

1

2

3

4

# The promise of disease gene discovery in South Asia

6

7

8

9

Nathan Nakatsuka<sup>1,2</sup>, Priya Moorjani<sup>1,3</sup>, Niraj Rai<sup>4</sup>, Biswanath Sarkar<sup>5</sup>, Arti Tandon<sup>1,6</sup>,  
Nick Patterson<sup>6</sup>, Lalji Singh<sup>4</sup>, David Reich<sup>1,6,7,\*</sup>, Kumarasamy Thangaraj<sup>4,\*</sup>

11

12

13

14

15

16

<sup>1</sup>Department of Genetics, Harvard Medical School, New Research Building, 77 Ave.  
Louis Pasteur, Boston, MA 02115, USA

17

18

<sup>2</sup>Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School,  
Boston, MA 02115, USA

19

20

<sup>3</sup>Department of Biological Sciences, Columbia University, 600 Fairchild Center, New  
York, NY 10027, USA

21

22

<sup>4</sup>CSIR-Centre for Cellular and Molecular Biology, Habsiguda, Hyderabad, Telangana  
500007, India

23

24

<sup>5</sup>Anthrological Survey of India, Kolkata, West Bengal 700016, India

25

<sup>6</sup>Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge,  
MA 02141, USA

26

27

<sup>7</sup>Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA

28 **It is tempting to think of the more than 1.5 billion people who live in South**  
29 **Asia as one large ethnic group, but in fact, South Asia is better viewed as**  
30 **comprised of very many small endogamous groups that usually marry within**  
31 **their own group (caste or tribe). To perform a high resolution assessment of**  
32 **South Asian demography, we assembled genome-wide data from over 2,000**  
33 **individuals from over 250 distinct South Asian groups, more than tripling the**  
34 **number of diverse India groups for which such data are available, and**  
35 **including tribe and caste groups sampled from every state in India. We**  
36 **document shared ancestry across groups that correlates with geography,**  
37 **language, and caste affiliation, and characterize the strength of the founder**  
38 **events that gave rise to many of these groups. Over a third of the groups—**  
39 **including eighteen with census sizes of more than a million—descend from**  
40 **founder events stronger than those in Ashkenazi Jews and Finns, both of**  
41 **which have high rates of recessive disease due to their histories of strong**  
42 **founder events. These results highlight a major and unappreciated**  
43 **opportunity for reducing the disease burden among South Asians through the**  
44 **discovery of and genetic testing for recessive disease genes.**

45

46 South Asia is a region of extraordinary cultural, linguistic, and genetic diversity, with  
47 a conservative estimate of over 4,600 anthropologically well-defined groups, many  
48 of which are endogamous communities with significant barriers to gene flow due to  
49 sociological and cultural factors that restrict intermarriage<sup>1</sup>. Of the small fraction of  
50 South Asian groups that have been characterized using genome-wide data, many  
51 exhibit large allele frequency differences from geographically proximal neighbors<sup>2-4</sup>,  
52 indicating that they have experienced strong founder events, whereby a small  
53 number of ancestors gave rise to the many descendants that exist today<sup>4</sup>. The  
54 evidence that a substantial fraction of groups in South Asia might descend from  
55 founder events represents a major opportunity for improving health. Detailed  
56 studies of founder populations of European ancestry, including Ashkenazi Jews,  
57 Finns, Amish, Hutterites, Sardinians, and French Canadians, have resulted in the  
58 discovery of dozens of rare recessive diseases in each group, allowing genetic

59 counseling that has helped reduce disease burden in each of these communities<sup>5</sup>.  
60 Opportunities for improving health through founder event disease mapping in India  
61 are even greater due to more widespread endogamy.

62

63 To characterize the medically relevant founder events in India, we carried out new  
64 genotyping of 890 samples from 206 endogamous groups in India on the Affymetrix  
65 Human Origins single nucleotide polymorphism (SNP) array<sup>6</sup>. Based on power  
66 calculations to determine the number of samples needed to confidently detect a  
67 founder event at least as strong as that in Ashkenazi Jews or Finns (Supplementary  
68 Figure 1), we aimed in most cases to genotype up to five individuals per group.  
69 Previous studies that sampled the genetic diversity of South Asia focused to a  
70 disproportionate extent on tribal groups and castes with small census sizes in order  
71 to capture the largest possible amount of anthropological diversity<sup>3,4,7-9</sup>. In this  
72 study, our sampling included many groups with large census sizes to investigate the  
73 prospects for future disease gene mapping. We combined the new data we collected  
74 with previously reported data, leading to four datasets (Figure 1a). The Affymetrix  
75 Human Origins SNP array data comprised 1,192 individuals from 231 groups in  
76 South Asia, to which we added 7 Ashkenazi Jews. The Affymetrix 6.0 SNP array data  
77 comprised 383 individuals from 52 groups in South Asia<sup>4,8</sup>. The Illumina SNP array  
78 data comprised 188 individuals from 21 groups in South Asia<sup>9</sup> and 21 Ashkenazi  
79 Jews<sup>9,10</sup>. The Illumina Omni SNP array data comprised 367 individuals from 20  
80 groups in South Asia<sup>7</sup>. We merged 1000 Genomes Phase 3 data (2,504 individuals  
81 from 26 different groups including 99 Finns) with each of these datasets. We  
82 performed quality control to remove SNPs and individuals with a high proportion of  
83 missing genotypes or those that were outliers in Principal Component Analysis  
84 (PCA). To remove close relatives, we also removed one individual from each pair  
85 that were outliers in their group for Identity-by-Descent (IBD) genomic segments,  
86 and we removed all IBD segments that were larger than 20 centimorgans (cM).

87

88 We performed PCA on each of the three different datasets along with European  
89 Americans (CEU), Han Chinese (CHB), and West Africans (YRI), and found that the

90 Siddi are strong outliers as previously reported (Supplementary Figure 2)<sup>4,11,12</sup>. We  
91 next removed YRI, Siddi and indigenous Andamanese (another known outlier) from  
92 the datasets and repeated PCA (Figure 1b, Supplementary Figure 3). The PCA  
93 documents three broad clusters<sup>4,8,7</sup>. First, almost all Indian groups speaking Indo-  
94 European and Dravidian languages lie along the “Indian Cline,” reflecting the fact  
95 that they are admixed, with different proportions of Ancestral Northern Indian  
96 (ANI) ancestry related to Europeans, Central Asians, and Near Easterners; and  
97 Ancestral Southern Indian (ASI) ancestry that is as different from ANI as Europeans  
98 and East Asians are from each other<sup>4</sup>. The second major cluster includes groups that  
99 speak Austroasiatic languages, as well as some non-Austroasiatic speaking groups  
100 that have similar ancestry possibly due to gene flow with Austro-Asiatic speaking  
101 neighbors or to a history of language shift. This set of groups cluster together near  
102 the ASI end of the Indian cline, likely reflecting a large proportion of ASI-like  
103 ancestry as well as a distinct ancestry that has some affinity to East Asians. The  
104 Tibeto-Burmese speaking groups and other groups with high proportions of East  
105 Asian ancestry such as the Austroasiatic speaking Khasi and Tharu form a gradient  
106 of ancestry relating them to East Asian groups such as Han Chinese. These three  
107 clusters are also evident in a neighbor-joining tree based on  $F_{ST}$  (Supplementary  
108 Figure 4). We confirmed the East Asian related admixture in some groups using the  
109 statistic  $f_3(\text{Test}; \text{Mala, Chinese})$ ; significantly negative values of this statistic provide  
110 unambiguous evidence that the *Test* population is admixed of populations related  
111 (perhaps distantly) to Mala (an Indian Cline group with high ASI ancestry) and  
112 Chinese<sup>6</sup> (Supplementary Table 1).

113  
114

115 For each pair of individuals, we used GERMLINE to detect segments of the genome  
116 where the individuals are likely to share a common ancestor within the last few  
117 dozen generations, that is, where they are IBD<sup>13</sup>. We used HaploScore to filter out  
118 segments that were likely to be false-positives<sup>14</sup>. After normalizing for sample size,  
119 we estimated the distribution of IBD genome-wide within each group with at least  
120 two samples (Figure 2, Online Data Table 1). We found systematic differences in the

121 inferred IBD on different platforms; however, by normalizing by the average IBD in  
122 each group by that detected in European Americans (CEU) (present in all three  
123 datasets), we were able to meaningfully compare groups across platforms  
124 (Supplementary Figure 5). We confirmed the accuracy of this method for detecting  
125 founder events by using two other methods that we found gave highly correlated  
126 results (correlation  $r=0.86-0.98$ ): first, we computed  $F_{ST}$  between each group and  
127 other groups with similar ancestry sources, and second, we fit a formal model of  
128 history using *qpGraph*<sup>6</sup> and measured the founder event as the population-specific  
129 genetic drift post-admixture (Supplementary Figure 6)(Online Data Table 1). These  
130 analyses suggest that over a third of the Indian groups we analyzed (111 in total)  
131 have stronger founder effects than those that occurred in both Finns and Ashkenazi  
132 Jews (Figure 3). These groups are geographically and anthropologically varied,  
133 include diverse tribe, caste, and different religious groups, and also include eighteen  
134 groups with census sizes of over a million (Figure 3; Table 1). However, the groups  
135 with smaller census sizes are also medically important, as the per-individual rate of  
136 recessive disease is expected to be higher in proportion to their IBD score. Study of  
137 groups with small population size such as Amish, Hutterites, and the people of the  
138 Saguenay Lac-St. Jean region have proven to be powerful, leading to the discovery of  
139 dozens of novel disease variants that are specific to each group.

140

141

142 To better understand the history of the groups in which we detected founder events,  
143 we computed IBD for all pairs of individuals across groups. After applying a cutoff  
144 on the IBD score corresponding to  $\sim 1/3$  of the founder event size of Ashkenazi Jews,  
145 most of the groups either have no matches or only match other individuals in their  
146 group. However, some groups also share IBD across groups, typically following a  
147 religious affiliation (e.g. Catholic Brahmins) or a distinctive linguistic affiliation  
148 (particularly Austroasiatic speakers) (Supplementary Table 3). These results point  
149 to recent gene flows among some of these pairs of groups.

150

151 The strong founder events offer major opportunities for improving health in South  
152 Asia. The first opportunity lies in targeted discovery of new genetic risk factors for  
153 disease. It is already known that one group we identified as having a strong founder  
154 event, the Vysya, has over a 100-fold higher rate of butyrylcholinesterase deficiency  
155 than other Indian groups, so that in India, Vysya ancestry is a known counter-  
156 indication for the use of common muscle relaxants such as succinylcholine or  
157 mivacurium that are given prior to anesthesia<sup>15</sup>. The Agarwal community in North  
158 India, though not present in our study, is also known to have founder mutations  
159 causing higher rates of Hereditary Fructose Intolerance<sup>16</sup> and Megalencephalic  
160 Leukoencephalopathy<sup>17</sup>. Systematic studies in the Vysya and other founder event  
161 groups—involving collaboration with clinical geneticists, local pediatricians,  
162 obstetricians, midwives, and social workers to identify congenital syndromes that  
163 are common in these communities—would discover many more examples.  
164 Identification of pathogenic mutations responsible for such syndromes is  
165 straightforward with present technology. All that is required is collection of DNA  
166 samples from a small number of affected individuals and their families, usually  
167 followed by whole-exome sequencing to discover the causal changes. While rare  
168 recessive diseases would be a prime target for gene mapping, the founder groups  
169 we have identified may also be of substantial importance for disease gene mapping  
170 studies of common disease, as rare variant association analyses are known to have  
171 enhanced power in such groups<sup>18,19</sup>.

172

173 Once group specific founder event disease mutations are discovered, they can be  
174 tested for prenatally, and indeed, much of the improvement in human health that  
175 has come from founder event disease gene mapping studies is due to prenatal  
176 testing. Another way that discovery of rare recessive disease genes is likely to be  
177 important, especially in India, is through pre-marriage counseling in traditional  
178 communities where arranged marriages are common. An example of the power of  
179 this approach is *Dor Yeshorim*, a community genetic testing program among  
180 orthodox Ashkenazi Jews in both the United States and Israel<sup>20</sup>. Matchmaking is the  
181 norm among the hundreds of thousands of traditional religious orthodox Ashkenazi

182 Jews in both the United States and Israel, and *Dor Yeshorim* has taken the approach  
183 of visiting schools, genetically screening students for common recessive disease  
184 causing mutations known to affect Ashkenazi Jews using an inexpensive test, and  
185 entering the results into a confidential database. Match-makers query the *Dor*  
186 *Yeshorim* database prior to making their suggestions to the families and receive  
187 feedback about whether the potential couple is “incompatible” in the sense of both  
188 being carriers for a recessive mutation at the same gene. The program is successful  
189 in both the United States and Israel, such that ~95% of community members whose  
190 marriages are arranged participate; as a result, recessive diseases like Tay Sachs  
191 have virtually disappeared from these communities. A similar approach should  
192 work as well in Indian communities where arranged marriages are common, and  
193 where there is already recognition of the power of clinical screening to affect birth  
194 outcomes. Given the potential for saving lives and ultimately financial and medical  
195 resources, this or similar kinds of research could serve as an important investment  
196 for future generations<sup>21</sup>.

197

198 This study of more than 250 distinct groups represents the first systematic survey  
199 for founder events in South Asia, and to our knowledge also presents the richest  
200 dataset of genome-wide data from anthropologically well-documented groups  
201 available for any region in the world. Despite the breadth of this data, the groups  
202 surveyed here represent only about 5% of the anthropologically well-defined  
203 groups in India. Extensions of the survey to all well-defined anthropological groups  
204 would make it possible to identify large numbers of additional founder groups  
205 susceptible to recessive diseases and to assess the extent to which the founder  
206 events we have already detected are localized to the specific regions from which our  
207 samples were drawn, or are shared across people of the same ethnic group across  
208 different regions in India. An important priority for future work is also to carry out  
209 pilot studies to find real disease genes in our groups, thereby proving by example  
210 the power of this approach for directing future disease mapping studies.

211

212 **Supplementary Data:**

213

214 Supplementary Data include an excel spreadsheet detailing all groups and their  
215 scores on the IBD, Fst, and Population-specific drift analyses. Also included are 6  
216 supplementary figures and 3 supplementary tables.

217

218 **Acknowledgements:**

219

220 We are grateful to the many Indian, Pakistani, Bangladeshi, and Nepalese  
221 communities and individuals who contributed the DNA samples analyzed here. We  
222 thank Raj Rajkumar (deceased) for his assistance in assembling the sample  
223 collection. We would also like to acknowledge Analabha Basu and Patha Majumdar  
224 for helping to share their data for analysis in this study. Funding for this project was  
225 provided by the NIGMS (T32GM007753) to NN. This work was supported by a  
226 Translational Seed Fund grant from the Dean's Office of Harvard Medical School to  
227 DR, who is also a member of the Howard Hughes Medical Institute. KT is supported  
228 by CSIR network project GENESIS (BSC0121). PM was supported by the National  
229 Institutes of Health (NIH) under a Ruth L. Kirschstein National Research Service  
230 Award F32 GM115006-01. Genotyping data for the samples collected for this study  
231 will be made available upon request from the corresponding authors.



## 232 **References:**

233

- 234 1. Mastana, S.S. Unity in diversity: an overview of the genomic anthropology of  
235 India. *Ann Hum Biol* **41**, 287-99 (2014).
- 236 2. Bamshad, M.J. *et al.* Female gene flow stratifies Hindu castes. *Nature* **395**,  
237 651-2 (1998).
- 238 3. Basu, A. *et al.* Ethnic India: a genomic view, with special reference to peopling  
239 and structure. *Genome Res* **13**, 2277-90 (2003).
- 240 4. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. Reconstructing  
241 Indian population history. *Nature* **461**, 489-94 (2009).
- 242 5. Lim, E.T. *et al.* Distribution and medical impact of loss-of-function variants in  
243 the Finnish founder population. *PLoS Genet* **10**, e1004494 (2014).
- 244 6. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-  
245 93 (2012).
- 246 7. Basu, A., Sarkar-Roy, N. & Majumder, P.P. Genomic reconstruction of the  
247 history of extant populations of India reveals five distinct ancestral  
248 components and a complex structure. *Proc Natl Acad Sci U S A* (2016).
- 249 8. Moorjani, P. *et al.* Genetic evidence for recent population mixture in India. *Am*  
250 *J Hum Genet* **93**, 422-38 (2013).
- 251 9. Metspalu, M. *et al.* Shared and unique components of human population  
252 structure and genome-wide signals of positive selection in South Asia. *Am J*  
253 *Hum Genet* **89**, 731-44 (2011).
- 254 10. Behar, D.M. *et al.* The genome-wide structure of the Jewish people. *Nature*  
255 **466**, 238-42 (2010).
- 256 11. Narang, A. *et al.* Recent admixture in an Indian population of African  
257 ancestry. *Am J Hum Genet* **89**, 111-20 (2011).
- 258 12. Shah, A.M. *et al.* Indian Siddis: African descendants with Indian admixture.  
259 *Am J Hum Genet* **89**, 154-61 (2011).
- 260 13. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden  
261 relatedness. *Genome Res* **19**, 318-26 (2009).

- 262 14. Durand, E.Y., Eriksson, N. & McLean, C.Y. Reducing pervasive false-positive  
263 identical-by-descent segments detected by large-scale pedigree analysis. *Mol*  
264 *Biol Evol* **31**, 2212-22 (2014).
- 265 15. Manoharan, I., Wieseler, S., Layer, P.G., Lockridge, O. & Boopathy, R. Naturally  
266 occurring mutation Leu307Pro of human butyrylcholinesterase in the Vysya  
267 community of India. *Pharmacogenet Genomics* **16**, 461-8 (2006).
- 268 16. Bijarnia-Mahay, S. *et al.* Molecular Diagnosis of Hereditary Fructose  
269 Intolerance: Founder Mutation in a Community from India. *JIMD Rep* **19**, 85-  
270 93 (2015).
- 271 17. Shukla, P. *et al.* Molecular genetic studies in Indian patients with  
272 megalencephalic leukoencephalopathy. *Pediatr Neurol* **44**, 450-8 (2011).
- 273 18. Wang, S.R. *et al.* Simulation of Finnish population history, guided by empirical  
274 genetic data, to assess power of rare-variant tests in Finland. *Am J Hum Genet*  
275 **94**, 710-20 (2014).
- 276 19. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a  
277 birth cohort from a founder population. *Nat Genet* **41**, 35-46 (2009).
- 278 20. Raz, A.E. Can population-based carrier screening be left to the community? *J*  
279 *Genet Couns* **18**, 114-8 (2009).
- 280 21. Rajasimha, H.K. *et al.* Organization for rare diseases India (ORDI) -  
281 addressing the challenges and opportunities for the Indian rare diseases'  
282 community. *Genet Res (Camb)* **96**, e009 (2014).
- 283 22. Sudmant, P.H. *et al.* Global diversity, population stratification, and selection  
284 of human copy-number variation. *Science* **349**, aab3761 (2015).
- 285 23. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human  
286 genomes. *Nature* **526**, 75-81 (2015).
- 287 24. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis.  
288 *PLoS Genet* **2**, e190 (2006).
- 289 25. Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*  
290 **5**, 164-166 (1989).

- 291 26. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and  
292 display of phylogenetic trees made easy. *Nucleic Acids Res* **39**, W475-8  
293 (2011).
- 294 27. Browning, B.L. & Browning, S.R. Improving the accuracy and efficiency of  
295 identity-by-descent detection in population data. *Genetics* **194**, 459-71  
296 (2013).
- 297 28. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and  
298 missing-data inference for whole-genome association studies by use of  
299 localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).
- 300  
301  
302  
303  
304  
305  
306  
307

Population	Sample Size	IBD Rank	F <sub>ST</sub> Rank	Drift Rank	Census Size	Location
Pattapu_Kapu	4	15	17	12	13,697,000	Andhra Pradesh
Kumhar	5	20	28	18	3,144,000	Uttar Pradesh
Rajbanshi	4	29	NA	NA	3,801,677	West Bengal
Kallar	5	38	37	29	2,426,929	Tamil Nadu
Arunthathiyar	5	40	48	34	1,084,162	Tamil Nadu
Yadav	13	41	66	44	20,000,000	Puducherry
Naga	4	51	NA	NA	1,667,712	Nagaland
Scheduled_Caste_Haryana	4	53	66	47	5,113,615	Haryana
Reddy	8	56	70	53	22,500,000	Andhra Pradesh
Baniyas	4	60	81	56	4,848,000	Uttar Pradesh
Dawoodi	2	62	41	40	4,592,854	Gujarat
Muslim_Bihar	4	64	62	44	13,722,048	Bihar
Vysya	10	73	30	21	3,200,000	Andhra Pradesh
Rajput	4	75	139	136	51,137,113	Haryana
Garasia	5	77	70	43	4,215,603	Gujarat
Satnami	5	85	96	65	4,200,000	Chhattisgarh
Shani	3	87	139	80	4,029,411	Bihar
Newar	6	95	NA	NA	1,523,000	Nepal

309

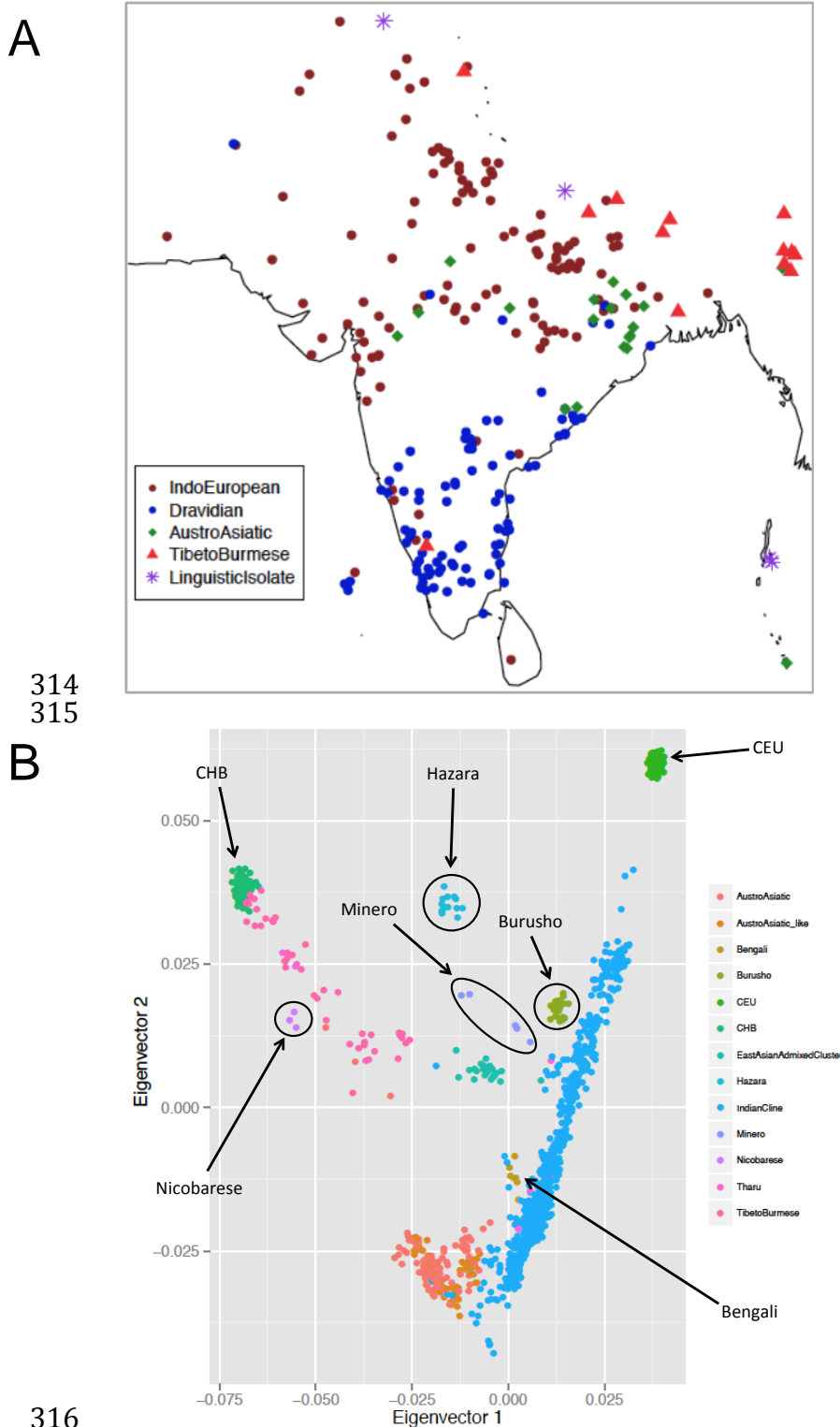
310

311

312

313

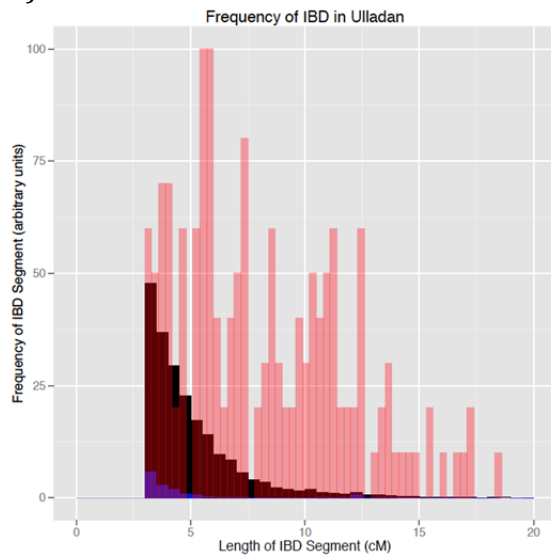
**Table 1. Indian groups with strong IBD scores.** Eighteen Indian groups with IBD scores higher than Ashkenazi Jews and census sizes over 1 million that are of particularly high interest for founder event disease gene mapping studies.



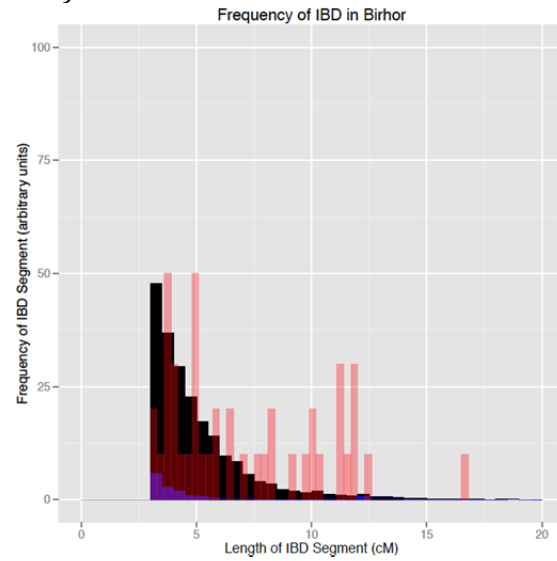
317 **Figure 1. Dataset overview.** (a) Sampling locations for all analyzed groups. Each  
318 point indicates a distinct group (random jitter added to help in visualization at  
319 locations where there are many groups). (b) PCA of Human Origins dataset along  
320 with European Americans (CEU) and Han Chinese (CHB).  
321

322

A)



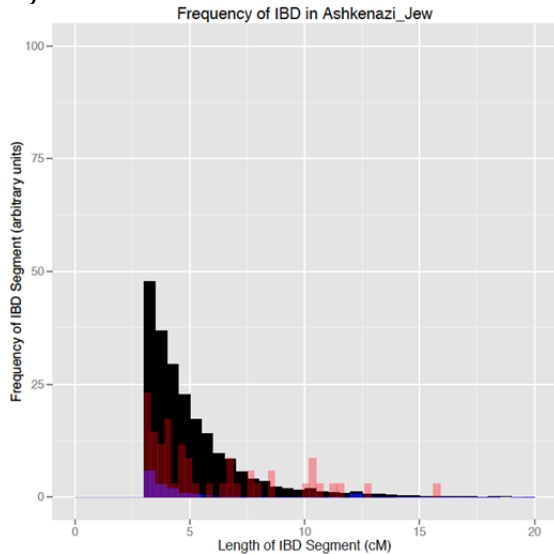
B)



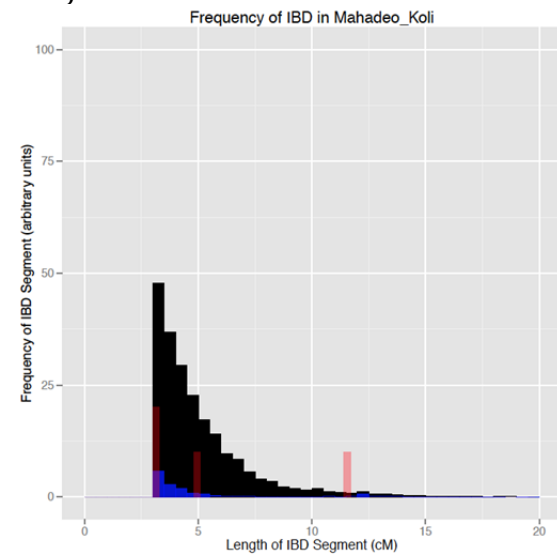
323

324

C)



D)



325

326

327

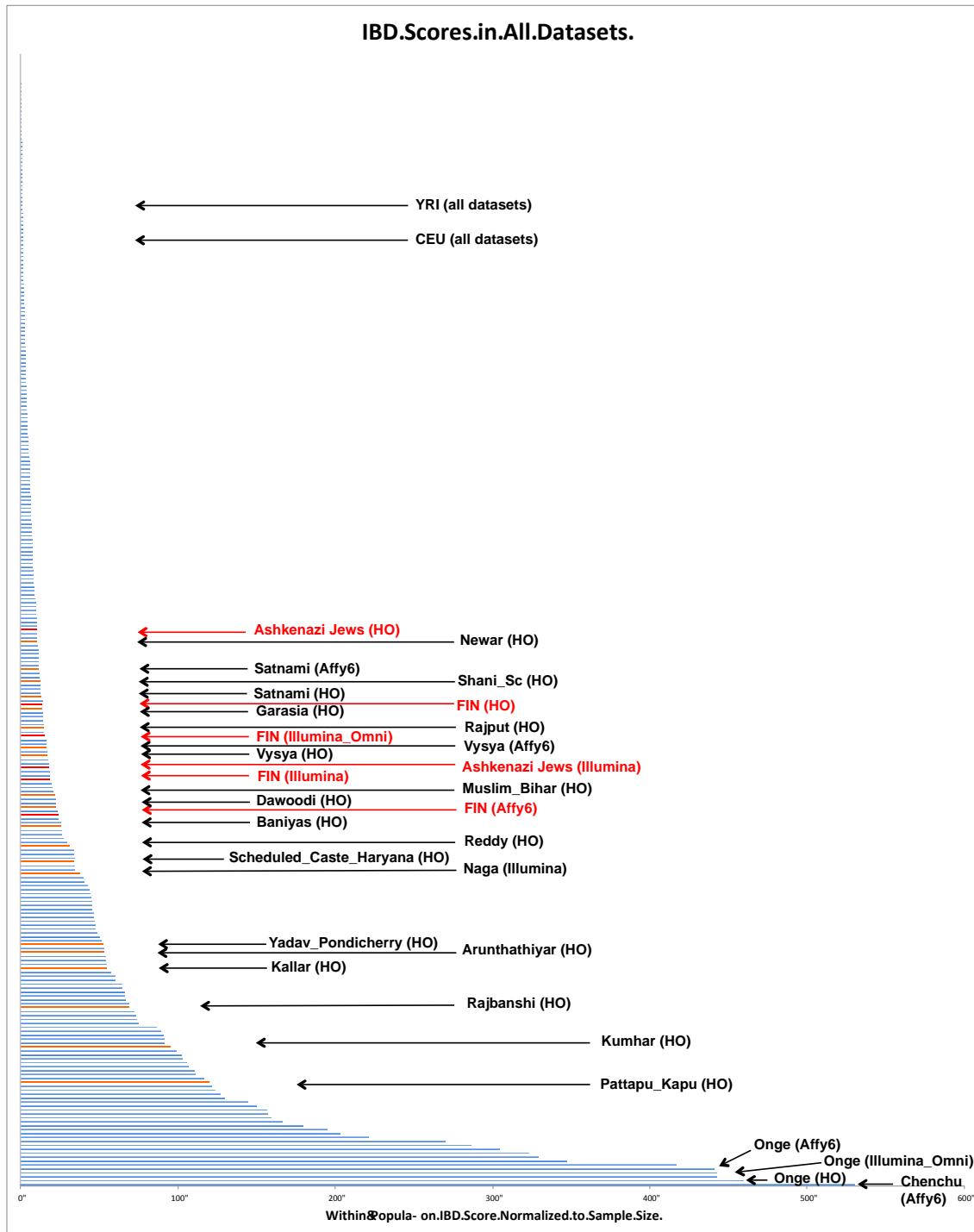
328

329

330

331

**Figure 2.** Histogram of IBD in groups with founder events of different magnitudes: (A) very large in Ulladan, (B) large in Birhor, (C) moderate in Ashkenazi Jews, and (D) small in Mahadeo\_Koli. In each plot, we showed for comparison the histogram of IBD for European Americans (CEU) with a negligible founder event in blue, and that in Finns (FIN) with a large founder event in black.



332  
333  
334  
335  
336  
337  
338  
339

**Figure 3. IBD scores normalized by that in European Americans (CEU).** Histogram ordered by IBD score, which is roughly proportional to the per-individual risk for recessive disease due to the founder event. (These results are also given quantitatively for each group in Online Table 1.) We restrict to groups with at least two samples, combining data from all four genotyping platforms onto one plot. Data from Ashkenazi Jews and Finns are highlighted in red, and from Indian groups with stronger founder events and census sizes of more than a million in orange.

## 340 **Online Methods:**

341

### 342 **Data Sets:**

343

344 We used genotyping array data from multiple sources. We assembled a dataset of  
345 1,182 individuals from 225 groups genotyped on the Affymetrix Human Origins  
346 array, of which data from 890 individuals from 206 groups is newly reported here  
347 (Figure 1a). We merged these data with a dataset published in Moorjani *et al.*<sup>8</sup>,  
348 which consisted of 332 individuals from 52 groups genotyped on the Affymetrix 6.0  
349 array. We also merged it with two additional datasets published in Metspalu *et al.*<sup>9</sup>,  
350 consisting of 151 individuals from 21 groups genotyped on Illumina 650K arrays as  
351 well as a dataset published in Basu *et al.*<sup>7</sup>, which consisted of 367 individuals from  
352 20 groups generated on Illumina Omni 1-Quad arrays. These groups came from  
353 India, Pakistan, Nepal, Sri Lanka, and Bangladesh, and we refer to all of them here as  
354 the “South Asia” dataset.

355

356 We analyzed two different Jewish datasets, one consisting of 21 Ashkenazi Jewish  
357 individuals genotyped on Illumina 610K and 660K bead arrays<sup>10</sup> and one consisting  
358 of 7 Ashkenazi Jewish individuals genotyped on Affymetrix Human Origins arrays.

359

360 Our “Affymetrix 6.0” dataset consisted of 332 individuals genotyped on 329,261  
361 SNPs, and our “Illumina\_Omni” dataset consisted of 367 individuals genotyped on  
362 750,919 SNPs. We merged the South Asia and Jewish data generated by the other  
363 Illumina arrays to create an “Illumina” dataset consisting of 172 individuals  
364 genotyped on 500,640 SNPs. Finally, we merged the data from the Affymetrix  
365 Human Origins arrays with the Ashkenazi Jewish data and data from the Simons  
366 Genome Diversity Project<sup>22</sup> to create a dataset with 1,225 individuals genotyped on  
367 512,615 SNPs. We analyzed the four datasets separately due to the small  
368 intersection of SNPs between them and the possible systematic differences across  
369 genotyping platforms. We merged in the 1000 Genomes Phase 3 data<sup>23</sup> (2504  
370 individuals from 26 different groups; notably, including 99 Finnish individuals) into



371 all of the datasets. We used genome reference sequence coordinates (hg19) for  
372 analyses.

373

374 **Quality Control:**

375

376 We filtered the data on both the SNP and individual level. On the SNP level, we  
377 required at least 95% genotyping completeness for each SNP (across all  
378 individuals). On the individual level, we required at least 95% genotyping  
379 completeness for each individual (across all SNPs).

380

381 To test for batch effects due to samples from the same group being genotyped on  
382 different array plates, we studied instances where samples from the same group  $A$   
383 were genotyped on both plates 1 and 2 and computed an allele frequency difference  
384 at each SNP,  $Diff_A^i = (Freq_{PopA,Plate1}^i - Freq_{PopA,Plate2}^i)$ . We then computed the  
385 product of these allele frequencies averaged over all SNPs for two groups  $A$  and  $B$   
386 genotyped on the same plates,  $\frac{1}{n} \sum_{i=1}^n (Diff_A^i)(Diff_B^i)$ , as well as a standard error  
387 from a Block Jackknife. This quantity should be consistent with zero within a few  
388 standard errors if there are no batch effects that cause systematic differences across  
389 the plates, as allele frequency differences between two samples of the same group  
390 should just be random fluctuations that have nothing to do with the array plates on  
391 which they are genotyped. This analysis found strong batch effects associated with  
392 one array plate, and we removed this from analysis.

393

394 We used EIGENSOFT 5.0.1 smartpca<sup>24</sup> on each group. We also developed a  
395 procedure to distinguish recent relatedness from founder effects so that we could  
396 remove recently related individuals. We first identified all duplicates or obvious  
397 close relatives by using Plink “genome” and removed all individuals who had both a  
398 PI\_HAT score greater than 0.45 and the presence of at least 1 IBD fragment greater  
399 than 30cM long. We used an iterative procedure of identifying any pairs within each  
400 group that had both total IBD and total long IBD (>20cM) that were greater than 2.5  
401 SDs and 1 SD, respectively, from the group mean. After each round we repeated the

402 process if the new IBD score was at least 30% lower than the prior IBD score. Due to  
403 their known very small census size (our sample consists of a substantial fraction of  
404 the entire population) and exceptional anthropological interest, we excluded Onge, a  
405 tribal population of the Andaman and Nicobar Islands, from this analysis. After data  
406 quality control and merging with the 1000 Genomes Project data, the Affymetrix 6.0  
407 dataset included 3,215 individuals genotyped on 326,181 SNPs, the Illumina dataset  
408 included 2,789 individuals genotyped on 484,293 SNPs, the Illumina Omni dataset  
409 included 2,834 individuals genotyped on 750,919 SNPs, and the Human Origins  
410 dataset included 3,696 individuals genotyped at 500,648 SNPs.

411

#### 412 **Distance-Based Phylogenetic Tree:**

413

414 We calculated genetic differentiation ( $F_{ST}$ ) between all pairs of groups using  
415 EIGENSOFT *smartpca* and created a neighbor-joining tree using PHYLIP<sup>25</sup> with  
416 Yoruba as the outgroup. We used Itol<sup>26</sup> to display the tree.

417

#### 418 **Power Calculations:**

419

420 We performed power calculations to determine the approximate number of samples  
421 required to detect founder events of a pre-specified strength. We used Beagle 3.3.2  
422 FastIBD<sup>27</sup> to identify all shared IBD segments between individuals within a group  
423 with parameters *missing=0; lowmem=true; gprobs=false; verbose=true; fastIBD=true;*  
424 *ibdscale=scale* (where  $scale = \sqrt{\#samples/100}$ ). We used the output of this  
425 program to calculate mean IBD sharing. We computed standard errors via jackknife  
426 resampling over individuals for each group. These analyses demonstrate that only  
427 3-5 individuals are needed to assess accurately the size of founder effects in groups  
428 with strong founder events (Supplementary Figure 1). Weaker founder effects are  
429 more difficult to detect, but these groups are of less interest from the perspective of  
430 founder event disease mapping, so we aimed to sample ~5 individuals per group in  
431 the new genotyping effort based on the Affymetrix Human Origins platform.

432

#### 433 **Phasing and IBD Detection:**

434

435 We phased all datasets using Beagle 3.3.2 with the settings *missing=0; lowmem=true;*  
436 *gprobs=false; verbose=true*<sup>28</sup>. We left all other settings at default. We determined IBD  
437 segments using the GERMLINE algorithm<sup>13</sup> with the parameters *-bits 75 -err\_hom 0 -*  
438 *err\_het 0 -min\_m 3*. We used the genotype extension mode to minimize the effect of  
439 any possible phasing heterogeneity amongst the different groups and used the  
440 HaploScore algorithm to remove false positive IBD fragments with the  
441 recommended genotype error and switch error parameters of 0.0075 and 0.003<sup>14</sup>.  
442 We chose a HaploScore threshold matrix based on calculations from Durand *et al.*  
443 for a “mean overlap” of 0.8, which corresponds to a precision of approximately 0.9  
444 for all genetic lengths from 2-10cM. In addition to the procedure we developed to  
445 remove close relatives (Quality Control section), we also removed segments longer  
446 than 20cM to ignore the effect of consanguinity and shorter than 3cM to minimize  
447 false positives and better differentiate groups with stronger founder effects from  
448 those with weaker effects. We treated all groups as subpopulations (e.g. Vysya and  
449 Ashkenazi Jews) and only retained IBD segments within the subpopulation (e.g. only  
450 IBD segments shared between Vysya individuals). We computed “founder effect  
451 size” as the total length of IBD segments between 3-20cM divided by  $((2 * (\# \text{ of}$   
452  $\text{individuals in sample})) \text{ choose } 2)$  to normalize for sample size. We also repeated  
453 these analyses with FastIBD<sup>27</sup> for the Affy 6.0 and Illumina datasets and observed  
454 that the results were highly correlated ( $r > 0.96$ ) (Supplementary Table 1). We chose  
455 GERMLINE for our main analyses, however, because the FastIBD algorithm required  
456 us to split the datasets into different groups, since it adapts to the relationships  
457 between LD and genetic distance in the data, and these relationships differ across  
458 groups. We used several different Jewish groups and all twenty-six 1000 Genomes  
459 groups to improve phasing and calibration of the IBD scores, but of these groups we  
460 only included results for two founder groups (Ashkenazi Jews and Finns for  
461 comparison with Indian groups), and two outbred populations (CEU and YRI for  
462 normalization) in the final IBD score ranking.

463

464 **Between-Group IBD Calculations:**

465

466 We determined IBD using GERMLINE as above. We collapsed individuals into  
467 respective groups and normalized for between-group IBD by dividing all IBD from  
468 each group by ( $2 \times$  # of individuals in the group). We normalized for within-group  
469 IBD as described above. We defined groups with high shared IBD as those with an  
470 IBD score greater than 3, corresponding to approximately three times the founder  
471 effect size of CEU (and  $\sim 1/3$  the effect size of Ashkenazi Jews).

472

#### 473 **$f_3$ -statistics:**

474

475 We used the  $f_3$ -statistic<sup>6</sup>  $f_3(\text{Test}; \text{Ref}_1, \text{Ref}_2)$  to determine if there was evidence that  
476 the *Test* group was derived from admixture of groups related to  $\text{Ref}_1$  and  $\text{Ref}_2$ . A  
477 significantly negative statistic provides unambiguous evidence of mixture in the  
478 *Test* group. We assessed the significance of the  $f_3$ -statistic using a Block Jackknife  
479 and a block size of 5 cM. We considered statistics over 3 standard errors below zero  
480 to be significant.

481

#### 482 **Calculating Group Specific Drift:**

483

484 We used ADMIXTUREGRAPH<sup>6</sup> to model each Indian population on the cline as a  
485 mixture of ANI and ASI ancestry. Within the limits of our resolution, this model (YRI,  
486 (Indian population, (Georgians, ANI)), [(ASI, Onge)]) proposed by Moorjani *et al.*<sup>8</sup> is  
487 a good fit to the data, in the sense that none of the  $f$ -statistics relating the groups are  
488 greater than three standard errors from expectation. This approach provides  
489 estimates for post-admixture drift in each group (Supplementary Figure 6), which is  
490 reflective of the strength of the founder event (high drift values imply stronger  
491 founder events). We only included groups on the Indian cline in this analysis, and  
492 we removed all groups with evidence of recent East Asian admixture.

493

#### 494 **PCA-Normalized $F_{ST}$ Calculations:**

495

496 To account for intermarriage across groups, we used clusters based on PCA to  
497 estimate the minimum  $F_{ST}$  for each South Asian population (Supplementary Figure  
498 6). Specifically, we calculated the  $F_{ST}$  between each group and the rest of the

499 individuals in their respective cluster based on EIGENSOFT *smartpca*. For these  
500 analyses we only included groups with Austroasiatic-related genetic patterns (i.e.  
501 those groups clustering near Austroasiatic speakers on the PCA) and those on the  
502 Indian cline; we excluded all groups with recent East Asian admixture. For  
503 Ashkenazi Jews and Finns, we used the minimum  $F_{ST}$  to their closest European  
504 neighbors.  
505  
506  
507