



30

## Abstract

31 Surrounded by speakers of Indo-European, Dravidian and Tibeto-Burman languages,  
32 around 11 million Munda (a branch of Austroasiatic language family) speakers live in  
33 the densely populated and genetically diverse South Asia. Their genetic makeup  
34 holds components characteristic of South Asians as well as Southeast Asians. The  
35 admixture time between these components has been previously estimated on the  
36 basis of archaeology, linguistics and uniparental markers. Using genome-wide  
37 genotype data of 102 Munda speakers and contextual data from South and  
38 Southeast Asia, we retrieved admixture dates between 2000 – 3800 years ago for  
39 different populations of Munda. The best modern proxies for the source populations  
40 for the admixture with proportions 0.78/0.22 are Lao people from Laos and Dravidian  
41 speakers from Kerala in India, while the South Asian population(s), with whom the  
42 incoming Southeast Asians intermixed, had a smaller proportion of West Eurasian  
43 component than contemporary proxies. Somewhat surprisingly Malaysian Peninsular  
44 tribes rather than the geographically closer Austroasiatic languages speakers like  
45 Vietnamese and Cambodians show highest sharing of IBD segments with the Munda.  
46 In addition, we affirmed that the grouping of the Munda speakers into North and  
47 South Munda based on linguistics is in concordance with genome-wide data.

48

## Introduction

49 Genetically diverse<sup>1-3</sup> South Asia is home to more than a billion people who belong  
50 to thousands of distinct socio-culturally or ethnically defined population groups.  
51 These groups speak languages of four major language families: Indo-European,  
52 Dravidian, Austroasiatic and Trans-Himalayan. Studies based on genome-wide  
53 genotype data have shown that the majority of present day populations of the Indian  
54 subcontinent derive their genetic ancestry to a large extent from two ancestral  
55 populations – ancestral northern and southern Indians – of which the former is  
56 genetically close to West Eurasian populations<sup>4-6</sup>. In addition to these two  
57 components, the Munda speakers of the Austroasiatic family share a minor  
58 proportion of their genetic ancestry with Southeast Asian populations<sup>7</sup>. Austroasiatic  
59 languages are spoken by more than 100 million people in Mainland Southeast Asia  
60 (MSEA) and >10 million Austroasiatic speakers<sup>8</sup> of Munda languages live in East and  
61 Central parts of India where they are surrounded by Indo-European, Dravidian and  
62 Trans-Himalayan languages speakers.

63 Considering the widespread sharing of words related to rice agriculture in all main  
64 branches of Austroasiatic, it has been proposed that this language family co-  
65 expanded with farming in MSEA and that the speakers of Munda languages spread  
66 to India as part of this farming expansion<sup>9,10</sup>. Alternatively, considering the deep splits  
67 of extant Munda and extinct Para-Munda languages and evidence for independent  
68 domestication of rice in India and in Southeast Asia, it has been proposed that  
69 Austroasiatic languages could have, instead, spread from India to Southeast Asia<sup>11</sup>.  
70 Given that about 25% of the genetic ancestry of Munda speakers has been shown to  
71 be shared with Southeast Asians, unlike in other Indian populations, and, reversely,  
72 because Austroasiatic speakers of Myanmar share some ancestry (~16%) with

73 Indian populations, it has been proposed that the expansion of rice farming may have  
74 involved bilateral movement of people<sup>7</sup>.

75 Studies analysing mtDNA and Y chromosome markers have revealed a sex-specific  
76 admixture pattern of admixture of Southeast and South Asian ancestry components  
77 for Munda speakers. While close to 100% of mtDNA lineages present in Mundas  
78 match those in other Indian populations, around 65% of their paternal genetic  
79 heritage is more closely related to Southeast Asian than South Asian variation<sup>7,12,13</sup>.

80 Such a contrasting distribution of maternal and paternal lineages among the Munda  
81 speakers is a classic example of 'father tongue hypothesis'<sup>14</sup>. However, the  
82 temporality of this expansion is contentious<sup>7,13,15,16</sup>. Based on Y-STR data the  
83 coalescent time of Indian O2a-M95 haplogroup was estimated to be >10 KYA<sup>7,13</sup>.

84 Recently, the reconstructed phylogeny of 8.8 Mb region of Y chromosome data  
85 showed that Indian O2a-M95 lineages coalesce within a clade nested within  
86 East/Southeast Asian within the last ~5-7 KYA<sup>17</sup>. This date estimate sets the upper  
87 boundary for the main episode of gene flow of Y chromosomes from Southeast Asia  
88 to India.

89 Previous autosomal study was limited to a single Austroasiatic population from  
90 Southeast Asia<sup>7</sup>, therefore in the present study, we generated and assembled large  
91 body of contextual genome-wide genotype data from Southeast Asia as well as from  
92 South Asia (Supplementary Table S1). We set out to affirm the signal of the  
93 admixture event in autosomal data and to address previously unresolved questions  
94 including: i) autosomal date of the South and Southeast Asian admixture event in  
95 Munda; ii) characteristics of the Indian ancestry component of the Mundas; iii) who  
96 are the closest living descendants of the source populations of the ancient admixture;

97 iv) and if the grouping of the Munda speakers into North and South Mundas based on  
98 some linguistic models is supported by genetic data.

99 To address these questions, we analysed 102 individual samples from Munda  
100 speaking populations (including 10 newly reported samples) in context of 978 other  
101 samples (including 46 newly reported samples) from 72 populations mainly from  
102 India, Southeast Asia and East Asia. The Munda speakers are divided into North  
103 Mundas (NM) and South Mundas (SM) based on linguistic affinities. List of all the  
104 populations, sample sizes, and some additional information on the dataset can be  
105 found in Supplementary Table S1.

## 106 **Results and Discussion**

### 107 ***The Munda speakers as an admixed population***

108 We first analysed Munda genomes with ADMIXTURE and PCA in context of other  
109 South and Southeast Asian populations and found that Munda share about three  
110 quarters of their genetic ancestry (k3 – k5 components in Figure 1) with Indian  
111 Dravidian and Indo-European speakers. Interestingly, Indian populations with the k3-  
112 k5 components have also a pink component (k2) which is widespread in European,  
113 West Eurasian, Near Eastern and Pakistani populations but missing in the Munda  
114 speakers. Roughly one quarter of the ancestral components in the Mundas' genome  
115 (k6 – k12) are shared with Southeast Asians. There are two populations with a similar  
116 genetic profile to the Mundas in Central India: Dravidian speaking Gond who are  
117 known to have received a substantial gene flow from the Munda speakers<sup>18</sup> and a  
118 linguistic isolate Nihali.

119 Principal component analysis (PCA) roughly reflects geographical locations of  
120 studied populations (see Supplementary Fig. S4). Based on the first two components

121 of PCA, the Mundas are genetically situated between South Asians and Southeast  
122 Asians and Oceanians. Furthermore, South and North Munda tribes are clearly  
123 different – South Mundas are genetically closer to Southeast Asians and Oceanians  
124 while North Mundas are closer to South Asians. In sum, the results of the  
125 ADMIXTURE and PCA are consistent with the model by which the genetic ancestry  
126 of Indian Munda speakers represents an admixture between Indian and Southeast  
127 Asian ancestries.

128 The scenario of independent evolution without admixture was rejected by 3-  
129 population formal test of admixture<sup>6</sup> for South Munda, Santhal (NM) and Ho (NM)  
130 speakers, as they yielded significantly negative  $f_3$  values (indicative of admixture)  
131 when tested together with populations from India and Southeast Asia  
132 (Supplementary Table S2). Birhor (NM) and Korwa (NM) speakers did not display  
133 significant admixture signal potentially because of the vast genetic drift they have  
134 gone through after the admixture event as they show the lowest average  
135 heterozygosity among the Munda speakers (Supplementary Table S3).

136 To understand further the position of Mundas in the genetic landscape of Indian  
137 populations, we plotted the second and third principal components from the global  
138 PCA analysis (see Supplementary Fig. S5). The Mundas were situated close to the  
139 Dravidian speaking southern Indian end of the gradient, near Pulliyar population from  
140 southwestern India, being stretched towards Southeast Asian populations, the  
141 closest ones being Bateq, Jehai, Kintaq and Mendriq from Malaysia.

#### 142 ***The best contemporary proxies for admixture sources***

143 Three populations that yield the highest outgroup- $f_3$  values as sources of Southeast  
144 Asian ancestry in Munda are Lao from Laos, Dai from China and Murut from Borneo.  
145 From South Asia, the populations that produce the highest  $f_3$  scores are Dravidian

146 speaking Paniya and Pulliyar from Kerala region of India. For North Mundas, among  
147 the top Indian populations is also Indo-European speaking Chamar, whereas for  
148 South Mundas, there are Jarawa and Onge from Andaman Islands (Supplementary  
149 Table S2). Consistently, the South Munda speakers are the biggest DNA chunk  
150 donors from India to the Andamanese populations based of fineSTRUCTURE<sup>19</sup>  
151 analysis (see Supplementary Fig. S7).

152 For a more detailed view of the temporary aspects of admixture, we assessed the  
153 sharing of DNA segments that are identical by descent between Munda speakers and  
154 other populations. Refined IBD analysis<sup>20</sup> showed that from India, Mundas share the  
155 highest number of DNA segments identical by descent (IBD) with Dravidian speaking  
156 Chenchus (1.68; CI: 1.46 – 1.91) and Indo-European speaking Chamar (1.63; CI:  
157 1.26 – 2.11) when disregarding Nihali and Gond tribes as Nihali, a language isolate,  
158 are possibly related to Munda and the Gond are reported to have received gene flow  
159 from the Mundas<sup>18</sup>. From Southeast Asia the sharing is highest with Mah Meri (2.04;  
160 CI: 1.79 – 2.33) and Temuan (1.93; CI: 1.67 – 2.24) tribes from Peninsular Malaysia,  
161 followed by Jakun and Che Wong from the same area (Figure 2, Supplementary  
162 Table S3). Surprisingly, the geographically closer Austroasiatic speakers from  
163 Southeast Asia, such as Cambodians and Vietnamese, do not share as many IBD  
164 segments with the Mundas. This effect could be caused by the fact that the mainland  
165 Southeast Asian populations have smaller proportions of the original Austroasiatic  
166 component in their genomes due to subsequent gene flow received from East Asia.  
167 Another explanation could be a more complex direction of gene flow in this area.  
168 Similar results were observed when using total lengths of shared IBD segments  
169 instead of their counts (Supplementary Figure S9).

170 When dividing the segments shared with the Mundas into two groups, short (<1 cM)  
171 and long (>1 cM), we noticed that the two sources, South Asian and Southeast Asian  
172 populations, clearly form two distinct groups based on shared segment length  
173 patterns (Figure 2). Both, mainland and island Southeast Asian populations share a  
174 high number of long IBD segments with the Mundas while Indian Dravidian and Indo-  
175 European speaking populations share plenty of short IBD segments. Surprisingly, no  
176 difference was found in Indian Dravidian and Indo-European speakers in context of  
177 sharing DNA with the Mundas (Welch's t-test; short IBD  $P = 0.5218$ ; long IBD  $P =$   
178  $0.5302$ ; all IBD  $P = 0.9305$ ). The formation of the two groups seen on Figure 2 could  
179 refer to different genetic distance between admixed populations and other  
180 populations from the corresponding areas; *i.e.*, the Southeast Asian share of the  
181 Munda speakers' genomes has diverged from present day Southeast Asians more  
182 recently than the South Asian part from present day South Asians. This result has to  
183 be taken with caution as we found correlation between the shared IBD segment  
184 lengths and the average heterozygosity in these populations (Supplementary Figure  
185 S8, Supplementary Table S3).

### 186 ***Admixture proportions suggest a novel scenario***

187 We used qpAdm<sup>21</sup> to determine the relative proportions of West, Southeast and  
188 South Asian ancestries in Munda speakers, using a number of modern and ancient  
189 West Asian populations, Lao, and Onge or Paniya as proxies for the three Asian  
190 components (Supplementary Table S4). Regardless of which West Asian population  
191 we used, we found that Munda speakers can be described on average as a mixture  
192 of ~19% Southeast Asian, 15% West Asian and 66% Onge (South Asian)  
193 components. Alternatively, the West and South Asian components of Munda could be  
194 modelled using a single South Asian population (Paniya), accounting on average to



195 77% of the Munda genome. When rescaling the West and South Asian (Onge)  
196 components to 1 to explore the Munda genetic composition prior to the introduction  
197 of the Southeast Asian component, we note that the West Asian component is lower  
198 (~19%) in Munda compared to Paniya (27%) (Supplementary Table S4:  
199 \*Average\_Lao=0). Consistently with qpGraph analyses in Narasimhan et al. (2018)<sup>22</sup>,  
200 this may point to an initial admixture of a Southeast Asian substrate with a South  
201 Asian substrate free of any West Asian component, followed by the encounter of the  
202 resulting admixed population with a Paniya-like population. Such a scenario would  
203 imply an inverse relationship between the Southeast and West Asian relative  
204 proportions in Munda or, in other words, the increase of Southeast Asian component  
205 should cause a greater reduction of the West Asian compared to the reduction in the  
206 South Asian component in Munda. However, we note that the scaled proportion of  
207 West and South Asian components in our North and South Munda are comparable  
208 (Supplementary Table S4: Average\_SM\_Lao=0 and Average\_NM\_Lao=0 both show  
209 ~18% West Asian and ~82% South Asian contributions) while the Southeast Asian  
210 component is higher in South than in North Munda. The independence between the  
211 amount of Southeast and West Asian components in our North and South Munda  
212 populations contradicts the expectations and therefore points to an opposite and  
213 simpler scenario: both South and North Munda could be modelled as an initial  
214 admixture between Southeast Asian populations and an autochthonous Indian group  
215 with a slightly lower West/South Asian composition compared to what observed in  
216 Paniya today. South Munda then kept isolated from additional gene flow, while North  
217 Munda received a longer admixture pulse from the local Indian groups, which caused  
218 the dilution of the newly arrived Southeast Asian components in North Munda,  
219 without affecting the relative proportions of West and South Asian components.

## 220 ***Dating the admixture event***

221 We used ALDER to test this scenario and to infer the admixture time that led to the  
222 genesis of the Mundas<sup>23</sup>. The admixture midpoint was 3846 (3235 – 4457) years ago  
223 for South Mundas, which may point to the time of arrival of the Southeast Asian  
224 component in the area, and 2867 (1751 – 4525) years ago for North Mundas (Figure  
225 3). The longer (1000 years) admixture time between North Munda and local Indian  
226 populations is consistent with the ADMIXTURE, PCA and qpAdm results where we  
227 saw North Mundas having a bigger proportion of Indian ancestry (made up,  
228 proportionally, by ~18% West and 82% South Asian) and a smaller Southeast Asian  
229 fraction than South Mundas (Supplementary Figure S3, Supplementary Figure S4,  
230 Supplementary Table S4).

231 While the ALDER dates that we obtained are, to our knowledge, the first estimates of  
232 the time of admixture of the Munda speakers based on genome-wide data, estimates  
233 from previous studies, based on other types of data, have yielded much earlier dates  
234 for the spread of Austroasiatic populations in India. Diamond and Bellwood<sup>24</sup> have  
235 estimated the age of the Munda speakers and cultivation of rice in India 5000 years  
236 old based on archaeological data. The Munda branch split from other Austroasiatic  
237 languages less than 7000 years ago based on Fuller's archeolinguistic  
238 reconstruction<sup>11,25</sup>. Recent Y chromosome studies, based on large scale  
239 resequencing of the whole Y chromosome, have estimated the age of haplogroup  
240 O2a, in which the East Asia component of the Munda Y chromosomes is nested  
241 within, to much more recent dates than the earlier estimates based on short tandem  
242 repeat variation<sup>7</sup>. The entire Southeast Asian Y chromosome variation within the  
243 clade O2a2 has been estimated to be only 5 965 (CI 5 312 – 7 013) years old<sup>17</sup>,  
244 while the variation within Munda speakers has been estimated to derive from a single

245 male ancestor who lived 4 300 (+- 200) years ago<sup>15</sup>. The latter date estimate is very  
246 similar to ours and implies a significant male-specific founder event as part of the  
247 admixture process.

248 In this study, we have replicated a result previously reported in Chaubey et al.  
249 (2011)<sup>7</sup> that the Mundas lack one ancestral component (k2) that is characteristic to  
250 Indian Indo-European and Dravidian speaking populations. If this component came to  
251 India through one of the Indo-Aryan migrations<sup>28</sup> then it would be fair to presume that  
252 the Munda admixture happened before this component reached India or at least  
253 before it spread all over the country. However, the admixture time computed here,  
254 falls in the exact same timeframe as the ANI-ASI mixture has been estimated to have  
255 happened in India<sup>5</sup> through which the k2 component probably spread. Therefore, we  
256 propose that if the Munda admixture happened at the same time, it is possible for it to  
257 have happened in the eastern part of the country, east of Bangladesh, and later  
258 when populations from East Asia moved to the area, the Mundas migrated towards  
259 central India. Such a scenario, which may be further clarified by ancient DNA  
260 analyses, seems to be further supported by the fact that Mundas harbor a smaller  
261 fraction of West Asian ancestry compared to contemporary Paniya (Supplementary  
262 Table S4) and cannot therefore be seen as a simple admixture product of Southern  
263 Indian populations with incoming Southeast Asian ancestries.

#### 264 ***Sex-biased admixture in Munda speakers***

265 In Chaubey et al. (2011)<sup>7</sup>, it was shown that the Munda speakers have high  
266 frequencies (19-95%) of East Asian chromosome Y haplogroup O2a at the  
267 background of almost no detectable East Asian mitochondrial DNA signal pointing to  
268 a sex-biased nature of admixture between Austroasiatic speakers and their local  
269 Indian neighbouring populations. We used outgroup f3 analysis to contrast allele

270 frequency patterns on the X chromosome versus those on the autosomal  
271 chromosomes to clarify the maternal side of this sex-biased admixture event. Our  
272 analysis revealed that on X chromosome, a Dravidian speaking group, North  
273 Kannadi, is relatively more similar to Munda speakers than on autosomes, while on  
274 autosomes Lao, Vietnamese and Burmese from Southeast Asia and Sino-Tibetan  
275 speaking Kuki from India have relatively higher  $f_3$  values than on X chromosome  
276 (Supplementary Figure S12). This relatively higher autosomal affinity to Southeast  
277 Asian populations, however, is detectable only when testing South Munda speakers.  
278 The fact that South Munda speakers show more evident signs of a sex-specific  
279 admixture on maternal side is in accordance with the Y chromosome results from  
280 Chaubey et al. (2011), where South Munda speakers have also higher (0.73)  
281 average frequency of haplogroup O2a than North Munda speakers (0.62)<sup>7</sup>. This  
282 finding is consistent with our proposed scenario where South Munda kept isolated  
283 after the admixture event, while North Munda received additional admixture from  
284 local Indian groups, which diluted Southeast Asian component and blurred the signs  
285 of the sex-specific nature of the admixture event as the latter admixture pulse in  
286 North Munda was not sex-specific anymore.

### 287 ***Linguistics is in concordance with genome-wide data***

288 Until now, we have presumed that the linguistic classification of the Mundas (North  
289 and South) is a suitable grouping criteria for genetic analyses. Here we take a glance  
290 at the genetic relationship between different North and South Munda populations.  
291 PCA of only Munda populations displayed North and South Mundas as separate  
292 groups, except one Juang and one Kharia individual fell together with North Mundas  
293 on first two principal components (see Supplementary Fig. S6). ADMIXTURE  
294 analysis showed that North Mundas have less of the combined k8 – k11 genetic

295 component than South Mundas (Wilcoxon rank sum test;  $N_1 = 75$ ;  $N_2 = 11$ ;  $P <$   
296  $0.0001$ ). These components were maximised in East and Southeast Asian samples.  
297 Smaller amount of Lao ancestry in North Mundas was also shown by qpAdm analysis  
298 (Supplementary Table S4). On the fineSTRUCTURE tree<sup>19</sup>, North and South Mundas  
299 clustered separately, except Kharia samples (South Munda) which clustered with  
300 Asur and Ho samples from North Munda (Figure 4). All these analyses showed that  
301 Kharia and Juang were the most similar population to North Mundas among South  
302 Munda populations. Refined IBD analysis infers that North Munda populations share  
303 more long and short IBD segments among each other than with South Munda  
304 populations (see Supplementary Fig. S10). Therefore, by and large, the linguistic  
305 classification justifies itself but Kharia and Juang do not fit in this simplification  
306 perfectly. Interestingly, although Diffloth's classification of the Munda languages into  
307 North and South Munda<sup>26</sup> is widely cited, in 2005, Diffloth changed the position of  
308 Kharia-Juang branch on the language tree from South Munda group to be a side  
309 branch of the group that was previously known as North Munda<sup>27</sup>. Hence, this is in  
310 accordance with our findings about Juang and Kharia genetic affinities.

## 311 **Methods**

### 312 ***Samples Collection and Genotyping***

313 The analyses were performed on a merged dataset of 56 new samples together with  
314 1024 previously published samples from different studies<sup>4,7,29-37</sup> (Supplementary  
315 Table S1). The new samples were collected from Laos (Lao  $N = 24$ ), Bangladesh  
316 (Santhal (NM)  $N = 10$ ), and East India (Hmar  $N = 4$ , Kom  $N = 2$ , Kuki  $N = 6$ , Mizo  $N =$   
317  $5$ , Naga  $N = 1$ , Nyishi  $N = 4$ ). DNA was extracted from blood samples collected from  
318 healthy adult donors who signed an informed consent form. New samples were  
319 genotyped using Illumina OmniExpress Bead Chips for 730k, 710k and 650k SNPs.

320 The study was approved by Research Ethics Committee of the University of Tartu. All  
321 genotyped data will be made publicly available on the [ebc.ee/free\\_data](http://ebc.ee/free_data) website.

### 322 ***Data Curation***

323 All the samples were filtered with plink v1.9<sup>38</sup>. Only SNPs on autosomal  
324 chromosomes with a minor allele frequency > 1% and genotyping success > 97%  
325 were used in the analyses. Only individuals with a genotyping success rate > 97%  
326 were left in the sample set. 245848 variants and 1072 people passed the filters; 8  
327 Gond were removed due to low genotyping success rate. For analyses that are  
328 affected by linkage disequilibrium (PCA, ADMIXTURE), dataset was further pruned  
329 by excluding SNPs with pairwise genotypic correlation  $r^2 > 0.4$  in a window of 200  
330 SNPs sliding the window by 25 SNPs at a time<sup>39</sup>. This left us 155743 SNPs.

### 331 ***Population Structure Analyses***

332 To capture genetic variability, we performed PCA using software EIGENSOFT 6.1.4<sup>40</sup>  
333 on pruned data of the whole filtered dataset (1072 individuals). To get some idea of  
334 the Munda speakers' genetic structure in context of other Asian populations, we ran  
335 ADMIXTURE 1.23 program<sup>41</sup> with random seed number generator on the LD pruned  
336 data set one hundred times at  $K = 2$  to  $K = 18$  (Supplementary Figure S1). Following  
337 an established procedure, we examined the log likelihood scores (LLs) of the  
338 individual runs and found that the highest  $K$  with stable (global maximum has been  
339 reached) LL values is  $K = 13$ . Based on cross-validation (CV) procedure, genetic  
340 structure of a sample set is best described choosing the value of  $K$  with the lowest  
341 CV error. In our dataset the lowest CV error was at  $K = 13$  (Supplementary Figure  
342 S2).

### 343 ***Tests Aimed at Providing Demographic Inferences***

344 To test the admixture, we ran three-population formal test of admixture<sup>6</sup> using  
345 popstats program by Skoglund et al. (2015)<sup>42</sup>. For f3 analysis, source 1 was South  
346 Asian or West Eurasian population and source 2 was Southeast Asian or East Asian  
347 population. Outcomes with  $|Z| > 3$  were considered significant. All the South Munda  
348 speaking tribes (Bonda, Gadaba, Juang, Kharia, Savara) were treated as one  
349 population due to small sample size. We ran outgroup f3 statistic as  $f3 =$   
350 (SouthMunda/Ho(NM), X, Yoruba) to find the closest modern populations from our  
351 data set for South and North Munda.

352 To retrieve the admixture proportions, we run the qpAdm software<sup>21</sup> testing the  
353 following South and North Munda populations (Bonda, Gadaba, Juang, Kharia,  
354 Savara, Asur, Birhor, Ho, Korwa, Mawasi, Santhal) as a three ways mixture of all  
355 possible combinations of West (Anatolia\_N, Armenia\_MLBA, Germans, Iran\_N,  
356 IranianLaz2016), East (Lao) and South (Onge, Paniya) Asian groups and using as  
357 outgroups the following groups (Natufian, WHG, Han, Kankanaey, Karitiana,  
358 MbutiLaz2016, Papuan, Ust\_Ishim, Yorubas)<sup>43,44</sup>.

359 We used ALDER<sup>23</sup> to infer admixture dates for South Munda, Ho (NM), Santhal (NM),  
360 Birhor (NM) and Korwa (NM). We used all the populations spanning from India to  
361 Europe from our data set as source 1 and all the populations from East and  
362 Southeast Asia as source 2. The population pairs to represent admixture times were  
363 chosen based on decay status and LD decay curve amplitude. Standard errors were  
364 estimated by jackknifing on chromosomes. We used generation length of 30 years<sup>45</sup>.

### 365 ***Haplotype-based Analyses***

366 To investigate the relationship between the Munda speakers and Andmanese, we  
367 used fineSTRUCTURE<sup>19</sup>. For this analysis, the data was previously phased with  
368 Beagle 3.3.2<sup>46</sup>. A co-ancestry matrix was constructed using ChromoPainter

369  $v_1^{19}$  with the default settings. From the co-ancestry matrix, the mean chunk lengths  
370 donated by Eurasian populations to Jarawa and Onge were extracted.

371 Beagle was also used in Refined IBD<sup>20</sup> analysis, where we studied the sharing of  
372 DNA segments of identity-by-descent (IBD) between the Munda speakers and  
373 other populations in our data set. From the results, we extracted the count of  
374 segments shared between every two individuals and found population medians.

375 We did the same with short (<1 cM) and long (>1 cM) segments, to find patterns.

376 We also compared total length of IBD segments shared between individuals from  
377 two different populations on average.

378 All the methods were performed in accordance with relevant guidelines and  
379 regulations.

380



381

## References

- 382 1. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142  
383 diverse populations. *Nature* **538**, 201–206 (2016).
- 384 2. Pagani, L. *et al.* Genomic analyses inform on migration events during the  
385 peopling of Eurasia. *Nature* **538**, 238–242 (2016).
- 386 3. Xing, J. *et al.* Genetic diversity in India and the inference of Eurasian population  
387 expansion. *Genome Biol.* **11**, R113 (2010).
- 388 4. Metspalu, M. *et al.* Shared and Unique Components of Human Population  
389 Structure and Genome-Wide Signals of Positive Selection in South Asia. *Am. J.*  
390 *Hum. Genet.* **89**, 731–744 (2011).
- 391 5. Moorjani, P. *et al.* Genetic Evidence for Recent Population Mixture in India. *Am.*  
392 *J. Hum. Genet.* **93**, 422–438 (2013).
- 393 6. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing  
394 Indian population history. *Nature* **461**, 489–494 (2009).
- 395 7. Chaubey, G. *et al.* Population Genetic Structure in Indian Austroasiatic Speakers:  
396 The Role of Landscape Barriers and Sex-Specific Admixture. *Mol. Biol. Evol.* **28**,  
397 1013–1024 (2011).
- 398 8. Census of India: Abstract of speakers' strength of languages and mother tongues  
399 –2001. Available at:  
400 [http://www.censusindia.gov.in/Census\\_Data\\_2001/Census\\_Data\\_Online/Langua](http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement1.aspx)  
401 [ge/Statement1.aspx](http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement1.aspx). (Accessed: 3rd May 2018)
- 402 9. Bellwood, P. First farmers. *Orig. Agric. Soc.* (2005).
- 403 10. Higham, C. *Chapter 18 Languages and Farming Dispersals: Austroasiatic*  
404 *Languages and Rice Cultivation.* (2003).

- 405 11. Fuller, D. Q. Non-human genetics, agricultural origins and historical linguistics in  
406 South Asia. in *The Evolution and History of Human Populations in South Asia*  
407 393–443 (Springer, Dordrecht, 2007). doi:10.1007/1-4020-5562-5\_18
- 408 12. Kumar, V. *et al.* Y-chromosome evidence suggests a common paternal heritage  
409 of Austro-Asiatic populations. *BMC Evol. Biol.* **7**, 47 (2007).
- 410 13. Zhang, X. *et al.* Y-chromosome diversity suggests southern origin and Paleolithic  
411 backwave migration of Austro-Asiatic speakers from eastern Asia to the Indian  
412 subcontinent. *Sci. Rep.* **5**, (2015).
- 413 14. Forster, P. & Renfrew, C. Mother Tongue and Y Chromosomes. *Science* **333**,  
414 1390–1391 (2011).
- 415 15. Arunkumar, G. *et al.* A late Neolithic expansion of Y chromosomal haplogroup  
416 O2a1-M95 from east to west: Late Neolithic expansion of O2a1-M95. *J. Syst.*  
417 *Evol.* **53**, 546–560 (2015).
- 418 16. Riccio, M. E. *et al.* The Austroasiatic Munda population from India and its  
419 enigmatic origin: a HLA diversity study. *Hum. Biol.* **83**, 405–435 (2011).
- 420 17. Karmin, M. *et al.* A recent bottleneck of Y chromosome diversity coincides with a  
421 global change in culture. *Genome Res.* **25**, 459–466 (2015).
- 422 18. Chaubey, G. *et al.* Reconstructing the population history of the largest tribe of  
423 India: the Dravidian speaking Gond. *Eur. J. Hum. Genet.* **25**, 493–498 (2017).
- 424 19. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population  
425 Structure using Dense Haplotype Data. *PLOS Genet.* **8**, e1002453 (2012).
- 426 20. Browning, B. L. & Browning, S. R. Improving the Accuracy and Efficiency of  
427 Identity-by-Descent Detection in Population Data. *Genetics* **194**, 459–471 (2013).
- 428 21. Patterson, N. *et al.* Ancient Admixture in Human History. *Genetics* **192**, 1065–  
429 1093 (2012).

- 430 22. Narasimhan, V. M. *et al.* The Genomic Formation of South and Central Asia.  
431 *bioRxiv* (2018). doi:10.1101/292581
- 432 23. Loh, P.-R. *et al.* Inferring Admixture Histories of Human Populations Using  
433 Linkage Disequilibrium. *Genetics* **193**, 1233–1254 (2013).
- 434 24. Diamond, J. Farmers and Their Languages: The First Expansions. *Science* **300**,  
435 597–603 (2003).
- 436 25. Fuller, D. Q. An agricultural perspective on Dravidian Historical Linguistics:  
437 Archaeological crop packages, livestock and Dravidian crop vocabulary. in  
438 *Assessing the Language/Farming Dispersal Hypothesis* (eds. Bellwood, P. &  
439 Renfrew, C.) 191–213 (McDonald Institute for Archaeological Research, 2003).
- 440 26. Diffloth, G. & Zide, N. Austro-asiatic languages. *Encycl. Br.* **2**, 480–484 (1974).
- 441 27. Sagart, L., Blench, R. & Sanchez-Mazas, A. *The peopling of East Asia putting*  
442 *together archaeology, linguistics and genetics.* (RoutledgeCurzon, 2005).
- 443 28. Silva, M. *et al.* A genetic chronology for the Indian Subcontinent points to heavily  
444 sex-biased dispersals. *BMC Evol. Biol.* **17**, (2017).
- 445 29. Aghakhanian, F. *et al.* Unravelling the Genetic History of Negritos and Indigenous  
446 Populations of Southeast Asia. *Genome Biol. Evol.* **7**, 1206–1215 (2015).
- 447 30. Basu, A., Sarkar-Roy, N. & Majumder, P. P. Genomic reconstruction of the history  
448 of extant populations of India reveals five distinct ancestral components and a  
449 complex structure. *Proc. Natl. Acad. Sci.* **113**, 1594–1599 (2016).
- 450 31. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466**,  
451 238–242 (2010).
- 452 32. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns  
453 of variation. *Science* **319**, 1100–1104 (2008).

- 454 33. Migliano, A. B. *et al.* Evolution of the pygmy phenotype: evidence of positive  
455 selection from genome-wide scans in African, Asian, and Melanesian pygmies.  
456 *Hum. Biol.* **85**, 251–284 (2013).
- 457 34. Mörseburg, A. *et al.* Multi-layered population structure in Island Southeast Asians.  
458 *Eur. J. Hum. Genet.* **24**, 1605–1611 (2016).
- 459 35. Pierron, D. *et al.* Genome-wide evidence of Austronesian-Bantu admixture and  
460 cultural reversion in a hunter-gatherer group of Madagascar. *Proc. Natl. Acad.*  
461 *Sci. U. S. A.* **111**, 936–941 (2014).
- 462 36. Yunusbayev, B. *et al.* The Caucasus as an Asymmetric Semipermeable Barrier to  
463 Ancient Human Migrations. *Mol. Biol. Evol.* **29**, 359–365 (2012).
- 464 37. Yunusbayev, B. *et al.* The Genetic Legacy of the Expansion of Turkic-Speaking  
465 Nomads across Eurasia. *PLOS Genet.* **11**, e1005068 (2015).
- 466 38. PLINK 1.9. Available at: <http://www.cog-genomics.org/plink/1.9/>. (Accessed: 3rd  
467 May 2018)
- 468 39. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger  
469 and richer datasets. *GigaScience* **4**, 7 (2015).
- 470 40. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis.  
471 *PLOS Genet.* **2**, e190 (2006).
- 472 41. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of  
473 ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
- 474 42. Skoglund, P. *et al.* Genetic evidence for two founding populations of the  
475 Americas. *Nature* **525**, 104–108 (2015).
- 476 43. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near  
477 East. *Nature* **536**, 419–424 (2016).

- 478 44. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-  
479 European languages in Europe. *Nature* **522**, 207–211 (2015).
- 480 45. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in  
481 genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–  
482 423 (2005).
- 483 46. Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and  
484 Missing-Data Inference for Whole-Genome Association Studies By Use of  
485 Localized Haplotype Clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
- 486
- 487

488

## Acknowledgements

489 Support was provided by the European Union through the European Regional  
490 Development Fund projects i) Centre of Excellence for Genomics and Translational  
491 Medicine Project No. 2014–2020.4.01.15-0012 (K.T., M.M., T.K.) and ii) Project No.  
492 2014-2020.4.01.16-0024, MOBTT53 (L.P.) and Estonian Institutional Research  
493 grants IUT 24-1 (K.T., T.K., A.P., D.M.B., M.M., and R.V.) and IUT 20-46 (S.K.). G.C.  
494 was supported by National Geographic explore grant HJ3-182R-18.

495

## Author Contributions

496 M.M., T.K. and G.C. devised and supervised the study. K.T. wrote the manuscript  
497 with input from M.M., T.K., L.P., G.C. and A.P. G.C., S.K., B.H.D., X.D.H., D.M.B.,  
498 Y.H., G.N.N.S., M.I.S., M.A., R.V. and M.M. performed anthropological work,  
499 sample collection and provided laboratory and computing facilities. Data analyses  
500 were performed by K.T., G.C. and L.P. Figures were prepared by K.T. All authors  
501 have reviewed the manuscript.

502

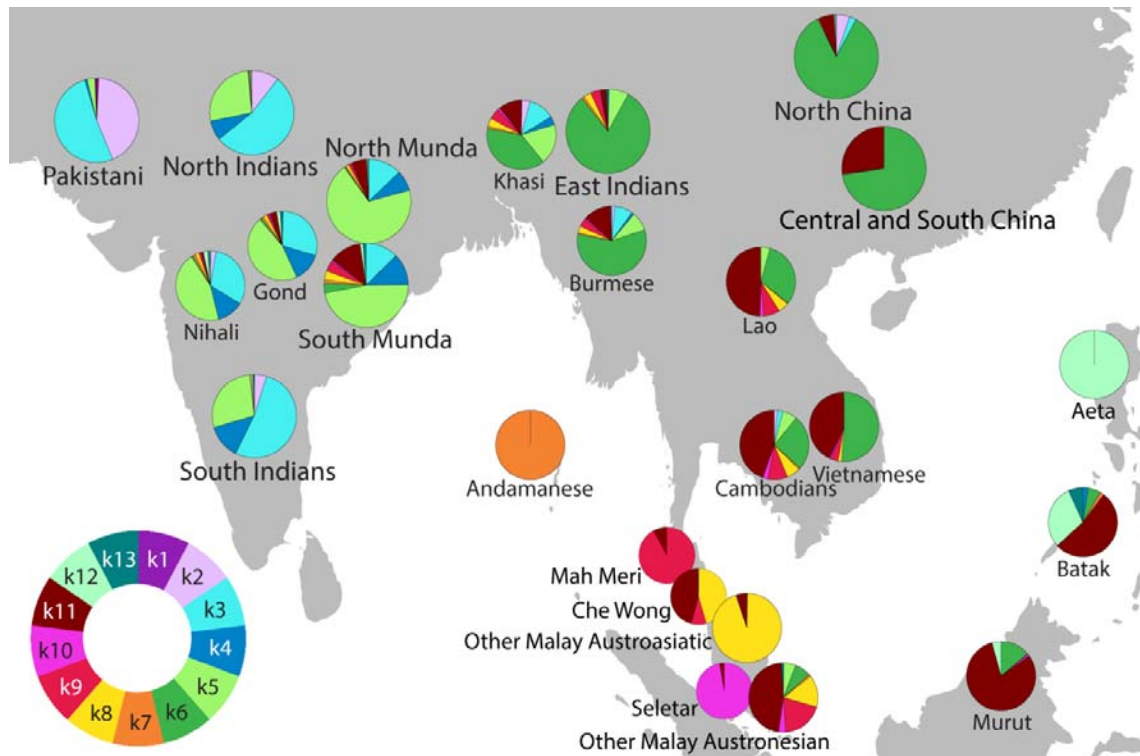
## Additional Information

503 The authors declare no competing interests.

504

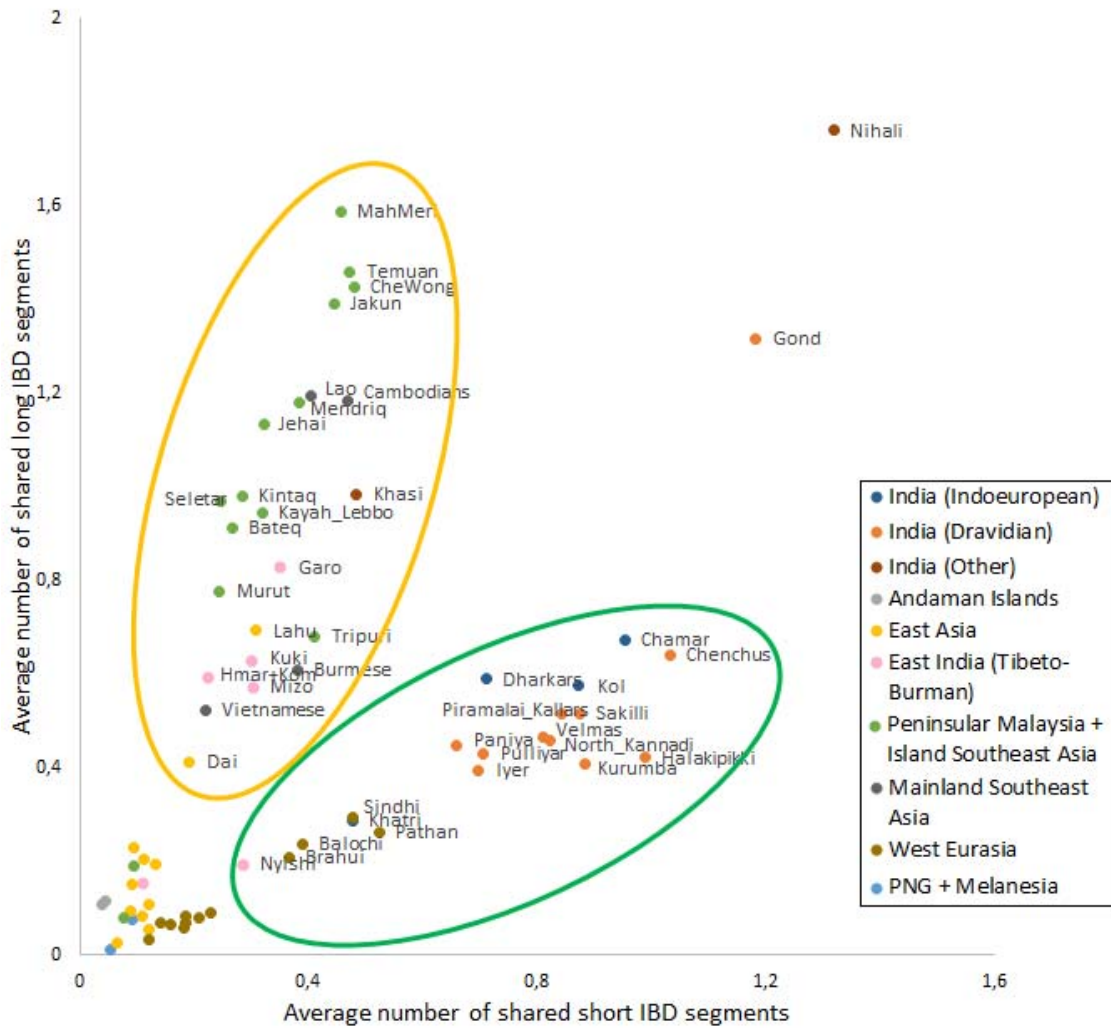
505

## Figures and Figure Titles



506

507 **Figure 1.** The distribution of genetic components (K=13) based on the global  
508 ADMIXTURE analysis (Supplementary Figure S1, S2, S3) for a subset of populations  
509 on a map of South and Southeast Asia. The circular legend in the bottom left corner  
510 shows the ancestral components corresponding to the colours on pie charts. The  
511 sector sizes correspond to population median.

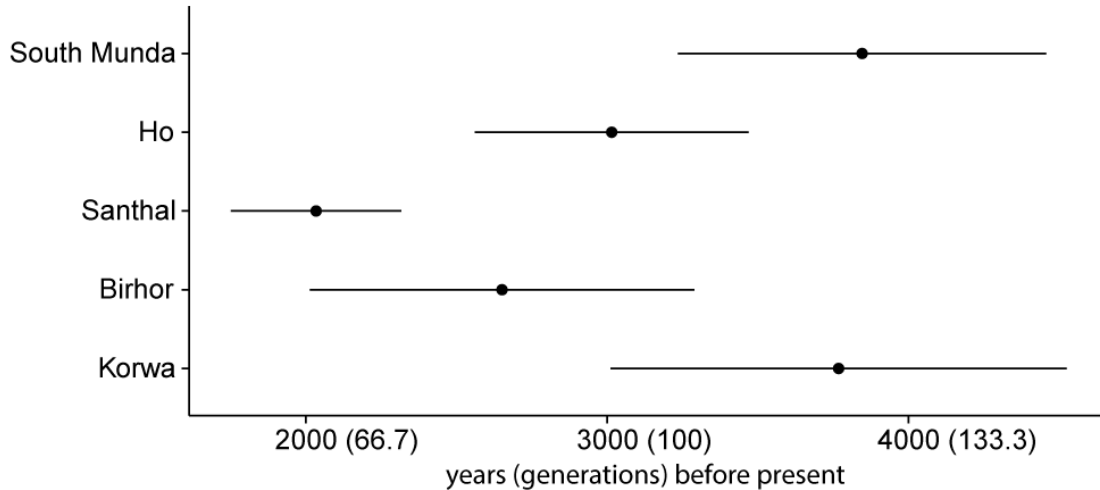


512

513 **Figure 2.** The plotted average counts of IBD segments up to 1 cM (short) and over 1  
514 cM (long) shared with the Munda speakers. The points are coloured based on  
515 linguistics and geography according to the legend on the right.

516





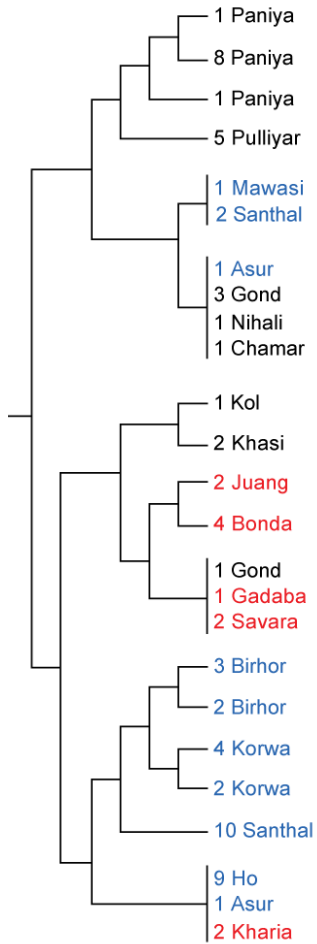
517

518 **Figure 3.** Admixture times as evaluated by ALDER. We let ALDER pair up  
519 populations from Southeast Asia and South Asia as several populations from either  
520 area were good proxies for the admixture event based on Refined IBD and  $f_3$   
521 analyses. For accuracy, North Munda speaking Santhal, Ho, Korwa and Birhor were  
522 addressed separately as admixed populations; due to a small sample size South  
523 Munda speakers were treated as one population. Reference population pair was  
524 chosen based on LD decay curve amplitude. Standard errors are estimated by  
525 jackknifing on chromosomes. Generation length is 30 years<sup>45</sup>. For all the pairs, see  
526 Supplementary Table S5.

527

528

529



**Figure 4.** A branch from a FineSTRUCTURE tree where all the Munda samples used in this analysis are situated on. Samples are coloured as follows: North Munda speakers – blue, South Munda speakers – red.