The promise of disease gene discovery in South Asia

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

[1]Department of Genetics, Harvard Medical School, New Research Building, 77 Ave. Louis Pasteur, Boston, MA 02115, USA
[2]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA
[3]Department of Biological Sciences, Columbia University, 600 Fairchild Center, New York, NY 10027, USA
[4]CSIR-Centre for Cellular and Molecular Biology, Habsiguda, Hyderabad, Telangana 500007, India
[5]Superintending Anthropologist (Physical) (Rtd.), Anthropological Survey of India, 27 Jawaharlal Nehru Road, Kolkata 700016, India
[6]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02141, USA
[7]Department of Medical Genetics, Kasturba Medical College, Manipal University, Manipal, India
[8]Department of Applied Zoology, Mangalore University, Mangalagangothri 574199, Mangalore, Karnataka, India
[9]Centre for Human Genetics, Biotech Park, Electronics City (Phase I), Bangalore 560100, India
[10]Amity Institute of Biotechnology, Amity University, Sector125, Noida 201303, India
[11]School of Life Sciences, Manipal University, Manipal 576104, India
[12]Fetal Care Research Foundation, 197 Dr. Natesan Road, Chennai 600004, India
[13]Present address: Genome Foundation, Hyderabad 500076, India
[14]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA
*co-senior authors

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

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

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

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

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

We expressed IBD scores for each group as a fraction of the IBD scores of the 1000 Genomes Project Finns merged into each respective dataset. Due to the fact that all the SNP arrays we analyzed included more SNPs ascertained in Europeans than in South Asians, the sensitivity of our methods to founder events is greater in Europeans than in South Asians, our estimates of founder event strengths in South Asian groups are conservative underestimates relative to that in Europeans (Supplementary Figure 1 demonstrates this effect empirically and shows that it is less of a bias for the strong founder events that are the focus of this study). We computed standard errors for these ratios by a weighted Block Jackknife across chromosomes and declared significance where the 95% confidence intervals did not overlap 1. We carried out computer simulations to validate our procedure. The simulations suggest that we are not substantially overestimating the magnitudes of modest founder events, since for a simulated founder event that is half the magnitude of that in Finns, we never infer the score to be significantly greater than in Finns. The simulations also suggest that our procedure is highly sensitive to detecting strong founder events, since for sample sizes of at least 5, the algorithm's sensitivity is greater than 95% for determining that a group with two times the bottleneck strength as Finns has an IBD score significantly greater than that of Finns (Supplementary Figure 2 and Supplementary Table 2).

We infer that 81 out of 263 unique groups (96 out of 327 groups if not considering the overlap of groups among datasets) have an IBD score greater than those of both Finns and Ashkenazi Jews (Figure 3). These results did not change when we added back the outlier samples that we removed in quality control. A total of 14 of these groups have estimated census sizes of over a million (Figure 3; Table 1). However, the groups with smaller census sizes are also very important – outside of South Asia, small census size groups with extremely strong founder events such as Amish, Hutterites, and people of the Saguenay Lac-St. Jean region have led to the discovery of dozens of novel disease variants. We also searched for IBD across groups – screening for cases in which the across-group IBD score is at least a third of the within-group IBD score of Ashkenazi Jews – and found 77 cases of clear IBD-sharing, which typically follow geography, religious affiliation, or linguistic grouping (particularly Austroasiatic speakers) (Supplementary Table 3). Pairs of groups with

135 high shared IBD and descent from a common founder event will share risk for the
136 same recessive disease. However, these cross-group IBD sharing patterns are not
137 driving our observations, as we still identify 68 unique sets of groups without high
138 IBD to other groups that have significantly higher estimated IBD scores than both
139 Finns and Ashkenazi Jews.
140
141 Our documentation that very strong founder events affect a large fraction of South
142 Asian groups presents an opportunity for decreasing disease burden in South Asia.
143 This source of risk for recessive diseases is very different from that due to marriages
144 among close relatives, which is also a major cause of recessive disease in South Asia.
145 To determine the relative impact of these factors, we computed F_{ST}, a measurement
146 of allele frequency differentiation, between each group in the dataset and a pool of
147 other South Asian groups chosen to be closest in terms of ancestry proportions. We
148 find that inbreeding is not driving many of these signals, as 89 unique groups have
149 higher F_{ST} scores than those of Ashkenazi Jews and Finns even after reducing the F_{ST}
150 score by the proportion of allele frequency differentiation due to inbreeding. These
151 results show that while most recessive disease gene mapping studies in South Asia
152 have focused on families that are the products of marriages between close relatives,
153 recessive diseases are also likely to occur at an elevated rate even in non-
154 consanguineous cases because of shared ancestors more distantly in time.
155
156 As an example of the promise of founder event disease gene mapping in South Asia,
157 we highlight the case of the Vysya, who have a census size of more than 3 million
158 and who we estimate have an IBD score about 1.2-fold higher than Finns (Figure 3).
159 The Vysya have an approximately 100-fold higher rate of butyrylcholinesterase
160 deficiency than other groups, and Vysya ancestry is a known counter-indication for
161 the use of muscle relaxants such as succinylcholine or mivacurium that are given
162 prior to surgery[13]. This disease is likely to occur at a higher rate due to the founder
163 event in Vysya's history, and we expect that, like Finns, Vysya likely have a higher
164 rate of many other diseases compared to other groups. Other examples of recessive
165 disease genes with a likely origin in founder events are known anecdotally in South
166 Asia, highlighting the importance of systematic studies to find them[14].
167
168 To demonstrate how a new recessive disease in a founder event group could be
169 mapped, we carried out whole genome SNP genotyping in 12 patients from southern
170 India who had progressive pseudorheumatoid dysplasia (PPD), a disease known to
171 be caused by mutations in the gene *WISP3*[15,16]. Of the 6 individuals with the
172 Cys78Tyr mutation in *WISP3,*[15,16] 5 were from non-consanguineous marriages, and
173 we found a much higher fraction of IBD at the disease mutation site than in the rest
174 of the genome in these individuals (Supplementary Figure 3a; Supplementary Figure
175 4a), consistent with the Cys78Tyr mutation that causes PPD in these patients owes
176 its origin to a founder event. This pattern contrasts with the 6 other patients with
177 different disease variants and 6 patients with a mutation causing a different disease
178 (mucopolysaccharidosis (MPS) type IVA) who were from primarily consanguineous
179 marriages, and who lacked significant IBD across their disease mutation sites,
180 implying that in the case of these groups the driver for the recessive diseases was

181 marriage between close relatives (Supplementary Text). This example highlights
182 how not only marriages of close relatives, but also founder events, are substantial
183 causes of rare recessive disease in South Asia.
184
185 The evidence of widespread strong founder events presents a major opportunity for
186 disease gene discovery and public health intervention in South Asia that is not
187 widely appreciated. The current paradigm for recessive disease gene mapping in
188 South Asia is to collect cases in tertiary medical centers and map diseases in
189 individuals with the same phenotype, often blinded to information about group
190 affiliation as in the case of the PPD study where we do not have access to the ethnic
191 group information. However, our results suggest that collecting information on
192 group affiliation could be highly valuable for improving the success of these studies.
193 A fruitful way to approach gene mapping would be to proactively survey
194 communities known to have strong founder events, searching for diseases that
195 occur at high rates in these communities. This approach was pioneered in the 1950s
196 in studies of the Old Order Amish in the U.S., a founder population of approximately
197 100,000 individuals in whom many dozens of recessive diseases were mapped, a
198 research program that was crucial to founding modern medical genetics and that
199 was of extraordinary health benefit. Our study suggests that the potential for
200 disease gene mapping in South Asia would be orders of magnitude greater.
201
202 Mapping of recessive diseases may be particularly important in communities
203 practicing arranged marriages, which are common in South Asia. An example of the
204 power of this approach is given by *Dor Yeshorim*, a community genetic testing
205 program among religious Ashkenazi Jews[17], which visits schools, screens students
206 for common recessive disease causing mutations previously identified to be
207 segregating at a higher frequency in the target group, and enters the results into a
208 confidential database. Matchmakers query the database prior to making suggestions
209 to the families and receive feedback about whether the potential couple is
210 "incompatible" in the sense of both being carriers for a recessive mutation at the
211 same gene. Given that approximately 95% of community members whose marriages
212 are arranged participate in this program, recessive diseases like Tay-Sachs have
213 virtually disappeared in these communities. A similar approach should work as well
214 in South Asians communities. Given the potential for saving lives, this or similar
215 kinds of research could be a valuable investment for future generations[18].
216

Supplementary Data:

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

Acknowledgements:

We are thankful to the many Indian, Pakistani, Bangladeshi, Sri Lankan, and Nepalese individuals who contributed the DNA samples analyzed here including the PPD and MPS patients. We are grateful to Analabha Basu and Partha P. Majumder for early sharing of data. Funding was provided by an NIGMS (GM007753) fellowship to NN, a Translational Seed Fund grant from the Dean's Office of Harvard Medical School to DR, Council of Scientific and Industrial Research, Government of India grant to KT, and an NIGMS grant 115006 to OPM. SS and SMJ acknowledge the funding from the Department of Biotechnology (BT/PR4224/MED/97/60/2011) and Department of Science and Technology (SR/WOS-A/LS-83/2011), Government of India. Funding for the mutation analysis of Indian patients with PPD was provided by Indian Council of Medical Research (BMS 54/2/2013) to KMG. DR is an Investigator of the Howard Hughes Medical Institute. Genotyping data for the samples collected for this study is available on request.

References:

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

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

Group	Sample Size	IBD Score	IBD Rank	F_{ST} 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

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

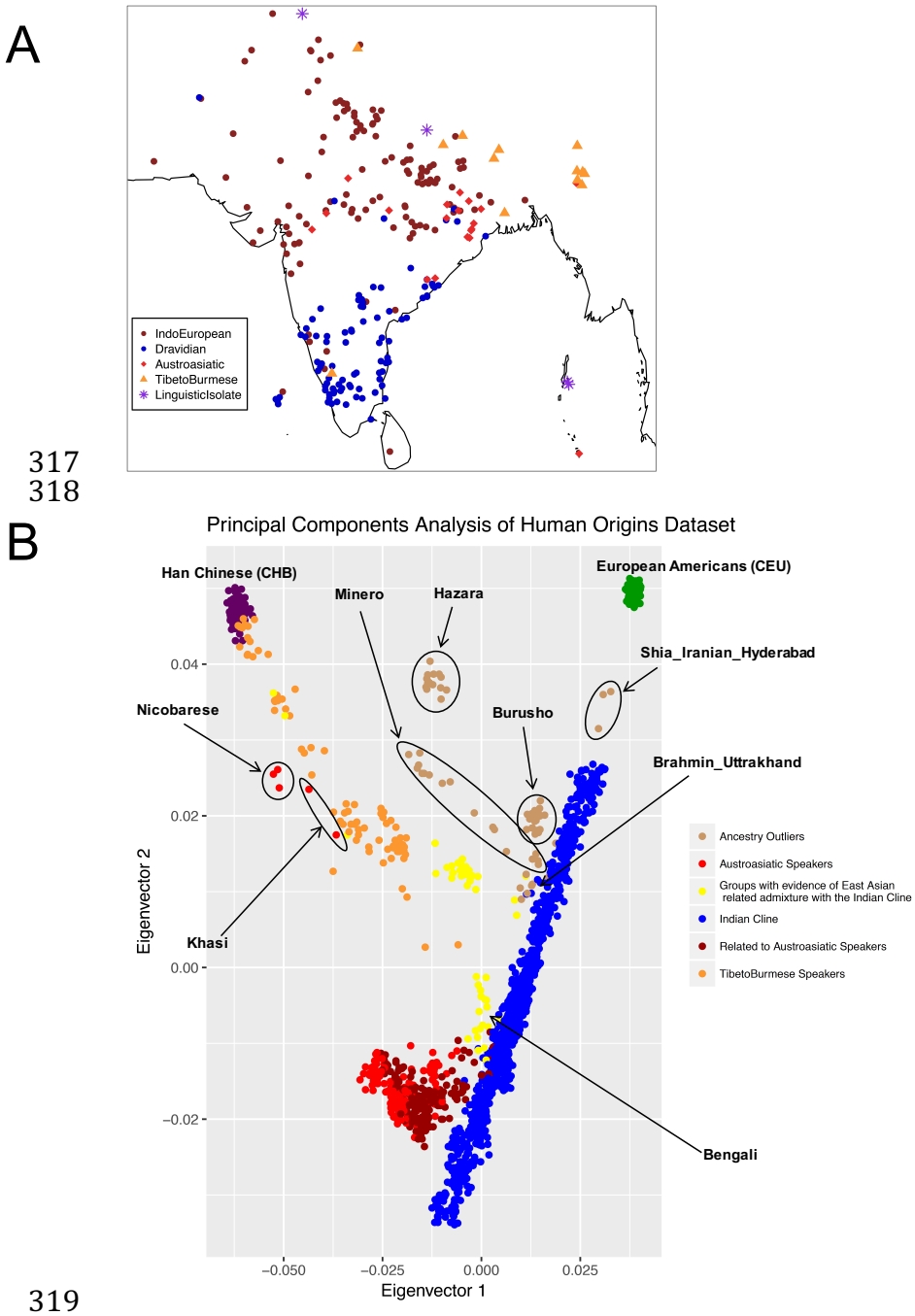

Figure 1. Dataset overview. (a) Sampling locations for all analyzed groups. Each point indicates a distinct group (random jitter was added to help in visualization at locations where there are many groups). (b) PCA of Human Origins dataset along with European Americans (CEU) and Han Chinese (CHB). There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

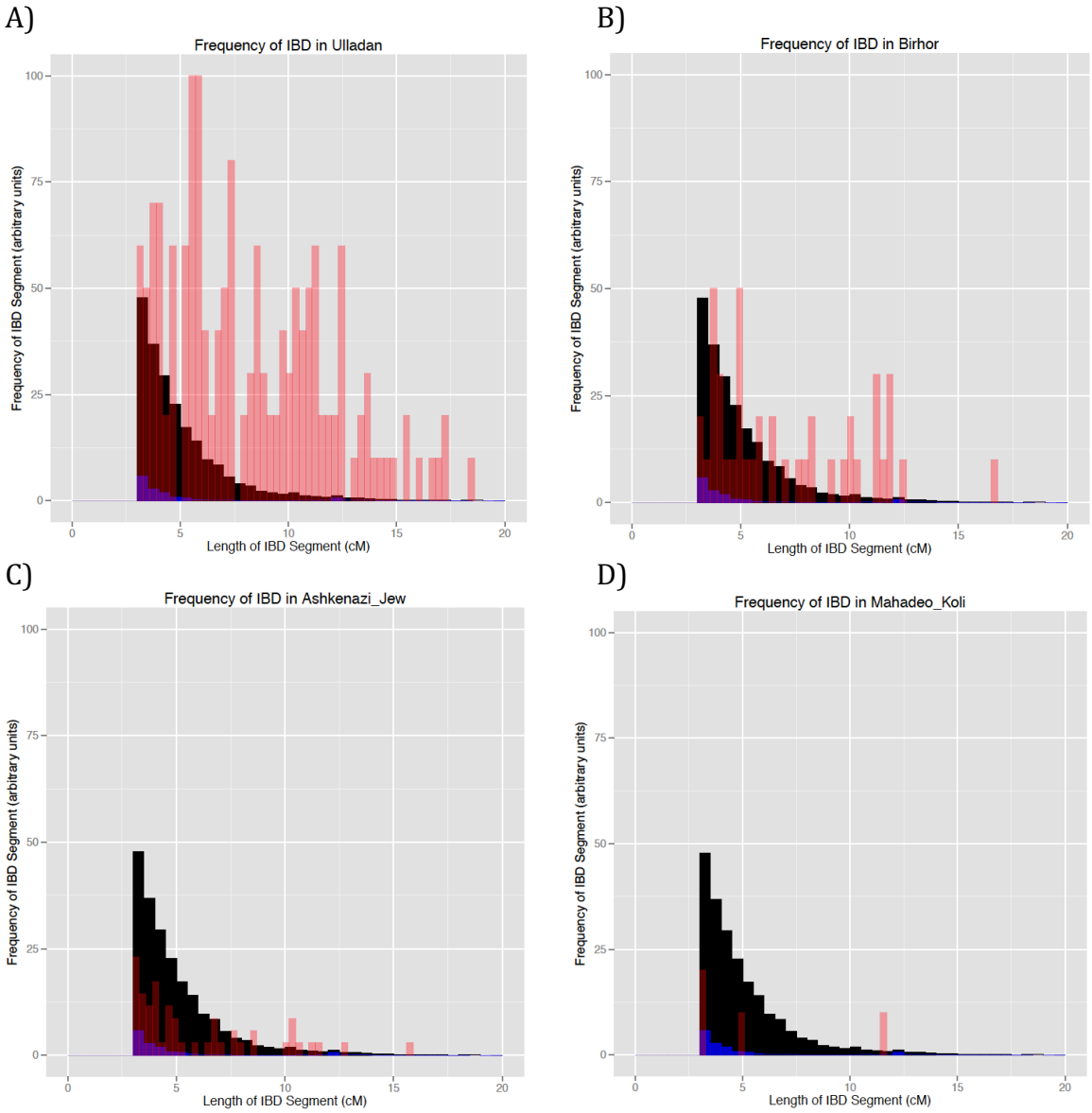

Figure 2. Example histograms of IBD segments to illustrate the differences between groups with founder events of different magnitudes: These histograms provide visual illustrations of differences between groups with different IBD scores. As a ratio relative to Finns (FIN; black), these groups (red) have 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 $\left\{\binom{2n}{2} - n\right\}$, where n is the number of individuals in the sample. All data for the figure are based on the Human Origins dataset.

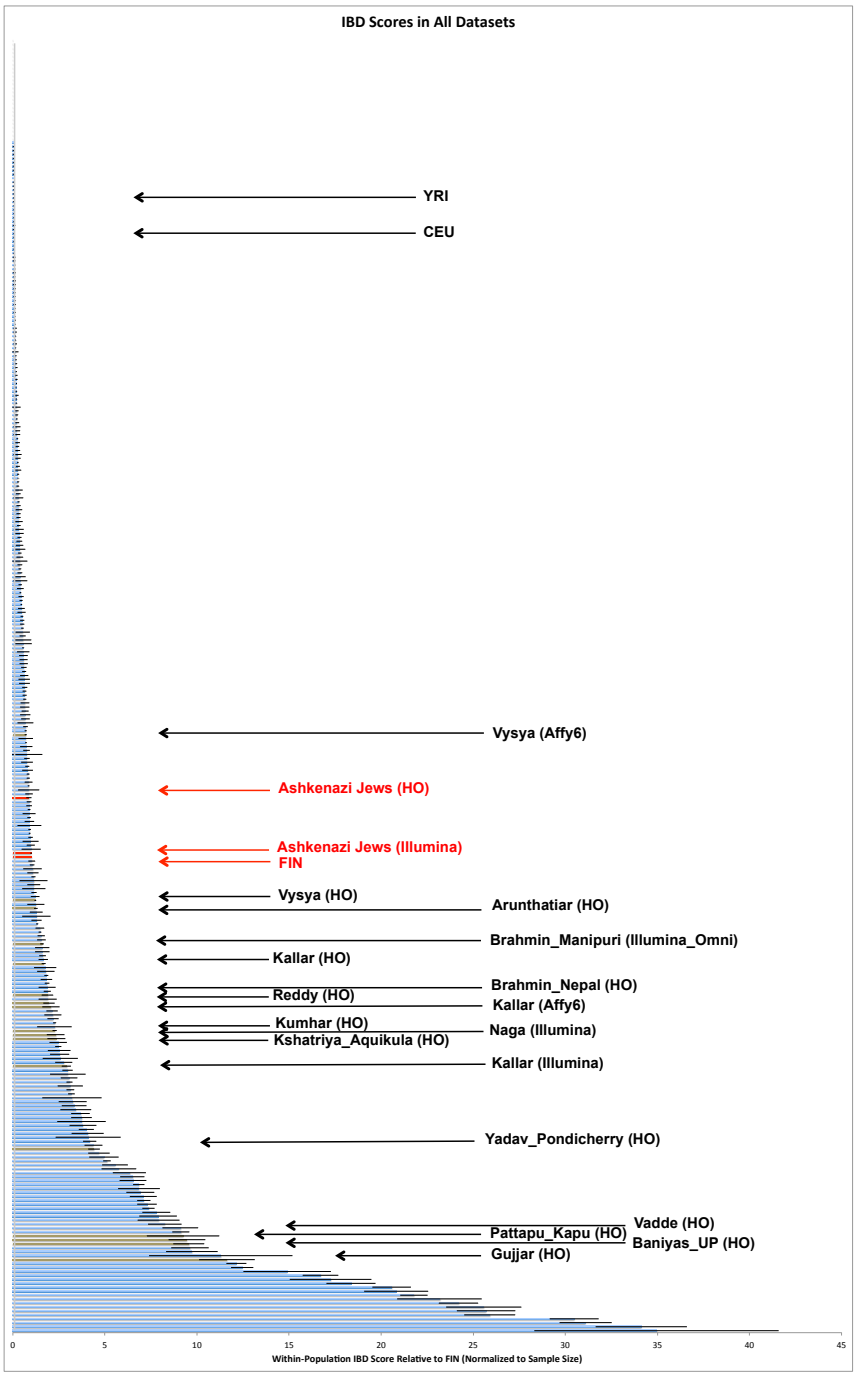

Figure 3. IBD scores relative to Finns (FIN). Histogram ordered by IBD score, roughly proportional to the per-individual risk for recessive disease due to the founder event. (These results are also given quantitatively for each group in Online Table 1.) We restrict to groups with at least two samples, 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.

Online Methods:

Data Sets:

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

We analyzed two different Ashkenazi Jewish datasets, one consisting of 21 individuals genotyped on Illumina 610K and 660K bead arrays[10] and one consisting of 7 individuals genotyped on Affymetrix Human Origins arrays.

Our "Affymetrix 6.0" dataset consists of 332 individuals genotyped on 329,261 SNPs, and our "Illumina_Omni" dataset consists of 367 individuals genotyped on 750,919 SNPs. We merged the South Asian and Ashkenazi Jewish data generated by the other Illumina arrays to create an "Illumina" dataset consisting of 172 individuals genotyped on 500,640 SNPs. We merged the data from the Affymetrix Human Origins arrays with the Ashkenazi Jewish data and data from the Simons Genome Diversity Project[19,20] to create a dataset with 4,402 individuals genotyped on 512,615 SNPs. We analyzed the four datasets separately due to the small intersection of SNPs between them. We merged in the 1000 Genomes Phase 3 data[21] (2,504 individuals from 26 different groups; notably, including 99 Finnish individuals) into all of the datasets. We used genome reference sequence coordinates (hg19) for analyses.

Quality Control:

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

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

400 from a weighted Block Jackknife across chromosomes. This quantity should be
401 consistent with zero within a few standard errors if there are no batch effects that
402 cause systematic differences across the plates, as allele frequency differences
403 between two samples of the same group should be random fluctuations that have
404 nothing to do with the array plates on which they are genotyped. This analysis
405 found strong batch effects associated with one array plate, and we removed these
406 samples from further analysis.
407
408 We used EIGENSOFT 5.0.1 smartpca[22] on each group to detect PCA outliers and
409 removed 51 samples. We also developed a procedure to distinguish recent
410 relatedness from founder events so that we could remove recently related
411 individuals. We first identified all duplicates or obvious close relatives by using
412 Plink[23] "genome" and GERMLINE[24] to compute IBD (described in more detail below)
413 and removed one individual from all pairs with a PI_HAT score greater than 0.45
414 and the presence of at least 1 IBD fragment greater than 30cM. We then used an
415 iterative procedure to identify additional recently related individuals. For sample
416 sizes above 5, we identified any pairs within each group that had both total IBD and
417 total long IBD (>20cM) that were greater than 2.5 SDs and 1 SD, respectively, from
418 the group mean. For sample sizes 5 or below, we used modified Z scores of
419 0.6745*(IBD_score - median(score))/MAD, where MAD is the median absolute
420 deviation, and identified all pairs with modified Z scores greater than 3.5 for both
421 total IBD and total long IBD as suggested by Iglewicz and Hoaglin[25]. After each
422 round, we repeated the process if the new IBD score was at least 30% lower than
423 the prior IBD score. Simulations showed that we were always able to remove a first
424 or second cousin in the dataset using this method (Supplementary Table 1).
425 Together these analyses removed 53 individuals from the Affymetrix 6.0 dataset, 21
426 individuals from the Illumina dataset, 43 individuals from the Illumina Omni
427 dataset, and 225 individuals from the Human Origins dataset.
428
429 After data quality control and merging with the 1000 Genomes Project data, the
430 Affymetrix 6.0 dataset included 2,842 individuals genotyped on 326,181 SNPs, the
431 Illumina dataset included 2,662 individuals genotyped on 484,293 SNPs, the
432 Illumina Omni dataset included 2,828 individuals genotyped on 750,919 SNPs, and
433 the Human Origins dataset included 4,177 individuals genotyped at 499,158 SNPs.
434
435 **Simulations to Test Relatedness Filtering and IBD Analyses**
436
437 We used ARGON[26] to simulate groups with different bottleneck strengths to test the
438 IBD analyses and relatedness filtering. We used ARGON's default settings, including
439 a mutation rate of $1.65*10^{-8}$ per base pair (bp) per generation and a recombination
440 rate of $1*10^{-8}$ per bp per generation and simulated 22 chromosomes of size 130 Mb
441 each. We pruned the output by randomly removing SNPs until there were 22,730
442 SNPs per chromosome to simulate the approximate number of positions in the
443 Affymetrix Human Origins array. For the IBD analyses, we simulated groups to have
444 descended from an ancestral group 1,800 years ago with N_e=50,000 and to have
445 formed two groups with N_e=25,000. These groups continued separately until 100

446 generations ago when they combined in equal proportions to form a group with
447 N_e=50,000. The group then split into 3 separate groups 72 generations ago that have
448 bottlenecks leading to N_e of either 400, 800, or 1600. The 3 groups then
449 exponentially expanded to a present size of N_e=50,000. We designed these
450 simulations to capture important features of demographic history typical of Indian
451 groups[4,8]. We chose the bottleneck sizes because they represent founder events with
452 approximately the strength of Finns (the bottleneck to 800), and twice as strong
453 (400) and half as strong (1600) as that group. We then performed the IBD analyses
454 described below with 99 individuals from the group with bottleneck strength
455 similar to that of Finns (198 haploid individuals were simulated and merged to
456 produce 99 diploid individuals) and different numbers of individuals from the other
457 groups. These analyses demonstrate that with only 4-5 individuals we can
458 accurately assess the strength of founder events in groups with strong founder
459 events (Supplementary Figure 2 and Supplementary Table 2). Weaker founder
460 events are more difficult to assess, but these groups are of less interest for founder
461 event disease mapping, so we aimed to sample ~5 individuals per group.
462
463 We wrote custom R scripts to simulate first and second cousin pairs. We took
464 individuals from the bottleneck of size 800 and performed "matings" by taking 2
465 individuals and recombining their haploid chromosomes assuming a rate of $1*10^{-8}$
466 per bp per generation across the chromosome and combining one chromosome
467 from each of these individuals to form a new diploid offspring. The matings were
468 performed to achieve first and second cousins. We then placed these back into the
469 group with group of size 800, and ran the relatedness filtering algorithms to
470 evaluate whether they would identify these individuals.
471
472 **Phasing, IBD Detection, and IBD Score Algorithm:**
473
474 We phased all datasets using Beagle 3.3.2 with the settings *missing=0; lowmem=true;*
475 *gprobs=false; verbose=true*[27]. We left all other settings at default. We determined IBD
476 segments using GERMLINE[24] with the parameters *-bits 75 -err_hom 0 -err_het 0 -*
477 *min_m 3*. We used the genotype extension mode to minimize the effect of any
478 possible phasing heterogeneity amongst the different groups and used the
479 HaploScore algorithm to remove false positive IBD fragments with the
480 recommended genotype error and switch error parameters of 0.0075 and 0.003[28].
481 We chose a HaploScore threshold matrix based on calculations from[28] for a "mean
482 overlap" of 0.8, which corresponds to a precision of approximately 0.9 for all genetic
483 lengths from 2-10cM. In addition to the procedure we developed to remove close
484 relatives (Quality Control section), we also removed segments longer than 20cM as
485 simulations showed that this increased sensitivity of the analyses (Supplementary
486 Table 2). We computed "IBD score" as the total length of IBD segments between 3-
487 20cM divided by $\left\{\binom{2n}{2} - n\right\}$ where n is the number of individuals in each group to
488 normalize for sample size. We then expressed each group's score as a ratio of their
489 IBD score to that of Finns and calculated standard errors for this score using a
490 weighted Block Jackknife over each chromosome with 95% confidence intervals
491 defined as IBD score ±1.96*s.e.

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

Disease patient analyses:

We use Affymetrix Human Origins arrays to successfully genotype 12 patients with progressive pseudorheumatoid dysplasia (PPD) and 6 patients with mucopolysaccharidosis (MPS) type IVA, all of whom had disease mutations previously determined[15,16,30] (3 of the surveyed MPS patients are newly reported here). A total of 6 of the PPD patients had Cys78Tyr mutations, 6 had Cys337Tyr mutations (all 6 of the MPS patients had Cys78Arg mutations). We measured IBD as described above and also detected homozygous segments within each individual by using GERMLINE with the parameters *-bits 75 -err_hom 2 -err_het 0 -min_m 0.5 -homoz-only*.

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

Between-Group IBD Calculations:

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

f_3-statistics:

We used the f_3-statistic[7] $f_3(Test; Ref_1, Ref_2)$ to determine if there was evidence that the *Test* group was derived from admixture of groups related to Ref_1 and Ref_2. A significantly negative statistic provides unambiguous evidence of mixture in the Test group. We determined the significance of the f_3-statistic using a Block Jackknife and a block size of 5 cM. We considered statistics over 3 standard errors below zero to be significant.

Computing Group Specific Drift:

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

PCA-Normalized F_{ST} Calculations:

To account for intermarriage across groups, we used clusters based on PCA (Supplementary Text) to estimate the minimum F_{ST} for each South Asian group (Supplementary Figure 6). For the Affymetrix 6.0, Illumina, and Illumina_Omni datasets, we split the Indian cline into two different clusters and combined the Austroasiatic speakers and those with ancestry related to Austroasiatic speakers (according to the PCA of Figure 1b) into one cluster for a total of three clusters (all other groups were ignored for this analysis). For the Human Origins dataset we split the Indian cline into three different clusters and combined the groups with ancestry related to the main cluster of Austroasiatic speakers into one cluster for a total of four clusters (Khasi and Nicobarese were ignored in this analysis, because they do not cluster with the other Austroasiatic speaking groups). We then computed the F_{ST} between each group and the rest of the individuals in their respective cluster based on EIGENSOFT *smartpca* with inbreed set to YES to correct for inbreeding. For Ashkenazi Jews and Finns, we used the minimum F_{ST} to other European groups.

Code Availability:

Code for all calculations available upon request.