

The promise of disease gene discovery in South Asia

Nathan Nakatsuka^{1,2}, Priya Moorjani^{3,6}, Niraj Rai⁴, Biswanath Sarkar⁵, Arti Tandon^{1,6}, Nick Patterson⁶, Gandham SriLakshmi Bhavani⁷, Katta Mohan Girisha⁷, Mohammed S Mustak⁸, Sudha Srinivasan⁹, Amit Kaushik¹⁰, Saadi Abdul Vahab¹¹, Sujatha M. Jagadeesh¹², Kapaettu Satyamoorthy¹¹, Lalji Singh^{4,13}, David Reich^{1,5,14,*}, Kumarasamy Thangaraj^{4,*}

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

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

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

⁴CSIR-Centre for Cellular and Molecular Biology, Habsiguda, Hyderabad, Telangana 500007, India

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

⁶Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02141, USA

⁷Department of Medical Genetics, Kasturba Medical College, Manipal University, Manipal, India

⁸Department of Applied Zoology, Mangalore University, Mangalagangothri 574199, Mangalore, Karnataka, India

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

¹⁰Amity Institute of Biotechnology, Amity University, Sector125, Noida 201303, India

¹¹School of Life Sciences, Manipal University, Manipal 576104, India

¹²Fetal Care Research Foundation, 197 Dr. Natesan Road, Chennai 600004, India

¹³Present address: Genome Foundation, Hyderabad 500076, India

¹⁴Howard 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 ethnic group, but as many small endogamous groups. We**
45 **assembled genome-wide data from over 2,800 individuals from over 275**
46 **distinct South Asian groups. We document shared ancestry across groups that**
47 **correlates with geography, language, and religious affiliation. We characterize**
48 **the strength of the founder events that gave rise to many of the groups, and**
49 **identify 14 groups with census sizes of more than a million that descend from**
50 **founder events significantly stronger than those in Ashkenazi Jews and Finns,**
51 **both of which have high rates of recessive disease due to founder events.**
52 **These results highlight a major and under-appreciated opportunity for**
53 **reducing disease burden among South Asians through the discovery of and**
54 **testing for recessive disease genes. As a proof-of-principle, we show that it is**
55 **possible to localize genes for progressive pseudorheumatoid dysplasia and**
56 **mucopolysaccharidosis type IVA, two recessive diseases prevalent in South**
57 **India, using the founder event disease gene mapping approach introduced**
58 **here.**

59
60 South Asia is a region of extraordinary cultural, linguistic, and genetic diversity, with
61 a conservative estimate of over 4,600 anthropologically well-defined groups, many
62 of which are endogamous communities with significant barriers to gene flow due to
63 sociological and cultural factors that restrict marriage between groups¹. Of the small
64 fraction of South Asian groups that have been characterized using genome-wide
65 data, many exhibit large allele frequency differences from geographically close
66 neighbors²⁻⁴, indicating that they have experienced strong founder events, whereby
67 a small number of ancestors gave rise to many descendants today⁴. The pervasive
68 founder events in South Asia present a major opportunity for reducing disease
69 burden in South Asia. The promise is highlighted by studies of founder groups of
70 European ancestry – including Ashkenazi Jews, Finns, Amish, Hutterites, Sardinians,
71 and French Canadians – which have resulted in the discovery of dozens of recessive
72 disease causing mutations in each group. Prenatal testing for these mutations has
73 substantially reduced recessive disease burden in all of these communities^{5,6}.

74
75 To characterize the medically relevant founder events in South Asia, we carried out
76 new genotyping of 1,663 samples from 230 endogamous groups in South Asia on the
77 Affymetrix Human Origins single nucleotide polymorphism (SNP) array⁷. We
78 developed an algorithm to measure the strength of an endogamous group's founder
79 event, aiming to genotype at least five samples per group as our power calculations
80 indicated that this was sufficient to detect a founder event stronger than that in
81 Finns with high confidence (Supplementary Figure 1 and Supplementary Table 1).
82 We combined the new data we collected with previously reported data, leading to
83 four datasets (Figure 1a). The Affymetrix Human Origins SNP array data comprised
84 1,955 individuals from 249 groups in South Asia, to which we added 7 Ashkenazi
85 Jews. The Affymetrix 6.0 SNP array data comprised 383 individuals from 52 groups
86 in South Asia^{4,8}. The Illumina SNP array data comprised 188 individuals from 21
87 groups in South Asia⁹ and 21 Ashkenazi Jews^{9,10}. The Illumina Omni SNP array data
88 comprised 367 individuals from 20 groups in South Asia¹¹. We merged 1000

89 Genomes Phase 3 data¹² (2,504 individuals from 26 different groups including 99
90 Finns) with each of these datasets. We performed quality control to remove SNPs
91 and individuals with a high proportion of missing genotypes or that were outliers in
92 Principal Component Analysis (PCA).

93

94 We performed PCA on each of the three different datasets along with European
95 Americans (CEU), Han Chinese (CHB), and West Africans (YRI), and found that the
96 Siddi are strong outliers as previously reported (Supplementary Figure 2)^{4,13,14}. We
97 next removed YRI, Siddi and indigenous Andamanese (another known outlier) from
98 the datasets and repeated PCA (Figure 1b, Supplementary Figure 3). Similar to past
99 studies, the PCA documents three broad genetic groupings^{4,8,11}. First, almost all
100 South Asian groups speaking Indo-European and Dravidian languages lie along the
101 “Indian Cline,” with different proportions of Ancestral North Indian (ANI) ancestry
102 related to Europeans, Central Asians, and Near Easterners, and Ancestral South
103 Indian (ASI) ancestry that is as different from ANI as Europeans and East Asians are
104 from each other⁴. The second major cluster includes groups that speak Austroasiatic
105 languages, as well as some non-Austroasiatic speaking groups that have similar
106 ancestry possibly due to gene flow with Austroasiatic speaking neighbors or due to
107 a history of language shift. This set of groups cluster together near the ASI end of the
108 Indian cline, likely reflecting a large proportion of ASI-like ancestry as well as a
109 distinct ancestry that has some affinity to East Asians. The Tibeto-Burman speaking
110 groups and other groups with high proportions of East Asian related ancestry such
111 as the Bengali and Austroasiatic speaking Khasi form a gradient of ancestry relating
112 them to East Asian groups such as Han Chinese. These groupings are also evident in
113 a neighbor-joining tree based on F_{ST} (Supplementary Figure 4). We confirmed the
114 East Asian related mixture in some groups by observing significantly negative
115 $f_3(\text{Test}; \text{Mala, Chinese})$ statistics⁷ (Supplementary Table 2).

116

117 We devised an algorithm to quantify the strength of the founder events in each
118 group based on Identity-by-Descent (IBD) segments, large stretches of DNA shared
119 from a common founder in the last approximately one hundred generations (Figure
120 2). We computed an “IBD score” as a measure for the strength of the founder event
121 in each group’s history: the average length of IBD segments between 3-20
122 centimorgans (cM) shared between two genomes normalized to sample size. Since
123 we are interested here in recessive diseases that do not owe their origin to
124 consanguineous marriages of close relatives, we filtered the data to minimize this
125 effect by computing IBD between all pairs of individuals in each group and removing
126 one individual from the pairs with outlying numbers of IBD segments. We validated
127 the effectiveness of this procedure for removing close relatives by simulation
128 (Supplementary Table 2; Online Methods).

129

130 We expressed IBD scores for each group as a fraction of the IBD scores of the 1000
131 Genomes Project Finns merged into each respective dataset. Due to the fact that all
132 the arrays we analyzed included more SNPs ascertained in Europeans than in South
133 Asians, the sensitivity of our methods to founder events is expected to be greater in
134 Europeans than in South Asians, and thus, our estimates of founder event strengths

135 in South Asian groups is likely to be a conservative underestimate relative to that in
136 Europeans (Supplementary Figure 5 demonstrates this effect empirically and shows
137 that it is less of a bias for the strong founder events that are the focus of this study).
138 We computed standard errors for these ratios by a weighted Block Jackknife across
139 chromosomes and declared significance where the 95% confidence intervals did not
140 overlap with 1. Our simulations suggest that for sample sizes of 4, the algorithm's
141 sensitivity was greater than 95% for determining that a group with two times the
142 bottleneck strength as Finns would have an IBD score significantly greater than that
143 of Finns, while specificity was perfect in all the simulations we performed
144 (Supplementary Figure 1 and Supplementary Table 1). We also used two other
145 methods for measuring founder events that did not require phasing or IBD
146 detection. First, we computed F_{ST} between each group and every other group with
147 similar ancestry sources. Second, for groups on the Indian Cline we fit a model of
148 population history using *qpGraph*⁷ and measuring the founder event as the group-
149 drift after admixture (Supplementary Figure 6 and Online Data Table 1). The results
150 of both methods were highly correlated to that of the IBD-based method for all cases
151 where a comparison was possible (Pearson correlation $r=0.82-0.98$).
152

153 The IBD analyses suggest that 29% of the South Asian groups we analyzed (96 out
154 of 327) have significantly stronger founder events than those in both Finns and
155 Ashkenazi Jews (Figure 3). The South Asian groups with evidence of strong founder
156 events include diverse tribe, caste, and religious groups, and our analysis identifies
157 14 groups with strong founder events census sizes of over a million (Figure 3; Table
158 1). The groups with smaller census sizes are also medically significant. Study of
159 small census size groups with extremely strong founder events such as Amish,
160 Hutterites, and the people of the Saguenay Lac-St. Jean region have led to the
161 discovery of dozens of novel disease variants specific to each group⁶, which
162 highlights the potential of similar studies in South Asian groups. In addition to these
163 analyses, we measured IBD across groups – searching for cases in which the across-
164 group IBD score is at least a third of the within-group IBD score of Ashkenazi Jews –
165 and found many cases of relatedness, which typically follow geography, religious
166 affiliation (e.g. Catholic Brahmins), or linguistic grouping (particularly Austroasiatic
167 speakers) (Supplementary Table 3).
168

169 Our documentation that medically significant founder events affect a large fraction
170 of South Asian groups presents an opportunity for decreasing disease burden. This
171 source of risk for recessive diseases is very different from that due to marriages
172 among close relatives, which is also a major cause of recessive disease, especially in
173 southern India. In the case of recessive diseases arising due to founder events, there
174 are mutations that occur recurrently across members of a group (due to deeply
175 shared founders), and these can be mapped at a group level and easily tested.
176

177 As proof-of-principle, we highlight two examples. The first concerns the Vysya, a
178 group with a census size of more than 3 million that we identified as having a
179 founder event about 1.2-fold stronger than that in Finns (Figure 3). The Vysya have
180 a 100-fold higher rate of butyrylcholinesterase deficiency than other Indian groups,

181 and Vysya ancestry is a known counter-indication for the use of muscle relaxants
182 such as succinylcholine or mivacurium that are given prior to surgery¹⁵.

183

184 The second proof-of-principle, we newly genotyped 18 patients from India, 12 of
185 whom had progressive pseudorheumatoid dysplasia (PPD), a disease known to be
186 caused by mutations in the gene *WISP3*^{16,17}, and 6 of whom had
187 mucopolysaccharidosis type IVA, known to be caused by mutations in the gene
188 *GALNS*¹⁸. Though we lacked ethnic group information for most of the 18 patients, 6
189 of the PPD patients carried Cys78Tyr mutations, and 5 of these 6 were from non-
190 consanguineous marriages. We found a much higher fraction of IBD at the disease
191 mutation site than in the rest of the genome (Supplementary Figure 7a). Thus,
192 Cys78Tyr in PPD is a mutation that owes its origin to a founder event^{16,17}. The 6
193 other PPD patients carried Cys337Tyr mutations, and 6 of 6 were from
194 consanguineous marriages, while the 6 patients with MPS carried Cys79Arg
195 mutations, and 4 of 6 were from consanguineous marriages. These patients did not
196 have IBD at the disease mutation site that was detectable using our conservative
197 settings, but we were still able to map the disease locus using homozygosity
198 mapping (Supplementary Figures 7b and 7c) similar to methods used by others^{19,20}.
199 When we examined the haplotypes at the disease loci, we found that each mutation
200 group had high sharing of unique haplotypes (Supplementary Figure 8), but the PPD
201 Cys337Tyr and MPS Cys79Arg haplotypes were smaller than the PPD Cys78Tyr
202 haplotypes. This suggests that these 2 mutations are at high frequency due to older
203 founder events than the one that occurred for the PPD Cys78Tyr mutations, which
204 could explain why they were not discovered by IBD (which is most sensitive for
205 young founder events) and also why they are present primarily in individuals
206 descending from consanguineous marriages (because they may be sufficiently rare
207 that they do not come together at an appreciable rate except in the context of a
208 consanguineous marriage). Beyond the new genotyping we performed here, another
209 study this year demonstrated that an Indian founder mutation in *ISCA1* mutation
210 causes predisposition to a severe mitochondrial dysfunction syndrome.²¹

211

212 These observations highlight how systematic studies of South Asian founder groups
213 are likely to be an effective approach for discovering mutations that cause recessive
214 disease. Identification of pathogenic mutations responsible for such founder
215 diseases is straightforward. All that is required is collection of DNA samples from a
216 small number of affected individuals and their families, usually followed by whole-
217 exome sequencing to discover the causal changes. Once group specific founder
218 event disease mutations are discovered, they can be tested for prenatally. Mapping
219 of recessive disease mutations may be particularly important in traditional
220 communities practicing arranged marriages, which is common in India. An example
221 of the power of this approach from outside India is given by *Dor Yeshorim*, a
222 community genetic testing program among religious Ashkenazi Jews²², which visits
223 schools, screens students for common recessive disease causing mutations
224 previously identified to be segregating at a higher frequency in the target group, and
225 enters the results into a confidential database. Match-makers query the database
226 prior to making suggestions to the families and receive feedback about whether the

227 potential couple is “incompatible” in the sense of both being carriers for a recessive
228 mutation at the same gene. Given that approximately 95% of community members
229 whose marriages are arranged participate in this program, recessive diseases like
230 Tay-Sachs have virtually disappeared in these communities. A similar approach
231 should work as well in Indian communities where arranged marriages are common.
232 Given the potential for saving lives, this or similar kinds of research could serve as
233 an important investment for future generations²³.

234

235 This study of more than 275 distinct groups represents the first systematic survey
236 for founder events in South Asia, and to our knowledge also presents the richest
237 dataset of genome-wide data from anthropologically well-documented groups
238 available from any region in the world. Despite the breadth of this data, the groups
239 surveyed here represent only about 5% of the well-documented endogamous
240 groups in South Asia, and extensions of the survey to all groups would make it
241 possible to identify large numbers of additional founder groups susceptible to
242 recessive diseases. To take advantage of the unique population structure of South
243 Asia to improve health, we propose to reverse the standard approach to recessive
244 disease gene discovery. Instead of focusing entirely on collecting cases in tertiary
245 medical centers and mapping diseases in a group of individuals found to have the
246 same phenotype while blinded to information about their caste or tribal status, we
247 propose that medical geneticists should adopt a parallel strategy of working with
248 community doctors, medical workers, and social workers to identify recessive
249 diseases that occur at a high rate in endogamous groups with founder events, and
250 then go out into the community to identify cases. Once a small number of cases are
251 sampled, it is straightforward to map the causal variants. This approach was
252 pioneered in 1950s by the work of Victor McKusick and his colleagues studying the
253 Old Order Amish in Pennsylvania U.S.A., a founder population of approximately
254 100,000 individuals in whom many dozens of recessive diseases were mapped, a
255 research program that was crucial to founding modern medical genetics and was of
256 extraordinary health benefit to that community²⁴. Our study suggests that the
257 potential for disease gene mapping in India would be orders of magnitude greater.

258

259
260 **Supplementary Data:**

261 Supplementary Data include an excel spreadsheet detailing all groups and their
262 scores on the IBD, F_{ST} , and group-specific drift analyses. Also included are 8
263 supplementary figures and 5 supplementary tables.
264

265
266 **Acknowledgements:**

267 We are thankful to the many Indian, Pakistani, Bangladeshi, and Nepalese
268 individuals who contributed the DNA samples analyzed here including in particular
269 the Indian PPD and MPS patients. We are grateful to Analabha Basu and Partha P.
270 Majumder for sharing data. Funding was provided by an NIGMS (GM007753)
271 fellowship to NN, a Translational Seed Fund grant from the Dean's Office of Harvard
272 Medical School to DR, CSIR network project GENESIS (BSC0121) to KT, and an
273 NIGMS grant 115006 to OPM. SS and SMJ acknowledge the funding from the
274 Department of Biotechnology (BT/PR4224/MED/97/60/2011) and Department of
275 Science and Technology (SR/WOS-A/LS-83/2011), Government of India. DR is an
276 Investigator of the Howard Hughes Medical Institute. Genotyping data for the
277 samples collected for this study will be made available upon request from the
278 corresponding authors.

References:

- 279
280
281 1. Mastana, S.S. Unity in diversity: an overview of the genomic anthropology of
282 India. *Ann Hum Biol* **41**, 287-99 (2014).
283 2. Bamshad, M.J. *et al.* Female gene flow stratifies Hindu castes. *Nature* **395**,
284 651-2 (1998).
285 3. Basu, A. *et al.* Ethnic India: a genomic view, with special reference to peopling
286 and structure. *Genome Res* **13**, 2277-90 (2003).
287 4. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. Reconstructing
288 Indian population history. *Nature* **461**, 489-94 (2009).
289 5. Lim, E.T. *et al.* Distribution and medical impact of loss-of-function variants in
290 the Finnish founder population. *PLoS Genet* **10**, e1004494 (2014).
291 6. Arcos-Burgos, M. & Muenke, M. Genetics of population isolates. *Clin Genet* **61**,
292 233-47 (2002).
293 7. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-
294 93 (2012).
295 8. Moorjani, P. *et al.* Genetic evidence for recent population mixture in India. *Am*
296 *J Hum Genet* **93**, 422-38 (2013).
297 9. Metspalu, M. *et al.* Shared and unique components of human population
298 structure and genome-wide signals of positive selection in South Asia. *Am J*
299 *Hum Genet* **89**, 731-44 (2011).
300 10. Behar, D.M. *et al.* The genome-wide structure of the Jewish people. *Nature*
301 **466**, 238-42 (2010).
302 11. Basu, A., Sarkar-Roy, N. & Majumder, P.P. Genomic reconstruction of the
303 history of extant populations of India reveals five distinct ancestral
304 components and a complex structure. *Proc Natl Acad Sci U S A* (2016).
305 12. Genomes Project, C. *et al.* A global reference for human genetic variation.
306 *Nature* **526**, 68-74 (2015).
307 13. Narang, A. *et al.* Recent admixture in an Indian population of African
308 ancestry. *Am J Hum Genet* **89**, 111-20 (2011).
309 14. Shah, A.M. *et al.* Indian Siddis: African descendants with Indian admixture.
310 *Am J Hum Genet* **89**, 154-61 (2011).
311 15. Manoharan, I., Wieseler, S., Layer, P.G., Lockridge, O. & Boopathy, R. Naturally
312 occurring mutation Leu307Pro of human butyrylcholinesterase in the Vysya
313 community of India. *Pharmacogenet Genomics* **16**, 461-8 (2006).
314 16. Dalal, A. *et al.* Analysis of the WISP3 gene in Indian families with progressive
315 pseudorheumatoid dysplasia. *Am J Med Genet A* **158A**, 2820-8 (2012).
316 17. Bhavani, G.S. *et al.* Novel and recurrent mutations in WISP3 and an atypical
317 phenotype. *Am J Med Genet A* **167A**, 2481-4 (2015).
318 18. Bidchol, A.M. *et al.* GALNS mutations in Indian patients with
319 mucopolysaccharidosis IVA. *Am J Med Genet A* **164A**, 2793-801 (2014).
320 19. Hildebrandt, F. *et al.* A systematic approach to mapping recessive disease
321 genes in individuals from outbred populations. *PLoS Genet* **5**, e1000353
322 (2009).

- 323 20. Hagiwara, K. *et al.* Homozygosity mapping on homozygosity haplotype
324 analysis to detect recessive disease-causing genes from a small number of
325 unrelated, outbred patients. *PLoS One* **6**, e25059 (2011).
- 326 21. Anju Shukla, M.H., Anshika Srivastava, Rajagopal Kadavigere, Priyanka
327 Upadhyai, Anil Kanthi, Oliver Brandau, Stephanie Bielas, Katta Girisha.
328 Homozygous c.259G>A variant in ISCA1 is associated with a new multiple
329 mitochondrial dysfunctions syndrome. *bioRxiv* (2016).
- 330 22. Raz, A.E. Can population-based carrier screening be left to the community? *J*
331 *Genet Couns* **18**, 114-8 (2009).
- 332 23. Rajasimha, H.K. *et al.* Organization for rare diseases India (ORDI) -
333 addressing the challenges and opportunities for the Indian rare diseases'
334 community. *Genet Res (Camb)* **96**, e009 (2014).
- 335 24. Francomano, K.R.D.a.C.A. Victor A. McKusick and the History of Medical
336 Genetics. *Springer Science & Business Media*, 119-130 (2012).
- 337 25. Sudmant, P.H. *et al.* Global diversity, population stratification, and selection
338 of human copy-number variation. *Science* **349**, aab3761 (2015).
- 339 26. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from
340 142 diverse populations. *Nature* **538**, 201-206 (2016).
- 341 27. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human
342 genomes. *Nature* **526**, 75-81 (2015).
- 343 28. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis.
344 *PLoS Genet* **2**, e190 (2006).
- 345 29. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden
346 relatedness. *Genome Res* **19**, 318-26 (2009).
- 347 30. Hoaglin, B.I.a.D. *How to Detect and Handle Outliers*, (1993).
- 348 31. Palamara, P.F. ARGON: fast, whole-genome simulation of the discrete time
349 Wright-fisher process. *Bioinformatics* (2016).
- 350 32. Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*
351 **5**, 164-166 (1989).
- 352 33. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and
353 display of phylogenetic trees made easy. *Nucleic Acids Res* **39**, W475-8
354 (2011).
- 355 34. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and
356 missing-data inference for whole-genome association studies by use of
357 localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).
- 358 35. Durand, E.Y., Eriksson, N. & McLean, C.Y. Reducing pervasive false-positive
359 identical-by-descent segments detected by large-scale pedigree analysis. *Mol*
360 *Biol Evol* **31**, 2212-22 (2014).
- 361 36. Browning, B.L. & Browning, S.R. Improving the accuracy and efficiency of
362 identity-by-descent detection in population data. *Genetics* **194**, 459-71
363 (2013).
- 364

365

Group	Sample Size	IBD Score	IBD Rank	F _{ST} Rank	Drift Rank	Census Size	Location
Gujjar	5	11.6	19	20	46	1,078,719	Jammu and Kashmir
Baniyas	7	9.6	24	24	18	4,200,000	Uttar Pradesh
Pattapu_Kapu	4	9.5	25	24	21	13,697,000	Andhra Pradesh
Vadde	3	9.2	26	34	26	3,695,000	Andhra Pradesh
Yadav	12	4.4	48	102	67	1,124,864	Puducherry
Kshatriya_Aqnikula	4	2.4	75	154	NA	12,809,000	Andhra Pradesh
Naga	4	2.3	76	NA	NA	1,834,483	Nagaland
Kumhar	27	2.3	77	222	197	3,144,000	Uttar Pradesh
Reddy	7	2.0	84	133	106	22,500,000	Telangana
Brahmin_Nepal	4	1.9	86	95	141	4,206,235	Nepal
Kallar	27	1.7	94	95	73	2,426,929	Tamil Nadu
Brahmin_Manipuri	17	1.6	99	NA	NA	1,544,296	Manipur
Arunthathiyar	18	1.3	108	133	81	1,192,578	Tamil Nadu
Vysya	39	1.2	110	55	35	3,200,000	Telangana

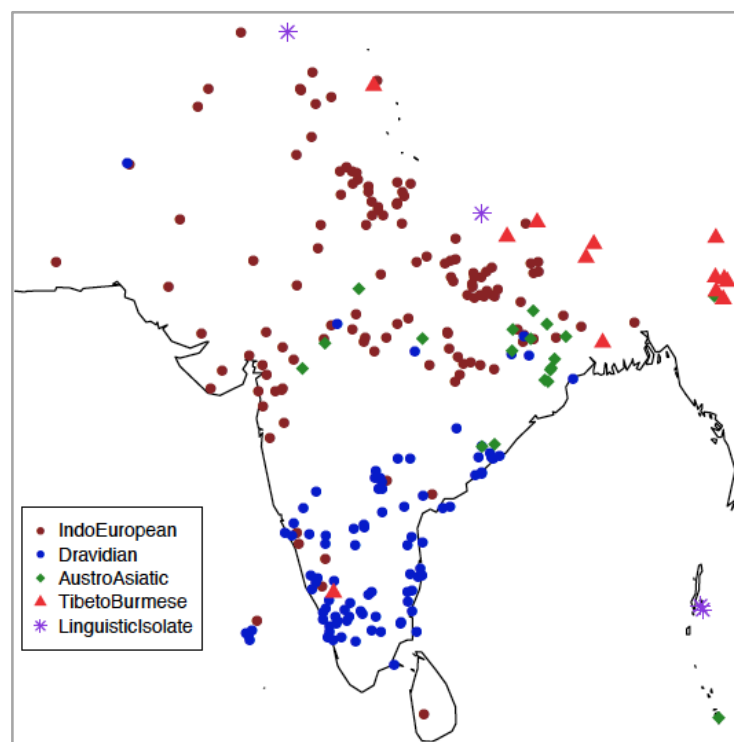
366

367 **Table 1. South Asian groups with census sizes over 1 million and IBD scores greater than those of Ashkenazi Jews**

368 **and Finns.** Fourteen South Asian groups with IBD scores significantly higher than that of Finns, census sizes over 1
369 million, and sample sizes of at least 3 that are of particularly high interest for founder event disease gene mapping studies.
370 For reference, Finns and Ashkenazi Jews (on Human Origins) would have IBD scores of 1.0 and 0.9, IBD ranks of 121 and
371 135, and F_{ST} ranks of 133 and 154, respectively (the group-specific drift is difficult to compare for groups with
372 significantly different histories, so they were not calculated for Finns or Ashkenazi Jews).

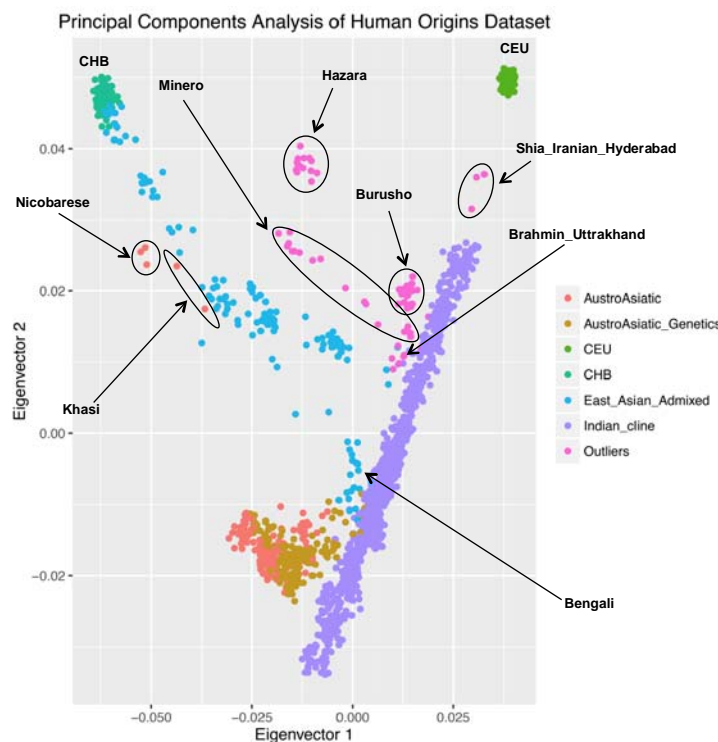
373

374
A



375
376

B

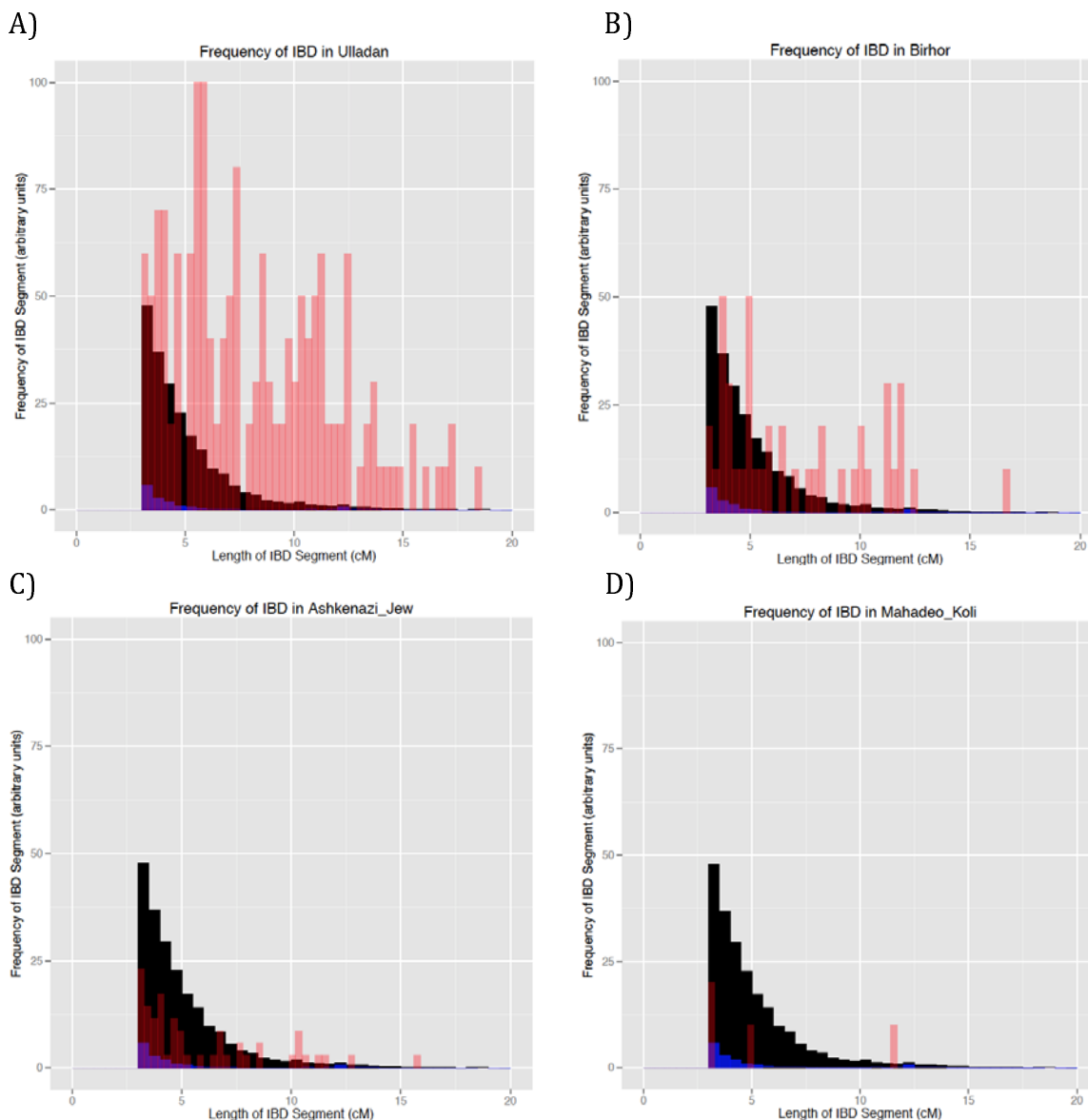


377

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

382

383



384

385

386

387

388

389

390

391

392

393

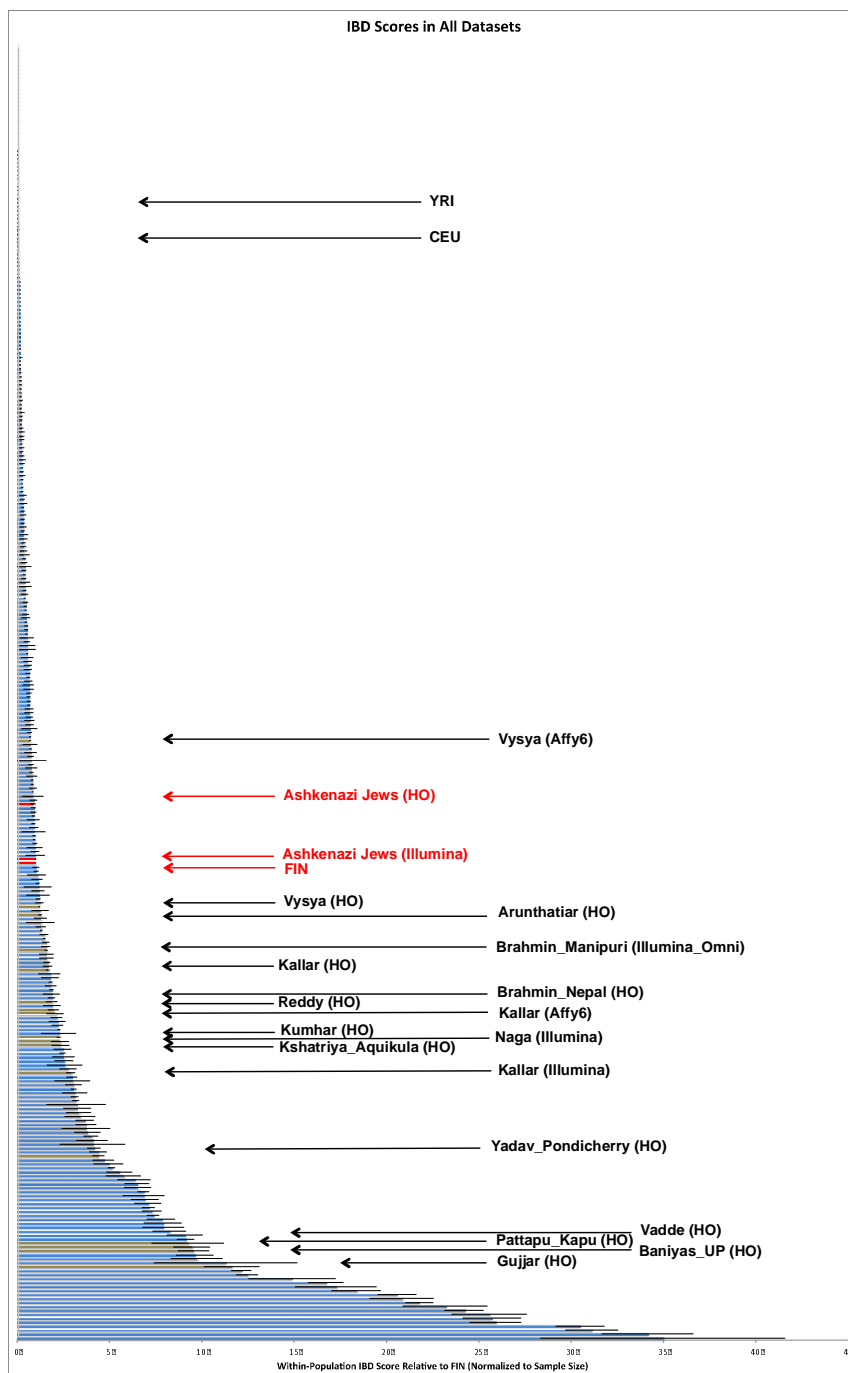
394

395

396

397

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 $\binom{2n}{2} - n$, where n is the number of individuals in the sample. All data for the figure are based on the Human Origins dataset.



398
399
400
401
402
403
404
405
406
407

Figure 3. IBD scores relative to Finns (FIN). 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 significantly stronger founder events than 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= Northern European.

408 **Online Methods:**

409 **Data Sets:**

410 We assembled a dataset of 1,955 individuals from 249 groups genotyped on the
411 Affymetrix Human Origins array, of which data from 1,663 individuals from 230
412 groups are newly reported here (Figure 1a). We merged these data with the dataset
413 published in Moorjani *et al.*⁸, which consisted of 332 individuals from 52 groups
414 genotyped on the Affymetrix 6.0 array. We also merged it with two additional
415 datasets published in Metspalu *et al.*⁹, consisting of 151 individuals from 21 groups
416 genotyped on Illumina 650K arrays as well as a dataset published in Basu *et al.*¹¹,
417 consisting of 367 individuals from 20 groups generated on Illumina Omni 1-Quad
418 arrays. These groups come from India, Pakistan, Nepal, Sri Lanka, and Bangladesh.
419 All samples were collected under the supervision of ethical review boards in India
420 with informed consent obtained from all subjects.
421
422

423
424 We analyzed two different Ashkenazi Jewish datasets, one consisting of 21
425 individuals genotyped on Illumina 610K and 660K bead arrays¹⁰ and one consisting
426 of 7 individuals genotyped on Affymetrix Human Origins arrays.
427

428 Our “Affymetrix 6.0” dataset consists of 332 individuals genotyped on 329,261 SNPs,
429 and our “Illumina_Omni” dataset consists of 367 individuals genotyped on 750,919
430 SNPs. We merged the South Asia and Jewish data generated by the other Illumina
431 arrays to create an “Illumina” dataset which consists of 172 individuals genotyped
432 on 500,640 SNPs. Finally, we merged the data from the Affymetrix Human Origins
433 arrays with the Ashkenazi Jewish data and data from the Simons Genome Diversity
434 Project^{25,26} to create a dataset with 4,402 individuals genotyped on 512,615 SNPs.
435 We analyzed the four datasets separately due to the small intersection of SNPs
436 between them and possible systematic differences across genotyping platforms. We
437 merged in the 1000 Genomes Phase 3 data²⁷ (2504 individuals from 26 different
438 groups; notably, including 99 Finnish individuals) into all of the datasets. We used
439 genome reference sequence coordinates (hg19) for analyses.
440

441 **Quality Control:**

442 We filtered the data at both the SNP and individual level. On the SNP level, we
443 required at least 95% genotyping completeness for each SNP (across all
444 individuals). On the individual level, we required at least 95% genotyping
445 completeness for each individual (across all SNPs).
446
447

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

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

461
462 We used EIGENSOFT 5.0.1 smartpca²⁸ on each group to detect PCA outliers. We also
463 developed a procedure to distinguish recent relatedness from founder events so
464 that we could remove recently related individuals. We first identified all duplicates
465 or obvious close relatives by using Plink “genome” and GERMLINE²⁹ to compute IBD
466 (described in more detail below) and removed one individual from all pairs with a
467 PI_HAT score greater than 0.45 and the presence of at least 1 IBD fragment greater
468 than 30cM. We then used an iterative procedure to identify additional recently
469 related individuals. For sample sizes above 5, we identified any pairs within each
470 group that had both total IBD and total long IBD (>20cM) that were greater than 2.5
471 SDs and 1 SD, respectively, from the group mean. For sample sizes 5 or below, we
472 used modified Z scores of $0.6745 * (\text{IBD_score} - \text{median}(\text{score})) / \text{MAD}$, where MAD is
473 the median absolute deviation, and identified all pairs with modified Z scores
474 greater than 3.5 for both total IBD and total long IBD as suggested by Iglewicz and
475 Hoaglin³⁰. After each round, we repeated the process if the new IBD score was at
476 least 30% lower than the prior IBD score. Simulations showed that we were always
477 able to remove a first or second cousin in the dataset using this method
478 (Supplementary Table 5). Together these analyses removed 53 individuals from the
479 Affymetrix 6.0 dataset, 21 individuals from the Illumina dataset, 43 individuals from
480 the Illumina Omni dataset, and 225 individuals from the Human Origins dataset.

481
482 After data quality control and merging with the 1000 Genomes Project data, the
483 Affymetrix 6.0 dataset included 2,842 individuals genotyped on 326,181 SNPs, the
484 Illumina dataset included 2,662 individuals genotyped on 484,293 SNPs, the
485 Illumina Omni dataset included 2,828 individuals genotyped on 750,919 SNPs, and
486 the Human Origins dataset included 4,177 individuals genotyped at 499,158 SNPs.

487 488 **Simulations to Test Relatedness Filtering and IBD Analyses**

489
490 We used ARGON³¹ to simulate groups with different bottleneck strengths to test the
491 IBD analyses, relatedness filtering, and founder event dating algorithms. We used
492 ARGON’s default settings, including mutation rate of $1.65 * 10^{-8}$ per base pair (bp) per
493 generation and a recombination rate of $1 * 10^{-8}$ per bp per generation and simulated
494 22 chromosomes of size 130 Mb each. We pruned the output by randomly removing
495 SNPs until there were 22,730 SNPs per chromosome to simulate the approximate
496 number of positions in the Affymetrix Human Origins array. For the IBD analyses,
497 we simulated groups to have descended from an ancestral group 1,800 years ago
498 with $N_e=50,000$ and to have formed two groups with $N_e=25,000$. These groups
499 continued separately until 100 generations ago when they combined in equal

500 proportions to form a group with $N_e=50,000$. The group then split into 3 separate
501 groups 72 generations ago that have bottlenecks leading to N_e of either 400, 800, or
502 1600. The 3 groups then exponentially expanded to a present size of $N_e=50,000$.
503 We designed these simulations to capture important features of demographic
504 history typical of Indian groups as detailed by Moorjani *et al.*⁸ and Reich *et al.*⁴ We
505 chose the bottleneck sizes because they represent founder events with
506 approximately the strength of Finns (the bottleneck to 800), and twice as strong
507 (400) and half as strong (1600) as that group. We then performed the IBD analyses
508 described below with 99 individuals from the group with bottleneck strength
509 similar to that of Finns (198 haploid individuals were simulated and merged to
510 produce 99 diploid individuals) and different number of individuals from the other
511 groups. These analyses demonstrate that with only 4-5 individuals we can
512 accurately assess the strength of founder events in groups with strong founder
513 events (Supplementary Figure 1 and Supplementary Table 1). Weaker founder
514 events are more difficult to assess, but these groups are of less interest for founder
515 event disease mapping, so we aimed to sample ~5 individuals per group.

516
517 We wrote custom R scripts to carry out forward simulations for creating first and
518 second cousins. We took individuals from the bottleneck of size 800 and performed
519 “matings” by taking 2 individuals and recombining their haploid chromosomes
520 assuming a rate of 1×10^{-8} per bp per generation across the chromosome and
521 combining one chromosome from each of these individuals to form a new diploid
522 offspring. The matings were performed to achieve first and second cousins. We then
523 placed these back into the group with group of size 800, and ran the relatedness
524 filtering algorithms to evaluate whether they would identify these individuals.

525 526 **Distance-Based Phylogenetic Tree:**

528 We calculated genetic differentiation (F_{ST}) between all pairs of groups using
529 EIGENSOFT *smartpca* and created a neighbor-joining tree using PHYLIP³² with
530 Yoruba as the outgroup. We used ItoI³³ to display the tree.

531 532 **Phasing, IBD Detection, and IBD Score Algorithm:**

533
534 We phased all datasets using Beagle 3.3.2 with the settings *missing=0; lowmem=true;*
535 *gprobs=false; verbose=true*³⁴. We left all other settings at default. We determined IBD
536 segments using GERMLINE²⁹ with the parameters *-bits 75 -err_hom 0 -err_het 0 -*
537 *min_m 3*. We used the genotype extension mode to minimize the effect of any
538 possible phasing heterogeneity amongst the different groups and used the
539 HaploScore algorithm to remove false positive IBD fragments with the
540 recommended genotype error and switch error parameters of 0.0075 and 0.003³⁵.
541 We chose a HaploScore threshold matrix based on calculations from Durand *et al.*
542 for a “mean overlap” of 0.8, which corresponds to a precision of approximately 0.9
543 for all genetic lengths from 2-10cM. In addition to the procedure we developed to
544 remove close relatives (Quality Control section), we also removed segments longer
545 than 20cM as simulations showed that this increased sensitivity of the analyses
546 (Supplementary Table 4). We computed “IBD score” as the total length of IBD

547 segments between 3-20cM divided by $\left\{ \binom{2n}{2} - n \right\}$ where n is the number of
548 individuals in each group to normalize for sample size. We then expressed each
549 group's score as a ratio of their IBD score to that of Finns and calculated standard
550 errors for this score using a weighted Block Jackknife over each chromosome with
551 95% confidence intervals defined as IBD score $\pm 1.96*s.e.$

552
553 We also repeated these analyses with FastIBD³⁶ for the Affymetrix 6.0 and Illumina
554 datasets and observed that the results were highly correlated ($r>0.96$) (data not
555 shown). We chose GERMLINE for our main analyses, however, because the FastIBD
556 algorithm required us to split the datasets into different groups, since it adapts to
557 the relationships between LD and genetic distance in the data, and these
558 relationships differ across groups. We used data from several different Jewish
559 groups and all twenty-six 1000 Genomes groups to improve phasing, but of these
560 groups we only included results for Ashkenazi Jews and two outbred groups (CEU
561 and YRI) in the final IBD score ranking.

562

563 Disease patient analyses:

564
565 We use Affymetrix Human Origins arrays to genotype 15 patients with progressive
566 pseudorheumatoid dysplasia (PPD) and 6 patients with mucopolysaccharidosis
567 (MPS) type IVA, all of which had disease mutations previously determined¹⁶⁻¹⁸, and
568 3 of which (MPS patients) are newly reported here. After quality control, 6 of the
569 PPD patients with Cys78Tyr mutations, 6 of the PPD patients with Cys337Tyr
570 mutations, and 6 of the MPS patients with Cys78Arg mutations remained. We
571 measured IBD as described above and also detected homozygous segments within
572 each individual by using GERMLINE with the parameters *-bits 75 -err_hom 2 -err_het*
573 *0 -min_m 0.5 -homoz-only*.

574

575 Haplotype sharing was assessed by analyzing phased genotypes for each mutation
576 group. At each SNP, we counted the number of identical genotypes for each allele
577 and calculated the fraction by dividing by the total number of possible haplotypes (2
578 times the number of individuals), then took the larger value of the two possible
579 alleles (thus the fraction range was 0.5-1). We then averaged these values over
580 blocks of 10 or 25 SNPs and plotted the averages around the relevant mutation site.

581

582 Between-Group IBD Calculations:

583
584 We determined IBD using GERMLINE as above. We collapsed individuals into
585 respective groups and normalized for between-group IBD by dividing all IBD from
586 each group by $\left\{ \binom{2n}{2} \right\}$ where n is the number of individuals in each group. We
587 normalized for within-group IBD as described above. We defined groups with high
588 shared IBD as those with an IBD score greater than three times the founder event
589 strength of CEU (and $\sim 1/3$ the event strength of Ashkenazi Jews).

590

591 f_3 -statistics:

592

593 We used the f_3 -statistic⁷ $f_3(\text{Test}; \text{Ref}_1, \text{Ref}_2)$ to determine if there was evidence that
594 the *Test* group was derived from admixture of groups related to *Ref*₁ and *Ref*₂. A
595 significantly negative statistic provides unambiguous evidence of mixture in the
596 Test group. We determined the significance of the f_3 -statistic using a Block Jackknife
597 and a block size of 5 cM. We considered statistics over 3 standard errors below zero
598 to be significant.

600 **Calculating Group Specific Drift:**

602 We used ADMIXTUREGRAPH⁷ to model each Indian group on the cline as a mixture
603 of ANI and ASI ancestry, using the model (YRI, (Indian group, (Georgians, ANI)),
604 [(ASI, Onge)]) proposed by Moorjani *et al.*⁸ This approach provides estimates for
605 post-admixture drift in each group (Supplementary Figure 6), which is reflective of
606 the strength of the founder event (high drift values imply stronger founder events).
607 We only included groups on the Indian cline in this analysis, and we removed all
608 groups with evidence of East Asian related admixture because this admixture is not
609 accommodated within the above model.

611 **PCA-Normalized F_{ST} Calculations:**

613 To account for intermarriage across groups, we used clusters based on PCA to
614 estimate the minimum F_{ST} for each South Asian group (Supplementary Figure 6).
615 Specifically, we computed the F_{ST} between each group and the rest of the individuals
616 in their respective cluster based on EIGENSOFT *smartpca*. For these analyses we
617 only included groups on the Indian Cline and those with Austroasiatic-related
618 genetic patterns (groups clustering near Austroasiatic speakers on the PCA). For
619 Ashkenazi Jews and Finns, we used the minimum F_{ST} to other European groups.

621 **Code Availability:**

623 Code for all calculations available upon request.