA* or *West_Siberian_Neolithic* related ancestry,
 260 providing evidence that these fortified steppe sites harbored people of diverse ancestries
 261 living side-by-side (Supplementary Text). Second, samples from three sites from the southern
 262 and eastern end of the Steppe, dated to xxxx-xxxx BCE and contemporaneous with the late
 263 BMAC, show evidence of significant admixture from Iranian farmer related populations
 264 demonstrating northward gene flow from Turan into the steppe in this period just as there was
 265 southward gene flow through Turan and into South Asia. Third, from xxxx-xxxx BCE we
 266 observe multiple sites that derive up to ~25% of their ancestry from a source related to
 267 present day East Asians and the rest from populations from *Steppe_EMBA* cluster. This type
 268 of ancestry became widespread in the region by the early Iron Age (Scytho-Sarmatian
 269 period)(22) and these samples provide a minimum date for when it arrived.

270

271 **Iranian Farmer Related Ancestry at Both Extremes of the Indian Cline**

272 We co-analyzed our newly reported ancient DNA data with data from diverse present-day
273 Indians(23) to gain insight into the deep ancestry sources of the “Indian Cline,”(24-26) the
274 primary driver of genetic variation across South Asia, represented here by 140 groups that
275 fall on this cline in a Principal Component Analysis (Supplementary Text). Previous work
276 has shown that the Indian Cline can be well modeled as having arisen from a mixture of two
277 statistically reconstructed ancestral populations (the ANI and the ASI) largely between 2000-
278 0 BCE.(24, 25) Ancient DNA analysis has furthermore shown that Indian Cline groups
279 descend more deeply from at least three ancestral populations: early Iranian farmers,
280 *Steppe_EMBA*, and an unsampled South Asian hunter-gatherer group related to indigenous
281 Andamanese (*Onge*)(11) (we refer to this deep Asian ancestry source as the “Ancient
282 Ancestral South Indians” (AASI), to distinguish it from the more Asian extreme of Indian
283 Cline, referred to in this study as the ASI). To shed light on the mixture events that
284 transformed this minimum of three ancestral populations into two (the ANI and ASI), we
285 used *qpAdm* to evaluate if each of the Indian Cline groups in turn, *Onge*, and diverse pairs of
286 ancient West Eurasian groups from the Copper Age onward, were consistent with descending
287 from just three ancestral populations relative to distantly related outgroups (Supplementary
288 Text). We obtain fits models for the great majority of Indian Cline groups when one source
289 was *Onge*, one was *Steppe_EMBA*, and third was a Turan Copper or Bronze Age group
290 (Figure 3, [Supplementary Text](#)). (We also obtained fits for a variety of northern populations
291 when one source was *SPGT*, a more proximal set of models we return to below.)

292

293 We found that the per-group *qpAdm* estimates for these three source populations are
294 statistically noisy, and we therefore developed new methodology that allows us to jointly fit
295 the data from all Indian Cline groups within a single hierarchical model. The analysis both
296 confirms that almost all groups on the Indian Cline can be jointly modeled as a mixture of
297 two populations (Supplementary Text), and produces estimates of the functional relationship
298 between the ancestry components: (*Steppe_EMBA*) = 62% – 82%(*Onge*-related), with
299 standard errors of $\pm 3\%$ (Supplementary Text). Setting *Onge*-related ancestry to 1 to
300 determine what would be expected if the ASI had no West Eurasian related ancestry, we infer
301 negative *Steppe_EMBA* ancestry, a nonsensical result. Setting *Steppe_EMBA*=0%, the
302 smallest proportion it can possible be, we find that the ASI must have had no less than
303 25 $\pm$ 3% Iranian farmer-related ancestry. In fact, our estimate in real Indian Cline groups

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

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

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

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

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

347

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

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

370 Taken together, these four lines of evidence provide a compelling case that *Andronovo_SE*
371 dispersed southward through Turan into the northern part of South Asia in the first half of the
372 2nd millennium BCE.

373

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

389

390 The history of the formation of ASI is less clear. Iranian farmer-related groups plausibly first
391 spread into South Asia from the Iranian plateau some time after 7000 BCE when there is the
392 first archaeologically attested spread of Iranian farming technology to the hills surrounding
393 the Indus Valley, although there is not yet any ancient DNA from the Indus Valley to test
394 this. However, the 2850-750 BCE date of admixture suggests that even if this ancestry came
395 into South Asia around this time, it did not immediately homogenize with the *AASI*, and
396 instead continued to admix with it well after the beginning of the Bronze Age. One plausible
397 scenario is that the *AASI*-Iranian admixture occurred as a result of the spread of agriculture
398 and pastoralism from the Indus Valley into peninsular South Asia 3000-2000 BCE(17, 20,
399 34-37) However, an alternative is that the southward movement of *Andronovo_SE* groups
400 precipitated the formation of both the ANI and the ASI: the ANI through admixture with
401 peoples with ancestry like *Gonur2* and *SPGT* who were plausibly similar in ancestry to the
402 IVC, and the ASI as unfavorable circumstances pushed these people into peninsular India
403 after ~1700 BCE to mix with the *AASI*.

Conclusions

These results shed light on the spread of Indo-European languages in two ways. First, our documentation of a large-scale southward spread of *Andronovo_SE* groups in the 2nd millennium provides a prime candidate for the spread of late proto-Indo-European languages southward into India from the Steppe where they were likely spread by Yamnaya pastoralists at the beginning of the 3rd millennium.(13) It has been argued—controversially—that material culture remains from *Andronovo_SE* groups have connections to early Vedic culture in India,(13) and our results of an ancestry link to later cultures to the South adds weight to this theory. Second, our genetic analysis also provides an entirely new line of evidence for a connection of steppe ancestry to Indo-European culture. When we used *qpAdm* to test if a mixture of ANI and ASI is a fit to the data for all 140 Indian Cline groups, we found five with poor fits and a significantly higher ratio of *Andronovo_SE* to *Gonur2*-like ancestry than other groups in the Indian Cline. These were all groups of priestly status (Brahmins or Bhumihars), with the strongest signal seen in *Brahmin_Tiwari* ($P < 2 \times 10^{-7}$) (Supplementary Text; Online Table 2). Thus, at the time the ANI began mixing with the ASI, there was likely a substructured meta-population including subgroups with relatively higher proportions of steppe ancestry that may have had a disproportional role in spreading early Vedic culture, and this stratification has been preserved even to this day by the strong endogamy rules in India. A working model for the formation of Indian Cline groups is shown in Figure 4.

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

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

446

447 **Data availability**
448 All sequencing data are available from the European Nucleotide Archive, accession number
449 XXXXXXXXX [to be made available on publication]. Genotype data obtained by random
450 sampling of sequences at approximately 1.24 million analyzed positions are available from
451 the Reich Lab website at [to be made available on publication].
452

453 **Acknowledgments**
454 We are grateful to the Minusinsk Regional Museum n.a. N. M. Martyanov for sharing some
455 of the skeletal samples analyzed in this study. N.P. carried out this work while a fellow at the
456 Radcliffe Institute for Advanced Study at Harvard University. T.C. and A.D. were supported
457 by the Russian Science Foundation (project no. 14-50-00036). D.R. was supported by the
458 U.S. National Science Foundation HOMINID grant BCS-1032255, the U.S. National
459 Institutes of Health grant GM100233, by an Allen Discovery Center grant, and is an
460 investigator of the Howard Hughes Medical Institute.

461
462

470

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

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

A (locations of analyzed groups)

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

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

C (Distal model of the Cline)

E (admixture graph relating data from diverse groups)

D (Proximal model of the Cline)

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

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

1 Online Methods

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

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

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

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

Commented [DR2]: Add sections here on laboratory methods

35 from a mixture of N ‘reference’ populations by exploiting (but not explicitly modeling) shared
36 genetic drift with a set of ‘Outgroup’ populations. We set the details: YES parameter, which reports
37 a normally distributed Z-score estimated with a block jackknife.
38

39 **Heirarchical model of the Indian cline:** We estimated ancestry proportions using *qpAdm* as
40 described above to obtain mixture proportions for the proportion of Steppe-related Iranian-related
41 and AASI-related ancestries and their relevant covariance matrices for each population on the
42 Indian cline. We then jointly modelled these estimates using a bi-variate normal model (since the
43 sum of the 3 proportions sum to 1) and estimate the mean and covariance of these proportions
44 across all samples on the Indian cline using maximum likelihood estimation. Then, using this
45 estimated matrix, we tested that the cline could be modelled by a mixture of two populations, the
46 ANI and the ASI in two ways. First we examined that the covariance matrix estimated is singular,
47 implying that knowledge of one estimated proportion of ancestry of one of the ancestry
48 components revealed knowledge of the other two, as expected in a two way mix. Second, after
49 showing that the first was true, we examined the difference between the expected and observed
50 ratios of the ancestry proportions of individual populations on this generative model that was
51 obtained from fitting all the populations simultaneously. This process resulted in several
52 populations deviating from expectation and we discuss their relevance in the main text and
53 Supplementary Text.
54

55 **Abbreviations**

56 We have used the following abbreviations in population labels: N, Neolithic; EN, Early Neolithic;
57 C, Copper Age; EMBA, Early to middle Bronze Age; MLBA, Middle to late Bronze Age; IA, Iron
58 Age.
59

60 References

- 61 1. J. Dabney *et al.*, Complete mitochondrial genome sequence of a Middle Pleistocene cave
62 bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy
63 of Sciences of the United States of America* **110**, 15758-15763 (2013).
- 64 2. N. Rohland, E. Harney, S. Mallick, S. Nordenfelt, D. Reich, Partial uracil-DNA-glycosylase
65 treatment for screening of ancient DNA. *Philosophical transactions of the Royal Society of
66 London. Series B, Biological sciences* **370**, 20130624 (2015).
- 67 3. Q. Fu *et al.*, DNA analysis of an early modern human from Tianyuan Cave, China.
68 *Proceedings of the National Academy of Sciences of the United States of America* **110**,
69 2223-2227 (2013).
- 70 4. Q. Fu *et al.*, An early modern human from Romania with a recent Neanderthal ancestor.
71 *Nature*, (2015).
- 72 5. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation
73 Sequencing Data. *BMC bioinformatics* **15**, 356 (2014).
- 74 6. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS genetics*
75 **2**, e190 (2006).
- 76 7. K. J. Galinsky *et al.*, Fast Principal-Component Analysis Reveals Convergent Evolution of
77 ADH1B in Europe and East Asia. *American journal of human genetics* **98**, 456-472 (2016).
- 78 8. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in
79 unrelated individuals. *Genome research* **19**, 1655-1664 (2009).
- 80 9. N. J. Patterson *et al.*, Ancient Admixture in Human History. *Genetics* **192**, 1065-1093
81 (2012).
- 82 10. W. Haak *et al.*, Massive migration from the steppe was a source for Indo-European
83 languages in Europe. *Nature* **522**, 207-211 (2015).
- 84 11. I. Lazaridis *et al.*, Genomic insights into the origin of farming in the ancient Near East.
85 *Nature* **536**, 419-424 (2016).
- 86 12. F. Broushaki *et al.*, Early Neolithic genomes from the eastern Fertile Crescent. *Science*,
87 (2016).
- 88 13. D. W. Anthony, *The horse, the wheel, and language : how bronze-age riders from the
89 Eurasian steppes shaped the modern world.* (Princeton University Press, Princeton, NJ,
90 2007), pp. xii, 553 p.
- 91 14. B. Cerasetti, *WALKING IN THE MURGHAB ALLUVIAL FAN (SOUTHERN
92 TURKMENISTAN): AN INTEGRATED APPROACH BETWEEN OLD AND NEW
93 PROVIDES NEW INTERPRETATIONS ABOUT THE INTERACTION BETWEEN
94 SETTLED AND NOMADIC PEOPLE.* (2014), vol. BAR International Series 2690, pp.
95 105-114.
- 96 15. G. L. Possehl, *The indus civilization : a contemporary perspective.* (AltaMira Press,
97 Walnut Creek, CA, 2002), pp. xi, 276 p.
- 98 16. N. A. Dubova, A. B. Saipov, S. M. Junusbayev, Interaction between Steppe and
99 Agricultural Tribes during the Bronze Age: Mophological Aspects. *International Journal of
100 Anthropology* **31**, 109-125 (2016).
- 101 17. C. J. Stevens *et al.*, Between China and South Asia: A Middle Asian corridor of crop
102 dispersal and agricultural innovation in the Bronze Age. *The Holocene* **26**, 1541-1555
103 (2016).
- 104 18. M. Raghavan *et al.*, Upper Palaeolithic Siberian genome reveals dual ancestry of Native
105 Americans. *Nature* **505**, 87-91 (2014).
- 106 19. M. E. Allentoft *et al.*, Population genomics of Bronze Age Eurasia. *Nature* **522**, 167-+
107 (2015).
- 108 20. I. Mathieson *et al.*, Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*
109 **528**, 499-503 (2015).

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