

# 1 **50,000 years of Evolutionary History of India: Insights from ~2,700** 2 **Whole Genome Sequences**

3 Elise Kerdoncuff<sup>1,\*,#</sup>, Laurits Skov<sup>1,\*,#</sup>, Nick Patterson<sup>2</sup>, Wei Zhao<sup>3,4</sup>, Yuk Yee Lueng<sup>5</sup>, Gerard D.  
4 Schellenberg<sup>5</sup>, Jennifer A. Smith<sup>3,4</sup>, Sharmistha Dey<sup>6</sup>, Andrea Ganna<sup>7</sup>, AB Dey<sup>8</sup>, Sharon L.R.  
5 Kardia<sup>4</sup>, Jinkook Lee<sup>9</sup>, Priya Moorjani<sup>1,10,#</sup>

6

7 <sup>1</sup>Department of Molecular and Cell Biology, University of California, Berkeley, United States of America

8 <sup>2</sup>Department of Human Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America

9 <sup>3</sup>Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, United States of America

10 <sup>4</sup>Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America

11 <sup>5</sup>Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, Perelman School of Medicine,

12 University of Pennsylvania, United States of America

13 <sup>6</sup>Department of Biophysics, All India Institute of Medical Sciences, New Delhi, India

14 <sup>7</sup>Institute for Molecular Medicine Finland, Helsinki, Finland

15 <sup>8</sup>Department of Geriatric Medicine, All India Institute of Medical Sciences, New Delhi, India

16 <sup>9</sup>Department of Economics, and Center for Economic & Social Research, University of Southern California, Los Angeles,  
17 California, United States of America

18 <sup>10</sup>Center for Computational Biology, University of California, Berkeley, United States of America

19 \* These authors contributed equally to the work.

20

21 #To whom correspondence should be addressed: Elise Kerdoncuff <[kerdoncuff@berkeley.edu](mailto:kerdoncuff@berkeley.edu)>, Laurits Skov

22 <[lauritsskov@berkeley.edu](mailto:lauritsskov@berkeley.edu)> and Priya Moorjani <[moorjani@berkeley.edu](mailto:moorjani@berkeley.edu)>

23

## 24 **Abstract**

25 India has been underrepresented in whole genome sequencing studies. We generated 2,762 high  
26 coverage genomes from India—including individuals from most geographic regions, speakers of  
27 all major languages, and tribal and caste groups—providing a comprehensive survey of genetic  
28 variation in India. With these data, we reconstruct the evolutionary history of India through space  
29 and time at fine scales. We show that most Indians derive ancestry from three ancestral groups  
30 related to ancient Iranian farmers, Eurasian Steppe pastoralists and South Asian hunter-gatherers.  
31 We uncover a common source of Iranian-related ancestry from early Neolithic cultures of Central  
32 Asia into the ancestors of Ancestral South Indians (ASI), Ancestral North Indians (ANI),  
33 Austro-asiatic-related and East Asian-related groups in India. Following these admixtures, India  
34 experienced a major demographic shift towards endogamy, resulting in extensive homozygosity  
35 and identity-by-descent sharing among individuals. At deep time scales, Indians derive around  
36 1-2% of their ancestry through gene flow from archaic hominins, Neanderthals and Denisovans.  
37 By assembling the surviving fragments of archaic ancestry in modern Indians, we recover ~1.5  
38 Gb (or 50%) of the introgressing Neanderthal and ~0.6 Gb (or 20%) of the introgressing  
39 Denisovan genomes, more than any other previous archaic ancestry study. Moreover, Indians  
40 have the largest variation in Neanderthal ancestry, as well as the highest amount of  
41 population-specific Neanderthal segments among worldwide groups. Finally, we demonstrate  
42 that most of the genetic variation in Indians stems from a single major migration out of Africa  
43 that occurred around 50,000 years ago, with minimal contribution from earlier migration waves.  
44 Together, these analyses provide a detailed view of the population history of India and  
45 underscore the value of expanding genomic surveys to diverse groups outside Europe.

## 46 Introduction

47

48 With more than 1.5 billion people and approximately 5,000 anthropologically well-defined  
49 ethno-linguistic and religious groups, India is a region of extraordinary diversity<sup>1</sup>. Yet, Indian  
50 populations are often underrepresented in genomic studies. Recent sequencing endeavors such as  
51 the 1000 Genomes Project (1000G)<sup>2</sup>, UK Biobank<sup>3</sup>, TopMed<sup>4</sup>, Simons Genome Diversity Panel<sup>5</sup>  
52 and GenomeAsia<sup>6,7</sup> have incorporated Indian populations. However, with the exception of  
53 GenomeAsia<sup>6,7</sup>, these efforts have either included very few individuals or primarily sampled  
54 expatriate communities outside of India, leading to a limited (and biased) representation of the  
55 genetic variation seen in India. As a result, many open questions remain about the population  
56 history of India: When did people first migrate to India from Africa—as part of the major  
57 migration out of Africa or at earlier times along the southern coastal route of migration? What is  
58 the contribution and legacy of archaic gene flow from Neanderthals and Denisovans to Indians?  
59 How have recent technological innovations like Neolithic farming and spread of languages  
60 impacted variation in India?

61

62 To obtain a more complete picture of human diversity in India, we generated deep coverage  
63 genome sequences of ~2,700 individuals from 18 states in India. Our samples are part of the  
64 Longitudinal Aging Study in India - Diagnostic Assessment of Dementia (LASI-DAD)<sup>8</sup> that is a  
65 population-based prospective cohort study that has collected nationally representative data of  
66 individuals that are 60 years or older. These data contain individuals from diverse geographic  
67 regions (including rural and urban areas), speakers for many language families (e.g.,  
68 Indo-European, Dravidian and Tibeto-Burman languages) and various ethno-linguistic and caste  
69 groups (e.g., self-reported castes recognized by the Indian government), providing the most  
70 comprehensive snapshot of genetic diversity in India.

71

## 72 Data and catalog of novel variants

73

74 A total of 2,762 LASI-DAD participants, including 22 trios (mother-father-child), were  
75 sequenced at MedGenome, Inc. (Bangalore, India) at an average read depth of 30x. Individuals  
76 were sampled from 18 different states across India (Fig 1A), with median sample size of 157  
77 individuals per state (Supplementary Note S1). The raw whole genome sequences were sent to  
78 the Genome Center for Alzheimer's Disease (GCAD) at the University of Pennsylvania for joint  
79 calling and quality control. A total of 2,679 samples and 73.2 million autosomal bi-allelic  
80 variants passed quality control filters, including 67.1 million single nucleotide variants (SNVs)  
81 and 6.04 million insertion-deletions (indels) (Supplementary Note S2). We identified 24 million  
82 novel SNVs and 2.2 million novel indels, underscoring the limitations of existing human genetic  
83 variation databases like the 1000G and Genome Aggregation Database (gnomAD)<sup>9</sup> in  
84 representing diverse populations. The vast majority (>99%) of the newly identified variants are  
85 rare, including 68% of singletons and less than 1% common variants (with greater than 1%

86 frequency) (Table S2.1). Genome phasing was conducted using SHAPEIT4<sup>10</sup>, and we estimated a  
87 low phase switch error rate of less than 1.15% in trios (Table S3.1).

88

89 Our dataset is representative of the population diversity in India. It includes individuals born in  
90 23 different states from both rural (63%) and urban (37%) areas. It comprises speakers of around  
91 26 different languages that belong to diverse caste groups as recognized by the Indian  
92 government: 4% from Scheduled Tribes, 18% from Scheduled Castes, and 44% from other  
93 backward class (OBC). Nearly equal numbers of males and females were recruited in the study,  
94 with our dataset constituting 52% of females. For many analyses, we categorized individuals  
95 based on their birth location into six major geographic regions: North ( $n=555$ ), West ( $n=385$ ),  
96 Central ( $n=373$ ), South ( $n=715$ ), North-East ( $n=73$ ), and East ( $n=530$ ). After performing quality  
97 control checks and excluding first-degree relatives, we used a sample of 2,620 individuals for  
98 most of our analyses described below, unless specified otherwise (see Methods, Supplementary  
99 Note S1-2).

100

## 101 **Population structure and admixture**

102

103 To study population relationships of Indians to other worldwide populations, we combined the  
104 LASI-DAD dataset with the 1000G<sup>11</sup> and applied Principal component analysis (PCA)<sup>12</sup>,  
105 ADMIXTURE<sup>13</sup> and  $f$ -statistics<sup>14</sup>. Consistent with previous reports<sup>15,16</sup>, we find that the  
106 population structure in India is related to individuals of West Eurasian-related ancestry (1000G  
107 EUR), with limited or no recent gene flow from populations related to sub-Saharan Africans (Fig  
108 1B, Fig S4.1). The population structure in India is correlated to geography (state of birth) and  
109 linguistic affiliation, with three main clusters—one cluster that includes the majority of the  
110 individuals from North and South of India who speak Indo-European and Dravidian languages  
111 and represents varying relatedness to West Eurasians, referred to as ‘Indian cline’ (Fig 1B, Fig  
112 S4.2-3). The Indian cline has previously been shown to reflect variable proportions of ancestry  
113 from two ancestral groups: the *Ancestral North Indians (ANI)* who harbor large proportions of  
114 ancestry related to West Eurasians, and the *Ancestral South Indians (ASI)* who are distantly  
115 related to West Eurasians<sup>15,16</sup>. Recent ancient DNA analysis have shown that both *ANI* and *ASI*  
116 are admixed and in turn, have ancestry from groups related to ancient Iranian farmers, ancient  
117 Eurasian Steppe pastoralists, and unsampled indigenous South Asians (*Ancient Ancestral South*  
118 *Indians (AASI)*) distantly related to Andamanese hunter-gatherers (*AHG*)<sup>17</sup>.

119

120 Beyond the Indian cline, we find two primary clusters of individuals ( $n=494$ ): a cluster towards  
121 the *ASI*-end of the cline, and another found closer to the center exhibiting clear relatedness to  
122 East Asian-related groups (1000G EAS) in PCA (Fig 1B). The former mainly includes  
123 individuals from Central and East India, with the majority from the state of Odisha where  
124 predominantly Indo-European and Austro-asiatic languages are spoken. The East Asian-related  
125 cluster includes individuals from East and North-East regions of India. West Bengal is the most  
126 representative state in this cluster, with almost 10% ancestry related to East Asians. Using  
127 ALDER<sup>18</sup>, we estimated the admixture related linkage disequilibrium related to EAS to infer that

128 this gene flow occurred 50 generations ago or around 520 AD, possibly related to the invasions  
129 of the Huna people to India after the collapse of the Gupta Empire (Fig S4.11)<sup>19,20</sup>. Another  
130 predominant group in the East Asian-related cluster is from Assam. This group exhibits  
131 significant heterogeneity, as individuals have varying degrees of relatedness to EAS, indicative  
132 of the recent gene flow possibly related to the recent migration of East Asian tea plantation  
133 workers to India in the last two centuries<sup>21</sup> (Fig 1B). Our ADMIXTURE<sup>13</sup> analysis mirrors the  
134 patterns seen in PCA (Fig S4.6).

135

## 136 **Ancestry Composition and Sources**

137

138 To model the ancestry in India, we used *qpAdm* that compares allele frequency correlations  
139 between a population of interest and a set of reference and outgroup populations<sup>14,22</sup>. First, we  
140 examined how well the three-way model with ancient Iranian farmer-related, Eurasian Steppe  
141 pastoralist-related, and *AHG*-related groups describes the ancestry of individuals on the Indian  
142 cline (Fig 1B). Following Narasimhan et al. 2019<sup>17</sup>, we used *Indus Periphery West* that is part of  
143 the *Indus Periphery Cline*—a heterogeneous group of 11 outlier samples from Bronze Age  
144 cultures of Shahr-i-Sokhta and Bactria Margiana Archaeological Complex—as the proxy for  
145 Iranian farmer-related ancestry, Central Steppe Middle to late Bronze age  
146 (*Central Steppe MLBA*) as the source for Yamnaya Steppe pastoralist-derived ancestry and  
147 *AHG*-related individuals to represent *AASI* ancestry<sup>17</sup>. We find the three-way model provides a  
148 good fit for the majority (>90%) of the individuals on the Indian cline, with some exceptions (we  
149 define ‘good fit’ as models with *qpAdm* *p*-value > 0.01, see Methods). Notably, we find 22  
150 individuals that can be fitted as a two-way mixture between ancient Iranian farmer-related and  
151 *AHG*-related ancestries without Steppe pastoralist-related ancestry (referred to as *ASI*  
152 henceforth).

153

154 The archaeological context of the *Indus Periphery Cline* and their relationship to ancient Indian  
155 civilizations (e.g., Indus Valley Civilization) is unclear as these were migrant samples from  
156 Bronze Age Central Asian cultures<sup>17</sup>. Thus, we examined fifteen ancient Iranian-related groups  
157 from the Neolithic to Iron Age as the potential source of the Iranian farmer-related ancestry for  
158 the 22 *ASI* individuals and *Indus Periphery West*. We obtain good fits for all 22 *ASI* individuals  
159 when the Iranian-related ancestry derives from early Neolithic and Copper Age individuals from  
160 Central Asian cultures of either *Sarazm\_EN* or *Namazga\_CA* or a group containing *Sarazm\_EN*  
161 and *Parkhai\_Ananu\_EN* that was previously suggested as the source for *Indus Periphery Cline*<sup>17</sup>.  
162 The latter two models also provide good fits for *Indus Periphery West*, though using *Sarazm\_EN*  
163 alone as the source does not yield a good fit (Table S4.2). Furthermore, a model with  
164 *Sarazm\_EN*, *AHG*-related and *Central Steppe MLBA* also provides a good fit for the vast  
165 majority (>95%) of individuals on the Indian cline (*p*-value in *qpAdm* > 0.01). In contrast,  
166 models with *Namazga\_CA* fail for >15% of individuals on the Indian cline, contrary to previous  
167 claims based on fewer samples<sup>23</sup>. Similarly, models with *Sarazm\_EN* and *Parkhai\_Ananu\_EN* do  
168 not work well for modern Indians and yield negative coefficients for *Parkhai\_Ananu\_EN* ancestry  
169 (Table S4.3).

170

171 Turning to the individuals that fall outside the Indian cline, we tried three models including  
172 *Sarazm\_EN*, *AHG*-related, and either (a) Steppe pastoralist-related (as the Indian cline model),  
173 (b) Austro-asiatic-related (using *Nicobarese*), or (c) East Asian-related (using *EAS*) ancestries.  
174 We also tested four-way models with addition of *Central\_Steppe\_MLBA* if models (b-c) failed.  
175 We obtain good fits for 91% of the individuals that fall outside the cline (Table S4.4). Notably,  
176 there are 91 individuals that can be modeled without Steppe pastoralist-related ancestry,  
177 including ~96% of the Austro-asiatic-related individuals (using model b). This suggests Iranian  
178 farmer-related ancestry likely did not come through Steppe pastoralist-related groups to India.

179

180 Archaeological studies have also documented trade connections between Sarazm and South Asia,  
181 including connections with agriculture sites of Mehrgarh and early Indus Valley Civilization<sup>24</sup>.  
182 Indeed, one of the two *Sarazm\_EN* individuals (*Sarazm\_EN\_1*) was found with shell bangles that  
183 are identical to ones found at sites in Pakistan and India such as Shahi-Tump, Makran and  
184 Surkotada, Gujarat<sup>25</sup> (*J. Mark Kenoyer*, personal communication). Surprisingly, when we applied  
185 *qpAdm*, we discovered that *Sarazm\_EN\_1* has substantial *AHG*-related ancestry (~15%), unlike  
186 the other individual from the *Sarazm\_EN* group (*Sarazm\_EN\_2*). Application of the three-way  
187 model with *Sarazm\_EN\_2*, *AHG*-related and *Central\_Steppe\_MLBA* continues to provide a good  
188 fit for most individuals (>96%) on the Indian cline, as well as off-cline individuals (Table  
189 S4.7-8). Moreover, the two-way model without Steppe Pastoralist-related ancestry works well  
190 for the 22 *ASI* individuals and *Indus Periphery West* (without need for additional ancestry from  
191 *Parkhai\_Anu\_EN*). Together, our data are consistent with a common source for the ancient  
192 Iranian-related ancestry in ANI, ASI, Austroasiatics-related and East Asian-related individuals in  
193 India, suggesting that the Iranian-related gene flow occurred well before the arrival of Steppe  
194 pastoralist-related ancestry in Bronze Age (~1900–1500 BCE<sup>17</sup>).

195

196 Using *AHG*-related, *Sarazm\_EN* and *Central\_Steppe\_MLBA* as reference populations, we  
197 inferred the genetic composition of individuals on the Indian cline. We find marked variation in  
198 ancestry proportions across India, with Iranian farmer-related ancestry varying between  
199 ~27–68%, *AHG*-related between ~19–69% and *Central\_Steppe\_MLBA* between ~0–45%.  
200 Among the three ancestry components, variation in *AHG*-related shows the strongest correlation  
201 to the ANI-ASI cline in PCA (Fig S4.10). *AHG*-related ancestry proportion is significantly  
202 associated with geography (e.g., highest in South and lowest in North of India), language (i.e.,  
203 higher in Dravidian vs. Indo-European language speakers) and caste affiliation (highest in  
204 Scheduled Castes, Scheduled Tribes and OBC compared to other groups) (Fig 1C, Extended  
205 Data Fig 1). This highlights that the ancient admixture events are related to the spread of  
206 languages and the history of the traditional caste system in India.

207

## 208 **Founder events increase homozygosity in India**

209

210 Previous studies have shown that many Indian groups have a history of strong founder events,  
211 due to endogamous and consanguineous marriages<sup>7,26,27</sup>. Founder events reduce genetic variation

212 and increase sharing of genomic regions that are inherited identical-by-descent (IBD) from a few  
213 common ancestors<sup>28</sup>. Descendants of consanguineous marriages (between close relatives) may  
214 inherit IBD segments from both parents, resulting in segments that are homozygous-by-descent  
215 (HBD). A founder event results in many, small HBD segments, while recent consanguinity  
216 results in fewer but longer HBD segments.

217

218 We identified IBD and HBD segments in LASI-DAD and 1000G datasets using a  
219 haplotype-based IBD detection method, *hap-IBD*<sup>29</sup>. To differentiate between the relative effects  
220 of founder events and recent consanguineous marriages, we stratified the HBD segments by  
221 length— long ( $> 8\text{cM}$ ) indicative of consanguinity and short ( $< 8\text{cM}$ ) mostly reflecting founder  
222 events. Indians, on average, have a larger fraction of their genome in HBD segments ( $\sim 29\text{ cM}$ )  
223 compared to 1000G EAS ( $\sim 6\text{ cM}$ ), EUR ( $\sim 6\text{ cM}$ ), and AFR ( $\sim 4\text{ cM}$ ) (Fig 2A). Within India,  
224 individuals from South have significantly higher homozygosity, both in terms of the total amount  
225 of their genome in HBD segments (on average,  $\sim 56\text{ cM}$  in South compared to  $\sim 19\text{ cM}$  in other  
226 regions,  $p\text{-value} < 10^{-16}$ ) and the fraction of long HBD segments (8.4% vs. 4.3%,  $p\text{-value} < 10^{-6}$ ),  
227 reflecting the higher prevalence of consanguineous marriages in the South of India<sup>30</sup> (Fig 2A, Fig  
228 S5.1-2). A majority ( $>90\%$ ) of the homozygosity stems from small HBD segments (rather than  
229 long HBD segments), suggesting a primary role of historical founder events rather than recent  
230 consanguinity as the source of homozygosity (Fig 2A, Fig S5.2). Similar results are obtained  
231 when we use a threshold of  $20\text{ cM}$  to define long HBD segments (Fig S5.1, Fig S5.2B).

232

233 Next, we investigated genome-wide IBD-sharing across individuals. We computed the fraction of  
234 individuals who find at least one close genetic relative within LASI-DAD and compared this  
235 proportion across worldwide populations in 1000G (see Methods, Fig S5.3). We infer that  
236  $\sim 51.0\%$  (38.4–59.2% across regions) of individuals in LASI-DAD find at least one genetic  
237 relative with expected IBD sharing equivalent to a 3rd degree cousin or closer relationship ( $\sim 53$   
238  $\text{cM}$ ) in LASI-DAD, which is markedly higher than 14.2% in SAS, 8.8% in EAS, 8.8% in EUR  
239 and 17.2% in AFR from 1000G (Fig 2B, Table S5.1) (note, a previous study identified  $\sim 5\text{--}10\%$   
240 of individuals are first and second-degree relatives in Gambians from Mandinka (GWD) and  
241 Esan in Nigeria (ESN) contributing to higher relatedness in AFR<sup>31</sup>). The higher IBD sharing in  
242 LASI-DAD, especially compared to 1000G SAS may stem from: (a) larger sample size of  
243 LASI-DAD, or (b) ascertainment bias in selecting individuals in either study. We examined each  
244 of these hypotheses in turn. We performed bootstrap resampling of equal numbers of individuals  
245 ( $n=500$ ) from LASI-DAD as 1000G SAS and inferred that the fraction of 3rd degree cousins  
246 decreased to 24.2% (95% CI: 19.4%–28.6%), yet significantly higher than 1000G SAS (Fig 2B,  
247 Table S5.1). In LASI-DAD, individuals were recruited using a stratified random sampling  
248 approach. First, Sampling Secondary Units (SSUs) (villages/urban census blocks) were chosen in  
249 each state and then within each SSU, individuals were selected randomly. To control for the  
250 impact of this ascertainment scheme, we considered pairwise cross-SSU comparisons among  
251 individuals (Supplementary Note S5). Using this approach and accounting for the sample size,  
252 we continue to find a significant shift in LASI-DAD compared to 1000G SAS, with  
253  $\sim 16.4\text{--}35.0\%$  of individuals sharing IBD equivalent to 3rd degree cousins (Fig S5.4). This

254 comparison highlights the limitations of the sampling of 1000G groups for representing genetic  
255 variation of India (with mainly a few groups from the subcontinent). Overall, we find that all  
256 individuals in LASI-DAD have at least one putative 4th degree cousin or closer relative (with  
257 IBD > 10 cM) in the dataset. The high level of relatedness in India is notable, as a similar level  
258 of IBD sharing is seen in Europeans with approximately 480,000 individuals (almost 200-fold  
259 higher sample size) in UK Biobank<sup>32</sup>.

260

261 The history of founder events predicts a high burden of deleterious variants and increased risk of  
262 recessive diseases, as seen in Finns and Ashkenazi Jews<sup>28,33</sup>. To assess the potential functional  
263 effects of founder events in India, we identified 385,985 missense and 20,319 putative loss of  
264 function (pLoF) variants (see Methods) (Table S5.2). Each individual carries ~10,344 (range:  
265 9,911–10,761) derived missense variants, and ~67 (46–96) pLoF variants on autosomes. Most  
266 (>90%) of these variants are rare (frequency below 1%) or singletons (62%). As expected, we  
267 observe strong correlation between the homozygous deleterious mutation burden (measured as  
268 sum of homozygous missense and pLoF variants carried by an individual) and the total sum of  
269 HBD per individual in India (Extended Data Fig 2). Among 18,451 protein-coding autosomal  
270 genes in the human genome (RefSeq database<sup>34</sup>), we find missense and pLoFs variants in 89.5%  
271 of the genes, ranging between 1–1,265 variants per gene. The top three genes with the highest  
272 number of pLoFs variants are mucin genes: MUC3A, MUC16 and MUC17, with respectively 52,  
273 42 and 41 pLoFs, including homozygous pLoFs in MUC17. As there is partial redundancy in the  
274 function of mucin genes, there may be greater tolerance for loss of function variants<sup>35</sup>.

275

276 Among the 406,304 SNVs, we find about half are South Asian-specific and a large fraction  
277 (40%) are absent in gnomAD or 1000G (Table S5.2). We find that ~4% of South-Asian specific  
278 non-ultra rare (frequency above 0.1%) missense/pLoF variants are present in the ClinVar  
279 database<sup>36</sup>, including 10 classified as ‘pathogenic’ variants (using ClinVar threshold of two-stars,  
280 Table S5.2). Among these, we find a South-Asian specific pathogenic variant in the *BHCE* gene  
281 that is present in 15 individuals (0.28%) in LASI-DAD (and not seen outside India). Patients  
282 with butyrylcholinesterase deficiency may experience prolonged neuromuscular blockade and  
283 muscle paralysis, in response to use of some muscle relaxants used during anesthesia. Previous  
284 studies have identified this variant in the founder community of Vysya from Andhra Pradesh  
285 where it has drifted to high frequency due to the history of founder events<sup>27,37</sup>. In LASI-DAD, 8  
286 of the 15 individuals are from Telangana, the neighboring state of Andhra Pradesh. Local  
287 community doctors use the Vysya ancestry as a counter-indicator before administering anesthetic  
288 drugs, highlighting the potential of reducing disease burden by understanding and documenting  
289 the effects of founder events in India.

290

## 291 **Gene flow from archaic hominins in India**

292

293 Most non-Africans, including Indians, derive ~1-2% of their ancestry from gene flow from  
294 archaic hominins, Neanderthals and Denisovans<sup>5,38</sup>. The functional impact and regional variation  
295 in archaic ancestry in India, however, remains unclear. We applied a reference-free hidden

296 Markov model, called *hmmix*<sup>33</sup>, to 2,679 phased individuals from India (to maximize our sample  
297 size, we retained first-degree relatives (except offspring of trios)). *hmmix* classifies genomic  
298 fragments into two states—‘modern human’ or ‘archaic’—by comparing the density of derived  
299 alleles that are not found in 490 sub-Saharan Africans (who have negligible amount of archaic  
300 ancestry<sup>25</sup>) (see Methods). We also applied *hmmix* to phased data from 2,309 individuals from  
301 1000G, 825 individuals from Human Genome Diversity Panel (HGDP), and used the published  
302 results for 27,566 Icelanders from deCODE genetics that were also analyzed using the same  
303 method<sup>26</sup>. Unless stated otherwise, we retained archaic ancestry segments with a posterior  
304 probability greater than 0.8 for subsequent analysis that translates to <4% false positive rate in  
305 simulations<sup>26</sup>.

306

307 We inferred that Indians have an average of 102.98 Mb or 2.07% of the callable genome (95%  
308 percentile range: 1.84–2.34%) of archaic ancestry. By comparing the putative archaic segments  
309 to sequenced Neanderthal and Denisovan genomes<sup>39,40,41,42</sup>, we inferred the source of the archaic  
310 ancestry based on measuring the number of shared derived archaic variants (DAV) present on  
311 archaic segments. We find that each individual has ~1.48% (95% percentile range: 1.30–1.69%)  
312 Neanderthal and ~0.14% (95% percentile range: 0.07–0.21%) Denisovan ancestry. The  
313 Neanderthal ancestry proportion in India is similar to Europeans (1.3%) and Americans (1.4%),  
314 though significantly lower than East Asians (~1.8%, Wilcoxon ranked test  $p$ -value <  $10^{-15}$ ). The  
315 highest Denisovan ancestry is inferred in Oceanians (~1.8%), while Americans, East Asians and  
316 South Asians have similar amounts (~0.1%) (Table S6.4-5).

317

318 By assembling non-overlapping archaic ancestry segments extracted from individuals in  
319 LASI-DAD, we reconstructed 1,524 Mb of the introgressing Neanderthal and 591 Mb of the  
320 introgressing Denisovan genome (Extended Data Fig 3). Notably, using individuals from all  
321 world-wide regions (from 1000G, HGDP and LASI-DAD), we reconstructed 1,679 Mb of the  
322 introgressed Neanderthal genome that is similar in size to the sequenced Neanderthal genomes  
323 (~1,650 Mb, Fig S6.8, Supplementary Note 6). Despite higher per individual Neanderthal  
324 ancestry in East Asians, we recover more Neanderthal sequence from Indians than East Asians  
325 even after controlling for the sample size (as seen in <sup>38</sup>, Table S6.5, S6.8). This is in part due to  
326 introgressed Neanderthal segments having a higher frequency in East Asia and thus being more  
327 likely to be shared across individuals (Fig S8.4)<sup>38,43</sup>. The largest study of archaic ancestry in  
328 27,566 Icelanders recovered 978 Mb of the introgressed Neanderthal and 112 Mb of the  
329 introgressed Denisovan genome (using posterior probability >0.9 in *hmmix*)<sup>44</sup>. Even with the  
330 more stringent posterior probability threshold, we recover >50% more Neanderthal ancestry  
331 segments from Indians (LASI-DAD) than from Icelanders (Fig 3A). Using all world-wide  
332 regions, we reconstructed 1,080 Mb of the introgressing Denisovan genome. The largest amount  
333 of this is recovered from Indians, though this is not significant after downsampling to the sample  
334 size of Oceanians ( $n=28$ ) (Fig S6.8).

335

336 Next, we calculated the amount of archaic sequence that is shared between Indians and other  
337 worldwide populations from 1000G and HGDP datasets. By sharing we refer to segments which



338 overlap the same genomic regions. We find that 81.2% of Neanderthal ancestry is shared  
339 between at least two global regions (Extended Data Fig 4). We find a total of ~11.7% (or 195.9  
340 Mb out of 1,679 Mb) of uniquely India-specific Neanderthal sequences. Strikingly, ~90.7% of  
341 worldwide Neanderthal sequences are seen in India (Extended Data Fig 5). Moreover, Oceanians  
342 and South Asians have large amounts of unique Denisovan ancestry sequences (Fig S6.6).  
343 Around 51% of Denisovan sequence (301.6 Mb out of 591 Mb) is unique to India (Fig S6.6).  
344 Even after downsampling to sample sizes to match the minimum sample sizes in 1000G ( $n=490$ )  
345 and HGDP ( $n=28$ ), we find significant enrichment for unique Denisovan sequences in Indians  
346 (Fig S6.8).

347

348 To infer the relationship of the introgressed archaic population to the sequenced archaic  
349 genomes, we estimated DAV SNP match rates for each introgressed segment to sequenced  
350 Neanderthals and Denisovan genomes. We find on average the introgressed Neanderthal  
351 segments share 83% of the DAVs with one of the three sequenced Neanderthal genomes, with  
352 the highest sharing with the Vindija Neanderthal (Table S6.10 and Fig S6.11). In contrast, the  
353 introgressed Denisovan genome only shares 47% of DAVs with the sequenced Denisovan  
354 genome, indicating the Denisovan ancestry primarily derives from a group that is distantly  
355 related to the sequenced Altai Denisovan. Using a similar approach as Browning et al. 2018, we  
356 replicate the finding of a single pulse of Neanderthal gene flow in India (Supplementary Note  
357 6)<sup>45</sup>. We find that a single Denisovan-related wave is consistent in most groups in India.  
358 Individuals in North-East and South of India, however, have evidence for two clusters of  
359 Denisovan-related sequences, one closely related to the sequenced Altai Denisovan genome  
360 (segments share on average 84% of DAV SNPs) and a more distantly related group (with  
361 46–50% of shared DAV SNPs) (Fig S6.9). Individuals in North-East India derive a large fraction  
362 of ancestry from recent East Asian-related groups (Fig 1B) that have previously been shown to  
363 have two pulses of Denisovan ancestry<sup>45</sup>. Beyond Neanderthal and Denisovan ancestry, we  
364 inferred 0.42% (95% percentile range: 0.37–0.48%) of archaic ancestry from an unknown source  
365 in Indians (Table S6.4-5). This proportion is similar across all non-Africans and potentially  
366 related to the difference between the sequenced archaic genome and the introgressing archaic  
367 individuals (Fig S6.3). Consequently, this suggests that there is no clear evidence for additional  
368 contribution from other unknown archaic hominins to Indians (at least not more than other  
369 worldwide populations), contrary to previous claims<sup>46</sup>.

370

371 Archaic ancestry varies across regions in India, with the highest archaic ancestry in the  
372 North-East and East of India and lowest in North India (Figs 3B and S6.3, Tables S6.4 and S6.6).  
373 To investigate how recent gene flow events have shaped the distribution of archaic ancestry in  
374 India, we examined the relationship between Neanderthal and Denisovan ancestry as a function  
375 of the three main ancestry components in India. Focussing on individuals on the Indian cline ( $n =$   
376 2,126), we find the *AHG*-related ancestry is positively correlated with both Denisovan ( $r = 0.46$ ,  
377  $p$ -value  $< 10^{-15}$ ) and Neanderthal ( $r = 0.24$ ,  $p$ -value  $< 10^{-15}$ ) ancestries (Fig 3B, Table S6.8).  
378 These results are robust to use of more stringent criteria for assigning archaic ancestry segments  
379 to Neanderthal and Denisovan origin, by focussing on sites where only one archaic group has a

380 derived allele that matches modern humans (see Table S6.8). This suggests that a large amount of  
381 the archaic ancestry seen in present-day Indians is inherited through *AHG*-related ancestry and in  
382 turn, groups with higher *AHG*-related ancestry in the South have higher archaic ancestry.

383

## 384 **Functional legacy of archaic ancestry in India**

385

386 Previous analyses have shown that archaic ancestry has played a major role in human adaptation  
387 and disease, however, few studies have evaluated its role in South Asian populations<sup>38,47</sup>. We  
388 examined the genome-wide distribution of archaic ancestry and identified regions of ‘high  
389 archaic frequency’ among Indians (defined as regions where the archaic frequency across  
390 individuals is two standard deviations above the genome-wide average) (Fig 3C). We identified  
391 1,590 and 818 candidate regions with high frequency of Neanderthal and Denisovan ancestry  
392 respectively. For Neanderthals, we replicated genes such as *FBP2* and *FYCO1* previously  
393 identified in other studies<sup>47–49</sup>, as well as identified *PCAT7* and *CXCR6* as new candidates. For  
394 Denisovans, we replicated signals in *WDFY2*, *CHD1L* and *HELZ2*<sup>47</sup> and identified several new  
395 candidates including *LINC00708* and *CDKN2B* (Supplementary Note 7, Extended Data Table  
396 S3). Performing a gene ontology (GO) enrichment analysis, we find 14 pathways enriched for  
397 Neanderthal and 22 pathways for Denisovan ancestry primarily related to immune function  
398 (Extended Data Table S4).

399

400 Next, we searched for regions that have a high number of derived alleles that are shared between  
401 modern humans and archaic groups, a signature previously observed for *EPAS1* and Denisovan  
402 ancestry in Tibetans<sup>50</sup>. Interestingly, we find certain regions of the genome have a  
403 disproportionately elevated number of variants with derived alleles that are uniquely shared  
404 between Denisovans and Indians; though no similar enrichment is seen for uniquely Neanderthal  
405 shared variants (Supplementary Note 7). Notably, we find that the *BTNL2* gene, part of the major  
406 histocompatibility complex (MHC), contains 78 uniquely derived Denisovan variants within a  
407 13.2-kilobase (kb) region with an exceptionally high Denisovan frequency in Indians of around  
408 10% (> 99.9th percentile). There are two Denisovan haplotypes in this region: a *short* haplotype  
409 of 55–65 kb and a *long* one of ~150 kb with 116.1 and 126.7 uniquely derived Denisovan  
410 variants respectively. The proportion of long haplotypes is lower in the North ( $Z = -2.26$ ) and  
411 higher in the West of India ( $Z = 2.57$ ) compared to all individuals in India (Fig S7.3–4). These  
412 Denisovan haplotypes are also present at high frequency in East Asians (~11.8%, >99.8  
413 percentile), but they are rare in Europeans (~0.4%) and notably, absent in Oceanians (Table  
414 S7.2). The haplotype length and number of shared derived alleles between Indians and  
415 Denisovans suggests this region is likely a product of gene flow from Denisovan or  
416 Denisovan-related populations, rather than ancestral lineage sorting ( $p$ -value <  $10^{-6}$  for the *long*  
417 haplotype;  $p$ -value=0.027 for the *short* haplotype). The MHC contains many genes associated  
418 with immune function and is most likely to be under balancing selection. Indeed, previous  
419 studies have identified *BTNL2* as a candidate for selection in East Asians<sup>51</sup>. Though simulations  
420 show that genetic drift generated by founder events alone can lead to high frequency of archaic

421 ancestry in a region, thus caution is warranted when interpreting high frequency archaic regions  
422 as candidates for selection or adaptive introgression in modern humans (Supplementary Note  
423 S8).

424

425 To identify Indian-specific enriched archaic segments, we computed the population branch  
426 statistic (PBS)<sup>52</sup>. The PBS measures the increase in frequency at a given locus in a population,  
427 since its divergence from the two reference populations. To this end, we apply PBS using Indians  
428 as the population of interest and East Asians and Europeans as reference groups using archaic  
429 allele frequency vs. genotype frequencies to identify candidate archaic enriched regions in India  
430 (see Methods). We identified ~10.7 Mb (or 235 genes) enriched for Neanderthal and ~5.5 Mb (or  
431 84 genes) for Denisovan ancestry (Extended Data Table S3). Denisovan ancestry regions are  
432 enriched for genes related to innate immune response, including several TRIM genes—TRIM26,  
433 TRIM31, TRIM15, TRIM10 and TRIM40—implicated in cellular processes related to entry (or  
434 exit) of virus into a host cell. Among the most significant candidate regions of Neanderthal  
435 ancestry is a gene cluster on chromosome 3 which has been previously associated to COVID  
436 susceptibility<sup>53,54</sup> ( $PBS_{\text{Neanderthal}} > 0.118$ , in the 99.99% percentile of genome-wide PBS scores). In  
437 turn, it was discovered that there are two main haplotypes introgressed from Neanderthals  
438 containing the risk variant: a *core* haplotype of 49.4 kb and a *long* haplotype of 333.8 kb. In  
439 LASI-DAD, both of these haplotypes fall outside the 99% tail of our genome-wide distribution  
440 of Neanderthal ancestry, though there is large variation in Neanderthal haplotypes in this region  
441 including some very long haplotypes that are greater than 1 Mb ( $p$ -value for *core* haplotype =  
442 0.00021,  $p$ -value for *long* haplotype = 0.0020, Fig S7.6A). Across India, the frequency of *core*  
443 haplotype ranges between 20.5% (in North-East) to 34.8% (in East India). The frequency of both  
444 the *core* and *long* haplotypes is significantly higher in the East of India compared to other  
445 regions (*core*: 34.8%,  $Z = 2.68$ , *long*: 23.2%,  $Z = 2.34$ ).

446

447 We also examined regions of the genome devoid of archaic ancestry in modern humans, referred  
448 to as ‘archaic deserts’<sup>44,48,55,56</sup>. We identified six Neanderthal deserts spanning a total of 87.1 Mb  
449 including five that were previously reported (Fig 3C, Fig S7.8, Table S7.4). The location of these  
450 five Neanderthal deserts remains similar with around 70% overlap with previously identified  
451 deserts in Europeans and other populations (Table S7.4). Interestingly, among these deserts is a  
452 region that includes the FOXP2 gene that is associated with language development in humans<sup>55</sup>.  
453 We also identified 13 Denisovan deserts in Indians, including one that overlaps with previously  
454 reported Neanderthal deserts (Fig 3C, Fig S7.9, Table S7.5). Given the low genome-wide  
455 proportion of Denisovan ancestry in Indians, we likely miss Denisovan ancestry in some regions  
456 and thus, over-call Denisovan-related deserts.

457

## 458 **First arrival of modern humans to the Indian subcontinent**

459

460 A central question in the peopling of India is when modern humans first arrived to the  
461 subcontinent from Africa. Archeological evidence suggests occupation in Northern India before  
462 and after the Toba eruption that occurred around 74,000 years ago<sup>57</sup>. It is unclear, however, if this

463 group contributed to the ancestry of present-day peoples in India. In order to test this hypothesis,  
464 we computed the minimum coalescence time of present-day Indians, East Asians, Europeans and  
465 Americans to sub-Saharan Africans. If there is a substantial contribution from the population  
466 who lived in India before the Toba eruption, it should be detectable as an increase in coalescence  
467 time of Indians compared to individuals from other worldwide regions. To estimate the  
468 coalescent time for each non-African individual to sub-Saharan Africans, we used the rate of  
469 emission in the modern human state of *hmmix* after controlling for bioinformatics effects  
470 (phasing errors and depletion of triallelic sites) and excluding individuals with more than 1%  
471 sub-Saharan African-related ancestry (see Methods). Theoretically, the emission parameter  
472 should be proportional to the minimum coalescence time between the test individual and  
473 sub-Saharan Africans, human mutation rate ( $0.45 \times 10^{-9}$  per base pair per year<sup>58</sup>, Fig S9.3) and the  
474 length of the genome surveyed.

475

476 We infer the minimum coalescence time between Indians and sub-Saharan Africans as 53,932  
477 (95% percentile range: 53,190–54,644) years ago (Table S9.2, Fig 4). We obtain qualitatively  
478 similar results for Europeans, East Asians and South Asians in the HGDP dataset. Moreover, by  
479 performing simulations, we show the observed emission parameter in India is consistent with  
480 variation stemming from 0–3% of ancestry from an earlier migration that occurred around  
481 74,000 years ago (Fig S9.5). Our results thus show that the majority of the ancestry of  
482 present-day Indians derives from a major migration event out of Africa that occurred 50,000  
483 years ago.

484

## 485 Discussion

486

487 India is a region of extraordinary genetic diversity, including largest variation in archaic ancestry  
488 among modern humans. Notably, a majority of Neanderthal ancestry that exists today in  
489 present-day individuals is found in India, while other worldwide populations retain only a subset  
490 of this variation (Extended Fig 5). Indians also harbor the most Denisovan ancestry among  
491 Eurasian populations. Moreover, some of the deepest mtDNA and Y-chromosome lineages are  
492 seen in people from Andaman Islands<sup>59</sup>. Interestingly, such large diversity is also reflected in the  
493 early Middle Paleolithic stone tool culture that shows overlap of distinct cultures—Acheulean  
494 hand-axe and Levallois technologies—for over 200,000 years, unlike in other regions of the  
495 world<sup>60,61</sup>. These findings raise important questions about the dispersal and settlement of humans  
496 outside Africa: Did the range of Neanderthals and Denisovans extend to South Asia? Did modern  
497 humans encounter Neanderthals, and to some extent Denisovans, further east in Eurasia rather  
498 than the Middle East as widely believed? These observations call for a re-evaluation of models  
499 of human origins, for both modern human and archaic hominins, in light of the complex diversity  
500 in India.

501

## 502 **Methods**

503

### 504 ***Samples***

505 We generated 2,762 high-coverage genomes as part of this project. These samples are a subset of  
506 the Longitudinal Aging Study in India (LASI) and are part of the Harmonized Diagnostic  
507 Assessment of Dementia of LASI (LASI-DAD)<sup>8</sup> (<https://lasi-dad.org>,  
508 [doi.org/10.25549/5hhx-s820](https://doi.org/10.25549/5hhx-s820)). Participants consented to give venous blood samples (VBS) for  
509 genomics analysis. They also have consented to detailed cognitive assessment and informational  
510 interviews. Details on the sequenced individuals and metadata (i.e., sampling location, sex,  
511 language, caste etc.) can be found in Supplementary Note S1.

512

### 513 ***Whole genome sequencing, variant calling and filtering***

514 Whole-genome sequencing libraries were processed using a PCR-free library preparation and  
515 sequenced on Illumina HiSeq X Ten machines at Medgenome, Bangalore, India. The samples  
516 were sequenced using 100 base pair paired-end sequencing. The raw sequence reads (fastq) from  
517 Medgenome were sent to the Genome Center for Alzheimer’s Disease (GCAD) at the University  
518 of Pennsylvania for genome mapping to the human reference genome (build GRCh38/hg38). We  
519 used Variant Calling Pipeline and data management tool (VCPA) developed by GCAD in  
520 collaboration with Alzheimer’s Disease Sequencing Project (ADSP) to call variants in a uniform  
521 way across other studies that are part of ADSP. The pipeline uses best practices from Genome  
522 Analysis Tool kit (GATK) to call variants. Details of the data processing are described in  
523 Supplementary Note S2. Overall, a total of 2,679 LASI-DAD samples passed sequencing metrics  
524 and quality control checks. Details of quality checks are described in Supplementary Note 2.

525

### 526 ***Identification of first-degree relative pairs***

527 We applied KING (v2.3.0)<sup>62</sup> and the “--ibdseg” option to identify first degree relatives.  
528 Following software guidelines, we applied the following filters: sample pairs without any long  
529 IBD segments (>10Mb) were excluded and short IBD segments (<3Mb) were not utilized to  
530 estimate the proportion of IBD sharing between two individuals. Parent-offspring pairs share  
531 50% of their genomes and siblings may share between 38-65% of their genome inherited IBD<sup>63</sup>.  
532 Thus, we use a minimum cutoff of 38% to identify first-degree relatives and consequently we  
533 flag 64 pairs of individuals . For each pair of first degree relatives, we removed the individual  
534 with the larger amount of missing data. In total, we removed 59 individuals (see details in  
535 Supplementary Note S2), leaving 2,620 individuals that were used for most downstream  
536 analyses.

537

### 538 ***Population structure analysis***

539 To learn about the population history of India and compare it to worldwide populations, we  
540 combined the LASI-DAD dataset with other published genomic datasets including present-day  
541 (1000G<sup>11</sup>, GenomeAsia<sup>6</sup>) and ancient DNA samples (Allen Ancient DNA Resource (AADR) v54  
542 <sup>64</sup>). GenomeAsia and AADR are available in hg19/GRCh37, we performed liftover to  
543 hg38/GRCh38 using liftOver (<https://liftover.broadinstitute.org/>). Then, we merged the datasets

544 using *mergeit* (with ‘strandcheck: YES’) from the EIGENSOFT package (v7.2.1)<sup>65,66</sup> which  
545 generates an intersection of the SNPs in the different datasets, keeping only variants present in  
546 all datasets. The number of individuals and variants for each merged dataset and the analyses  
547 they are used in are reported in Table S4.1.

548

#### 549 ***Principal component analysis (PCA) and ADMIXTURE***

550 To perform PCA and *ADMIXTURE*, we excluded SNPs in linkage disequilibrium (LD) using  
551 PLINK with the option ‘--indep-pairwise 50 10 0.5’ that removes , one variant in each pair of  
552 SNPs in a window of 50 SNPs, if the LD is greater than 0.5. We further excluded variants with a  
553 MAF<0.05. We performed PCA using *smartpca* from the EIGENSOFT package (v7.2.1)<sup>65,66</sup>. We  
554 also applied unsupervised hierarchical clustering of individuals using the maximum likelihood  
555 method implemented in the *ADMIXTURE* software (v1.3.0)<sup>13</sup>. Following program  
556 documentation, we varied the number of clusters (K) between 2–6 and performed cross  
557 validation ten times (option: --cv=10). We stopped the algorithm when the change in  
558 log-likelihood between iterations was less than 0.1 (option: -C 0.1).

559

#### 560 ***qpAdm***

561 We used the qpAdm<sup>14,22</sup> package in ADMIXTOOLS (v7.0.2) to identify the best fitting model  
562 and estimate ancestry proportions in a population of interest that is modeled as a mixture of *n*  
563 ‘reference’ populations using a set of ‘Outgroup’ populations (reference (*left*) and outgroup  
564 (*right*)) populations for each analysis are listed in Supplementary Note S4). We set the  
565 parameters as ‘allsnps: NO’ and ‘details: YES’, which reports a normally distributed *Z* score for  
566 the fitted model. We computed coefficient estimations, standard deviations and p-values through  
567 block jackknife resampling. We considered a model to be a good fit if *p*-value > 0.01 and all  
568 coefficients are positive.

569

#### 570 ***ALDER***

571 To infer the date of East Asian admixture and ancestry proportion in Bengalis (East of India), we  
572 used ALDER (v1.04)<sup>18</sup>. We used the ‘one-reference’ model (*runmode*: 1) with East Asians  
573 (*CHB.DG* from AADR v54) as the reference population with the following parameters: *binsize*:  
574 0.001 Morgans; *maximum distance*: 1.0 Morgans; *zdipcorrmode*: YES; *jackknife*: YES. To  
575 convert the dates of admixture from generations to years, we assume the mean human generation  
576 time was 28 years<sup>67</sup>.

577

#### 578 ***IBD and HBD sharing***

579 We identified IBD and HBD segments using hap-IBD<sup>29</sup> with the following parameters: min-seed:  
580 0.5; max-gap: 1000; min-extend: 0.5; min-output: 1.0; min-markers: 100; min-mac: 2;  
581 nthreads: 1. We used the HapMap genetic maps. To minimize false positives, we only considered  
582 shared segments with length greater at 2cM. Then, we filtered out segments that overlapped  
583 centromeres (using the GRCh38/hg38 annotation from genome.ucsc.edu/cgi-bin/hgTables). To  
584 infer the putative degree of relatedness between two individuals, we computed the total IBD  
585 sharing for *k*th degree cousins using  $2G(1/2)^{2(k+1)}$ , where *G* = 6,782cM is the total diploid

586 autosomal genome size<sup>68</sup> and  $k$  represents the degree of cousin relationship<sup>69</sup>. We note, however,  
587 the expected values assume a random mating population and a history of founder events could  
588 lead to increased genomic sharing and thus these values should be interpreted with caution.

589

#### 590 ***Loss of function (LoF)/missense variants***

591 To quantify the mutational burden in India, we used the Variant Effect Predictor (VEP; version  
592 105)<sup>70</sup> and LOFTEE (v1.0.3)<sup>9</sup> to identify missense and predicted loss-of-function (pLoF) single  
593 nucleotide variants (SNVs). VEP annotates each SNV according to its functional effect on gene  
594 transcripts. We used GENCODE<sup>71</sup> as the transcript annotation reference and focused our analysis  
595 on the most severe functional effect per SNV across different transcripts. Besides the functional  
596 annotations directly obtained from VEP, we identified pLoF SNVs by coupling VEP with  
597 LOFTEE<sup>9</sup>. LOFTEE further assesses stop-gained, splice-site-disrupting, or frameshift SNVs  
598 identified by VEP and implements a set of filters to infer if a SNV should be considered a  
599 pLoF. We intersect the list of pLoF/missense variants with the RefSeq database<sup>34</sup> and the ClinVar  
600 database<sup>36</sup> (data release of 2023-12-17) to infer the nearest gene and any disease associations  
601 respectively. We consider ClinVar status for variants with a review of at least two stars.  
602 Information for each of the pLoF/missense variants is available in Extended Data Table S1.

603

#### 604 ***Inference of archaic ancestry***

605 To learn about the genomic landscape and regional variation in archaic ancestry in Indians and  
606 compare it to worldwide populations, we applied *hmmix*<sup>72</sup> to 2,679 phased individuals from India  
607 (we retain first-degree relatives (except offspring of trios) as they may have archaic ancestry in  
608 different positions). This method uses an outgroup who have negligible amount of archaic  
609 ancestry. We used 426 individuals from the 1000G<sup>11</sup> including Yoruba in Ibadan, Nigeria, Mende  
610 in Sierra Leone (YRI), Esan in Nigeria (ESN) and 64 Africans from HGDP<sup>73</sup>, who have less than  
611 1% West Eurasian admixture, including Bantu South Africa, Biaka Pygmy, Mbuti Pygmy, San  
612 and Yoruba. We estimated the number of callable sites, the single-nucleotide polymorphism  
613 density (as a proxy for per-window mutation rate) and the number of private variants with  
614 respect to the outgroup individuals in 1-kb windows across the genome. We obtained regions  
615 identified as 'archaic' and compared them to the four published high coverage archaic  
616 genomes—Altai Neanderthal<sup>39</sup>, Chagyrskaya Neanderthal<sup>40</sup>, Vindija Neanderthal<sup>41</sup> and Altai  
617 Denisovan<sup>42</sup> to identify the source of the archaic ancestry (see details in Supplementary Note S6).  
618 We further compared archaic segments previously published for 27,566 individuals from  
619 Iceland<sup>44</sup> that were also inferred using *hmmix*. The datasets and number of individuals per  
620 population used for the analysis of archaic ancestry in non-Africans are reported in Table S6.1.

621

#### 622 ***Inferring the timing of Out-of-African migration (OOA)***

623 We infer the minimum coalescence time for non-African individuals with Sub-Saharan African  
624 individuals from the outgroup ( $n=490$ ). Any systematic difference might indicate a difference in  
625 the timing of the out of Africa migration (OOA) for different populations.

626 *hmmix* classifies the genome into 'modern human' and 'archaic' states. The emission parameters  
627 for the human state is informative about the minimum coalescence time between non-African

628 individuals and Sub-Saharan African individuals.

629 We merge HGDP, 1000G and LASI-DAD dataset and subset to SNPs found in 1240K array<sup>64</sup> and  
630 use ADMIXTURE (v1.3.0)<sup>13</sup> in unsupervised-mode ( $k=2$ ) to estimate Sub-Saharan ancestry. We  
631 remove all individuals with  $> 1\%$  Sub-Saharan ancestry to minimize the effect of recent  
632 gene-flow on the minimum coalescence time estimate. To minimize the effect of archaic ancestry  
633 on the emission parameters for the human state we correct for the amount of high confidence  
634 archaic segments (posterior probability  $> 0.9$ ). To compare coalescence times between HGDP  
635 and LASI-DAD we correct for phasing drop-out rate and the removal of multi-allelic sites.  
636 Assuming a mutation rate of  $0.45e-9$ <sup>58</sup> the emission parameter for the human state can be  
637 converted into a coalescence time.

### 638 **Ethics statement**

639 The Longitudinal Aging Study in India (LASI, <https://lasi-india.org>) is a joint effort by the  
640 Harvard T.H. Chan School of Public Health (HSPH), the International Institute for Population  
641 Sciences (IIPS) in India, and the University of Southern California (USC). Longitudinal Aging  
642 Study in India - Diagnostic Assessment of Dementia (LASI-DAD) is an in-depth study of  
643 late-life cognition and dementia, drawing a subsample of the LASI. Principal Investigators teams  
644 are located at USC and All India Institute Of Medical Sciences (AIIMS). Interviews and  
645 sampling were conducted in collaboration with the Regional Geriatric Centers (RGCs) at the  
646 respondents homes or at the participating hospitals, reaching out to both rural and urban areas in  
647 18 states across the country, representing the nation-wide diversity. The AIIMs in New Delhi,  
648 India coordinated field work across RGCs to recruit interviewers and provide training and  
649 logistical support to uniformly perform phenotyping across diverse regions across India. The lists  
650 of partner hospitals and field team members are accessible at <https://lasi-dad.org/teams>.

651 Ethics approval was obtained from the Indian Council of Medical Research and all collaborating  
652 institutions. The study was approved by Institutional Review Boards at the University of  
653 Southern California and the University of Michigan. Informed consent was obtained from all  
654 participants or their legal representative. As most individuals in this study are 60 years or older,  
655 some participants were cognitively impaired, in which case we obtained informed consent from a  
656 close family member, such as a spouse or adult child who was the legal representative of the  
657 participant. The consent materials were translated into as many local languages as necessary.  
658 Informed consent and interviews were collected and conducted in the respondent's language. If  
659 the participant was unable to read the consent forms, the interviewer would verbally relay the  
660 information in the consent form. Participants who were unable to sign the consent forms had the  
661 option to use their thumb impression in place of a signature (a common practice in India).

662

663 DNA extraction and whole genome sequencing was performed at MedGenome, Bangalore,  
664 India. Anonymised data is available for the larger research community through a secured website  
665 hosted by the Gateway to Global Aging Data platform. Research findings from the LASI-DAD



666 team are disseminated through journal publications and presentations at professional  
667 conferences.

## 668 **Data availability**

669 All data is available through the National Institute on Aging Genetics of Alzheimer's Disease  
670 Data Storage Site (NIAGADS) under the accession NG00148.v1. The post-qc vcf file is  
671 distributed by the Genome Center for Alzheimer's Disease (GCAD) at the University of  
672 Pennsylvania and can be obtained by following the data request instructions available:  
673 <https://dss.niagads.org/documentation/data-application-and-submission/application-instructions/>

## 674 **Acknowledgements**

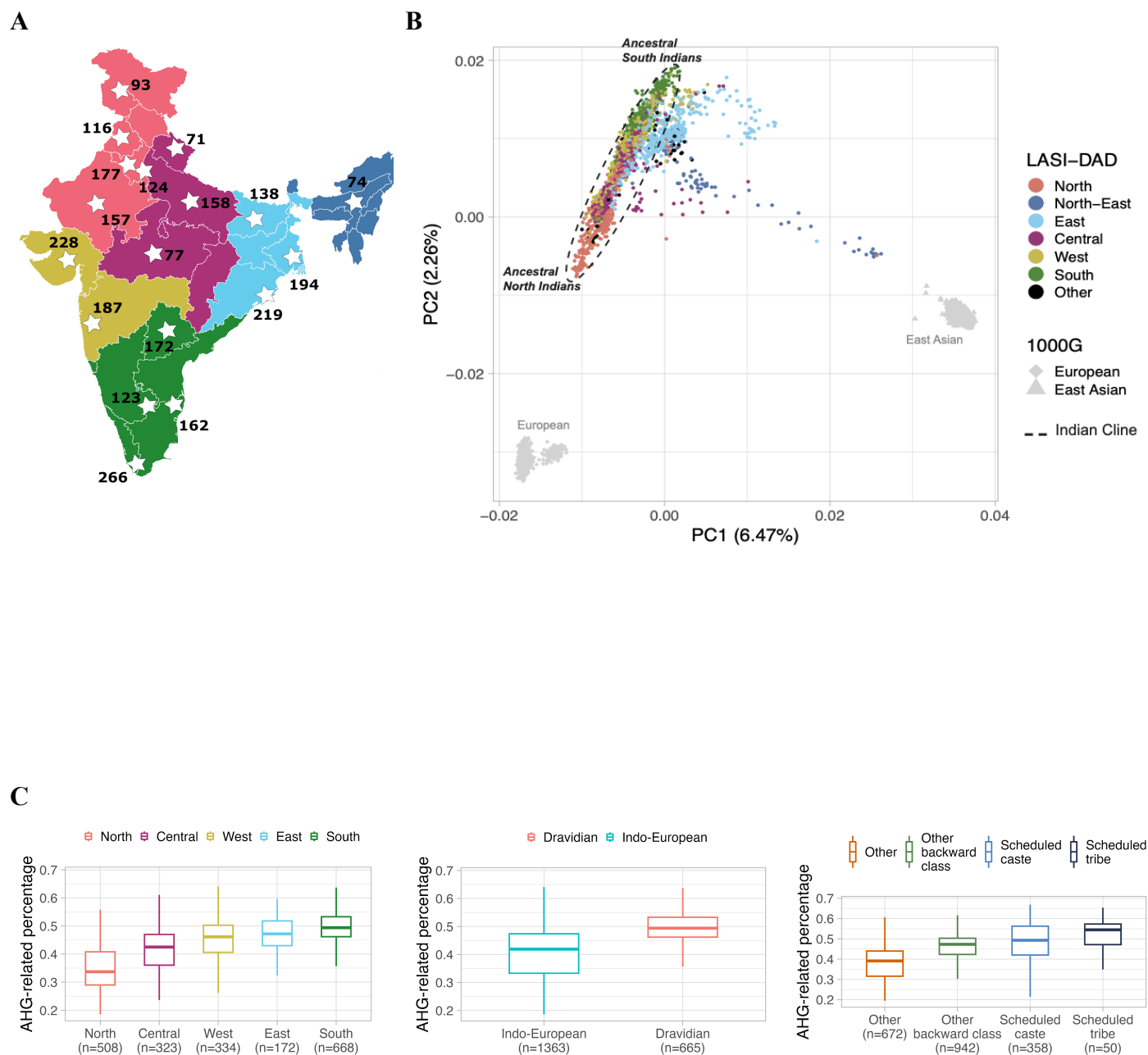
675 We thank Vagheesh Narasimhan, Michael Frachetti and J. Mark Kenoyer for helpful discussion  
676 about the archaeological connections between Sarazm, Iran and South Asia. We also thank  
677 members of the LASI-DAD advising committee and Moorjani lab for helpful feedback on the  
678 analysis and results throughout the project. We thank Yulin Zhang, Elena Zavala, Monty Slatkin,  
679 Ben Peter, Vagheesh Narasimhan and Shai Carmi for helpful comments on the manuscript. The  
680 Longitudinal Aging Study in India, Diagnostic Assessment of Dementia data  
681 ([doi.org/10.25549/5hhx-s820](https://doi.org/10.25549/5hhx-s820)) is sponsored by the National Institute on Aging (grant number  
682 R01AG051125, RF1AG055273, U01AG065958) and is conducted by the University of Southern  
683 California. PM was supported by U01AG065958 and NIH R35GM142978.

684

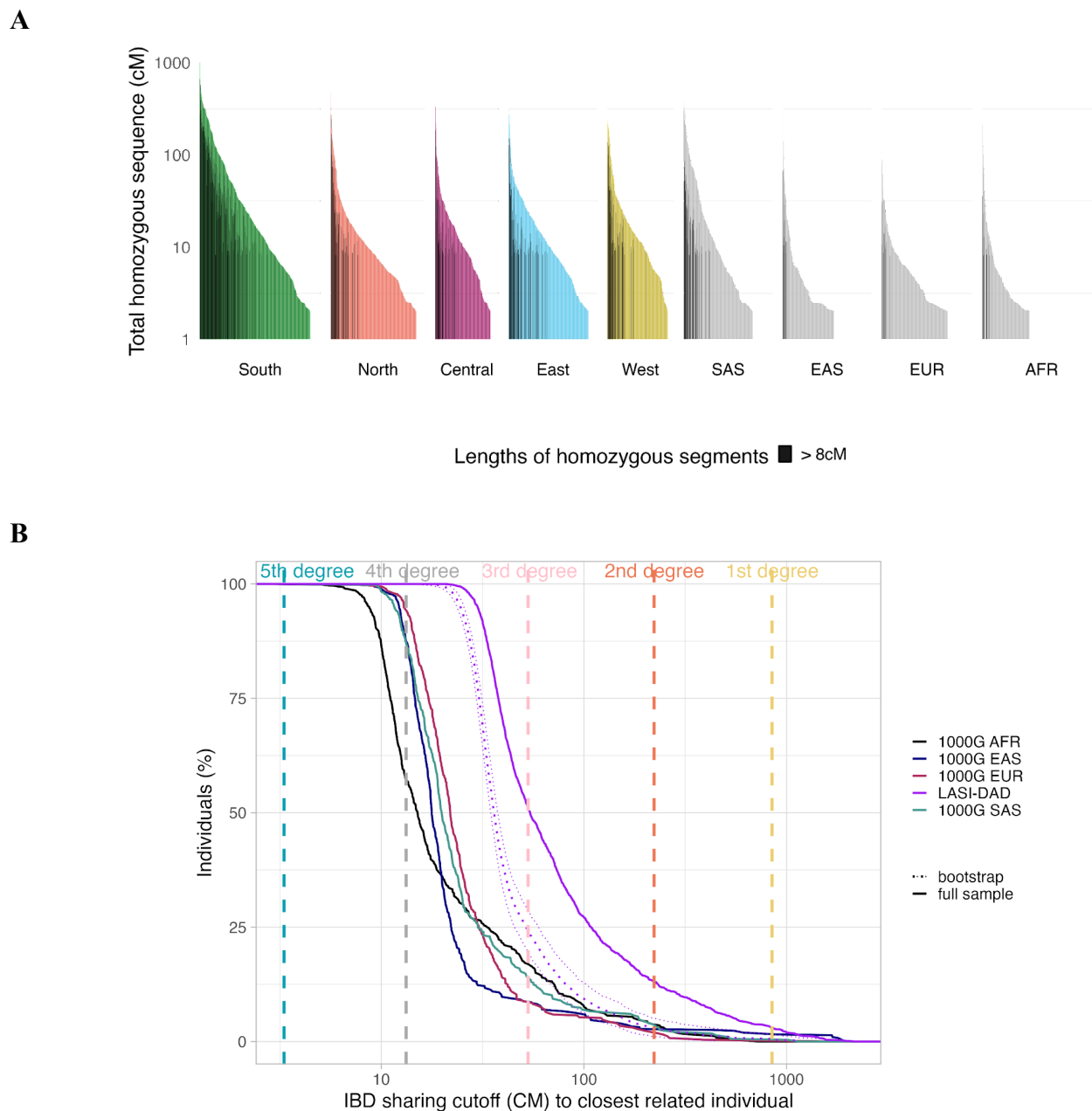
## 685 **Competing interests**

686

687 The authors declare no competing interests.

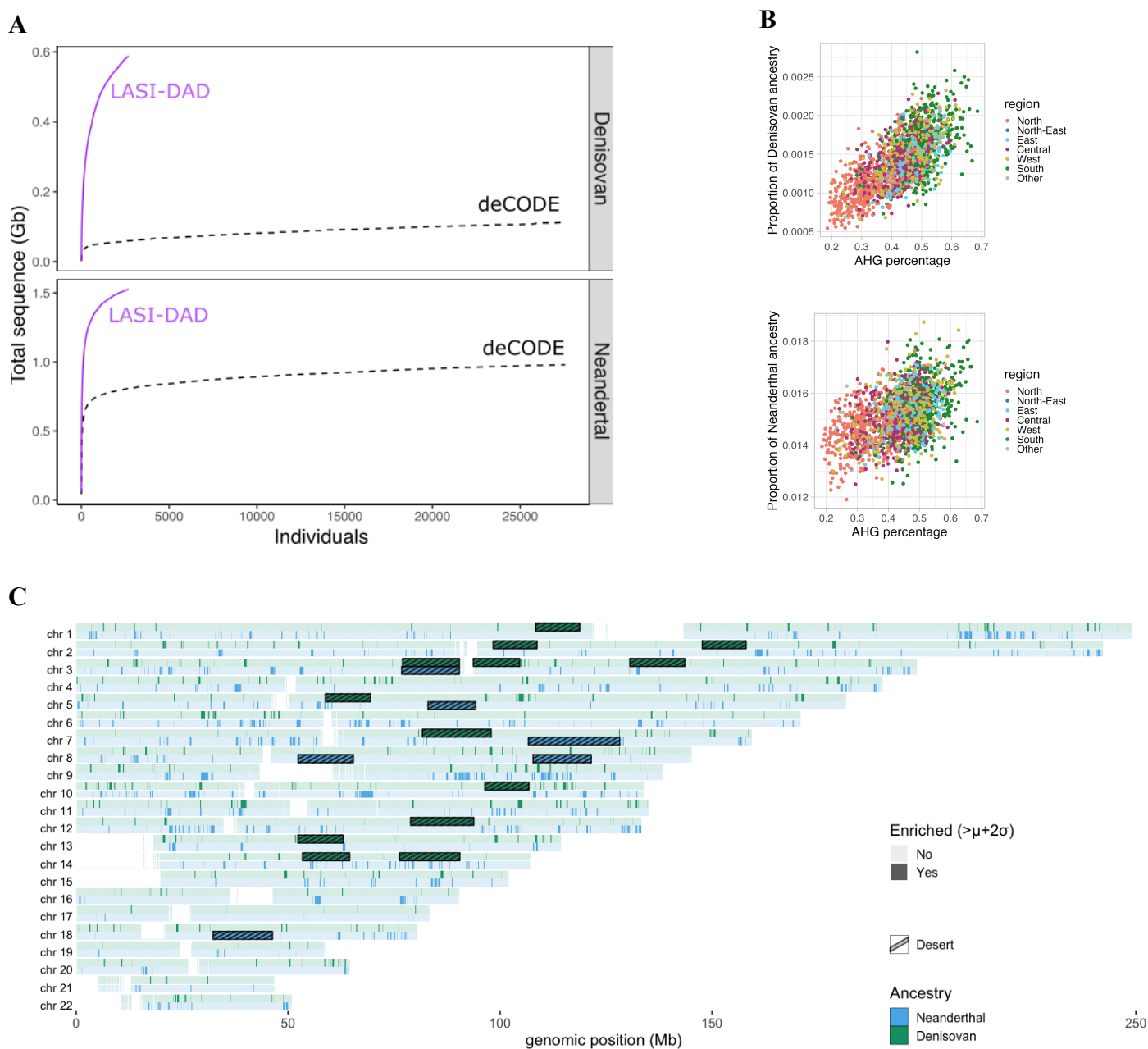


**Figure 1 Population structure and admixture in India.** (A) We show the sampling locations of individuals in the LASI-DAD study. States are colored by region (North, North-east, Central, South, East and West) used for analysis. (B) We ran Principal component analysis (PCA) for Indians in LASI-DAD and 1000G individuals of European (EUR), East Asian (EAS) and South Asian (SAS) ancestry. We show the projection of the first two principal components, colored by region of birth. (C) Using *qpAdm*, we inferred the ancestry proportions for each individual on the ‘Indian cline’ using *Sarazm\_EN* as a proxy for Iranian farmer-related, *Central\_Steppe\_MLBA* as a proxy for Steppe pastoralist-related and *AHG (Onge)* as a proxy for *AASI*-related ancestry. We compared *AHG*-related ancestry proportion by region (left), language family (middle), and caste group (right) of each individual.

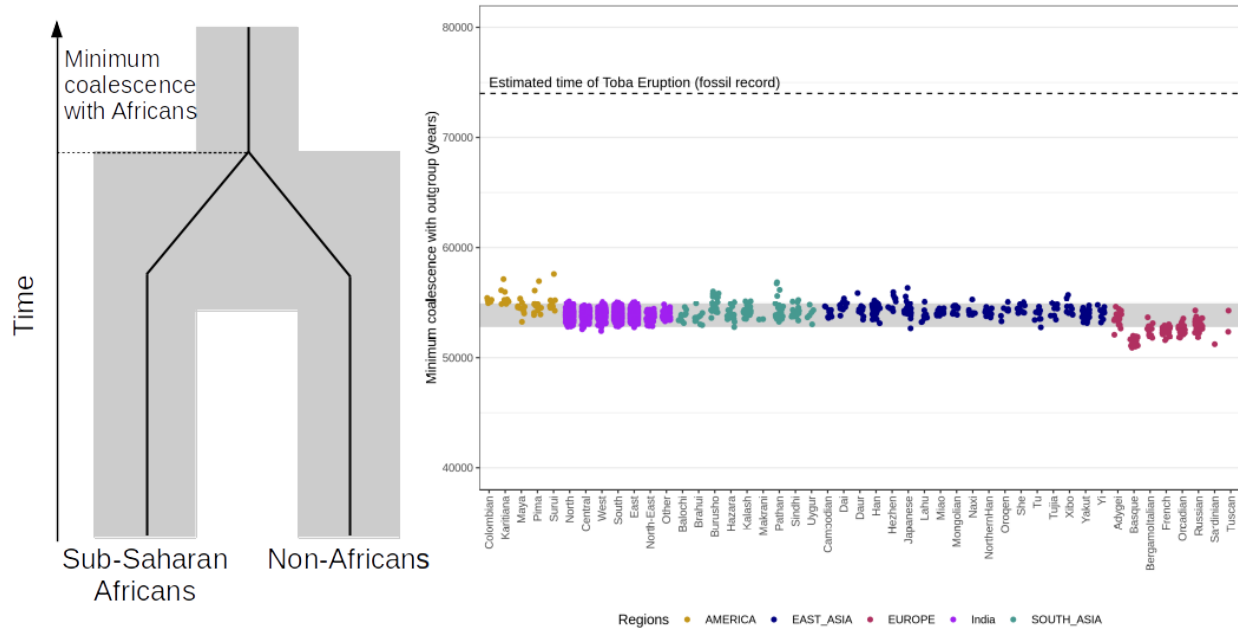


**Figure 2 Founder events and consanguinity leads to high rates of homozygosity and relatedness in Indians.** (A) We applied hap-IBD to infer genome-wide homozygosity in LASI-DAD samples grouped per region and compared with other world-wide groups: East Asian, European, and South Asian populations from 1000G. Black lines show the total amount of homozygous segments longer than 8cM per individual, and colored lines the total amount of homozygous segments shorter than 8cM. (B) For each of the 2,620 Indian samples and AFR, EAS, EUR and SAS individuals in 1000G, we detected the individual sharing the largest total amount (in cM) of genome IBD, referred to as ‘closest individual’. For each value  $x$  of total shared genome (in cM) on the  $X$ -axis, we report the percentage of samples ( $Y$ -axis) that share  $x$  or more with their closest related individual. For LASI-DAD individuals, we also detect the closest individuals while bootstrapping to 500 individuals (dashed lines representing mean and 95% CI). The horizontal dashed lines indicate the expected value of the total IBD sharing for  $k$ th degree cousins. This figure was adapted from <sup>32</sup>.

689



**Figure 3 History of archaic gene flows in India.** (A) Cumulative amount of unique sequence (in Gb) that is either Denisovan (top) or Neanderthal (bottom) as a function of number of individuals, in Indians from LASI-DAD (in purple) and Icelanders from deCODE (in black, dashed). (B) Correlation between *AHG*-related ancestry on the x-axis and total proportion of archaic sequence per individual. Individuals are colored according to which region of origin. We show the correlation for Denisovan (top,  $r=0.49$ ,  $p$ -value  $< 10^{-15}$ ) and Neanderthal (bottom,  $r=0.23$ ,  $p$ -value  $< 10^{-15}$ ). (C) Distribution of archaic ancestry regions across the genome. We computed the mean archaic frequency along the genome of LASI-DAD individuals and considered segments with an archaic frequency higher than the mean ( $\mu$ ) + two standard deviations ( $\sigma$ ) as enriched. We detected 117.28 Mb enriched in Neanderthal ancestry (in blue) and 61.52 Mb enriched in Denisovan ancestry (in green). We also show the location of archaic ancestry deserts: regions with  $< 0.1\%$  archaic ancestry over 10 Mb (striped rectangles in blue for Neanderthal and green for Denisovan).

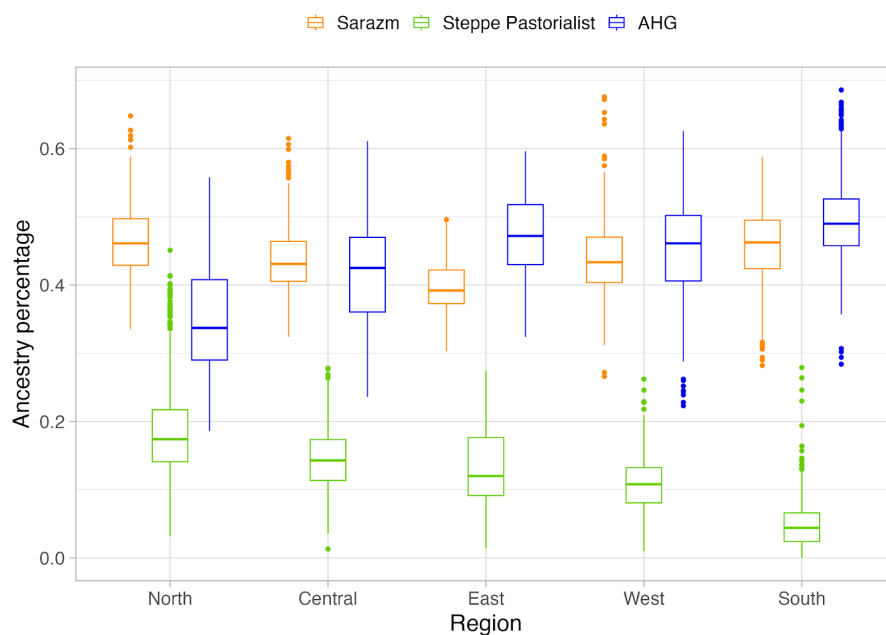


690

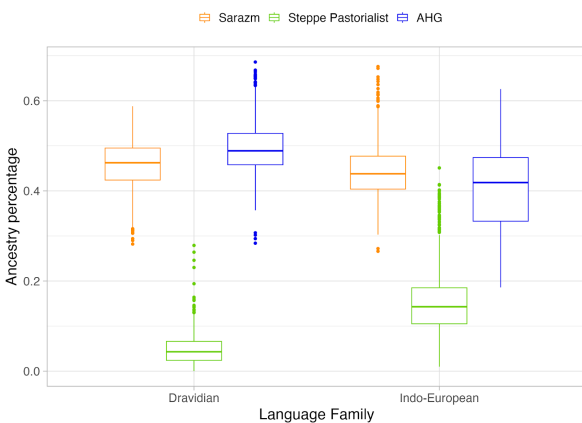
691 **Figure 4 Minimum coalescence time with Sub-Saharan African populations.** Each dot represents the  
692 minimum coalescence time with Sub-Saharan Africans estimated from the emission parameters of the  
693 human state using *hmmix*. The X-axis shows the population the individual belongs to and the color  
694 represents the region. The gray area represents 95% of the coalescence times for all non-African  
695 individuals. The dotted line shows the timing of the Toba eruption 74,000 years ago<sup>57</sup> which provides a  
696 minimum bound for the Southern Dispersal out of Africa.

## 697 Extended Data Figures

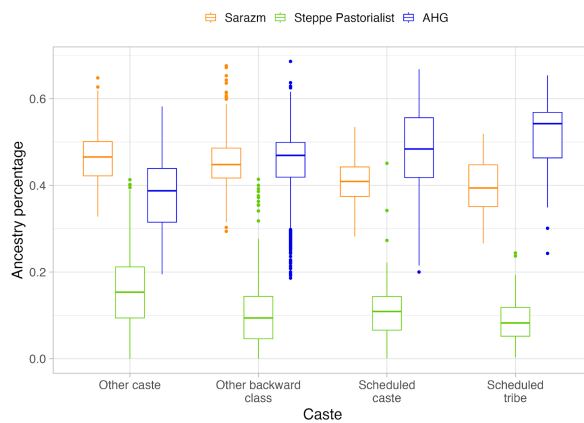
A



B

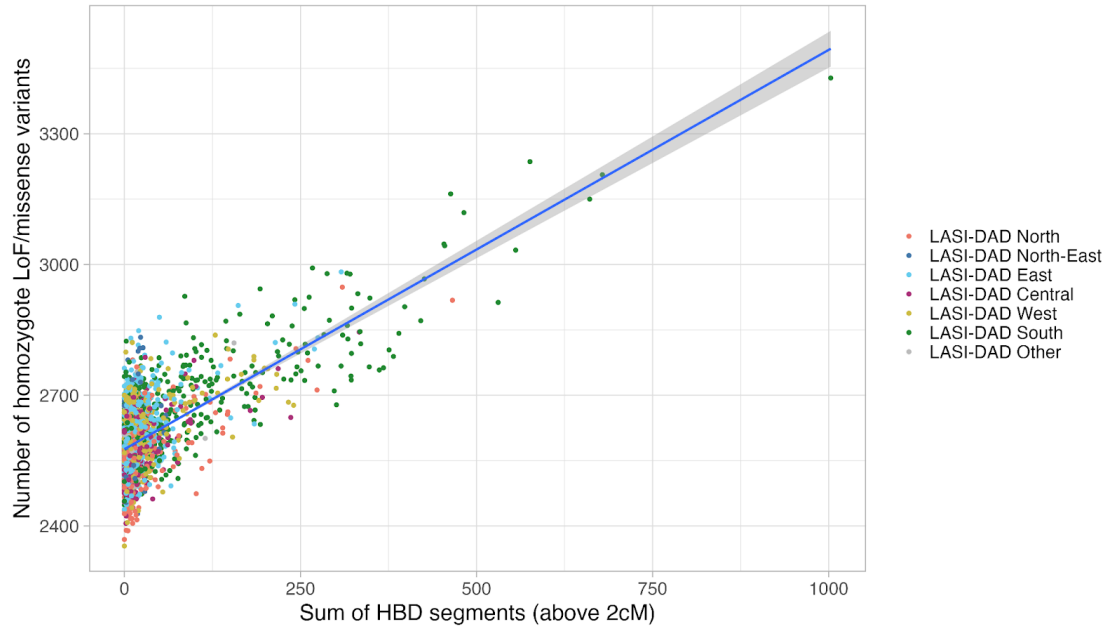


C



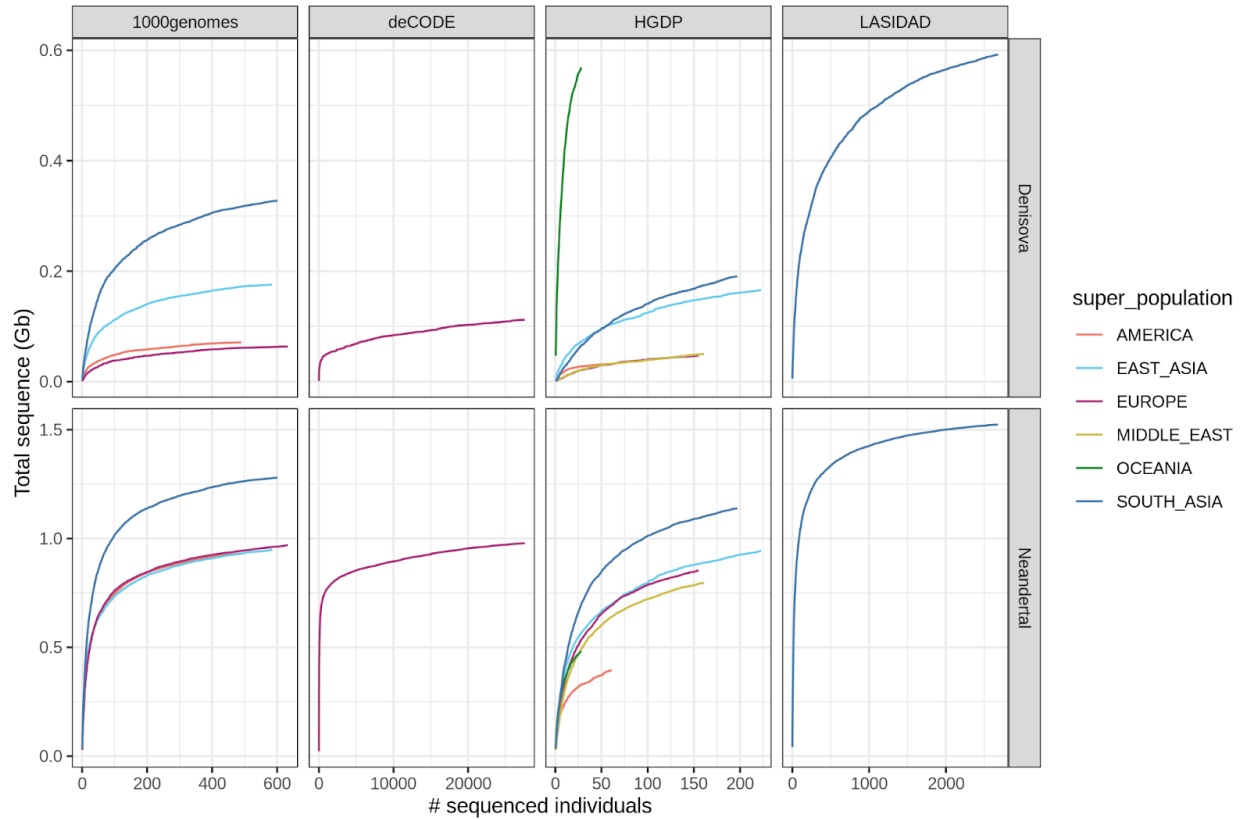
**Extended Data Figure 1.** Ancestral population-related coefficients using the revised model. Inferred coefficients based on qpAdm using the three-way model with Sarazm\_EN, Central\_Steppe\_MLBA and AHG-related groups shown by (A) region, (B) language family and (C) caste group. We show only results for 1,942 individuals for whom the three-way model was a good fit ( $p$ -value > 0.01 and inferred ancestry proportions were non-negative).

698



**Extended Data Figure 2.** Relationship between the number of homozygous derived missense/pLoFs and the total sum of HBD segments per individual. Individuals are colored by region of birth. We fit a regression using generalized linear model (glm) and obtain the following fit:  $y = 2576 + 0.916 \cdot x$ .

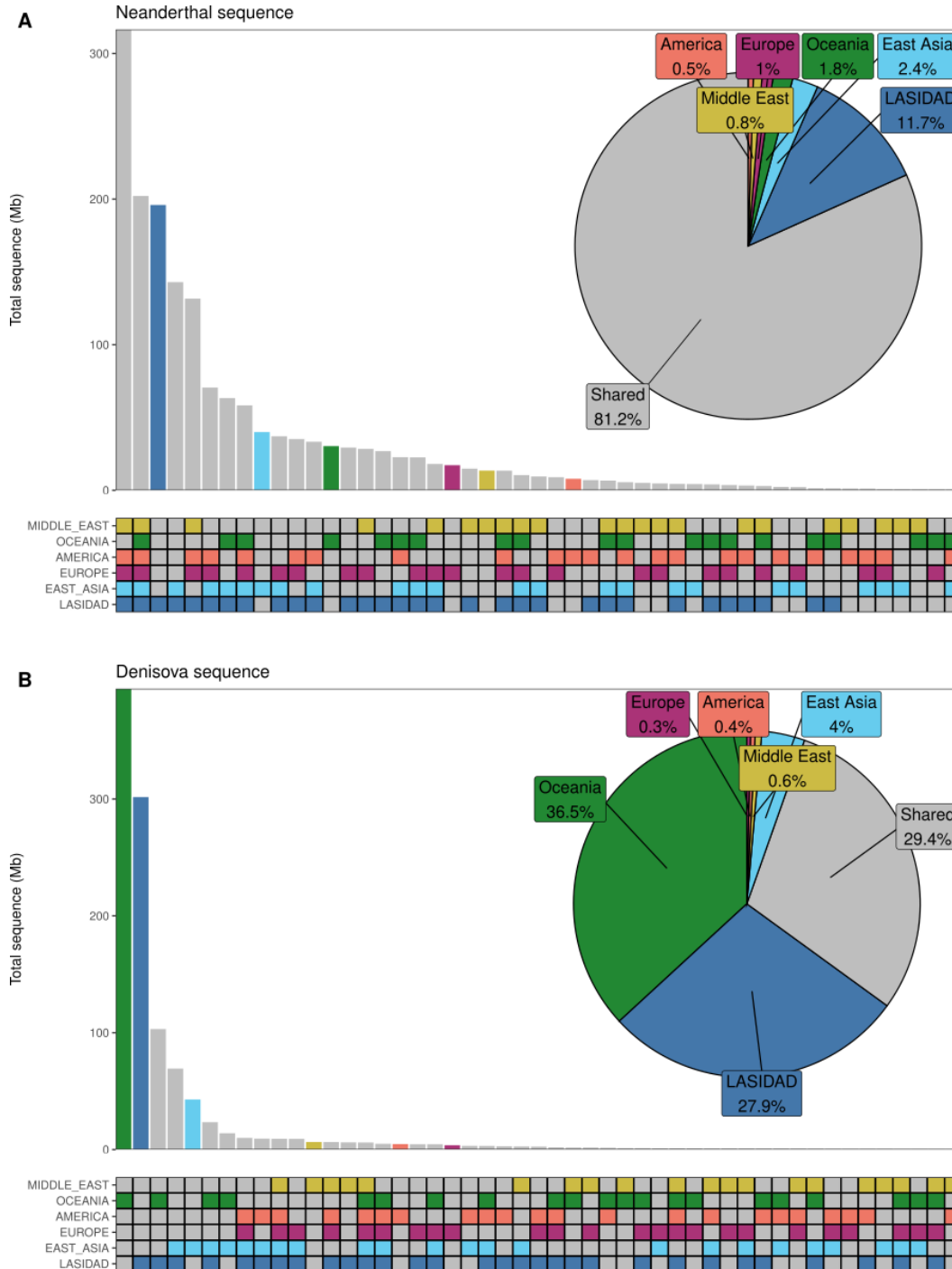
699



700

701 **Extended Data Figure 3. Amount of unique archaic sequence in worldwide populations.** For  
702 Denisovan (top) and Neanderthal (bottom) as a function of the analyzed number of individuals in  
703 four different datasets (at a posterior cutoff of 0.9).

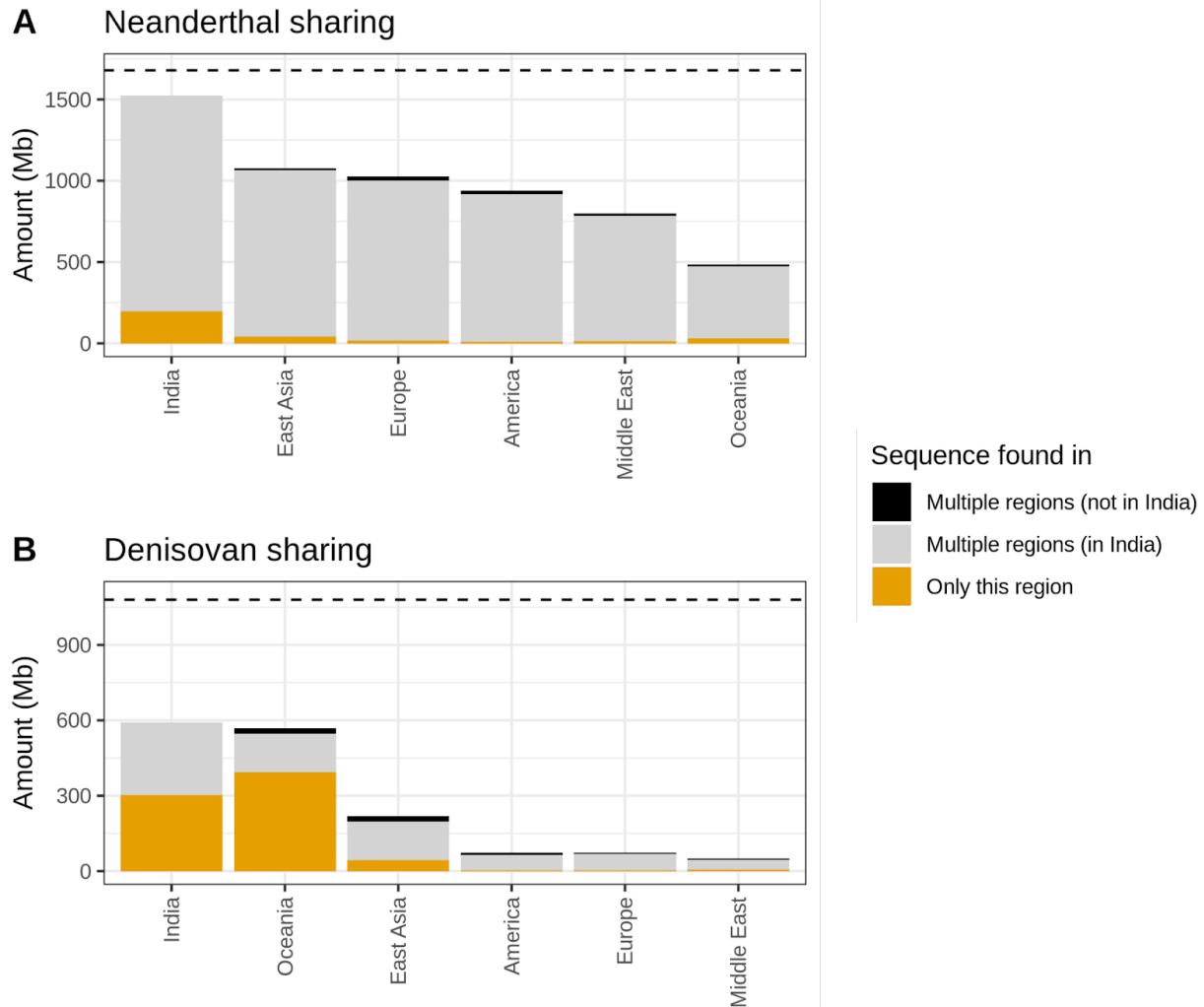




704

705 **Extended Data Figure 4. Sharing of Neanderthal and Denisova sequence.** **A)** Upset plot of  
 706 Neanderthal sequence found at a posterior probability cutoff at 0.8 (y-axis) that is shared  
 707 between any combinations of regions (x-axis). Sequence that is unique to one region is colored  
 708 according to which population it is found in while sequence that is found in at least 2 populations  
 709 (shared) is colored in grey. In the pie chart the total amount of shared and unique are denoted in  
 710 percent. **B)** same as **A)** but for Denisovan sequence.

711



712

713

714 **Extended Data Figure 5. Neanderthal and Denisova sequence found in world-wide regions.**

715 **A)** Amount of Neanderthal sequence found at a posterior probability cutoff at 0.8 (y-axis) that is  
716 unique to any region, found in multiple regions (at least two) where one includes India  
717 (LASI-DAD dataset) or found in multiple regions (at least two) where India is not included. **B)**  
718 same as **A)** but for Denisovan sequence. Horizontal lines indicate the total length of the  
719 assembled Neanderthal and Denisova genome using LASI-DAD, HGDP and 1000G datasets.

## 720 References

721

- 722 1. Mastana, S. S. Unity in diversity: an overview of the genomic anthropology of India. *Ann. Hum.*  
723 *Biol.* **41**, 287–299 (2014).
- 724 2. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- 725 3. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**,  
726 203–209 (2018).
- 727 4. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*  
728 **590**, 290–299 (2021).
- 729 5. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations.  
730 *Nature* **538**, 201–206 (2016).
- 731 6. GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across  
732 Asia. *Nature* **576**, 106–111 (2019).
- 733 7. Wall, J. D. *et al.* South Asian medical cohorts reveal strong founder effects and high rates of  
734 homozygosity. *Nat. Commun.* **14**, 3377 (2023).
- 735 8. Lee, J., Banerjee, J., Khobragade, P. Y., Angrisani, M. & Dey, A. B. LASI-DAD study: a protocol for  
736 a prospective cohort study of late-life cognition and dementia in India. *BMJ Open* **9**, e030300  
737 (2019).
- 738 9. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456  
739 humans. *Nature* **581**, 434–443 (2020).
- 740 10. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable  
741 and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
- 742 11. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes  
743 Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
- 744 12. Reich, D., Price, A. L. & Patterson, N. Principal component analysis of genetic data. *Nature genetics*  
745 vol. 40 491–492 (2008).
- 746 13. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated

- 747 individuals. *Genome Res.* **19**, 1655–1664 (2009).
- 748 14. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
- 749 15. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population  
750 history. *Nature* **461**, 489–494 (2009).
- 751 16. Moorjani, P. *et al.* Genetic evidence for recent population mixture in India. *Am. J. Hum. Genet.* **93**,  
752 422–438 (2013).
- 753 17. Narasimhan, V. M. *et al.* The formation of human populations in South and Central Asia. *Science*  
754 **365**, (2019).
- 755 18. Loh, P.-R. *et al.* Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium.  
756 *Genetics* **193**, 1233–1254 (2013).
- 757 19. Kumar, A. The Huns (‘Hūnas’) in India: A review. *Stud. People’s Hist.* **8**, 182–196 (2021).
- 758 20. Karlsson, E. K. *et al.* Natural selection in a bangladeshi population from the cholera-endemic ganges  
759 river delta. *Sci. Transl. Med.* **5**, 192ra86 (2013).
- 760 21. Das, B. M., Das, P. B., Das, R., Walter, H. & Danker-Hopfe, H. Anthropological studies in Assam,  
761 India. 7. Socio-biological observations on Assamese populations. *Anthropol. Anz.* **47**, 129–143  
762 (1989).
- 763 22. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in  
764 Europe. *Nature* **522**, 207–211 (2015).
- 765 23. de Barros Damgaard, P. *et al.* The first horse herders and the impact of early Bronze Age steppe  
766 expansions into Asia. *Science* **360**, (2018).
- 767 24. Spengler, R. N. & Willcox, G. Archaeobotanical results from Sarazm, Tajikistan, an Early Bronze  
768 Age Settlement on the edge: Agriculture and exchange. *Environ. Archaeol.* **18**, 211–221 (2013).
- 769 25. Frachetti, M. D. Multiregional Emergence of Mobile Pastoralism and Nonuniform Institutional  
770 Complexity across Eurasia. *Curr. Anthropol.* (2012) doi:10.1086/663692.
- 771 26. Tournebise, R., Chu, G. & Moorjani, P. Reconstructing the history of founder events using  
772 genome-wide patterns of allele sharing across individuals. *PLoS Genet.* **18**, e1010243 (2022).

- 773 27. Nakatsuka, N. *et al.* The promise of discovering population-specific disease-associated genes in  
774 South Asia. *Nat. Genet.* **49**, 1403–1407 (2017).
- 775 28. Slatkin, M. A population-genetic test of founder effects and implications for Ashkenazi Jewish  
776 diseases. *Am. J. Hum. Genet.* **75**, 282–293 (2004).
- 777 29. Zhou, Y., Browning, S. R. & Browning, B. L. A Fast and Simple Method for Detecting  
778 Identity-by-Descent Segments in Large-Scale Data. *Am. J. Hum. Genet.* **106**, 426–437 (2020).
- 779 30. Acharya, S. & Sahoo, H. Consanguineous Marriages in India: Prevalence and Determinants. *J.*  
780 *Health Manag.* (2021) doi:10.1177/09720634211050458.
- 781 31. Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E. & Leutenegger, A.-L. High level of inbreeding in  
782 final phase of 1000 Genomes Project. *Sci. Rep.* **5**, 17453 (2015).
- 783 32. Nait Saada, J. *et al.* Identity-by-descent detection across 487,409 British samples reveals fine scale  
784 population structure and ultra-rare variant associations. *Nat. Commun.* **11**, 6130 (2020).
- 785 33. Lim, E. T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder  
786 population. *PLoS Genet.* **10**, e1004494 (2014).
- 787 34. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic  
788 expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).
- 789 35. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding  
790 genes. *Science* **335**, 823–828 (2012).
- 791 36. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence.  
792 *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
- 793 37. Manoharan, I., Wieseler, S., Layer, P. G., Lockridge, O. & Boopathy, R. Naturally occurring  
794 mutation Leu307Pro of human butyrylcholinesterase in the Vysya community of India.  
795 *Pharmacogenet. Genomics* **16**, 461–468 (2006).
- 796 38. Witt, K. E., Villanea, F., Loughran, E., Zhang, X. & Huerta-Sanchez, E. Apportioning archaic  
797 variants among modern populations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **377**, 20200411 (2022).
- 798 39. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*

- 799 **505**, 43–49 (2014).
- 800 40. Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl. Acad.*  
801 *Sci. U. S. A.* **117**, 15132–15136 (2020).
- 802 41. Prüfer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**,  
803 655–658 (2017).
- 804 42. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science*  
805 **338**, 222–226 (2012).
- 806 43. Coll Macià, M., Skov, L., Peter, B. M. & Schierup, M. H. Different historical generation intervals in  
807 human populations inferred from Neanderthal fragment lengths and mutation signatures. *Nat.*  
808 *Commun.* **12**, 5317 (2021).
- 809 44. Skov, L. *et al.* The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes.  
810 *Nature* **582**, 78–83 (2020).
- 811 45. Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of Human Sequence  
812 Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173**, 53–61.e9 (2018).
- 813 46. Mondal, M. *et al.* Genomic analysis of Andamanese provides insights into ancient human migration  
814 into Asia and adaptation. *Nat. Genet.* **48**, 1066–1070 (2016).
- 815 47. Racimo, F., Marnetto, D. & Huerta-Sánchez, E. Signatures of Archaic Adaptive Introgression in  
816 Present-Day Human Populations. *Mol. Biol. Evol.* **34**, 296–317 (2017).
- 817 48. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The Combined Landscape of Denisovan  
818 and Neanderthal Ancestry in Present-Day Humans. *Curr. Biol.* **26**, 1241–1247 (2016).
- 819 49. Witt, K. E., Funk, A., Añorve-Garibay, V., Fang, L. L. & Huerta-Sánchez, E. The Impact of Modern  
820 Admixture on Archaic Human Ancestry in Human Populations. *Genome Biol. Evol.* **15**, (2023).
- 821 50. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like  
822 DNA. *Nature* **512**, 194–197 (2014).
- 823 51. Pagani, L. *et al.* Genomic analyses inform on migration events during the peopling of Eurasia.  
824 *Nature* **538**, 238–242 (2016).

- 825 52. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78  
826 (2010).
- 827 53. Zeberg, H. & Pääbo, S. A genomic region associated with protection against severe COVID-19 is  
828 inherited from Neandertals. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
- 829 54. Wang, F. *et al.* COVID-19 Host Genetics Initiative The COVID-19 Host Genetics Initiative, a global  
830 initiative to elucidate the role of host genetic factors in susceptibility and severity of the  
831 SARS-CoV-2 virus pandemic. *Eur. J. Hum. Genet.* **28**, 715–718 (2020).
- 832 55. Vernot, B. *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian  
833 individuals. *Science* **352**, 235–239 (2016).
- 834 56. Chen, L., Wolf, A. B., Fu, W., Li, L. & Akey, J. M. Identifying and Interpreting Apparent  
835 Neanderthal Ancestry in African Individuals. *Cell* **180**, 677–687.e16 (2020).
- 836 57. Clarkson, C. *et al.* Human occupation of northern India spans the Toba super-eruption ~74,000 years  
837 ago. *Nat. Commun.* **11**, 961 (2020).
- 838 58. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from  
839 Iceland. *Nature* **549**, 519–522 (2017).
- 840 59. Thangaraj, K. *et al.* Reconstructing the origin of Andaman Islanders. *Science* **308**, 996 (2005).
- 841 60. Akhilesh, K. *et al.* Early Middle Palaeolithic culture in India around 385–172 ka reframes Out of  
842 Africa models. *Nature* **554**, 97–101 (2018).
- 843 61. Blinkhorn, J., Achyuthan, H., Durcan, J., Roberts, P. & Ilgner, J. Constraining the chronology and  
844 ecology of Late Acheulean and Middle Palaeolithic occupations at the margins of the monsoon. *Sci.*  
845 *Rep.* **11**, 19665 (2021).
- 846 62. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies.  
847 *Bioinformatics* **26**, 2867–2873 (2010).
- 848 63. Visscher, P. M. *et al.* Assumption-free estimation of heritability from genome-wide  
849 identity-by-descent sharing between full siblings. *PLoS Genet.* **2**, e41 (2006).
- 850 64. Mallick, S. *et al.* The Allen Ancient DNA Resource (AADR): A curated compendium of ancient

- 851 human genomes. *bioRxiv* (2023) doi:10.1101/2023.04.06.535797.
- 852 65. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide  
853 association studies. *Nat. Genet.* **38**, 904–909 (2006).
- 854 66. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLoS Genet.* **2**, e190  
855 (2006).
- 856 67. Moorjani, P. *et al.* A genetic method for dating ancient genomes provides a direct estimate of human  
857 generation interval in the last 45,000 years. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 5652–5657 (2016).
- 858 68. Halldorsson, B. V. *et al.* Characterizing mutagenic effects of recombination through a sequence-level  
859 genetic map. *Science* **363**, (2019).
- 860 69. Gusev, A. *et al.* The architecture of long-range haplotypes shared within and across populations.  
861 *Mol. Biol. Evol.* **29**, 473–486 (2012).
- 862 70. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
- 863 71. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project.  
864 *Genome Res.* **22**, 1760–1774 (2012).
- 865 72. Skov, L. *et al.* Detecting archaic introgression using an unadmixed outgroup. *PLoS Genet.* **14**,  
866 e1007641 (2018).
- 867 73. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse  
868 genomes. *Science* **367**, (2020).

869