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

South Asia is a region of extraordinary cultural, linguistic, and genetic diversity, with a conservative estimate of over 4,600 anthropologically well-defined groups, many of which are endogamous communities with significant barriers to gene flow due to sociological and cultural factors that restrict marriage between groups[1]. Of the small fraction of South Asian groups that have been characterized using genome-wide data, many exhibit large allele frequency differences from geographically close neighbors[2-4], indicating that they have experienced 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 major 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].

To characterize the medically relevant founder events in South Asia, 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 developed an algorithm to measure the strength of an endogamous group's founder event, aiming to genotype at least five samples per group as our power calculations indicated that this was sufficient to detect a founder event stronger than that in Finns with high confidence (Supplementary Figure 1 and Supplementary Table 1). We combined the new data we 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 performed quality control to remove SNPs and individuals with a high proportion of missing genotypes or that were outliers in Principal Component Analysis (PCA).

We performed PCA on each of the three different datasets along with European Americans (CEU), Han Chinese (CHB), and West Africans (YRI), and found that the Siddi are strong outliers as previously reported (Supplementary Figure 2)[4,13,14]. We next removed YRI, Siddi and indigenous Andamanese (another known outlier) from the datasets and repeated PCA (Figure 1b, Supplementary Figure 3). Similar to past studies, the PCA documents three broad genetic groupings[4,8,11]. First, almost all South Asian groups speaking Indo-European and Dravidian languages lie along the "Indian Cline," with different proportions of Ancestral North Indian (ANI) ancestry related to Europeans, Central Asians, and Near Easterners, and Ancestral South Indian (ASI) ancestry that is as different from ANI as Europeans and East Asians are from each other[4]. The second major cluster includes groups that speak Austroasiatic languages, as well as some non-Austroasiatic speaking groups that have similar ancestry possibly due to gene flow with Austroasiatic speaking neighbors or due to a history of language shift. This set of groups cluster together near the ASI end of the Indian cline, likely reflecting a large proportion of ASI-like ancestry as well as a distinct ancestry that has some affinity to East Asians. The Tibeto-Burman speaking groups and other groups with high proportions of East Asian related ancestry such as the Bengali and Austroasiatic speaking Khasi form a gradient of ancestry relating them to East Asian groups such as Han Chinese. These groupings are also evident in a neighbor-joining tree based on F_{ST} (Supplementary Figure 4). We confirmed the East Asian related mixture in some groups by observing significantly negative f_3(Test; Mala, Chinese) statistics[7] (Supplementary Table 2).

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 here in recessive diseases that do not owe their origin to consanguineous marriages of close relatives, we filtered the data to minimize this effect by computing IBD between all pairs of individuals in each group and removing one individual from the pairs with outlying numbers of IBD segments. We validated the effectiveness of this procedure for removing close relatives by simulation (Supplementary Table 2; 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 arrays we analyzed included more SNPs ascertained in Europeans than in South Asians, the sensitivity of our methods to founder events is expected to be greater in Europeans than in South Asians, and thus, our estimates of founder event strengths in South Asian groups is likely to be a conservative underestimate relative to that in Europeans (Supplementary Figure 5 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 with 1. Our simulations suggest that for sample sizes of 4, the algorithm's sensitivity was greater than 95% for determining that a group with two times the bottleneck strength as Finns would have an IBD score significantly greater than that of Finns, while specificity was perfect in all the simulations we performed (Supplementary Figure 1 and Supplementary Table 1). We also used two other methods for measuring founder events that did not require phasing or IBD detection. First, we computed F_{ST} between each group and every other group with similar ancestry sources. Second, for groups on the Indian Cline we fit a model of population history using *qpGraph*[7] and measuring the founder event as the group-drift after admixture (Supplementary Figure 6 and Online Data Table 1). The results of both methods were highly correlated to that of the IBD-based method for all cases where a comparison was possible (Pearson correlation r=0.82-0.98).

The IBD analyses suggest that 29% of the South Asian groups we analyzed (96 out of 327) have significantly stronger founder events than those in both Finns and Ashkenazi Jews (Figure 3). The South Asian groups with evidence of strong founder events include diverse tribe, caste, and religious groups, and our analysis identifies 14 groups with strong founder events census sizes of over a million (Figure 3; Table 1). The groups with smaller census sizes are also medically significant. Study of small census size groups with extremely strong founder events such as Amish, Hutterites, and the people of the Saguenay Lac-St. Jean region have led to the discovery of dozens of novel disease variants specific to each group[6], which highlights the potential of similar studies in South Asian groups. In addition to these analyses, we measured IBD across groups – searching 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 many cases of relatedness, which typically follow geography, religious affiliation (e.g. Catholic Brahmins), or linguistic grouping (particularly Austroasiatic speakers) (Supplementary Table 3).

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

As proof-of-principle, we highlight two examples. The first concerns the Vysya, a group with a census size of more than 3 million that we identified as having a founder event about 1.2-fold stronger than that in Finns (Figure 3). The Vysya have a 100-fold higher rate of butyrylcholinesterase deficiency than other Indian groups, and Vysya ancestry is a known counter-indication for the use of muscle relaxants such as succinylcholine or mivacurium that are given prior to surgery[15].

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

These observations highlight how systematic studies of South Asian founder groups are likely to be an effective approach for discovering mutations that cause recessive disease. Identification of pathogenic mutations responsible for such founder diseases is straightforward. All that is required is collection of DNA samples from a small number of affected individuals and their families, usually followed by whole-exome sequencing to discover the causal changes. Once group specific founder event disease mutations are discovered, they can be tested for prenatally. Mapping of recessive disease mutations may be particularly important in traditional communities practicing arranged marriages, which is common in India. An example of the power of this approach from outside India is given by *Dor Yeshorim*, a community genetic testing program among religious Ashkenazi Jews[22], which visits schools, screens students for common recessive disease causing mutations previously identified to be segregating at a higher frequency in the target group, and enters the results into a confidential database. Match-makers query the database prior to making suggestions to the families and receive feedback about whether the potential couple is “incompatible” in the sense of both being carriers for a recessive mutation at the same gene. Given that approximately 95% of community members whose marriages are arranged participate in this program, recessive diseases like Tay-Sachs have virtually disappeared in these communities. A similar approach should work as well in Indian communities where arranged marriages are common. Given the potential for saving lives, this or similar kinds of research could serve as an important investment for future generations[23].

This study of more than 275 distinct groups represents the first systematic survey for founder events in South Asia, and to our knowledge also presents the richest dataset of genome-wide data from anthropologically well-documented groups available from any region in the world. Despite the breadth of this data, the groups surveyed here represent only about 5% of the well-documented endogamous groups in South Asia, and extensions of the survey to all groups would make it possible to identify large numbers of additional founder groups susceptible to recessive diseases. To take advantage of the unique population structure of South Asia to improve health, we propose to reverse the standard approach to recessive disease gene discovery. Instead of focusing entirely on collecting cases in tertiary medical centers and mapping diseases in a group of individuals found to have the same phenotype while blinded to information about their caste or tribal status, we propose that medical geneticists should adopt a parallel strategy of working with community doctors, medical workers, and social workers to identify recessive diseases that occur at a high rate in endogamous groups with founder events, and then go out into the community to identify cases. Once a small number of cases are sampled, it is straightforward to map the causal variants. This approach was pioneered in 1950s by the work of Victor McKusick and his colleagues studying the Old Order Amish in Pennsylvania U.S.A., a founder population of approximately 100,000 individuals in whom many dozens of recessive diseases were mapped, a research program that was crucial to founding modern medical genetics and was of extraordinary health benefit to that community[24]. Our study suggests that the potential for disease gene mapping in India would be orders of magnitude greater.

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 8 supplementary figures and 5 supplementary tables.

Acknowledgements:

We are thankful to the many Indian, Pakistani, Bangladeshi, and Nepalese individuals who contributed the DNA samples analyzed here including in particular the Indian PPD and MPS patients. We are grateful to Analabha Basu and Partha P. Majumder for sharing 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, CSIR network project GENESIS (BSC0121) 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. DR is an Investigator of the Howard Hughes Medical Institute. Genotyping data for the samples collected for this study will be made available upon request from the corresponding authors.

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. Narang, A. *et al.* Recent admixture in an Indian population of African ancestry. *Am J Hum Genet* **89**, 111-20 (2011).
14. Shah, A.M. *et al.* Indian Siddis: African descendants with Indian admixture. *Am J Hum Genet* **89**, 154-61 (2011).
15. 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).
16. 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).
17. Bhavani, G.S. *et al.* Novel and recurrent mutations in WISP3 and an atypical phenotype. *Am J Med Genet A* **167A**, 2481-4 (2015).
18. Bidchol, A.M. *et al.* GALNS mutations in Indian patients with mucopolysaccharidosis IVA. *Am J Med Genet A* **164A**, 2793-801 (2014).
19. Hildebrandt, F. *et al.* A systematic approach to mapping recessive disease genes in individuals from outbred populations. *PLoS Genet* **5**, e1000353 (2009).

20. Hagiwara, K. *et al.* Homozygosity mapping on homozygosity haplotype analysis to detect recessive disease-causing genes from a small number of unrelated, outbred patients. *PLoS One* **6**, e25059 (2011).

21. 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).

22. Raz, A.E. Can population-based carrier screening be left to the community? *J Genet Couns* **18**, 114-8 (2009).

23. 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).

24. Francomano, K.R.D.a.C.A. Victor A. McKusick and the History of Medical Genetics. *Springer Science & Business Media*, 119-130 (2012).

25. Sudmant, P.H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).

26. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201-206 (2016).

27. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).

28. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).

29. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res* **19**, 318-26 (2009).

30. Hoaglin, B.I.a.D. *How to Detect and Handle Outliers*, (1993).

31. Palamara, P.F. ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process. *Bioinformatics* (2016).

32. Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166 (1989).

33. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39**, W475-8 (2011).

34. 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).

35. 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).

36. Browning, B.L. & Browning, S.R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459-71 (2013).

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

Table 1. South Asian groups with census sizes over 1 million and IBD scores 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 Human Origins) would have IBD scores of 1.0 and 0.9, IBD ranks of 121 and 135, and F_{ST} ranks of 133 and 154, 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).

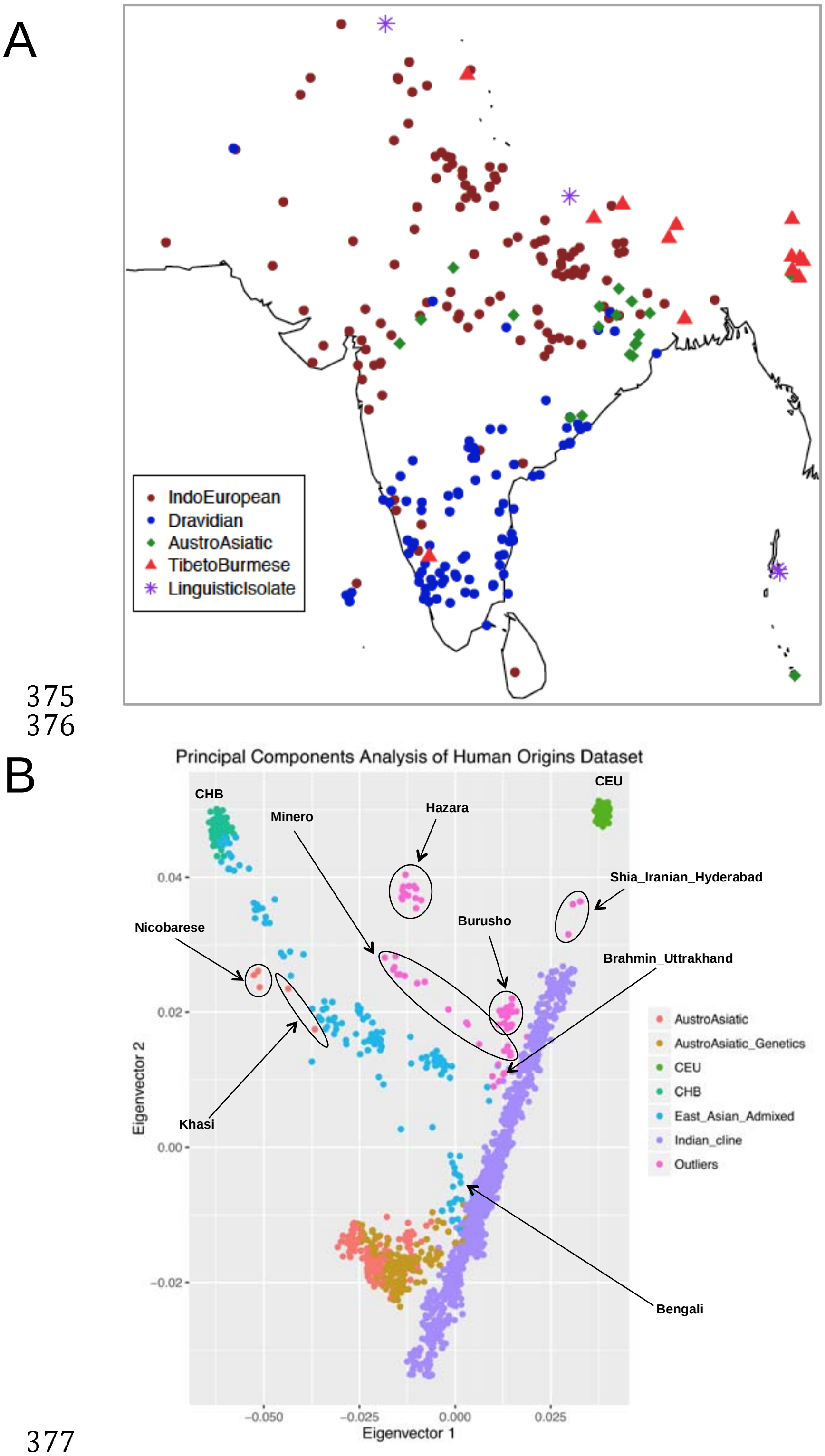

Figure 1. Dataset overview. (a) Sampling locations for all analyzed groups. Each point indicates a distinct group (random jitter was added to help in visualization at locations where there are many groups). (b) PCA of Human Origins dataset along with European Americans (CEU) and Han Chinese (CHB).

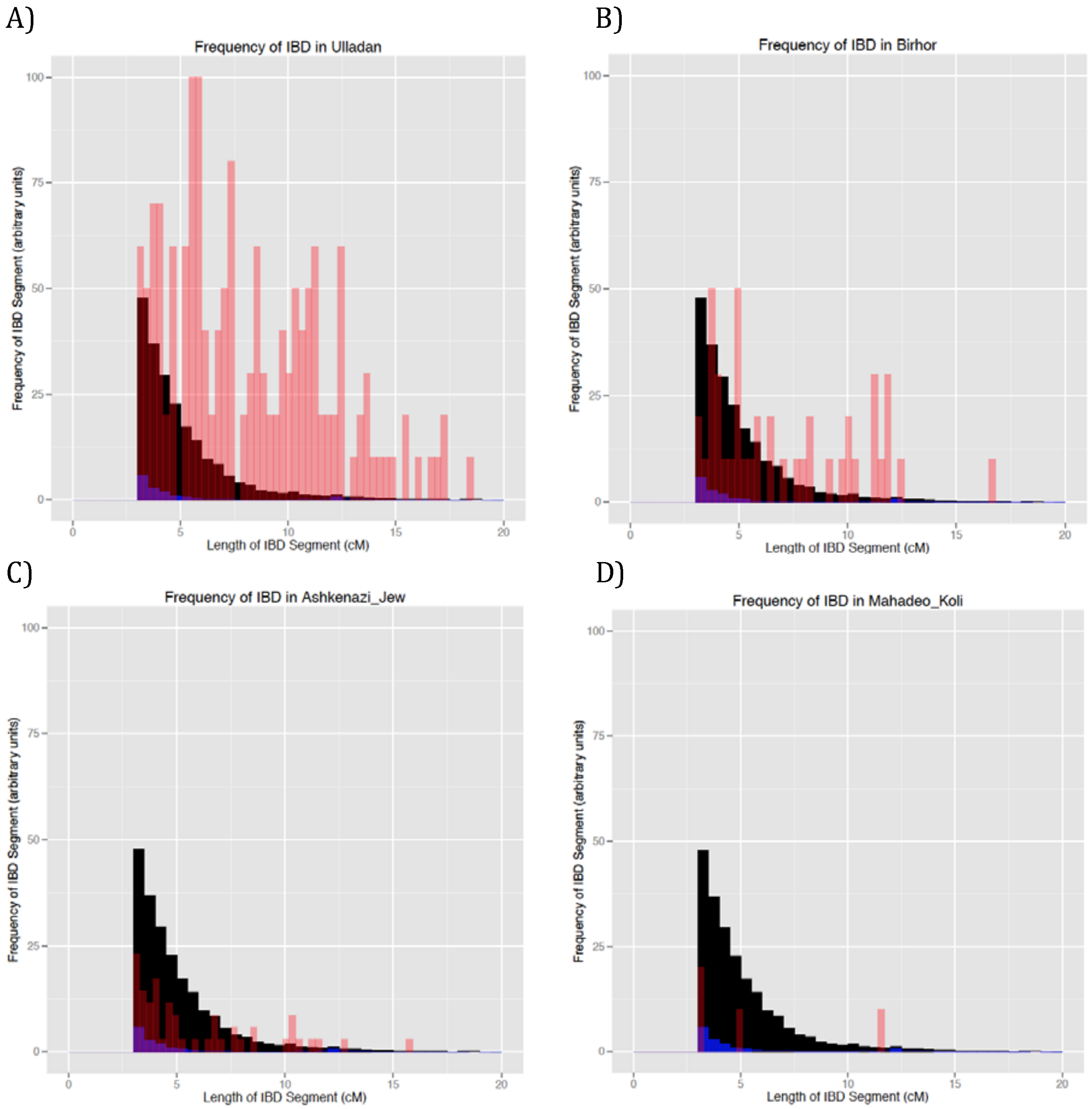

Figure 2. Example histograms of IBD segments to illustrate the differences between groups with founder events of different magnitudes: These histograms provide visual illustrations of differences between groups with different IBD scores. As a ratio relative to Finns (FIN; black), these groups (red) have 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.

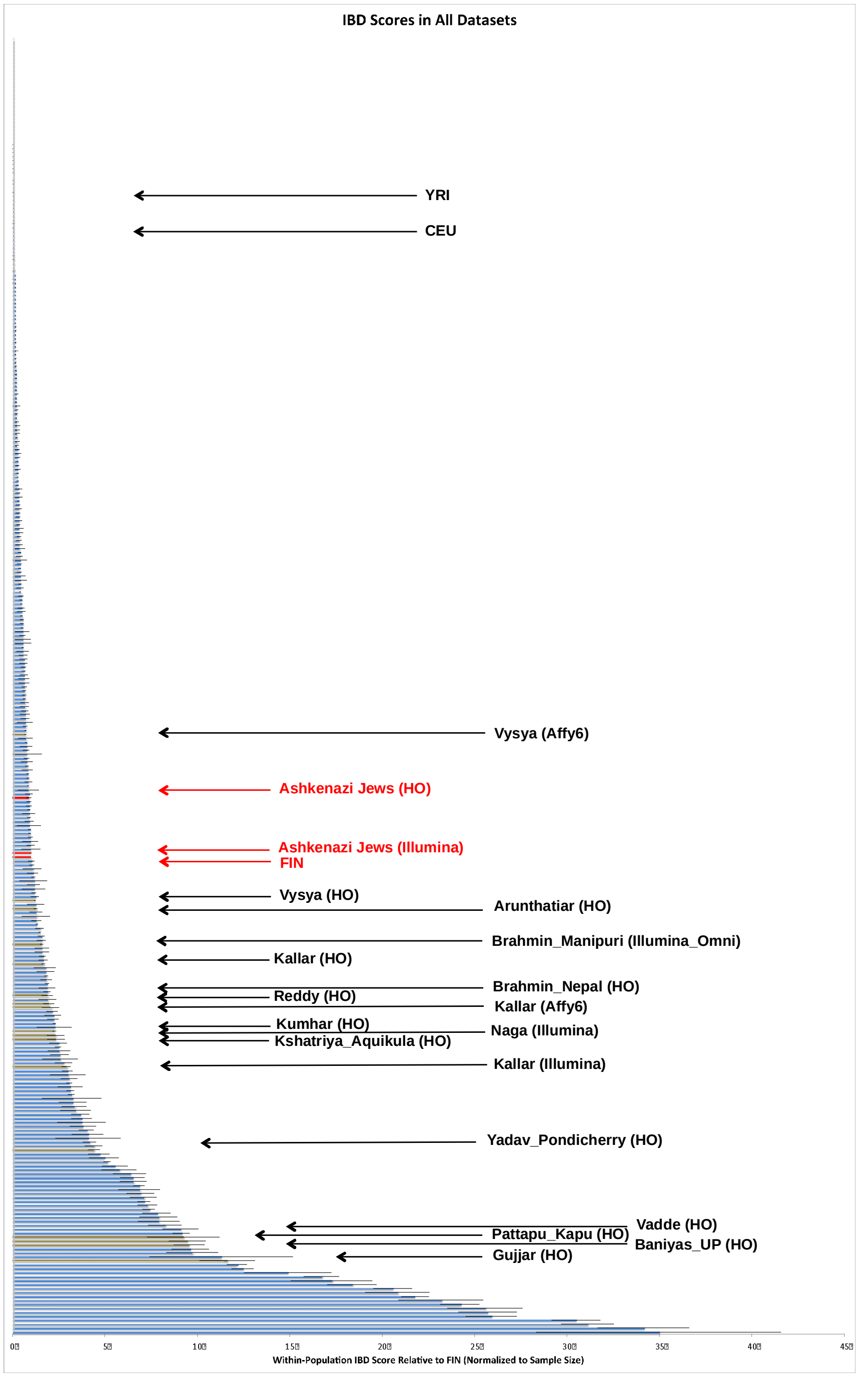

Figure 3. IBD scores relative to Finns (FIN). Histogram ordered by IBD score, which is roughly proportional to the per-individual risk for recessive disease due to the founder event. (These results are also given quantitatively for each group in Online Table 1.) We restrict to groups with at least two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

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 Basu *et al.*[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 Asia and Jewish data generated by the other Illumina arrays to create an "Illumina" dataset which consists of 172 individuals genotyped on 500,640 SNPs. Finally, we merged the data from the Affymetrix Human Origins arrays with the Ashkenazi Jewish data and data from the Simons Genome Diversity Project[25,26] 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 and possible systematic differences across genotyping platforms. We merged in the 1000 Genomes Phase 3 data[27] (2504 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, $Diff_A^i = \left(Freq^i_{PopA,Plate1} - Freq^i_{PopA,Plate2}\right)$. 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 from a weighted Block Jackknife across chromosomes. This quantity should be consistent with zero within a few standard errors if there are no batch effects that cause systematic differences across the plates, as allele frequency differences between two samples of the same group should be random fluctuations that have nothing to do with the array plates on which they are genotyped. This analysis found strong batch effects associated with one array plate, and we removed these samples from further analysis.

We used EIGENSOFT 5.0.1 smartpca[28] on each group to detect PCA outliers. We also developed a procedure to distinguish recent relatedness from founder events so that we could remove recently related individuals. We first identified all duplicates or obvious close relatives by using Plink "genome" and GERMLINE[29] to compute IBD (described in more detail below) and removed one individual from all pairs with a PI_HAT score greater than 0.45 and the presence of at least 1 IBD fragment greater than 30cM. We then used an iterative procedure to identify additional recently related individuals. For sample sizes above 5, we identified any pairs within each group that had both total IBD and total long IBD (>20cM) that were greater than 2.5 SDs and 1 SD, respectively, from the group mean. For sample sizes 5 or below, we used modified Z scores of 0.6745*(IBD_score - median(score))/MAD, where MAD is the median absolute deviation, and identified all pairs with modified Z scores greater than 3.5 for both total IBD and total long IBD as suggested by Iglewicz and Hoaglin[30]. After each round, we repeated the process if the new IBD score was at least 30% lower than the prior IBD score. Simulations showed that we were always able to remove a first or second cousin in the dataset using this method (Supplementary Table 5). Together these analyses removed 53 individuals from the Affymetrix 6.0 dataset, 21 individuals from the Illumina dataset, 43 individuals from the Illumina Omni dataset, and 225 individuals from the Human Origins dataset.

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

Simulations to Test Relatedness Filtering and IBD Analyses

We used ARGON[31] to simulate groups with different bottleneck strengths to test the IBD analyses, relatedness filtering, and founder event dating algorithms. We used ARGON's default settings, including mutation rate of $1.65*10^{-8}$ per base pair (bp) per generation and a recombination rate of $1*10^{-8}$ per bp per generation and simulated 22 chromosomes of size 130 Mb each. We pruned the output by randomly removing SNPs until there were 22,730 SNPs per chromosome to simulate the approximate number of positions in the Affymetrix Human Origins array. For the IBD analyses, we simulated groups to have descended from an ancestral group 1,800 years ago with N_e=50,000 and to have formed two groups with N_e=25,000. These groups continued separately until 100 generations ago when they combined in equal proportions to form a group with N_e=50,000. The group then split into 3 separate groups 72 generations ago that have bottlenecks leading to N_e of either 400, 800, or 1600. The 3 groups then exponentially expanded to a present size of N_e=50,000. We designed these simulations to capture important features of demographic history typical of Indian groups as detailed by Moorjani *et al.*[8] and Reich *et al.*[4] We chose the bottleneck sizes because they represent founder events with approximately the strength of Finns (the bottleneck to 800), and twice as strong (400) and half as strong (1600) as that group. We then performed the IBD analyses described below with 99 individuals from the group with bottleneck strength similar to that of Finns (198 haploid individuals were simulated and merged to produce 99 diploid individuals) and different number of individuals from the other groups. These analyses demonstrate that with only 4-5 individuals we can accurately assess the strength of founder events in groups with strong founder events (Supplementary Figure 1 and Supplementary Table 1). Weaker founder events are more difficult to assess, but these groups are of less interest for founder event disease mapping, so we aimed to sample ~5 individuals per group.

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

Distance-Based Phylogenetic Tree:

We calculated genetic differentiation (F_{ST}) between all pairs of groups using EIGENSOFT *smartpca* and created a neighbor-joining tree using PHYLIP[32] with Yoruba as the outgroup. We used Itol[33] to display the tree.

Phasing, IBD Detection, and IBD Score Algorithm:

We phased all datasets using Beagle 3.3.2 with the settings *missing=0; lowmem=true; gprobs=false; verbose=true*[34]. We left all other settings at default. We determined IBD segments using GERMLINE[29] with the parameters *-bits 75 -err_hom 0 -err_het 0 -min_m 3*. We used the genotype extension mode to minimize the effect of any possible phasing heterogeneity amongst the different groups and used the HaploScore algorithm to remove false positive IBD fragments with the recommended genotype error and switch error parameters of 0.0075 and 0.003[35]. We chose a HaploScore threshold matrix based on calculations from Durand *et al*. for a "mean overlap" of 0.8, which corresponds to a precision of approximately 0.9 for all genetic lengths from 2-10cM. In addition to the procedure we developed to remove close relatives (Quality Control section), we also removed segments longer than 20cM as simulations showed that this increased sensitivity of the analyses (Supplementary Table 4). We computed "IBD score" as the total length of IBD segments between 3-20cM divided by $\left\{\binom{2n}{2} - n\right\}$ where n is the number of individuals in each group to normalize for sample size. We then expressed each group's score as a ratio of their IBD score to that of Finns and calculated standard errors for this score using a weighted Block Jackknife over each chromosome with 95% confidence intervals defined as IBD score ±1.96*s.e.

We also repeated these analyses with FastIBD[36] 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 genotype 15 patients with progressive pseudorheumatoid dysplasia (PPD) and 6 patients with mucopolysaccharidosis (MPS) type IVA, all of which had disease mutations previously determined[16-18], and 3 of which (MPS patients) are newly reported here. After quality control, 6 of the PPD patients with Cys78Tyr mutations, 6 of the PPD patients with Cys337Tyr mutations, and 6 of the MPS patients with Cys78Arg mutations remained. 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 calculated the fraction by dividing by the total number of possible haplotypes (2 times the number of individuals), then took the larger value of the two possible alleles (thus the fraction range was 0.5-1). We then 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.

Calculating Group Specific Drift:

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

Code Availability:

Code for all calculations available upon request.