

<https://www.biorxiv.org/content/10.1101/2024.03.31.587466v3.full.pdf>

1 **Novel 4,400-year-old ancestral component in a tribe speaking a Dravidian language**

2 Jaison Jeevan Sequeira ², Swathy Krishna ¹, George van Driem ^{3*}, Mohammed Shafiul Mustak ^{2*},

3 Ranajit Das ^{1*}

4 ¹ Yenepoya Research Centre, Yenepoya (deemed-to-be) University, Mangalore, Karnāṭaka, India

5 ² Department of Applied Zoology, Mangalore University, Mangalagangothri, Karnāṭaka, India

6 ³ Institut für Sprachwissenschaft, Universität Bern, Länggassstrasse, Bern, Switzerland

7 ***Correspondence:**

8 Ranajit Das

9 das.ranjit@gmail.com

10 Mohammed Shafiul Mustak

11 msmustak@gmail.com

12 George van Driem

13 george.vandriem@unibe.ch

14 **Keywords**

15 Koraga, North Dravidian, Dravidian, Elamo-Dravidian, Indus Valley, Harappan civilisation, ANI,

16 ASI

17 **Abstract**

18 Research has shown that the present-day population on the Indian subcontinent derives its ancestry from

19 at least three components identified with pre-Indo-Iranian agriculturalists once inhabiting the Iranian

20 plateau, pastoralists originating from the Pontic-Caspian steppe and ancient hunter-gatherer related to

21 the Andamanese Islanders. The present-day Indian gene pool represents a gradient of mixtures from

22 these three sources. However, with more sequences of ancient and modern genomes and fine structure

23 analyses, we can expect a more complex picture of ancestry to emerge. In this study, we focus on

24 Dravidian linguistic groups to propose a fourth putative source which may have branched out from the

25 basal Middle Eastern component that gave rise to the Iranian plateau farmer related ancestry. The
26 Elamo-Dravidian theory and the linguistic phylogeny of the Dravidian family tree provide
27 chronological fits for the genetic findings presented here. Our findings show a correlation between the
28 linguistic and genetic lineages in language communities speaking Dravidian languages when they are
29 modelled together. We suggest that this source, which we shall call ‘Proto- Dravidian’ ancestry,
30 emerged around the dawn of the Indus Valley civilisation. This ancestry is distinct from all other sources
31 described so far, and its plausible origin not later than 4,400 years ago on the region between the Iranian
32 plateau and the Indus valley supports a Dravidian heartland before the arrival of Indo-European
33 languages on the Indian subcontinent. Admixture analysis shows that this Proto-Dravidian ancestry is
34 still carried by most modern inhabitants of the Indian subcontinent other than the tribal populations.
35 This momentous finding underscores the importance of population-specific fine structure studies. We
36 also recommend informed sampling strategies for biobanks and to avoid oversimplification of ancestral
37 reconstruction. Achieving this requires interdisciplinary collaboration.

38

39 **Introduction**

40 Numerous studies have highlighted the genetic complexity of modern Indian populations, which include
41 primitive hunter-gatherer tribes, dry and wet land agriculturist communities, pastoralists, warrior clans,
42 trading communities, artisans and priestly castes ¹⁻⁵. In the last two decades, we have seen a transition
43 from SNP chip technology to whole genome sequencing. Currently, with the increased number of
44 ancient and modern genomes, population genetics has been transformed into an interdisciplinary
45 enterprise involving archaeologists, historians, linguists and ethnographers, thereby inviting public
46 attention and debate around complex questions of our past. The genetic source of present-day Indians
47 remains a topic of utmost interest. Numerous studies have used genotypic information to reveal the
48 complex ancestral components in the gene pool of the Indian subcontinent. Reich et al. (2009) ⁶
49 proposed two broad ancestral components, i.e. Ancestral North Indians and Ancestral South Indians,
50 with varying proportions of Andamanese ancestry and steppe pastoralist ancestry. Subsequently,
51 attempts were made to understand the correlation of these ancestral components with language families
52 and geography ³. The most recent work involving 2,700 Indian genomes suggests three putative
53 ancestral sources for modern Indians that have been labelled ‘Iranian plateau farmer related’, ‘Pontic-
54 Caspian steppe pastoralist related’ and ‘Andamanese hunter-gatherer related’ ⁷, but we suggest that this
55 is an oversimplification.

56 Interdisciplinary studies investigating local ancestry using adaptive markers, linguistic affinity, cultural
57 similarities and social affiliation testify to the complexity of genetic structure in present-day
58 populations. A recent study suggests that the U1 macrohaplogroup in the Koraga tribe could be a
59 correlate of Dravidian linguistic lineage ⁸.

60 The Koraga are an indigenous tribal community who reside on the southwestern coast of India,
61 primarily in Dakṣiṇa Kannaḍa and Uḍupi districts of Karnāṭaka and Kāsaragoḍ district of Kerala. They
62 represent one of the most marginalised and impoverished populations in southern Karnāṭaka, and their
63 livelihood mainly depends on activities such as basket weaving, collecting firewood and honey from
64 nearby forests, and working as seasonal daily wage labourers. Although the Koraga language has been
65 influenced for centuries by surrounding Tuḷu speakers, and many Koraga are bilingual in Tuḷu, Bhat ⁹

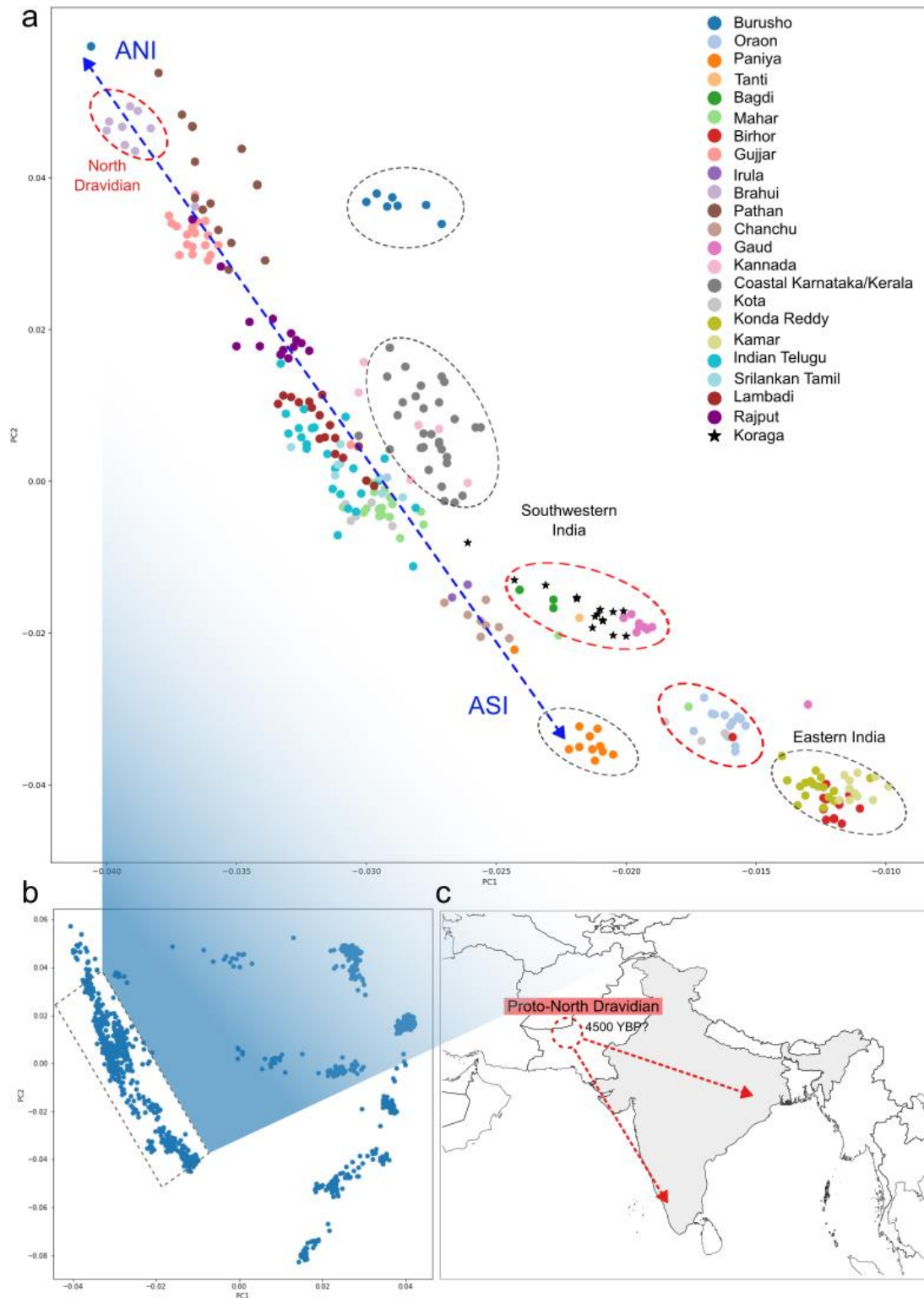
66 (1971) and McAlpin¹⁰ (1981) grouped Koraga together with Kurukh and Malto under the North
67 Dravidian branch. North Dravidian language communities are mainly distributed geographically across
68 the northwest and north of the Indian subcontinent. Krishnamurti¹¹⁻¹³ classified the Brahui language,
69 spoken in Belochistan in the far northwest, as North Dravidian, whereas McAlpin¹⁰ (1981) and Zvelebil
70¹⁴ (1990) placed Brahui under a separate node as a distinct branch of its own within Dravidian or Elamo-
71 Dravidian, intermediate between Elamite and mainstream Dravidian. Zvelebil¹⁴ (1990) likewise
72 proposed to treat Koraga as an independent branch of Dravidian under its own node in the tree. So,
73 whereas these languages are sometimes lumped together under the label Northern Dravidian *sensu lato*,
74 linguists have also seen Brahui, Northern Dravidian *strictu sensu* and Koraga as possibly representing
75 three primary branches of the language family. By contrast, a fourth branch, dubbed ‘mainstream
76 Dravidian’ by McAlpin¹⁰ (1981), encompasses most modern Dravidian languages and consists of a
77 southern subgroup (i.e. Tamil, Malayalam, Iruḷa, Tōḍa, Kota, Koḍagu, Kuṛumba, Baḍaga, Kannāḍa,
78 Tuḷu), a south-central subgroup (i.e. Telugu, Gondi, Koṇḍa, Maṇḍa, Pengo, Kuvi, Kui) and a central
79 subgroup (i.e. Kolami, Naikri, Naiki, Gadaba, Paṛji). The objective of this study was to determine the
80 ancestral origin of the Koraga tribe in this context and therefore the roots of the Dravidian family. In
81 order to achieve this, we utilised single nucleotide polymorphisms (SNP). We genotyped 29 unrelated
82 Koraga individuals using the Infinium Global Screening Array-24 v3.0 (GSA v3.0) BeadChip platform
83 and utilised population genetic tools to trace their ancestry.

84 **Results**

85 **Population structure in the Indian cline**

86 We performed PCA analysis to observe the position of Koraga in the Indian cline. Koraga samples
87 clustered with Gauḍa, Bāgdi and Tānti samples (Figure 1 a). Geographically these castes are found in
88 the northern and eastern zones. This cluster appears to have drifted away from the Ancestral North
89 Indian to Ancestral South Indian cline, parallel to the southwestern and eastern Indian clusters.
90 Interestingly, the clusters with North Dravidian speaking tribes at the Ancestral South Indian end
91 appeared closer to each other, whereas the Brahui samples clustered at the Ancestral North Indian end.
92 The geographical locus of the primary branches of the Dravidian language family lies in the northwest,

93 where the Brahui language community has survived as a remnant population, whereas today most
94 modern Dravidian language speakers live in South India ^{14,15}. It is evident from the PCA plot that these
95 groups have drifted away with time (Figure 1 c).



96

97 Figure 1 Population structure in the Indian cline. (a) PCA plot show a clustering pattern along the
 98 ANI-ASI cline. Red ellipses represent clusters which also include North Dravidian language
 99 communities sensu lato: Brahui, Koraga, Gauda, Tanti, Bagdi, Oraon, Kota, Birhor and Mahar. Kota
 100 and Mahar are found in the ANI-ASI cline as well. Grey ellipses represent clusters that have drifted
 101 away from the ANI-ASI cline, which includes both Dravidian and other groups: Burusho, Kannada
 102 and coastal populations of Karnataka and Kerala, Paniya, Konda Reddy, Kamar and one Birhor

103 sample. (b) Overall PCA plot for GenomeAsia100K including Koraga samples. (c) Putative migration
104 routes for Dravidian tribes.

105 **Admixture profile and ancestry of the Koraga tribe**

106 In order to estimate admixture dates between the geographically separated tribes speaking Dravidian
107 languages, we performed ALDER analysis. This test computes LD decay curves and performs curve-
108 fitting to estimate the admixture dates. Admixture between major North Dravidian tribes *sensu lato*,
109 viz. Koraga, Brahui and Oraon (Kurukh), happened between 5,988 and 2,800 years ago (Table 1). The
110 median falls around 2,370 BC, coinciding with the Mature Harappan period. Notably, the Oraon and
111 Birhor populations are presently found in eastern India, and Brahui is found in Belochistan.
112 Linguistically, both the Koraga and Brahui languages belong to the North Dravidian lineage, whilst
113 North Dravidian language communities are found in eastern India as well. Our findings substantiate the
114 hypothesis of a Dravidian homeland before the arrival of Indo-European languages into the Indian
115 subcontinent.

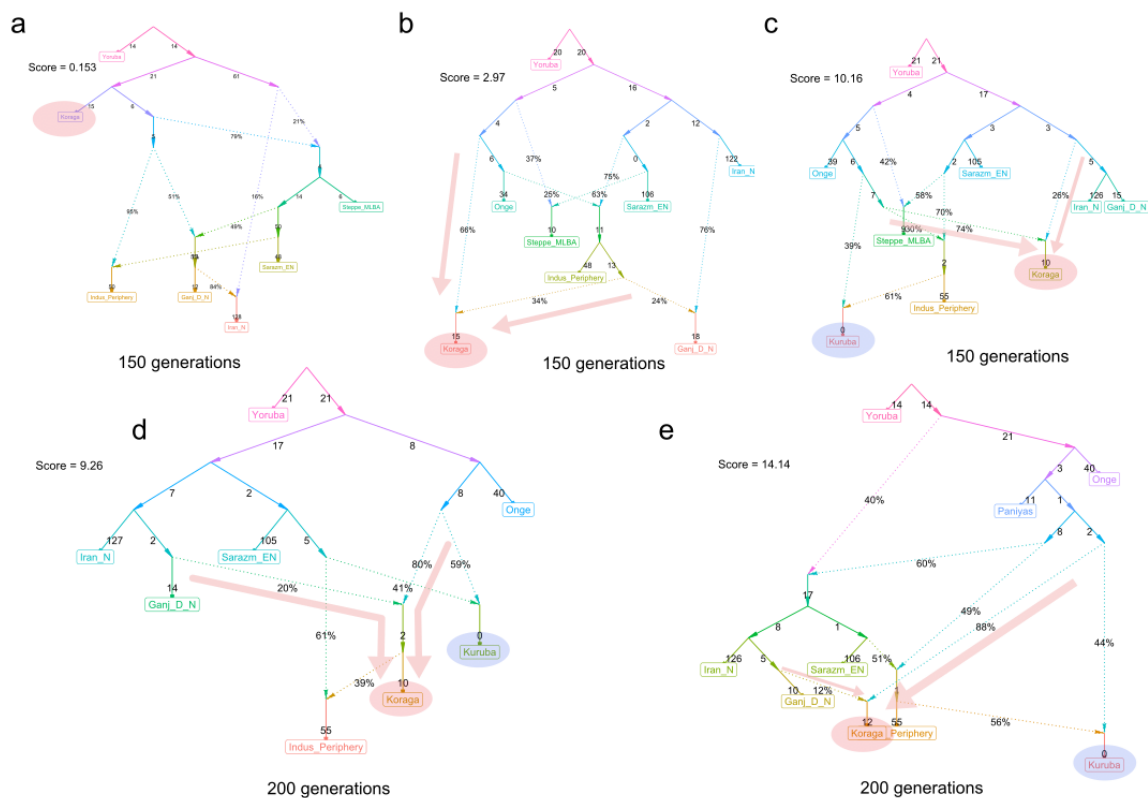
116 Table 1. Date estimates for admixture between Koraga, Brahui and other populations

Koraga vs refpops	Admixture date estimate (generations ago)	Z score and p-value	Admixture date estimate considering 30 years/generation (Years ago)
Brahui;Oraon	146.48 +/- 53.12	z=2.65, p=0.0081	4394.4 +/- 1593.6
Brahui;Birhor	153.55 +/- 70.38	z=2.18, p=0.029	4606.5 +/- 2111.4
Brahui;Tānti	59.15 +/- 18.89	z=2.56, p=0.011	1774.5 +/- 566.7
Brahui;Mahar	78.57 +/- 25.67	z=2.14, p=0.032	2357.1 +/- 770.1
Brahui;Bāgdi	79.74 +/- 32.93	z=2.13, p=0.033	2392.2 +/- 987.9

117

118 Furthermore, to confirm admixture events between the North Dravidian tribes, Admixtools 2 produced
119 the best fit when Koraga, Brahui and Oraon shared ancestry (Supplementary Figure 3). Admixture
120 graphs are more meaningful when models include ancient and modern samples. Therefore, we repeated
121 the same analysis using Dataset 2 containing 4605 modern and ancient samples. In spite of a lesser
122 number of SNPs, the admixture graph produced the model which fit best, at a score of 2.97, when the
123 ancestral Koraga diverged from the Andamanese ancestor carrying 66% of this ancestry (Figure 2 b).
124 The remaining 34% was derived from the ancestor of the 10,000-year-old Neolithic sample from Ganj

125 Dareh, in the Zagros mountains of what today is western Iran. Notably, the Indus Periphery samples
 126 did not share any ancestry with the Koraga tribe (Fig 2 a, b, c, e). “Indus Periphery” is the label given
 127 to ancient DNA samples dating from the 4th–3rd millennium BC from Gonur in what today is
 128 Turkmenistan and Shahr-i-Sokhtah in what today is eastern Iran. To corroborate this key finding, we
 129 included Paniya, a hunter-gatherer tribe, and Kuruba, a pastoralist tribe, in the model. Whilst Paniya
 130 formed a separate branch (Figure 2 e), the pastoralist tribe Kuruba turn out to have received their
 131 ancestral components from an Andamanese-ancestor-related population and the ancestor of the Indus
 132 Periphery samples (Figure 2 c, d, e). Koraga, on the other hand, remains unrelated to the Indus
 133 Periphery, although at a deeper level the Koraga ancestor shared 39% of his ancestry with the Indus
 134 Periphery lineage (Figure 2 d).



135
 136 Figure 2 Admixture graphs show different models for Koraga ancestry. Koraga is represented in the red
 137 ellipse, and Kuruba is represented in the blue ellipse. Arrows show the most recent admixture event for
 138 the Koraga tribe. N=Neolithic, EN=Eneolithic.

139 We used f3 statistics to distinguish between known ancestries and to compare them with the Koraga
 140 ancestry. We used Hân and Yoruba as outgroups. A significant negative z-score indicates that admixture

141 has taken place between the three populations tested. It has been established that Önge and the “Indus
 142 Periphery” represent two of the major ancestral components in populations of the Indian subcontinent
 143 ¹⁶. So, we compared their significant z-scores with those of Koraga population. Notably, the significant
 144 z-scores, for which higher values indicate a lower probability of admixture, are comparable between
 145 Önge, Indus Periphery and Koraga (Supplementary Table S1). However, the genetic distance between
 146 Koraga (~0.015) and other populations is not as great as between Önge (~0.041) and the Indus Periphery
 147 (~0.060) and other ancient and modern populations.

148 The statistical tool qpAdm, implemented in Admix Tools v5.1 ¹⁷, was employed to estimate the ancestry
 149 proportions in the Koraga genomes originating from a mixture of reference populations. The Koraga
 150 were modeled as a combination of three source populations, namely Önge, Indus Periphery and a third
 151 population from Eurasia and Africa (See Materials and Methods). We found that the Koraga can be best
 152 modeled as the genomic admixture of Önge, the Indus periphery and modern-day Iranians (Table 2).
 153 Whilst discernible ancestry fractions from various countries across the Middle East were identified in
 154 the Koraga, the Mid to Late Bronze Age steppe ancestry (MLBA) and African (Yoruba) ancestry
 155 fractions were found to be the least prominent (Table 2). This finding further refutes any possibility of
 156 an African origin for the Koraga. Our findings align with the recent mitochondrial DNA results ⁸,
 157 affirming that the maternal ancestry of the Koraga can be traced back to the Middle East, with
 158 discernible fractions of West Eurasian ancestry in their genomes.

159 **Table 2:** Ancestry fractions (%) in Koraga genomes. The Koraga were modeled as a
 160 combination of three source populations, namely Önge, the Indus periphery and a third
 161 population from Eurasia and Africa.

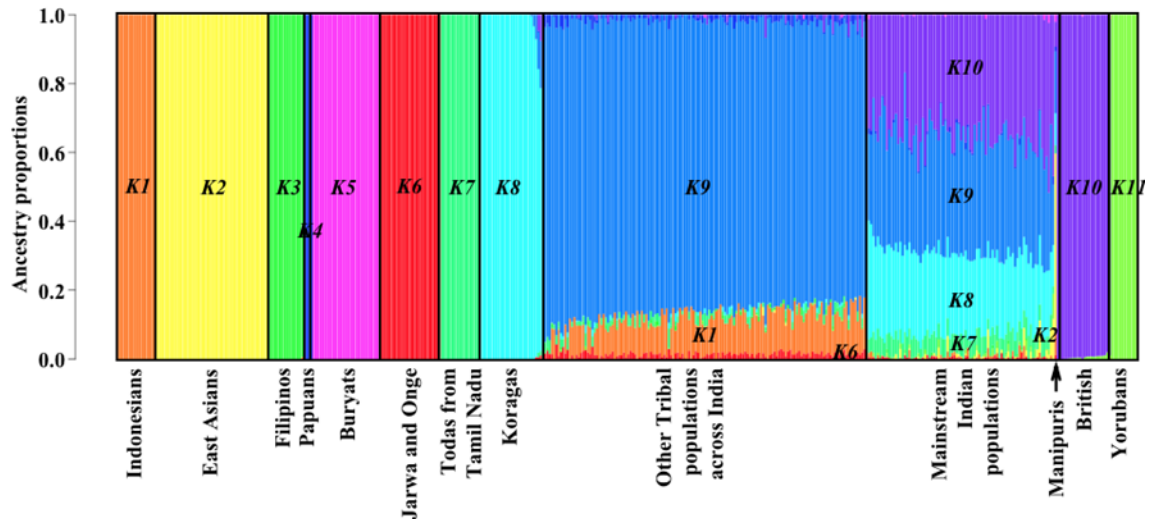
Önge	Indus Periphery	Third source	Third source
54.3	37.2	8.5	Iranian
54.4	37.8	7.8	Syrian
54	38.7	7.4	Lebanese
54.2	38.4	7.3	Saudi Arabian
53.9	39.1	7	Bedouin
54.4	38.6	7	Druze
53.9	39.1	7	Jordanian
54.1	39.1	6.8	Palestinian
53.4	39.9	6.7	Turkish
54.5	38.8	6.7	Armenian

53.5	39.9	6.6	Egyptian
51.5	47.9	0.6	Yoruba
51.6	48.2	0.1	Steppe MLBA

162

163 We then performed a model-based population structure analysis, using ADMIXTURE v1.3. Dataset 1
164 was used for this purpose, as this set covered 5,53,102 SNPs, thus providing greater reliability. The
165 lowest cross-validation error (CVE) was observed for K = 11 (Supplemental Figure 2). At K = 11,
166 Indonesians (K1, orange), East Asians (K2, yellow), Filipinos (K3, green), Papuans (K4, navy blue),
167 Buryats (K5, purple), the Jarawa and Önge (K6, red), the Tōḍa from Tamiḷ Nāḍu (K7, light green),
168 Koraga (K8, cyan), British (K10, violet) and Yorubans from Africa (K11, dark green) were assigned to
169 distinct clusters (Figure 3). The tribal populations exhibited higher proportions of an indigenous Indian
170 component (K9, blue), along with moderate proportions of K1 and minor fractions of K6, indicating
171 their ancestral origin associated with these two populations. Consistent with prior studies^{2,3,18,19}, the
172 admixture plot revealed that mainstream Indian populations were comprised of variable proportions of
173 light green, cyan and blue (K7, K8 and K9, indigenous Indian components, likely ASI-related), violet
174 (K10, West-Eurasian ancestry fraction), and minor fractions of red (K6, likely derived from Ancient
175 Ancestral South Indians: AASI populations) and yellow (K2, East Asian ancestry fractions). In
176 particular, the Manipuri genome employed in this plot revealed a high fraction of yellow (K2) with
177 minor fractions of K8, K9 and K10, potentially linked to a common origin and admixture history.

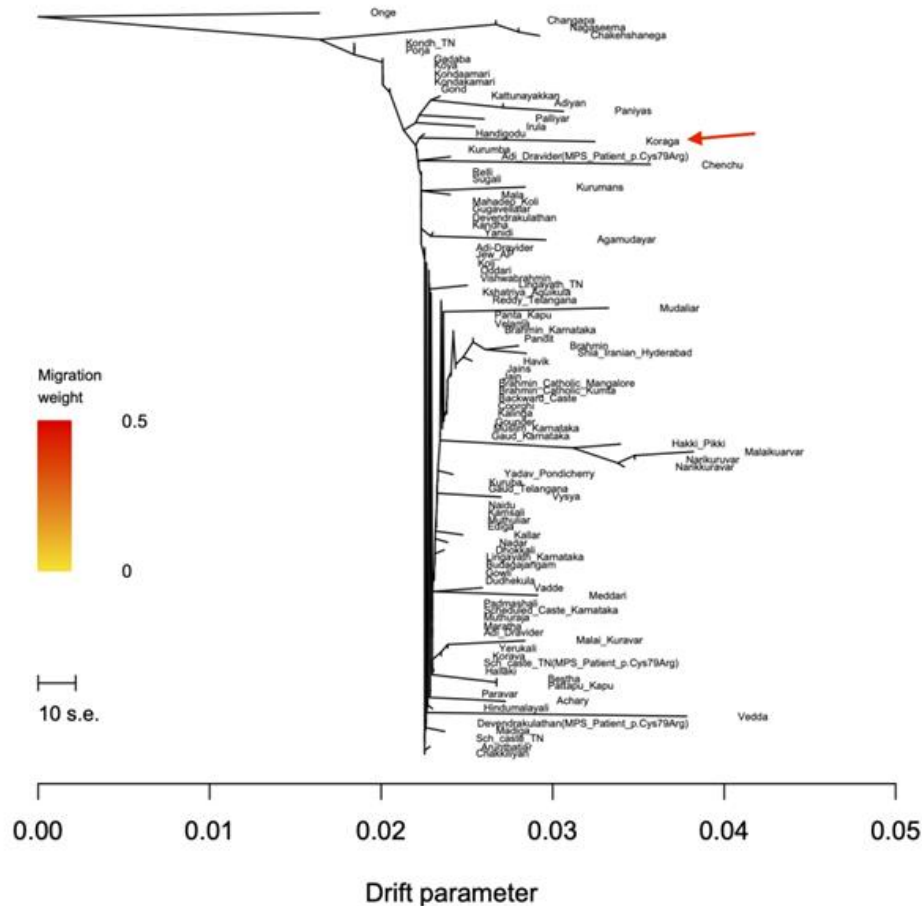
178 Except for three samples, the majority of Koraga genomes analysed in this study do not exhibit K9 and
179 K10, suggesting an early divergence from mainstream Indian populations. This finding supports our
180 hypothesis that Koraga ancestry diverged from other ancestries found on the Indian subcontinent.
181 Moreover, the complete absence of the K11 component reflects the likely absence of ancestral links
182 between the Koraga and African populations.



183

184 Figure 3 Admixture analysis of 1,279 individuals from Dataset 1. The admixture plot illustrates the
185 ancestry components of the samples across the globe. Admixture proportions were obtained through
186 unsupervised analysis at $K = 11$, using ADMIXTURE v1.3 and visualised in R v3.5.1. Each individual
187 is depicted by a vertical line, segmented into coloured sections. The lengths of these segments are
188 proportionate to the contributions of ancestral components to an individual's genome.

189 We utilised TreeMix v.1.13 to explore the patterns of population splits and admixtures amongst selected
190 South Asian populations, using Dataset 3. The maximum likelihood (ML) tree generated by TreeMix
191 using Dataset 1, revealed a high degree of genetic relatedness between the Koraga and Handigōḍu
192 people from the Śivamogga (Shimoga) district of Karnāṭaka. The Handigōḍu, much like the Koraga,
193 are a small isolated South Indian population with a high prevalence of Handigōḍu syndrome, a rare
194 autosomal dominant form of spondyloepimetaphyseal dysplasia (PMID: 7886470). Similar to the
195 Koraga, many Handigōḍu die at a very early age (25-30 years). The ML tree also indicated a genetic
196 affinity of the Koraga with other primitive indigenous South Indian groups, including the Paniya, Iruḷa,
197 Kuruba, and Adi Drāviḍa from Tamiḷ Nāḍu.



198

199 Figure 4 Maximum Likelihood (ML) tree showing the genetic relatedness between the Koraga and
 200 selected South Asian populations present in Dataset 3, using TreeMix v1.13. The tree was rooted with
 201 Önge, an Andamanese population. The horizontal axis represents the drift parameter, and the scale bar
 202 indicates ten times the average standard error of the entries in the sample covariance matrix. The ML
 203 tree depicted notable genetic relatedness between the Koraga and Handigōḍu people of Karnāṭaka.
 204 Additionally, the ML tree highlighted their genetic similarities with various indigenous South Indian
 205 populations, including Paniya, Iruḷa and Kuruba.

206 Founder event in the Koraga tribe

207 Previous studies have revealed that many Indian tribes exhibit high levels of founder effect and
 208 inbreeding^{20,21}. We utilised ASCEND software to determine the date of the founding event of the
 209 Koraga tribe. Our analysis revealed that the most recent founder event occurred between 750 and 1020
 210 years ago, with an intensity of 6.7% to 8.2%. This intensity is approximately five times stronger than
 211 that observed in the Ashkenazi Jews²⁰. During this time, the Kadamba dynasty ruled the region, but
 212 their reign was disrupted by the rule of the Rāṣṭrakūṭa, Hoysāḷa and Western Cālukya dynasties, whilst
 213 the Perumāḷ dynasty ruled in Kerala to the south. From temple records and inscriptions, it is evident

214 that Brahmins, who had settled along the southwestern coast, had begun to exert significant influence
215 under the patronage of local rulers²². Imposition of Brahmanism and more rigid observance of the caste
216 system may have increased social exclusion due to the practice of untouchability, which led to the
217 genetic isolation of the tribal populations. These developments may have contributed to the gradual
218 reduction of the Koraga population and left genetic consequences. Our research, along with the findings
219 of a recent study on the maternal ancestry⁸, indicates that the formation of the Koraga gene pool began
220 not later than 2,000 years ago, followed by a founder event about 1,000 years ago. Consequently, this
221 intact gene pool can serve as an appropriate proxy in the absence of ancient DNA in southern India.

222 **Discussion**

223 Earlier studies on the Koraga tribe using uniparental markers revealed unusually high frequencies of
224 the U1 and H1 haplogroups in the maternal and paternal ancestry respectively^{8,23}. This finding was
225 considered unusual because the maternal U1 haplogroup is West Asian, whereas the paternal H1a is
226 Indian-specific. Such a contrast is either the result of a complete turnover or a bottleneck event. Sequeira
227 et al. 2024 concluded that the U1 haplogroup is a correlate for the Koraga language, which belongs to
228 the North Dravidian branch *sensu lato* of the Dravidian language family. Against this backdrop, we
229 have now investigated the ancestral origin of the Koraga tribe using autosomal SNPs.

230 The exceptional nature of the Koraga gene pool was evident in the overall F_{st} as well. The Koraga
231 clustered away from most of the Northern and Southern populations. Whereas in the PCA plot, Koraga
232 samples scattered in a pattern similar to that of another North Dravidian speaking tribe from the eastern
233 frontier. Both these tribes clustered away from Brahui, a North Dravidian speaking remnant population
234 that through genetic assimilation now clusters with Indo-European speaking groups. The geographical
235 distribution of North Dravidian language communities has been pointed out by linguists for over 200
236 years. Our findings are an attempt to understand the geographical range of North Dravidian based on
237 admixture models and dating. Using ALDER admixture date estimation, we demonstrate that the three
238 major North Dravidian speaking populations shared a common ancestor about 4,500 years ago. Both
239 f_3 -statistics and admixture graphs show that the Brahui and Oraon (Kurukh) underwent different
240 demographic changes as compared with the Koraga tribe, whose gene pool remained relatively intact

241 until experiencing a strong bottleneck 1,000 years ago, possibly as a result of social exclusion. Ancestral
242 connections between the Koraga and the 10,000-year-old Neolithic sample from Ganj Dareh in the
243 Zagros mountains of eastern Iran points to a shared ancestry at the putative time depth ascribed by
244 linguists to Elamo-Dravidian, a proto-language ancestral to both the Elamite language and the Proto-
245 Dravidian language hypothetically spoken by the inhabitants of the Mehrgarh Neolithic that represented
246 the direct antecedents to the Indus Valley civilization^{5,15,24,25}. Both the Koraga and the Early Neolithic
247 Ganj Dareh sample share a common ancestor, and we observe the Koraga component (K8 in the
248 ADMIXTURE bar plot) in most of the present-day non-tribal populations alongside and in addition to
249 the Iranian plateau farmer related, Pontic-Caspian steppe pastoralist related and Andamanese hunter-
250 gatherer related components. The bias in the contribution of a West Asian maternal ancestry to the
251 Koraga gene pool and its correlation with the Koraga mother tongue, as well as the geographical spread
252 of Dravidian languages allows us to hypothetically identify this novel ancestry tentatively as Proto-
253 Dravidian. The descendants bearing this ancestry dispersed throughout the Indian subcontinent as
254 pastoralists and farmers and gave rise to the populations that formed today's Dravidian language
255 communities. In this context, the question must be addressed as to what relationship obtains with the
256 population that peopled the Indus Valley civilisation.

257 A recent study suggests that archaic DNA (Sarazm_EN) dating from the 4th millennium BC from what
258 today is Tajikistan as the best proxy for Iranian plateau farmer-related ancestry⁷. We do not find any
259 direct ancestry sharing between our hypothetical Proto-Dravidian and Sarazm_EN or with the Indus
260 Periphery cline. However, the f3-statistics for our hypothetical Proto-Dravidian are similar to those of
261 Sarazm_EN, the Indus Periphery cline as well as the ancient DNA dating from 9th–8th millennium BC
262 in the Zagros mountains, i.e. Iran_N and Ganj Dareh_N, suggesting a pre-Neolithic common ancestor
263 related to the ancient Caucasus hunter-gatherer component that diverged from the Andamanese hunter-
264 gatherer lineage in the Late Pleistocene²⁶. Our putative Proto-Dravidian ancestry therefore evidently
265 constituted a separate entity that existed alongside the Iranian plateau farmer related ancestry since the
266 Neolithic period through the Chalcolithic in the vicinity of Indus Valley civilisation. The Elamo-
267 Dravidian theory and the linguistic phylogeny of the Dravidian family tree provide ideal chronological

268 fits for the genetic findings presented here. The time depth of the shared ancestry between the Koraga
269 and Early Neolithic Ganj Dareh 10,000 years ago coincides with the time ascribed by linguists to the
270 hypothetical Elamo-Dravidian linguistic phylum in the Early Holocene and matches geographically
271 with the Elamo-Dravidian homeland in the Zagros mountains, as proposed by McAlpin¹⁰ (1981). The
272 dating of the ‘Proto-Dravidian’ ancestry component matches the flourishing of the Indus Valley
273 civilisation before the demise and break-up of Harappan civilisation. Subsequently, this ancestral
274 component diffused throughout the Indian subcontinent except into tribal populations, whose
275 indigenous ancestry in the Indian subcontinent antedates the time of the Dravidian diffusion.

276 **Materials and Methods**

277 **Sample collection**

278 Saliva samples were collected in compliance with the institutional ethical guidelines of the Yenepoya
279 Ethical Committee 1 (YEC-1/2021/052), affiliated with the Yenepoya deemed-to-be University in
280 Mangalore with written informed consent from 29 unrelated individuals from the Koraga community
281 in the Belṭaṅgaḍi and Mangalore taluks of Dakṣiṇa Kannaḍa district in Karnāṭaka. DNA was extracted
282 by the non-invasive MagStable DNA Saliva Collection Kit (MagGenome Technologies Pvt. Ltd.,
283 Cochin). Genotyping was performed using the Infinium Global Screening Array-24 v3.0 (GSA v3.0)
284 BeadChip Illumina Inc, California. This platform comprised 648,465 Single Nucleotide Polymorphism
285 (SNP) markers. The study participants included 17 males and 12 females, specifically from the Kuṅṭu
286 (N=7) and Tappu (N=22) Koraga clans, ranging from 20 to 70 years of age.

287 **Data curation and dataset generation**

288 Four datasets were created for downstream analysis. In Dataset 1 (N=1,279), the genotype data from 29
289 Koraga samples were merged with 87 non-Koraga Indian samples, genotyped on the GSA v3.0
290 platform, available in our in-house database and 1,163 samples from the GenomeAsia100K Consortium
291 (PMID: 31802016), assessing 5,53,102 SNPs. For Dataset 2 (N=4,605), the genotype data from 29
292 Koraga samples were merged with 4,576 ancient and modern genomes across the globe, including the
293 recently published Harappan genome from Rākhīgaḍhī, assessing 14,539 SNPs^{2,16,21,27,28}. Furthermore,

294 a subset of Dataset 2 was generated (Dataset 3), comprised of 29 Koraga genomes and 693 other South
295 Asian genomes, assessing 14,539 SNPs. File conversions and manipulations were performed using
296 VCFtools v.0.1.13²⁹, PLINK v1.9³⁰ and EIGENSOFT v7.2.1^{31,32}. Dataset 4 included 313 individuals
297 from GenomeAsia100K, covering 190515 SNPs.

298 **Population structure and admixture analysis**

299 To examine the population structure of the Koraga tribe and their relationship to other populations, we
300 conducted Principal Component Analysis (PCA) using smartpca from the EIGENSOFT package
301 (version 18140). We analysed Dataset 1, which included 1,279 individuals and 553,102 SNPs. The top
302 two principal components were plotted using in-house python script.

303 To model an admixture graph, we used Dataset 2 with 12,379 SNPs from modern and ancient
304 populations that were pruned for LD (`-indep-pairwise 50 10 0.1`) and fed it into ADMIXTOOLS 2³³.
305 We began by calculating pairwise f2 statistics between the groups using the “`extract_f2`” function. Then,
306 we used “`f2_from_precomp`” to extract allele frequency products from the computed f2 blocks. For
307 each scenario in which we were interested, we sought the best-fitting admixture graph using
308 “`find_graphs`”. We selected the graph with the lowest score. First, we started with no migrations, and
309 then gradually added migrations until we found the best-fitting graph for that scenario. We tested
310 models for up to 200 generations.

311 In order to examine the genetic relationship and potential for admixture between the Koraga population
312 and other modern and ancient populations, we utilised the f3 function in the ADMIXTOOLS 2 package
313 within R to perform outgroup f3-statistics. Negative f3-statistics indicate admixture within the test
314 population, whilst positive statistics suggest an unadmixed population. We designated the Yoruba and
315 Hân as our outgroups.

316 The genetic ancestry of all individuals was determined utilising an unsupervised clustering algorithm,
317 ADMIXTURE v1.3³⁴. We used Dataset 2 for this purpose. The optimal number of ancestral
318 components (K) was identified by minimising the cross-validation error using the `-cv` flag in the

319 ADMIXTURE command line. The analysis revealed the lowest cross-validation error when K was set
320 to 11 (Supplementary Figure 2).

321 To uncover ancient splits and genetic relationships among populations, TreeMix v1.13³⁵ was employed.
322 The Önge genomes were utilised to establish the root of the maximum likelihood tree generated by
323 TreeMix. Dataset 3 was used for plotting the maximum likelihood tree.

324 The statistical tool qpAdm³⁶ implemented in AdmixTools v5.1¹⁷ was employed to estimate ancestry
325 proportions in the Koraga genomes originating from a mixture of reference populations by utilising
326 shared genetic drift with a set of outgroup populations. Dataset 2, comprised of ancient and modern
327 genomes, was employed for qpAdm and qpWave analysis. In the qpAdm analysis, the designation
328 ‘Indus Periphery’ was applied to four ancient samples, specifically Rākhīgadhī_BA_Harappan, Shahr-
329 i-Soktha_MLBA2, Shahr-i-Soktha_MLBA3 and Gonur2_BA^{16,27}. The Koraga were modeled as a
330 combination of three source populations namely: Önge, Indus Periphery and a third population from
331 Eurasia and Africa. The Mbuti from Africa, Scandinavian hunter-gatherers (SHG), Eastern hunter-
332 gatherers (EHG), the Neolithic samples from Ganj Dareh in what today is western Iran, Neolithic
333 Anatolians, Neolithic western Siberians, Hān Chinese and the Karitiana from Brazil were used as the
334 ‘Right’ outgroup populations (O8).

335 **Determination of the time of admixture**

336 ALDER v.1.02³⁷ was used to compute a weighted linkage disequilibrium (LD) analysis to infer the
337 likely date of last admixture between the Koraga and other populations, considering a generation time
338 of 30 years. The Koraga were modelled as the ‘admixpop’ (admixed population), and the remaining
339 South Asian populations were considered as the ‘refpops’ (reference populations).

340 **Estimation of founder age using ASCEND**

341 ASCEND v10.1.1 was used to estimate the founder age and intensity of the founder event in the Koraga
342 population. The analysis was performed using two datasets for reliability. The first dataset included
343 14,539 SNPs, and the second dataset included 5,35,017 SNPs. Yoruba samples were considered as the
344 outgroup. The founder age was calculated assuming a generation time of 25 and 30 years.

345 **Declarations and statements**

346 **Conflict of Interest**

347 The authors declare that the research was conducted in the absence of any commercial or financial
348 relationships that could be construed as a potential conflict of interest.

349 **Ethics Approval**

350 This study was performed in line with the principles of the Declaration of Helsinki. Approval was
351 granted by Yenepoya Ethical Committee 1 (YEC-1/2021/052), affiliated with the Yenepoya deemed-
352 to-be University.

353 **Author Contributions**

354 JJS, RD and GvD contributed in conceptualisation. RD, MSM and SK collected the samples and
355 performed data curation. JJS and RD performed statistical analysis and wrote the first draft of the
356 manuscript. GvD reviewed, revised and redacted the entire manuscript draft. RD and MSM reviewed
357 the final manuscript. All authors contributed to the manuscript and approved the submitted version.

358 **Funding**

359 This work was supported by Yenepoya (Deemed to be University) Seed Grant (YU/Seed Grant/093-
360 2020).

361 **Acknowledgments**

362 The authors acknowledge the participants, lab technicians and research staff involved in the study.

363 **Data Availability Statement**

364 The raw genotype data of the participants cannot be shared for ethical reasons. However, the secondary
365 data in the form of analysed files are available with the corresponding author and will be shared upon
366 request.

367 **References**

- 368 1. ArunKumar, G., Soria-Hernanz, D.F., Kavitha, V.J., Arun, V.S., Syama, A., Ashokan, K.S.,
369 Gandhirajan, K.T., Vijayakumar, K., Narayanan, M., Jayalakshmi, M., et al. (2012).
370 Population Differentiation of Southern Indian Male Lineages Correlates with Agricultural
371 Expansions Predating the Caste System. *PLoS One* 7, e50269.
- 372 2. Moorjani, P., Thangaraj, K., Patterson, N., Lipson, M., Loh, P.-R., Govindaraj, P., Berger, B.,
373 Reich, D., and Singh, L. (2013). Genetic evidence for recent population mixture in India. *Am J*
374 *Hum Genet* 93, 422–438. [10.1016/j.ajhg.2013.07.006](https://doi.org/10.1016/j.ajhg.2013.07.006).
- 375 3. Basu, A., Sarkar-Roy, N., and Majumder, P.P. (2016). Genomic reconstruction of the history
376 of extant populations of India reveals five distinct ancestral components and a complex
377 structure. *Proc Natl Acad Sci U S A* 113, 1594–1599. [10.1073/pnas.1513197113](https://doi.org/10.1073/pnas.1513197113).
- 378 4. Kumar, L., Chowdhari, A., Sequeira, J.J., Mustak, M.S., Banerjee, M., and Thangaraj, K.
379 (2023). Genetic Affinities and Adaptation of the South-West Coast Populations of India.
380 *Genome Biol Evol* 15, evad225. [10.1093/gbe/evad225](https://doi.org/10.1093/gbe/evad225).
- 381 5. van Driem, G.L. (2021). *Ethnolinguistic Prehistory* (Brill) [10.1163/9789004448377](https://doi.org/10.1163/9789004448377).
- 382 6. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. (2009). Reconstructing
383 Indian population history. *Nature* 461, 489–494. [10.1038/nature08365](https://doi.org/10.1038/nature08365).
- 384 7. Kerdoncuff, E., Skov, L., Patterson, N., Zhao, W., Lueng, Y.Y., Schellenberg, G.D., Smith,
385 J.A., Dey, S., Ganna, A., Dey, A.B., et al. (2024). 50,000 years of Evolutionary History of
386 India: Insights from ~2,700 Whole Genome Sequences. *bioRxiv*, 2024.02.15.580575.
387 [10.1101/2024.02.15.580575](https://doi.org/10.1101/2024.02.15.580575).
- 388 8. Sequeira, J.J., Vinuthalakshmi, K., Das, R., van Driem, G., and Mustak, M.S. (2024). The
389 maternal U1 haplogroup in the Koraga tribe as a correlate of their North Dravidian linguistic
390 affinity. *Front Genet* 14.
- 391 9. Bhat [i.e. Bhaṭṭa], D.N.Ś. (1971). *The Koraga Language* (Deccan College Postgraduate and
392 Research Institute).

- 393 10. McAlpin, D.W. (1981). Proto Elamo Dravidian: The Evidence and Its Implications. In
394 Transactions of the American Philosophical Society, v. 71, part 3 (American Philosophical
395 Society).
- 396 11. Krishnamurti, B. (1978). Areal and lexical diffusion of sound change. *Language (Baltim)* 54,
397 1–20.
- 398 12. Krishnamurti, B. (1998). ‘Patterns of sound change in Dravidian’, pp. 63-79 in Rajendra
399 Singh, ed., *The Yearbook of South Asian Languages and Linguistics 1998*. (Sage
400 Publications).
- 401 13. Krishnamurti, B. (2003). *The Dravidian Languages* (Cambridge University Press).
- 402 14. Zvelebil, K. (1990). *Dravidian Linguistics: An Introduction* (Pondicherry Institute of
403 Linguistics and Culture).
- 404 15. van Driem, G. (2001). *Languages of the Himalayas: An Ethnolinguistic Handbook of the
405 Greater Himalayan Region, containing an Introduction to the Symbiotic Theory of Language*
406 (Brill).
- 407 16. Narasimhan, V.M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S.,
408 Lazaridis, I., Nakatsuka, N., Olalde, I., Lipson, M., et al. (2019). The formation of human
409 populations in South and Central Asia. *Science (80-)* 365, eaat7487. 10.1126/science.aat7487.
- 410 17. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T.,
411 Webster, T., and Reich, D. (2012). Ancient Admixture in Human History. *Genetics* 192, 1065–
412 1093. 10.1534/genetics.112.145037.
- 413 18. Das, R., and Upadhyai, P. (2018). An ancestry informative marker set which recapitulates the
414 known fine structure of populations in South Asia. *Genome Biol Evol* 10, 2408–2416.
415 10.1093/gbe/evy182.
- 416 19. Das, R., Ivanisenko, V.A., Anashkina, A.A., and Upadhyai, P. (2020). The story of the lost
417 twins: decoding the genetic identities of the Kumhar and Kurcha populations from the Indian

- 418 subcontinent. *BMC Genet* 21, 117. 10.1186/s12863-020-00919-2.
- 419 20. Tournebize, R., Chu, G., and Moorjani, P. (2022). Reconstructing the history of founder events
420 using genome-wide patterns of allele sharing across individuals. *PLOS Genet* 18, e1010243.
- 421 21. Nakatsuka, N., Moorjani, P., Rai, N., Sarkar, B., Tandon, A., Patterson, N., Bhavani, G.S.,
422 Girisha, K.M., Mustak, M.S., Srinivasan, S., et al. (2017). The promise of discovering
423 population-specific disease-associated genes in South Asia. *Nat Genet* 49, 1403–1407.
424 10.1038/ng.3917.
- 425 22. Sturrock, J. (1894). *Madras District Manuals - South Canara Volume I* (Government Press).
- 426 23. Anthropological Survey of India (2021). Y-Chromosome Phylogeny in Indian Population. In
427 *Genomic Diversity in People of India* (Springer Singapore), pp. 145–169. 10.1007/978-981-
428 16-0163-7_6.
- 429 24. Parpola, A. (1994). *Deciphering the Indus Script* (Cambridge University Press).
- 430 25. Parpola, A. (2015). *The Roots of Hinduism: The Early Aryans and The Indus Civilisation*
431 (Oxford University Press).
- 432 26. Jones, E.R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R.,
433 McLaughlin, R.L., Gallego Llorente, M., Cassidy, L.M., Gamba, C., et al. (2015). Upper
434 Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat Commun* 6, 8912.
435 10.1038/ncomms9912.
- 436 27. Shinde, V., Narasimhan, V.M., Rohland, N., Mallick, S., Mah, M., Lipson, M., Nakatsuka, N.,
437 Adamski, N., Broomandkoshbacht, N., Ferry, M., et al. (2019). An Ancient Harappan
438 Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers. *Cell* 179, 729-735.e10.
439 10.1016/j.cell.2019.08.048.
- 440 28. Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D.,
441 Novak, M., Gamarra, B., Sirak, K., et al. (2016). Genomic insights into the origin of farming
442 in the ancient Near East. *Nature* 536, 419–424. 10.1038/nature19310.

- 443 29. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker,
444 R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and
445 VCFtools. *Bioinformatics* 27, 2156–2158. 10.1093/bioinformatics/btr330.
- 446 30. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J.,
447 Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome
448 association and population-based linkage analyses. *Am J Hum Genet* 81, 559–575.
449 10.1086/519795.
- 450 31. Patterson, N., Price, A.L., and Reich, D. (2006). Population Structure and Eigenanalysis.
451 *PLOS Genet* 2, e190.
- 452 32. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D.
453 (2006). Principal components analysis corrects for stratification in genome-wide association
454 studies. *Nat Genet* 38, 904–909. 10.1038/ng1847.
- 455 33. Maier, R., Flegontov, P., Flegontova, O., Işıldak, U., Changmai, P., and Reich, D. (2023). On
456 the limits of fitting complex models of population history to f-statistics. *Elife* 12, e85492.
457 10.7554/eLife.85492.
- 458 34. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of
459 ancestry in unrelated individuals. *Genome Res* 19, 1655–1664. 10.1101/gr.094052.109.
- 460 35. Pickrell, J.K., and Pritchard, J.K. (2012). Inference of Population Splits and Mixtures from
461 Genome-Wide Allele Frequency Data. *PLOS Genet* 8, e1002967.
- 462 36. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G.,
463 Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe
464 was a source for Indo-European languages in Europe. *Nature* 522, 207–211.
465 10.1038/nature14317.
- 466 37. Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., and Berger, B.
467 (2013). Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium.

468 *Genetics* 193, 1233 LP – 1254. 10.1534/genetics.112.147330.

469