

# The promise of disease gene discovery in South Asia

Nathan Nakatsuka<sup>1,2</sup>, Priya Moorjani<sup>3,6</sup>, Niraj Rai<sup>4</sup>, Biswanath Sarkar<sup>5</sup>, Arti Tandon<sup>1,6</sup>, Nick Patterson<sup>6</sup>, Gandham SriLakshmi Bhavani<sup>7</sup>, Katta Mohan Girisha<sup>7</sup>, Mohammed S Mustak<sup>8</sup>, Sudha Srinivasan<sup>9</sup>, Amit Kaushik<sup>10</sup>, Saadi Abdul Vahab<sup>11</sup>, Sujatha M. Jagadeesh<sup>12</sup>, Kapaettu Satyamoorthy<sup>11</sup>, Lalji Singh<sup>4,13</sup>, David Reich<sup>1,5,14,\*</sup>, Kumarasamy Thangaraj<sup>4,\*</sup>

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

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

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

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

<sup>5</sup>Superintending Anthropologist (Physical) (Rtd.), Anthropological Survey of India, 27 Jawaharlal Nehru Road, Kolkata 700016, India

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

<sup>7</sup>Department of Medical Genetics, Kasturba Medical College, Manipal University, Manipal, India

<sup>8</sup>Department of Applied Zoology, Mangalore University, Mangalagangothri 574199, Mangalore, Karnataka, India

<sup>9</sup>Centre for Human Genetics, Biotech Park, Electronics City (Phase I), Bangalore 560100, India

<sup>10</sup>Amity Institute of Biotechnology, Amity University, Sector125, Noida 201303, India

<sup>11</sup>School of Life Sciences, Manipal University, Manipal 576104, India

<sup>12</sup>Fetal Care Research Foundation, 197 Dr. Natesan Road, Chennai 600004, India

<sup>13</sup>Present address: Genome Foundation, Hyderabad 500076, India

<sup>14</sup>Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA

\*co-senior authors

43 **The more than 1.5 billion people who live in South Asia are correctly viewed**  
44 **not as a single large population, but as many small endogamous groups. We**  
45 **assembled genome-wide data from over 2,800 individuals from over 260**  
46 **distinct South Asian groups. We identify 81 unique groups, of which 14 have**  
47 **estimated census sizes of more than a million, that descend from founder**  
48 **events more extreme than those in Ashkenazi Jews and Finns, both of which**  
49 **have high rates of recessive disease due to founder events. We identify**  
50 **multiple examples of recessive diseases in South Asia that are the result of**  
51 **such founder events. This study highlights an under-appreciated opportunity**  
52 **for reducing disease burden among South Asians through the discovery of and**  
53 **testing for recessive disease genes.**

54  
55 South Asia is a region of extraordinary diversity, containing over 4,600  
56 anthropologically well-defined groups many of which are endogamous communities  
57 with significant barriers to gene flow due to cultural practices that restrict marriage  
58 between groups<sup>1</sup>. Of the tiny fraction of South Asian groups that have been  
59 characterized using genome-wide data, many exhibit large allele frequency  
60 differences from close neighbors<sup>2-4</sup>, consistent with strong founder events whereby  
61 a small number of ancestors gave rise to many descendants today<sup>4</sup>. The pervasive  
62 founder events in South Asia present a potential opportunity for reducing disease  
63 burden in South Asia. The promise is highlighted by studies of founder groups of  
64 European ancestry – including Ashkenazi Jews, Finns, Amish, Hutterites, Sardinians,  
65 and French Canadians – which have resulted in the discovery of dozens of recessive  
66 disease causing mutations in each group. Prenatal testing for these mutations has  
67 substantially reduced recessive disease burden in all of these communities<sup>5,6</sup>.

68  
69 We carried out new genotyping of 1,663 samples from 230 endogamous groups in  
70 South Asia on the Affymetrix Human Origins single nucleotide polymorphism (SNP)  
71 array<sup>7</sup>. We combined the data we newly collected with previously reported data,  
72 leading to four datasets (Figure 1a). The Affymetrix Human Origins SNP array data  
73 comprised 1,955 individuals from 249 groups in South Asia, to which we added 7  
74 Ashkenazi Jews. The Affymetrix 6.0 SNP array data comprised 383 individuals from  
75 52 groups in South Asia<sup>4,8</sup>. The Illumina SNP array data comprised 188 individuals  
76 from 21 groups in South Asia<sup>9</sup> and 21 Ashkenazi Jews<sup>9,10</sup>. The Illumina Omni SNP  
77 array data comprised 367 individuals from 20 groups in South Asia<sup>11</sup>. We merged  
78 1000 Genomes Phase 3 data<sup>12</sup> (2,504 individuals from 26 different groups including  
79 99 Finns) with each of these datasets. We removed SNPs and individuals with a high  
80 proportion of missing genotypes or that were outliers in Principal Components  
81 Analysis (PCA) (Figure 1b; Supplementary Text). The total number of unique groups  
82 analyzed in this study is 263 after accounting for groups represented in multiple  
83 datasets. To our knowledge, this represents the richest set of genome-wide data  
84 from anthropologically well-documented groups from any region in the world.

85  
86 We devised an algorithm to quantify the strength of the founder events in each  
87 group based on Identity-by-Descent (IBD) segments, large stretches of DNA shared  
88 from a common founder in the last approximately one hundred generations (Figure

89 2). We computed an “IBD score” as a measure for the strength of the founder event  
90 in each group’s history: the average length of IBD segments between 3-20  
91 centimorgans (cM) shared between two genomes normalized to sample size. Since  
92 we are interested in characterizing the impact of recessive diseases that do not owe  
93 their origin to consanguineous marriages of close relatives, we ignored self-matches  
94 (internal homozygosity) in IBD calculations. We removed all individuals that had  
95 evidence of recent relatedness (within a few generations) to others in the dataset by  
96 computing IBD between all pairs of individuals in each group and removing one  
97 individual from the pairs with outlying numbers of IBD segments (our focus on  
98 founder events rather than recent relatedness also explains our choice to exclude  
99 IBD segments of greater than 20 cM in size). We validated the effectiveness of this  
100 procedure by simulation (Supplementary Table 1; Online Methods).

101  
102 We expressed IBD scores for each group as a fraction of the IBD scores of the 1000  
103 Genomes Project Finns merged into each respective dataset. Due to the fact that all  
104 the SNP arrays we analyzed included more SNPs ascertained in Europeans than in  
105 South Asians, the sensitivity of our methods to founder events is greater in  
106 Europeans than in South Asians, and thus our estimates of founder event strengths  
107 in South Asian groups are conservative underestimates relative to that in Europeans  
108 (Supplementary Figure 1 demonstrates this effect empirically and shows that it is  
109 less of a bias for the strong founder events that are the focus of this study). We  
110 computed standard errors for these ratios by a weighted Block Jackknife across  
111 chromosomes and declared significance where the 95% confidence intervals did not  
112 overlap 1. We carried out computer simulations to validate our procedure. The  
113 simulations suggest that we are not substantially overestimating the magnitudes of  
114 modest founder events, since for a simulated founder event that is half the  
115 magnitude of that in Finns, we never infer the score to be significantly greater than  
116 in Finns. The simulations also suggest that our procedure is highly sensitive to  
117 detecting strong founder events, since for sample sizes of at least 5, the algorithm’s  
118 sensitivity is greater than 95% for determining that a group with two times the  
119 bottleneck strength as Finns has an IBD score significantly greater than that of Finns  
120 (Supplementary Figure 2 and Supplementary Table 2). We also used two additional  
121 non-IBD based methods to measure the strength of founder events and in cases  
122 where a comparison was possible found high correlation of these results with our  
123 IBD scores (Supplementary Text and Supplementary Table 3).

124  
125 We infer that 81 out of 263 unique groups (96 out of 327 groups if not considering  
126 the overlap of groups among datasets) have an IBD score greater than those of both  
127 Finns and Ashkenazi Jews (Figure 3). These results did not change when we added  
128 back the outlier samples that we removed in quality control. A total of 14 of these  
129 groups have estimated census sizes of over a million (Figure 3; Table 1). However,  
130 the groups with smaller census sizes are also very important – outside of South Asia,  
131 small census size groups with extremely strong founder events such as Amish,  
132 Hutterites, and people of the Saguenay Lac-St. Jean region have led to the discovery  
133 of dozens of novel disease causing variants. We also searched for IBD across groups  
134 – screening for cases in which the across-group IBD score is at least a third of the

135 within-group IBD score of Ashkenazi Jews – and found 77 cases of clear IBD-sharing,  
136 which typically follow geography, religious affiliation, or linguistic grouping  
137 (particularly Austroasiatic speakers) (Supplementary Table 4). Pairs of groups with  
138 high shared IBD and descent from a common founder event will share risk for the  
139 same recessive disease. However, these cross-group IBD sharing patterns are not  
140 driving our observations, as we still identify 68 unique sets of groups without high  
141 IBD to other groups that have significantly higher estimated IBD scores than both  
142 Finns and Ashkenazi Jews.

143

144 Our documentation that very strong founder events affect a large fraction of South  
145 Asian groups presents an opportunity for decreasing disease burden in South Asia.  
146 This source of risk for recessive diseases is very different from that due to marriages  
147 among close relatives, which is also a major cause of recessive disease in South Asia.  
148 To determine the relative impact of these factors, we computed  $F_{ST}$ , a measurement  
149 of allele frequency differentiation, between each group in the dataset and a pool of  
150 other South Asian groups chosen to be closest in terms of ancestry proportions. We  
151 find that inbreeding is not driving many of these signals, as 89 unique groups have  
152 higher  $F_{ST}$  scores than those of Ashkenazi Jews and Finns even after reducing the  $F_{ST}$   
153 score by the proportion of allele frequency differentiation due to inbreeding. These  
154 results show that while most recessive disease gene mapping studies in South Asia  
155 have focused on families that are the products of marriages between close relatives,  
156 recessive diseases are also likely to occur at an elevated rate even in non-  
157 consanguineous cases because of shared ancestors more distantly in time.

158

159 As an example of the promise of founder event disease gene mapping in South Asia,  
160 we highlight the case of the Vysya, who have a census size of more than 3 million  
161 and who we estimate have an IBD score about 1.2-fold higher than Finns (Figure 3).  
162 The Vysya have an approximately 100-fold higher rate of butyrylcholinesterase  
163 deficiency than other groups, and Vysya ancestry is a known counter-indication for  
164 the use of muscle relaxants such as succinylcholine or mivacurium that are given  
165 prior to surgery<sup>13</sup>. This disease is likely to occur at a higher rate due to the founder  
166 event in Vysya's history, and we expect that, like Finns, Vysya likely have a higher  
167 rate of many other diseases compared to other groups. Other examples of recessive  
168 disease genes with a likely origin in founder events are known anecdotally in South  
169 Asia, highlighting the importance of systematic studies to find them<sup>14</sup>.

170

171 To demonstrate how a new recessive disease in a founder event group could be  
172 mapped, we carried out whole genome SNP genotyping in 12 patients from southern  
173 India who had progressive pseudorheumatoid dysplasia (PPD), a disease known to  
174 be caused by mutations in the gene *WISP3*<sup>15,16</sup>. Of the 6 individuals with the  
175 Cys78Tyr mutation in *WISP3*,<sup>15,16</sup> 5 were from non-consanguineous marriages, and  
176 we found a much higher fraction of IBD at the disease mutation site than in the rest  
177 of the genome in these individuals (Supplementary Figure 3a; Supplementary Figure  
178 4a), consistent with the Cys78Tyr mutation that causes PPD in these patients owing  
179 its origin to a founder event. This pattern contrasts with the 6 other patients with  
180 different disease variants and 6 patients with a mutation causing a different disease

181 (mucopolysaccharidosis (MPS) type IVA) who were from primarily consanguineous  
182 marriages, and who lacked significant IBD across their disease mutation sites,  
183 implying that in the case of these groups the driver for the recessive diseases was  
184 marriage between close relatives (Supplementary Text). This example highlights  
185 how not only marriages of close relatives, but also founder events, are substantial  
186 causes of rare recessive disease in South Asia.

187

188 The evidence of widespread strong founder events presents a major opportunity for  
189 disease gene discovery and public health intervention in South Asia that is not  
190 widely appreciated. The current paradigm for recessive disease gene mapping in  
191 South Asia is to collect cases in tertiary medical centers and map diseases in  
192 individuals with the same phenotype, often blinded to information about group  
193 affiliation as in the case of the PPD study where we do not have access to the ethnic  
194 group information. However, our results suggest that collecting information on  
195 group affiliation could be greatly strengthen the power of these studies. A fruitful  
196 way to approach gene mapping would be to proactively survey communities known  
197 to have strong founder events, searching for diseases that occur at high rates in  
198 these communities. This approach was pioneered in the 1950s in studies of the Old  
199 Order Amish in the U.S., a founder population of approximately 100,000 individuals  
200 in whom many dozens of recessive diseases were mapped, a research program that  
201 was crucial to founding modern medical genetics and that was of extraordinary  
202 health benefit. Our study suggests that the potential for disease gene mapping in  
203 South Asia would be orders of magnitude greater.

204

205 Mapping of recessive diseases may be particularly important in communities  
206 practicing arranged marriages, which are common in South Asia. An example of the  
207 power of this approach is given by *Dor Yeshorim*, a community genetic testing  
208 program among religious Ashkenazi Jews<sup>17</sup>, which visits schools, screens students  
209 for common recessive disease causing mutations previously identified to be  
210 segregating at a higher frequency in the target group, and enters the results into a  
211 confidential database. Matchmakers query the database prior to making suggestions  
212 to the families and receive feedback about whether the potential couple is  
213 “incompatible” in the sense of both being carriers for a recessive mutation at the  
214 same gene. Given that approximately 95% of community members whose marriages  
215 are arranged participate in this program, recessive diseases like Tay-Sachs have  
216 virtually disappeared in these communities. A similar approach should work as well  
217 in South Asian communities. Given the potential for saving lives, this or similar  
218 kinds of research could be a valuable investment for future generations<sup>18</sup>.

219



## 220 **Supplementary Data:**

221  
222 Supplementary Data include an Excel spreadsheet detailing all groups and their  
223 scores on the IBD,  $F_{ST}$ , and group-specific drift analyses. Also included are 7  
224 supplementary figures and 5 supplementary tables.  
225

## 226 **Acknowledgements:**

227  
228 We are thankful to the many Indian, Pakistani, Bangladeshi, Sri Lankan, and  
229 Nepalese individuals who contributed the DNA samples analyzed here including the  
230 PPD and MPS patients. We are grateful to Analabha Basu and Partha P. Majumder  
231 for early sharing of data. Funding was provided by an NIGMS (GM007753)  
232 fellowship to NN, a Translational Seed Fund grant from the Dean's Office of Harvard  
233 Medical School and an HG006399 to DR, Council of Scientific and Industrial  
234 Research, Government of India grant to KT, and an NIGMS grant 115006 to PM. SS  
235 and SMJ acknowledge the funding from the Department of Biotechnology  
236 (BT/PR4224/MED/97/60/2011) and Department of Science and Technology  
237 (SR/WOS-A/LS-83/2011), Government of India. Funding for the mutation analysis  
238 of Indian patients with PPD was provided by Indian Council of Medical Research  
239 (BMS 54/2/2013) to KMG. DR is an Investigator of the Howard Hughes Medical  
240 Institute. The informed consents and permits associated with the newly reported  
241 data are not consistent with fully public release. Therefore, researchers who wish to  
242 analyze the data should send the corresponding authors a PDF of a signed letter  
243 containing the following language: "(a) We will not distribute the data outside my  
244 collaboration, (b) We will not post data publicly, (c) We will make no attempt to  
245 connect the genetic data to personal identifiers, (d) We will not use the data for  
246 commercial purposes."  
247

## 248 **Author Contributions:**

249  
250 N.N., P.M., D.R., and K.T. conceived the study. N.N., P.M., N.R., B.S., A.T., N.P. and D.R.  
251 performed analysis. G.B., K.M.G., M.S.M., S.S. A.K., S.A.V., S.M.J., K.S., L.S. and K.T.  
252 collected data. N.N., D.R., and K.T. wrote the manuscript with the help of all co-  
253 authors.  
254

## 255 **Competing Financial Interests:**

256  
257 The authors declare no competing financial interests.  
258

259 Reprints and permissions information is available online at  
260 <http://www.nature.com/reprints/index.html>.

261  
262  
263

## References:

264  
265

- 266 1. Mastana, S.S. Unity in diversity: an overview of the genomic anthropology of  
267 India. *Ann Hum Biol* **41**, 287-99 (2014).
- 268 2. Bamshad, M.J. *et al.* Female gene flow stratifies Hindu castes. *Nature* **395**,  
269 651-2 (1998).
- 270 3. Basu, A. *et al.* Ethnic India: a genomic view, with special reference to peopling  
271 and structure. *Genome Res* **13**, 2277-90 (2003).
- 272 4. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. Reconstructing  
273 Indian population history. *Nature* **461**, 489-94 (2009).
- 274 5. Lim, E.T. *et al.* Distribution and medical impact of loss-of-function variants in  
275 the Finnish founder population. *PLoS Genet* **10**, e1004494 (2014).
- 276 6. Arcos-Burgos, M. & Muenke, M. Genetics of population isolates. *Clin Genet* **61**,  
277 233-47 (2002).
- 278 7. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-  
279 93 (2012).
- 280 8. Moorjani, P. *et al.* Genetic evidence for recent population mixture in India. *Am*  
281 *J Hum Genet* **93**, 422-38 (2013).
- 282 9. Metspalu, M. *et al.* Shared and unique components of human population  
283 structure and genome-wide signals of positive selection in South Asia. *Am J*  
284 *Hum Genet* **89**, 731-44 (2011).
- 285 10. Behar, D.M. *et al.* The genome-wide structure of the Jewish people. *Nature*  
286 **466**, 238-42 (2010).
- 287 11. Basu, A., Sarkar-Roy, N. & Majumder, P.P. Genomic reconstruction of the  
288 history of extant populations of India reveals five distinct ancestral  
289 components and a complex structure. *Proc Natl Acad Sci U S A* (2016).
- 290 12. Genomes Project, C. *et al.* A global reference for human genetic variation.  
291 *Nature* **526**, 68-74 (2015).
- 292 13. Manoharan, I., Wieseler, S., Layer, P.G., Lockridge, O. & Boopathy, R. Naturally  
293 occurring mutation Leu307Pro of human butyrylcholinesterase in the Vysya  
294 community of India. *Pharmacogenet Genomics* **16**, 461-8 (2006).
- 295 14. Anju Shukla, M.H., Anshika Srivastava, Rajagopal Kadavigere, Priyanka  
296 Upadhyai, Anil Kanthi, Oliver Brandau, Stephanie Bielas, Katta Girisha.  
297 Homozygous c.259G>A variant in ISCA1 is associated with a new multiple  
298 mitochondrial dysfunctions syndrome. *bioRxiv* (2016).
- 299 15. Dalal, A. *et al.* Analysis of the WISP3 gene in Indian families with progressive  
300 pseudorheumatoid dysplasia. *Am J Med Genet A* **158A**, 2820-8 (2012).
- 301 16. Bhavani, G.S. *et al.* Novel and recurrent mutations in WISP3 and an atypical  
302 phenotype. *Am J Med Genet A* **167A**, 2481-4 (2015).
- 303 17. Raz, A.E. Can population-based carrier screening be left to the community? *J*  
304 *Genet Couns* **18**, 114-8 (2009).
- 305 18. Rajasimha, H.K. *et al.* Organization for rare diseases India (ORDI) -  
306 addressing the challenges and opportunities for the Indian rare diseases'  
307 community. *Genet Res (Camb)* **96**, e009 (2014).
- 308 19. Sudmant, P.H. *et al.* Global diversity, population stratification, and selection  
309 of human copy-number variation. *Science* **349**, aab3761 (2015).

- 310 20. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from  
311 142 diverse populations. *Nature* **538**, 201-206 (2016).
- 312 21. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human  
313 genomes. *Nature* **526**, 75-81 (2015).
- 314 22. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis.  
315 *PLoS Genet* **2**, e190 (2006).
- 316 23. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger  
317 and richer datasets. *Gigascience* **4**, 7 (2015).
- 318 24. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden  
319 relatedness. *Genome Res* **19**, 318-26 (2009).
- 320 25. Hoaglin, B.I.a.D. *How to Detect and Handle Outliers*, (1993).
- 321 26. Palamara, P.F. ARGON: fast, whole-genome simulation of the discrete time  
322 Wright-fisher process. *Bioinformatics* (2016).
- 323 27. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and  
324 missing-data inference for whole-genome association studies by use of  
325 localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).
- 326 28. Durand, E.Y., Eriksson, N. & McLean, C.Y. Reducing pervasive false-positive  
327 identical-by-descent segments detected by large-scale pedigree analysis. *Mol*  
328 *Biol Evol* **31**, 2212-22 (2014).
- 329 29. Browning, B.L. & Browning, S.R. Improving the accuracy and efficiency of  
330 identity-by-descent detection in population data. *Genetics* **194**, 459-71  
331 (2013).
- 332 30. Bidchol, A.M. *et al.* GALNS mutations in Indian patients with  
333 mucopolysaccharidosis IVA. *Am J Med Genet A* **164A**, 2793-801 (2014).
- 334



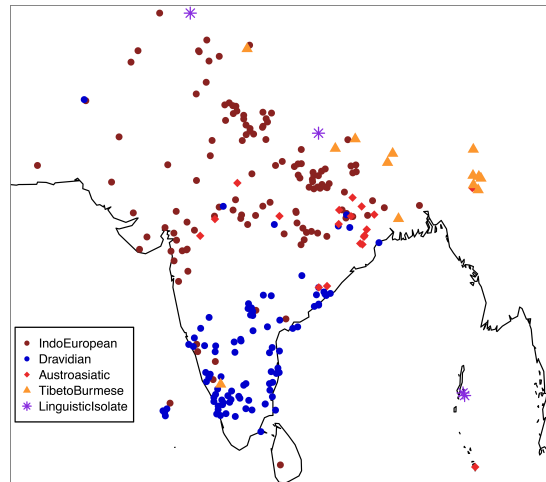
335

Group	Sample Size	IBD Score	IBD Rank	F <sub>ST</sub> Rank	Drift Rank	Census Size	Location
Gujjar	5	11.6	19	33	46	1,078,719	Jammu and Kashmir
Baniyas	7	9.6	24	22	18	4,200,000	Uttar Pradesh
Pattapu_Kapu	4	9.5	25	24	21	13,697,000	Andhra Pradesh
Vadde	3	9.2	26	30	26	3,695,000	Andhra Pradesh
Yadav	12	4.4	48	87	67	1,124,864	Puducherry
Kshatriya_Aqnikula	4	2.4	75	109	NA	12,809,000	Andhra Pradesh
Naga	4	2.3	76	NA	NA	1,834,483	Nagaland
Kumhar	27	2.3	77	35	197	3,144,000	Uttar Pradesh
Reddy	7	2.0	84	129	106	22,500,000	Telangana
Brahmin_Nepal	4	1.9	86	63	141	4,206,235	Nepal
Kallar	27	1.7	94	87	73	2,426,929	Tamil Nadu
Brahmin_Manipuri	17	1.6	99	NA	NA	1,544,296	Manipur
Arunthathiyar	18	1.3	108	109	81	1,192,578	Tamil Nadu
Vysya	39	1.2	110	46	35	3,200,000	Telangana

336

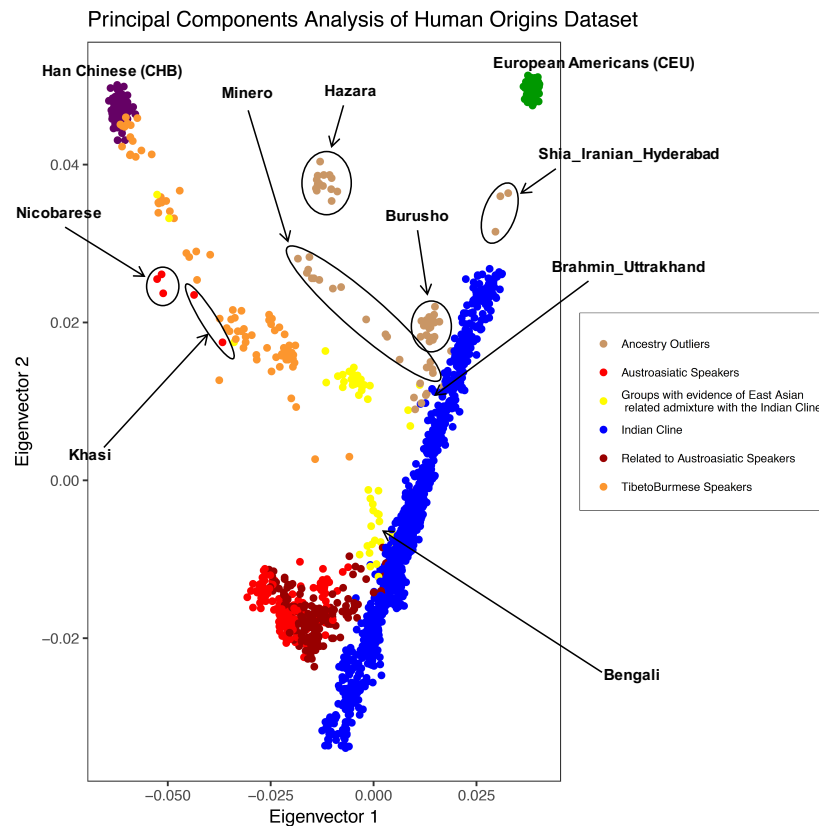
337 **Table 1. South Asian groups with estimated census sizes over 1 million and IBD scores significantly greater than**  
338 **those of Ashkenazi Jews and Finns.** Fourteen South Asian groups with IBD scores significantly higher than that of Finns,  
339 census sizes over 1 million, and sample sizes of at least 3 that are of particularly high interest for founder event disease  
340 gene mapping studies. For reference, Finns and Ashkenazi Jews (on the Human Origins array) would have IBD scores of  
341 1.0 and 0.9, IBD ranks of 121 and 135, and F<sub>ST</sub> ranks of 109 and 129, respectively (the group-specific drift is difficult to  
342 compare for groups with significantly different histories, so they were not calculated for Finns or Ashkenazi Jews).

A



343  
344

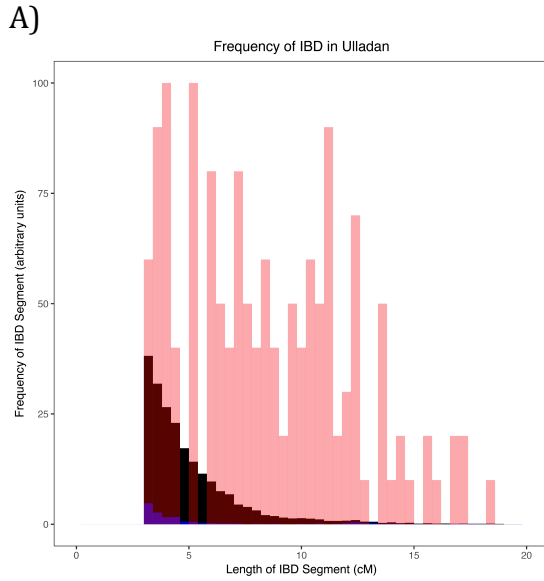
B



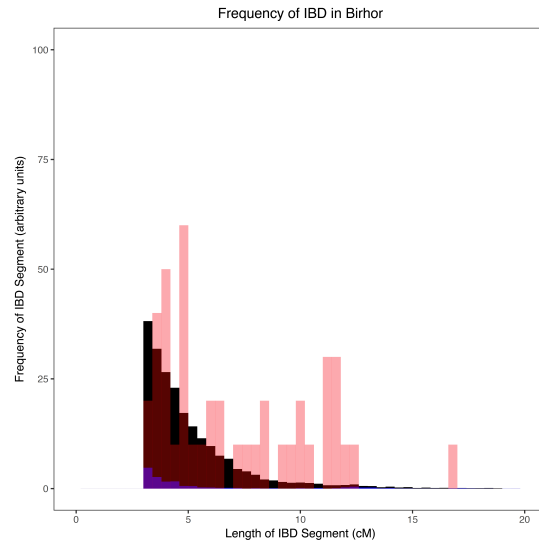
345

346 **Figure 1. Dataset overview.** (a) Sampling locations for all analyzed groups. Each  
347 point indicates a distinct group (random jitter was added to help in visualization at  
348 locations where there are many groups). (b) PCA of Human Origins dataset along  
349 with European Americans (CEU) and Han Chinese (CHB). There is a large cluster  
350 (blue) of IndoEuropean and Dravidian speaking groups that stretch out along a line  
351 in the plot and that are well-modeled as a mixture of two highly divergent ancestral  
352 populations (the “Indian Cline”). There is another larger cluster of Austroasiatic  
353 speakers (light red) and groups that cluster with them genetically (dark red).  
354 Finally, there are groups with genetic affinity to East Asians that include Tibeto-  
355 Burman speakers (orange) and those that speak other languages (yellow).

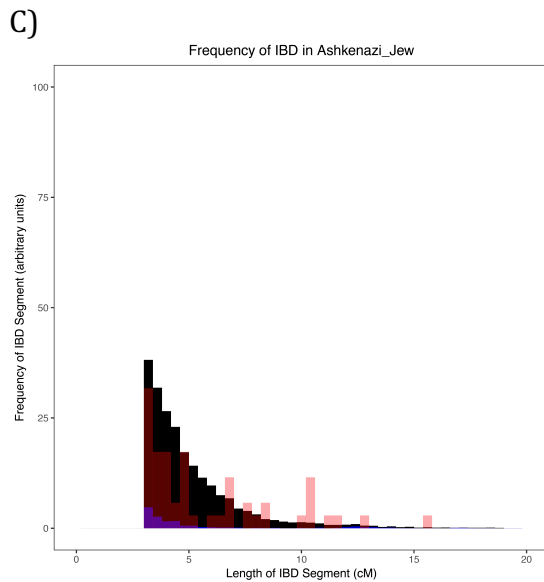
356



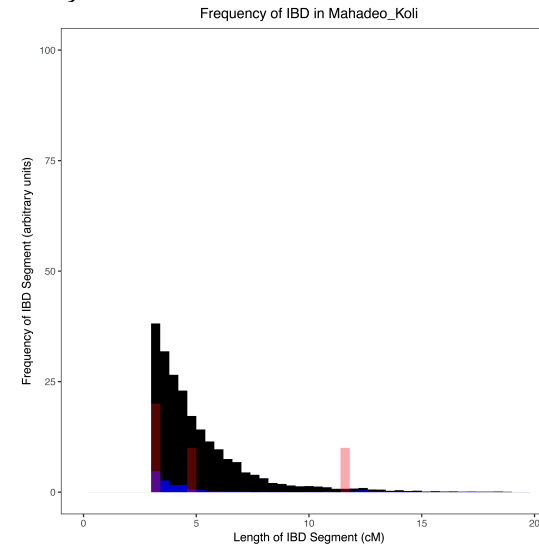
B)



357  
358



D)



359

360

361

362

363

364

365

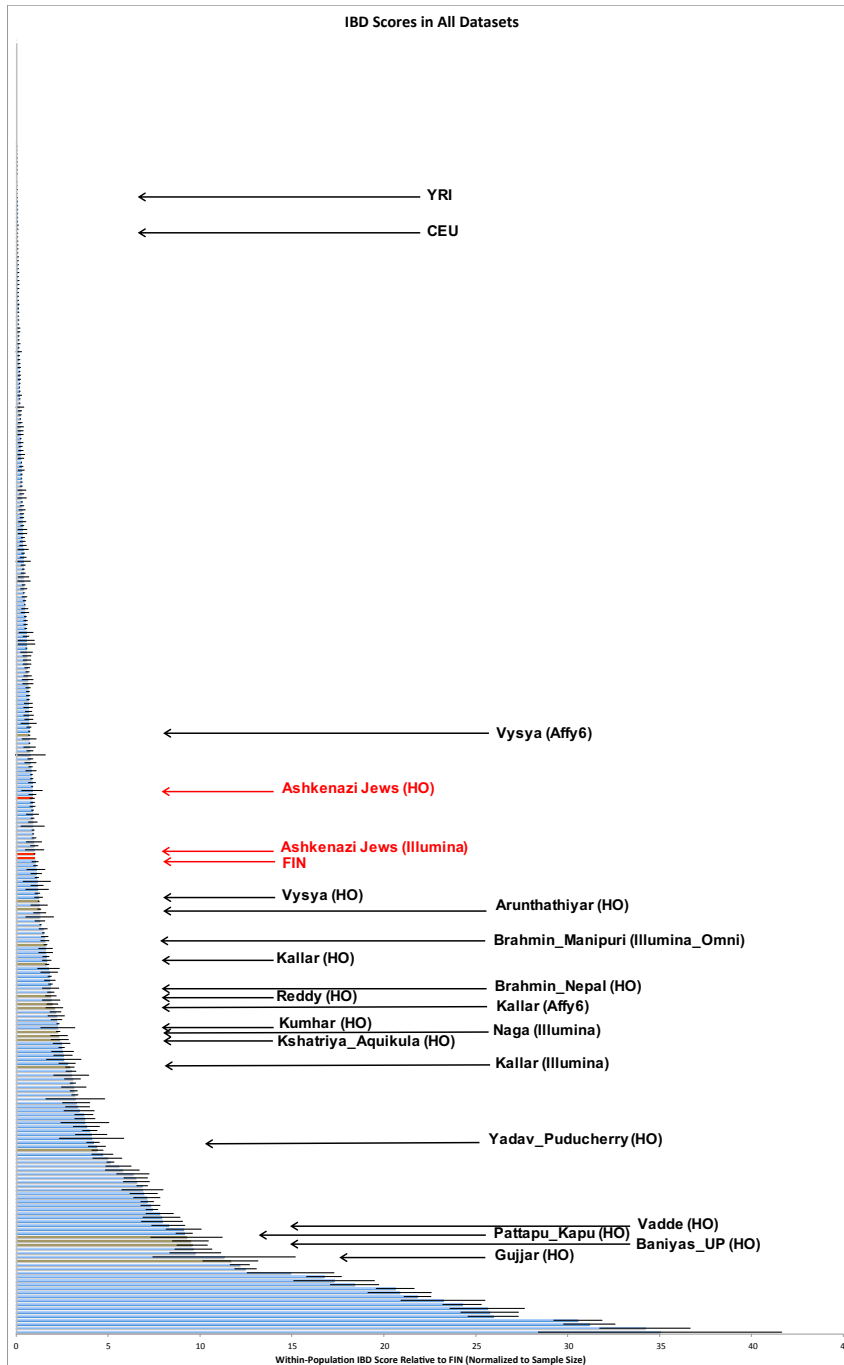
366

367

368

369

**Figure 2. Example histograms of IBD segments to illustrate the differences between groups with founder events of different magnitudes:** These histograms provide visual illustrations of differences between groups with different IBD scores. As a ratio relative to Finns (FIN; black), these groups (red) have IBD scores of: (A) ~26 in Ulladan, (B) ~3 in Birhor, (C) ~0.9 in Ashkenazi Jews, and (D) ~0.1 in Mahadeo\_Koli. In each plot, we also show European Americans (CEU) with a negligible founder event in blue. Quantification of these founder events is shown in Figure 3 and Online Table 1. The IBD histograms were normalized for sample size by dividing their frequency by  $\binom{2^n}{2} - n$ , where  $n$  is the number of individuals in the sample. All data for the figure are based on the Human Origins dataset.



370  
371  
372  
373  
374  
375  
376  
377  
378  
379

**Figure 3. IBD scores relative to Finns (FIN).** Histogram ordered by IBD score, 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 South Asian groups with significantly higher IBD scores than that of Finns and census sizes of more than a million in brown. Error bars for each IBD score are standard errors calculated by weighted block jackknife over each chromosome. YRI= Yoruba (West African); CEU= European American.

## 380 **Online Methods:**

### 381 **Data Sets:**

382  
383  
384 We assembled a dataset of 1,955 individuals from 249 groups genotyped on the  
385 Affymetrix Human Origins array, of which data from 1,663 individuals from 230  
386 groups are newly reported here (Figure 1a). We merged these data with the dataset  
387 published in Moorjani *et al.*<sup>8</sup>, which consisted of 332 individuals from 52 groups  
388 genotyped on the Affymetrix 6.0 array. We also merged it with two additional  
389 datasets published in Metspalu *et al.*<sup>9</sup>, consisting of 151 individuals from 21 groups  
390 genotyped on Illumina 650K arrays as well as a dataset published in Basu *et al.*<sup>11</sup>,  
391 consisting of 367 individuals from 20 groups generated on Illumina Omni 1-Quad  
392 arrays. These groups come from India, Pakistan, Nepal, Sri Lanka, and Bangladesh.  
393 All samples were collected under the supervision of ethical review boards in India  
394 with informed consent obtained from all subjects.

395  
396 We analyzed two different Ashkenazi Jewish datasets, one consisting of 21  
397 individuals genotyped on Illumina 610K and 660K bead arrays<sup>10</sup> and one consisting  
398 of 7 individuals genotyped on Affymetrix Human Origins arrays.

399  
400 Our “Affymetrix 6.0” dataset consists of 332 individuals genotyped on 329,261 SNPs,  
401 and our “Illumina\_Omni” dataset consists of 367 individuals genotyped on 750,919  
402 SNPs. We merged the South Asian and Ashkenazi Jewish data generated by the other  
403 Illumina arrays to create an “Illumina” dataset consisting of 172 individuals  
404 genotyped on 500,640 SNPs. We merged the data from the Affymetrix Human  
405 Origins arrays with the Ashkenazi Jewish data and data from the Simons Genome  
406 Diversity Project<sup>19,20</sup> to create a dataset with 4,402 individuals genotyped on  
407 512,615 SNPs. We analyzed the four datasets separately due to the small  
408 intersection of SNPs between them. We merged in the 1000 Genomes Phase 3 data<sup>21</sup>  
409 (2,504 individuals from 26 different groups; notably, including 99 Finnish  
410 individuals) into all of the datasets. We used genome reference sequence  
411 coordinates (hg19) for analyses.

### 412 **Quality Control:**

413  
414  
415 We filtered the data at both the SNP and individual level. On the SNP level, we  
416 required at least 95% genotyping completeness for each SNP (across all  
417 individuals). On the individual level, we required at least 95% genotyping  
418 completeness for each individual (across all SNPs).

419  
420 To test for batch effects due to samples from the same group being genotyped on  
421 different array plates, we studied instances where samples from the same group  $A$   
422 were genotyped on both plates 1 and 2 and computed an allele frequency difference  
423 at each SNP,  $Diff_A^i = (Freq_{PopA,Plate1}^i - Freq_{PopA,Plate2}^i)$ . We then computed the  
424 product of these allele frequencies averaged over all SNPs for two groups  $A$  and  $B$   
425 genotyped on the same plates,  $\frac{1}{n} \sum_{i=1}^n (Diff_A^i)(Diff_B^i)$ , as well as a standard error



426 from a weighted Block Jackknife across chromosomes. This quantity should be  
427 consistent with zero within a few standard errors if there are no batch effects that  
428 cause systematic differences across the plates, as allele frequency differences  
429 between two samples of the same group should be random fluctuations that have  
430 nothing to do with the array plates on which they are genotyped. This analysis  
431 found strong batch effects associated with one array plate, and we removed these  
432 samples from further analysis.

433

434 We used EIGENSOFT 5.0.1 smartpca<sup>22</sup> on each group to detect PCA outliers and  
435 removed 51 samples. We also developed a procedure to distinguish recent  
436 relatedness from founder events so that we could remove recently related  
437 individuals. We first identified all duplicates or obvious close relatives by using  
438 Plink<sup>23</sup> “genome” and GERMLINE<sup>24</sup> to compute IBD (described in more detail below)  
439 and removed one individual from all pairs with a PI\_HAT score greater than 0.45  
440 and the presence of at least 1 IBD fragment greater than 30cM. We then used an  
441 iterative procedure to identify additional recently related individuals. For sample  
442 sizes above 5, we identified any pairs within each group that had both total IBD and  
443 total long IBD (>20cM) that were greater than 2.5 SDs and 1 SD, respectively, from  
444 the group mean. For sample sizes 5 or below, we used modified Z scores of  
445  $0.6745 * (\text{IBD\_score} - \text{median}(\text{score})) / \text{MAD}$ , where MAD is the median absolute  
446 deviation, and identified all pairs with modified Z scores greater than 3.5 for both  
447 total IBD and total long IBD as suggested by Iglewicz and Hoaglin<sup>25</sup>. After each  
448 round, we repeated the process if the new IBD score was at least 30% lower than  
449 the prior IBD score. Simulations showed that we were always able to remove a first  
450 or second cousin in the dataset using this method (Supplementary Table 1).  
451 Together these analyses removed 53 individuals from the Affymetrix 6.0 dataset, 21  
452 individuals from the Illumina dataset, 43 individuals from the Illumina Omni  
453 dataset, and 225 individuals from the Human Origins dataset.

454

455 After data quality control and merging with the 1000 Genomes Project data, the  
456 Affymetrix 6.0 dataset included 2,842 individuals genotyped on 326,181 SNPs, the  
457 Illumina dataset included 2,662 individuals genotyped on 484,293 SNPs, the  
458 Illumina Omni dataset included 2,828 individuals genotyped on 750,919 SNPs, and  
459 the Human Origins dataset included 4,177 individuals genotyped at 499,158 SNPs.

460

#### 461 **Simulations to Test Relatedness Filtering and IBD Analyses**

462

463 We used ARGON<sup>26</sup> to simulate groups with different bottleneck strengths to test the  
464 IBD analyses and relatedness filtering. We used ARGON’s default settings, including  
465 a mutation rate of  $1.65 * 10^{-8}$  per base pair (bp) per generation and a recombination  
466 rate of  $1 * 10^{-8}$  per bp per generation and simulated 22 chromosomes of size 130 Mb  
467 each. We pruned the output by randomly removing SNPs until there were 22,730  
468 SNPs per chromosome to simulate the approximate number of positions in the  
469 Affymetrix Human Origins array. For the IBD analyses, we simulated groups to have  
470 descended from an ancestral group 1,800 years ago with  $N_e=50,000$  and to have  
471 formed two groups with  $N_e=25,000$ . These groups continued separately until 100

472 generations ago when they combined in equal proportions to form a group with  
473  $N_e=50,000$ . The group then split into 3 separate groups 72 generations ago that have  
474 bottlenecks leading to  $N_e$  of either 400, 800, or 1600. The 3 groups then  
475 exponentially expanded to a present size of  $N_e=50,000$ . We designed these  
476 simulations to capture important features of demographic history typical of Indian  
477 groups<sup>4,8</sup>. We chose the bottleneck sizes because they represent founder events with  
478 approximately the strength of Finns (the bottleneck to 800), and twice as strong  
479 (400) and half as strong (1600) as that group. We then performed the IBD analyses  
480 described below with 99 individuals from the group with bottleneck strength  
481 similar to that of Finns (198 haploid individuals were simulated and merged to  
482 produce 99 diploid individuals) and different numbers of individuals from the other  
483 groups. These analyses demonstrate that with only 4-5 individuals we can  
484 accurately assess the strength of founder events in groups with strong founder  
485 events (Supplementary Figure 2 and Supplementary Table 2). Weaker founder  
486 events are more difficult to assess, but these groups are of less interest for founder  
487 event disease mapping, so we aimed to sample  $\sim 5$  individuals per group.  
488

489 We wrote custom R scripts to simulate first and second cousin pairs. We took  
490 individuals from the bottleneck of size 800 and performed “matings” by taking 2  
491 individuals and recombining their haploid chromosomes assuming a rate of  $1 \times 10^{-8}$   
492 per bp per generation across the chromosome and combining one chromosome  
493 from each of these individuals to form a new diploid offspring. The matings were  
494 performed to achieve first and second cousins. We then placed these back into the  
495 group with group of size 800, and ran the relatedness filtering algorithms to  
496 evaluate whether they would identify these individuals.  
497

#### 498 **Phasing, IBD Detection, and IBD Score Algorithm:**

500 We phased all datasets using Beagle 3.3.2 with the settings *missing=0; lowmem=true;*  
501 *gprobs=false; verbose=true*<sup>27</sup>. We left all other settings at default. We determined IBD  
502 segments using GERMLINE<sup>24</sup> with the parameters *-bits 75 -err\_hom 0 -err\_het 0 -*  
503 *min\_m 3*. We used the genotype extension mode to minimize the effect of any  
504 possible phasing heterogeneity amongst the different groups and used the  
505 HaploScore algorithm to remove false positive IBD fragments with the  
506 recommended genotype error and switch error parameters of 0.0075 and 0.003<sup>28</sup>.  
507 We chose a HaploScore threshold matrix based on calculations from Durand *et al.*<sup>28</sup>  
508 for a “mean overlap” of 0.8, which corresponds to a precision of approximately 0.9  
509 for all genetic lengths from 2-10cM. It can sometimes be difficult to measure IBD in  
510 admixed populations due to differential proportions of the divergent ancestries  
511 amongst different individuals in the same group, but we found that in both the  
512 simulated and real data we were able to detect IBD at the expected amounts.  
513

514 In addition to the procedure we developed to remove close relatives (Quality  
515 Control section), we also removed segments longer than 20cM as simulations  
516 showed that this increased sensitivity of the analyses (Supplementary Table 2). We  
517 computed “IBD score” as the total length of IBD segments between 3-20cM divided

518 by  $\left\{\binom{2n}{2} - n\right\}$  where  $n$  is the number of individuals in each group to normalize for  
519 sample size. We then expressed each group's score as a ratio of their IBD score to  
520 that of Finns and calculated standard errors for this score using a weighted Block  
521 Jackknife over each chromosome with 95% confidence intervals defined as IBD  
522 score  $\pm 1.96 * s.e.$

523

524 We repeated these analyses with FastIBD<sup>29</sup> for the Affymetrix 6.0 and Illumina  
525 datasets and observed that the results were highly correlated ( $r > 0.96$ ) (data not  
526 shown). We chose GERMLINE for our main analyses, however, because the FastIBD  
527 algorithm required us to split the datasets into different groups, since it adapts to  
528 the relationships between LD and genetic distance in the data, and these  
529 relationships differ across groups. We used data from several different Jewish  
530 groups and all twenty-six 1000 Genomes groups to improve phasing, but of these  
531 groups we only included results for Ashkenazi Jews and two outbred groups (CEU  
532 and YRI) in the final IBD score ranking.

533

### **Disease patient analyses:**

534

535 We use Affymetrix Human Origins arrays to successfully genotype 12 patients with  
536 progressive pseudorheumatoid dysplasia (PPD) and 6 patients with  
537 mucopolysaccharidosis (MPS) type IVA, all of whom had disease mutations  
538 previously determined<sup>15,16,30</sup> (3 of the surveyed MPS patients are newly reported  
539 here). A total of 6 of the PPD patients had Cys78Tyr mutations, 6 had Cys337Tyr  
540 mutations (all 6 of the MPS patients had Cys78Arg mutations). We measured IBD as  
541 described above and also detected homozygous segments within each individual by  
542 using GERMLINE with the parameters *-bits 75 -err\_hom 2 -err\_het 0 -min\_m 0.5 -*  
543 *homoz-only*.

544

545 Haplotype sharing was assessed by analyzing phased genotypes for each mutation  
546 group. At each SNP, we counted the number of identical genotypes for each allele  
547 and computed the fraction by dividing by the total number of possible haplotypes (2  
548 times the number of individuals). We took the larger value of the two possible  
549 alleles (thus the fraction range was 0.5-1). We averaged these values over blocks of  
550 10 or 25 SNPs and plotted the averages around the relevant mutation site.

551

### **Between-Group IBD Calculations:**

552

553 We determined IBD using GERMLINE as above. We collapsed individuals into  
554 respective groups and normalized for between-group IBD by dividing all IBD from  
555 each group by  $\left\{\binom{2n}{2}\right\}$  where  $n$  is the number of individuals in each group. We  
556 normalized for within-group IBD as described above. We defined groups with high  
557 shared IBD as those with an IBD score greater than three times the founder event  
558 strength of CEU (and  $\sim 1/3$  the event strength of Ashkenazi Jews).

559

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

560

561

562

563

564 We used the  $f_3$ -statistic<sup>7</sup>  $f_3(\text{Test}; \text{Ref}_1, \text{Ref}_2)$  to determine if there was evidence that  
565 the *Test* group was derived from admixture of groups related to *Ref*<sub>1</sub> and *Ref*<sub>2</sub>. A  
566 significantly negative statistic provides unambiguous evidence of mixture in the  
567 *Test* group. We determined the significance of the  $f_3$ -statistic using a Block Jackknife  
568 and a block size of 5 cM. We considered statistics over 3 standard errors below zero  
569 to be significant.

570

#### 571 **Computing Group Specific Drift:**

572

573 We used qpGraph<sup>7</sup> to model each Indian group on the cline as a mixture of ANI and  
574 ASI ancestry, using the model (YRI, (Indian group, (Georgians, ANI)), [(ASI, Onge)])  
575 proposed by Moorjani *et al.*<sup>8</sup> This approach provides estimates for post-admixture  
576 drift in each group (Supplementary Figure 5), which is reflective of the strength of  
577 the founder event (high drift values imply stronger founder events). We only  
578 included groups on the Indian cline in this analysis, and we removed all groups with  
579 evidence of East Asian related admixture (Figure 1b and Supplementary Table 5)  
580 because this admixture is not accommodated within the above model.

581

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

583

584 As a third method to measure strength of founder events, we estimated the  
585 minimum  $F_{ST}$  between each South Asian group (Supplementary Figure 6) and their  
586 closest clusters based on PCA (Supplementary Text) (the clusters were used to  
587 account for intermarriage across groups that would otherwise produce a downward  
588 bias in the minimum  $F_{ST}$ ). For the Affymetrix 6.0, Illumina, and Illumina\_Omni  
589 datasets, we split the Indian cline into two different clusters and combined the  
590 Austroasiatic speakers and those with ancestry related to Austroasiatic speakers  
591 (according to the PCA of Figure 1b) into one cluster for a total of three clusters (all  
592 other groups were ignored for this analysis). For the Human Origins dataset we split  
593 the Indian cline into three different clusters and combined the groups with ancestry  
594 related to the main cluster of Austroasiatic speakers into one cluster for a total of  
595 four clusters (Khasi and Nicobarese were ignored in this analysis, because they do  
596 not cluster with the other Austroasiatic speaking groups). We then computed the  $F_{ST}$   
597 between each group and the rest of the individuals in their respective cluster based  
598 on EIGENSOFT *smartpca* with Inbreed set to YES to correct for inbreeding. For  
599 Ashkenazi Jews and Finns, we used the minimum  $F_{ST}$  to other European groups.

600

#### 601 **$F_{ST}$ Calculations to Determine Overlapping Groups:**

602

603 Overlapping groups between the datasets were determined in the first place based  
604 on anthropological information (Online Table 1). We further tested empirically for  
605 overlap by computing  $F_{ST}$  between different groups across all datasets for groups  
606 with significantly stronger IBD scores than those of Finns (we could not perform  
607 this analysis for groups with less strong founder events, because they would have  
608 low  $F_{ST}$  to each other even if they were truly distinct groups). We considered pairs  
609 with  $F_{ST}$  less than 0.004 to be overlapping. These included all groups known to be  
610 overlapping based on anthropological information as well as 3 additional pairs of  
611 groups that might be genetically similar due to recent mixing (e.g. Kanjars and

612 Dharkar are distinct nomadic groups that live near each other but intermarry,  
613 leading to low  $F_{ST}$  between them).

614

615 **Code Availability:**

616  
617 Code for all calculations available upon request.