nature genetics

The promise of discovering population-specific disease-associated genes in South Asia

Nathan Nakatsuka[1,2], Priya Moorjani[3,4], Niraj Rai[5,15], Biswanath Sarkar[6], Arti Tandon[1,4], Nick Patterson[4], 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[13], David Reich[1,4,14,16] & Kumarasamy Thangaraj[5,16]

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 identified 81 unique groups, 14 of which had estimated census sizes of more than 1 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 identified multiple examples of recessive diseases in South Asia that are the result of such founder events. This study highlights an underappreciated opportunity for decreasing disease burden among South Asians through discovery of and testing for recessive disease-associated genes.

South Asia is a region of extraordinary diversity, containing over approximately 5,000 anthropologically well-defined groups, many of which are endogamous communities with substantial barriers to gene flow, owing to cultural practices that restrict marriage between groups[1]. Of the tiny fraction of South Asian groups that have been characterized with genome-wide data, many exhibit large allele-frequency differences from their close neighbors[2–4], reflecting strong founder events whereby a small number of ancestors gave rise to many descendants[4]. The pervasive founder events in South Asia present a potential opportunity for decreasing disease burden in South Asia, as 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 decreased recessive disease burden in all of these communities[5,6].

Here, we carried out new genotyping of 1,663 samples from 230 endogamous groups in South Asia by using the Affymetrix Human Origins SNP array[7]. We combined the newly collected data with previously reported data, thus yielding four data sets (**Fig. 1a**). The Affymetrix Human Origins SNP array data comprised 1,955 individuals from 249 groups in South Asia, to which we added data for 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, including 99 Finns, from 26 different groups) with each of these data sets. We removed SNPs and individuals that had a high proportion of missing genotypes or were outliers in principal component analysis (PCA) (**Fig. 1b** and **Supplementary Note**). The total number of unique groups analyzed in this study was 263, after accounting for groups represented in multiple data sets. To our knowledge, this work provides 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 on the basis of identity by descent (IBD) segments, large stretches of DNA originating from a common founder in the last approximately 100 generations (**Fig. 2**). We computed an IBD score (the average length of IBD segments between 3 and 20 cM detected between two genomes, normalized to sample size) as a measure of the strength of the founder event in each group's history. Because we were interested in characterizing the effects of recessive diseases that did not originate from consanguineous marriages of close relatives, we ignored self-matches (internal homozygosity) in IBD calculations. We removed all individuals with evidence of recent

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [2]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Biological Sciences, Columbia University, New York, New York, USA. [4]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India. [6]Anthropological Survey of India, Kolkata, India. [7]Department of Medical Genetics, Kasturba Medical College, Manipal University, Manipal, India. [8]Department of Applied Zoology, Mangalore University, Mangalore, India. [9]Centre for Human Genetics, Bangalore, India. [10]Amity Institute of Biotechnology, Amity University, Noida, India. [11]School of Life Sciences, Manipal University, Manipal, India. [12]Fetal Care Research Foundation, Chennai, India. [13]Genome Foundation, Hyderabad, India. [14]Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts, USA. [15]Present address: Birbal Sahni Institute of Palaeosciences, Lucknow, India. [16]These authors jointly directed this work. Correspondence should be addressed to D.R. (reich@genetics.med.harvard.edu) or K.T. (thangs@ccmb.res.in).

Received 24 February 2016; accepted 21 June 2017; published online 17 July 2017; doi:10.1038/ng.3917

relatedness (within several generations) to others in the data set by computing the IBD between all pairs of individuals in each group and removing one individual from each pair with an outlying number of IBD segments. (Our focus on founder events rather than recent relatedness also explains our choice to exclude IBD segments >20 cM in size.) We validated the effectiveness of this procedure by simulation (**Supplementary Table 1** and 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 data set. Because all the SNP arrays analyzed included more SNPs ascertained in Europeans than in South Asians, the sensitivity of our methods to founder events was greater in Europeans than in South Asians, and thus our estimates of founder-event strengths in South Asian groups are conservative underestimates relative to those in Europeans. (**Supplementary Fig. 1** demonstrates this effect empirically and shows that it results in less bias for the strong founder events that were the focus of this study.) We computed standard errors for these ratios by using a weighted block jackknife across chromosomes and concluded significance when the 95% confidence intervals did not overlap 1. We further carried out computer simulations to validate our procedure. The simulations suggested that we did not substantially overestimate the magnitudes of modest founder events, because for a simulated founder event with half the magnitude of that in Finns, we never inferred the score to be significantly greater than that in Finns. The simulations also suggested that our procedure was highly sensitive to detecting strong founder events, because for sample sizes of at least 5, the algorithm's sensitivity was >95% for determining that a group with twice the bottleneck strength as that of Finns had an IBD score significantly greater than that of Finns (**Supplementary Fig. 2** and **Supplementary Table 2**). We also used two additional non-IBD-based methods to measure the strength of founder events and, in cases in which a comparison was possible, found that these results were highly correlated with our IBD scores (**Supplementary Note** and **Supplementary Table 3**).

We inferred that 81 out of 263 unique groups (96 out of 327 groups if not considering the overlap of groups among data sets) had an IBD score greater than those of both Finns and Ashkenazi Jews (**Fig. 3**). These results did not change when we added back the outlier samples that we had removed in quality control. A total of 14 of these groups had estimated census sizes of over 1 million (**Fig. 3** and **Table 1**). However, the groups with smaller census sizes are also important: outside of South Asia, groups with small census sizes and extremely strong founder events, such as the Amish, Hutterites, and people of the Saguenay–Lac Saint-Jean region have led to the discovery of dozens of novel disease-causing variants. We also searched for IBD across groups—screening for cases in which the across-group IBD score was at least one-third of the within-group IBD score of Ashkenazi Jews—and found 77 cases of clear IBD sharing, which typically followed geography, religious affiliation, or linguistic grouping (particularly for Austroasiatic speakers) (**Supplementary Table 4**). Pairs of groups with high shared IBD and descent from a common founder event probably share risk for the same recessive diseases. However, these cross-group IBD sharing patterns did not drive our observations, because we identified 68 unique sets of groups without high IBD to other groups with significantly higher estimated IBD scores than both Finns and Ashkenazi Jews.

Our evidence that very strong founder events affect a large fraction of South Asian groups presents an opportunity to decrease disease burden in South Asia. This source of risk of recessive diseases is very different from risk due to marriages among close relatives, which is also a major cause of recessive disease in South Asia.

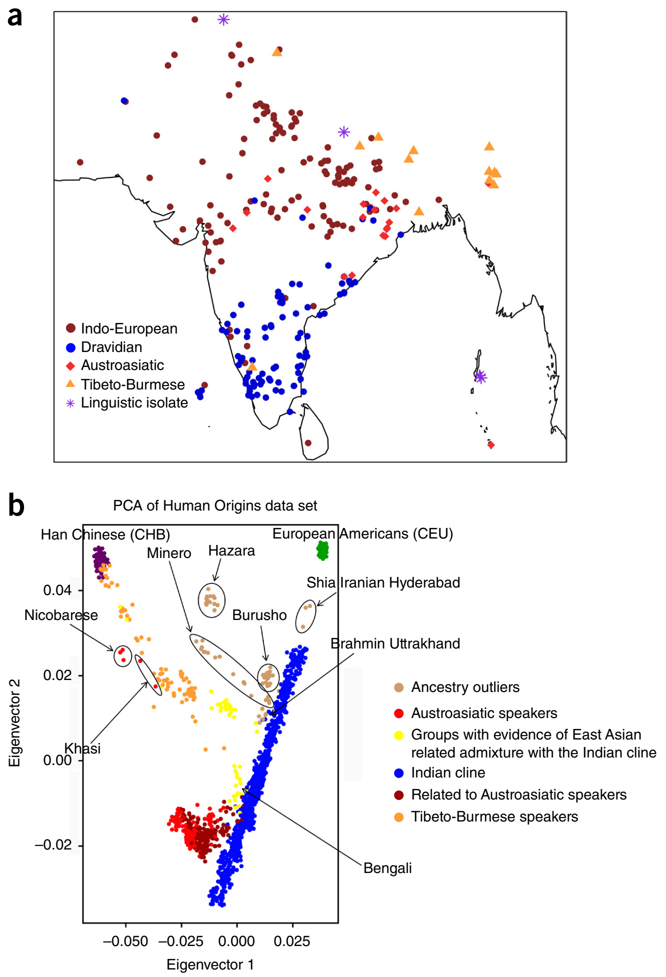

Figure 1 Data set overview. (**a**) Sampling locations for all analyzed groups. Each point indicates a distinct group (random jitter was added to aid in visualization at locations where there are many groups). (**b**) PCA of the Human Origins data set along with European Americans (CEU) and Han Chinese (CHB). There is a large cluster (blue) of Indo-European 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 genetically cluster with them (dark red). Finally, there are groups with genetic affinity to East Asians, including Tibeto-Burmese speakers (orange) and those that speak other languages (yellow).

To determine the relative effects of these risk factors, we computed F_{ST}, a measurement of allele-frequency differentiation, between each group in the data set and a pool of other South Asian groups chosen to be closest in terms of ancestry proportions. We found that inbreeding was not driving many of these signals, because 89 unique groups had higher F_{ST} scores than those of Ashkenazi Jews and Finns, even after the F_{ST} score was decreased by the proportion of allele-frequency differentiation due to inbreeding. These results show that although most studies mapping recessive disease-associated genes in South Asia have focused on families that are the products of marriages between close relatives, recessive diseases are also likely to occur at an elevated rate, even in nonconsanguineous cases, because of recent shared ancestors.

As an example of the promise of founder-event mapping of disease-associated genes in South Asia, we highlight the case of the Vysya, who

Table 1 South Asian groups with estimated census sizes over 1 million and IBD scores significantly greater than those of Ashkenazi Jews and Finns

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

14 South Asian groups with IBD scores significantly higher than those of Finns, census sizes over 1 million, and sample sizes of at least 3 that are of particularly high interest for founder-event mapping studies of disease-associated genes. For reference, Finns and Ashkenazi Jews (on the Human Origins array) 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). NA, not available.

have a census size of more than 3 million and an estimated IBD score approximately 1.2-fold higher than that of Finns (**Fig. 3**). The Vysya have an approximately 100-fold higher rate of butyrylcholinesterase deficiency than other groups, and Vysya ancestry is a known counterindication for the use of muscle relaxants, such as succinylcholine or mivacurium, that are given before surgery[13]. Butyrylcholinesterase deficiency is likely to occur at an elevated rate, owing to the founder event in the Vysya's history, and we expect that, like Finns, the Vysya probably exhibit a higher rate of many other diseases than do other groups. Other examples of recessive disease-associated genes with a likely origin in founder events are known anecdotally in South Asia, thus highlighting the importance of systematic studies to identify these genes[14].

To demonstrate how a new recessive disease in a founder-event group can be mapped, we carried out genome-wide SNP genotyping in 12 patients from southern India who had progressive pseudorheumatoid dysplasia (PPD), a disease known to be caused by mutations in *WISP3* (refs. 15,16). Of the six individuals with the WISP3 p.Cys78Tyr substitution[15,16], five were from nonconsanguineous marriages, and we found a much higher fraction of IBD at the disease-mutation site than in the rest of the genome in these individuals (**Supplementary Fig. 3a** and **Supplementary Fig. 4a**), in agreement with the p.Cys78Tyr substitution originating from a founder event and causing PPD in these patients. This pattern contrasted with those in the six other patients with different disease variants as well as those in six patients who carried a mutation in *GALNS* causing a different disease (mucopolysaccharidosis IVA (MPS IVA)), who were from primarily consanguineous marriages and who lacked substantial IBD across their disease mutation sites. Thus, these results suggested that, in these groups, the driver of the recessive diseases was marriage between close relatives (**Supplementary Note**). This example highlights how not only marriages between close relatives but also founder events are substantial causes of rare recessive diseases in South Asia.

The evidence of widespread strong founder events presents a major opportunity for discovering disease-associated genes and implementing public-health interventions in South Asia that is not widely appreciated (**Supplementary Table 5**). The current paradigm for mapping recessive disease-associated genes in South Asia is to collect cases in tertiary medical centers and map diseases in individuals with the same phenotype, a procedure often carried out by experimenters blinded to information about group affiliation, as was the case in our PPD study, in which we did not have access to the identity of the ancestral groups. However, our results suggested that collecting information on group affiliation may greatly strengthen the power of these studies.

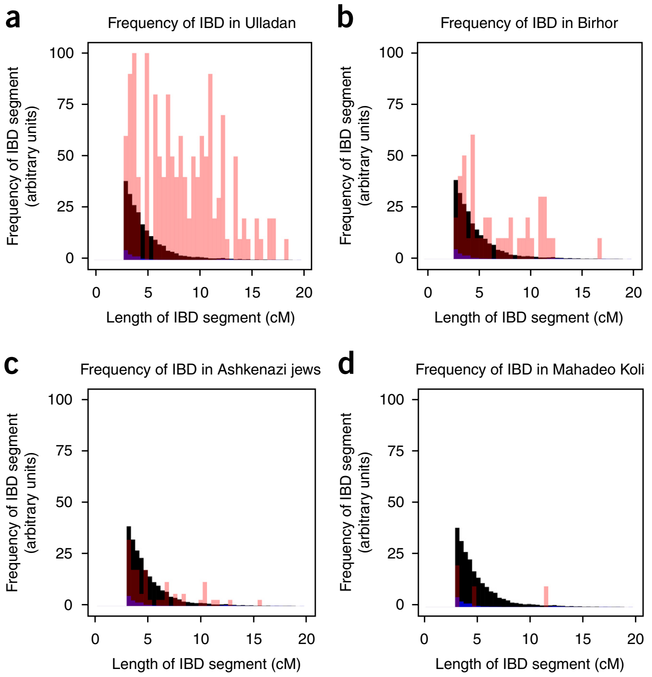

Figure 2 Example histograms of IBD segments, illustrating the differences between groups with founder events of different magnitudes. (**a**–**d**) Histograms providing visual illustrations of differences between groups with different IBD scores. As a ratio relative to Finns (FIN; black), these groups (red) have IBD scores of ~26 in Ulladan (**a**), ~3 in Birhor (**b**), ~0.9 in Ashkenazi Jews (**c**), and ~0.1 in Mahadeo Koli (**d**). 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 **Supplementary Table 5**. The IBD histograms were normalized to 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 data set.

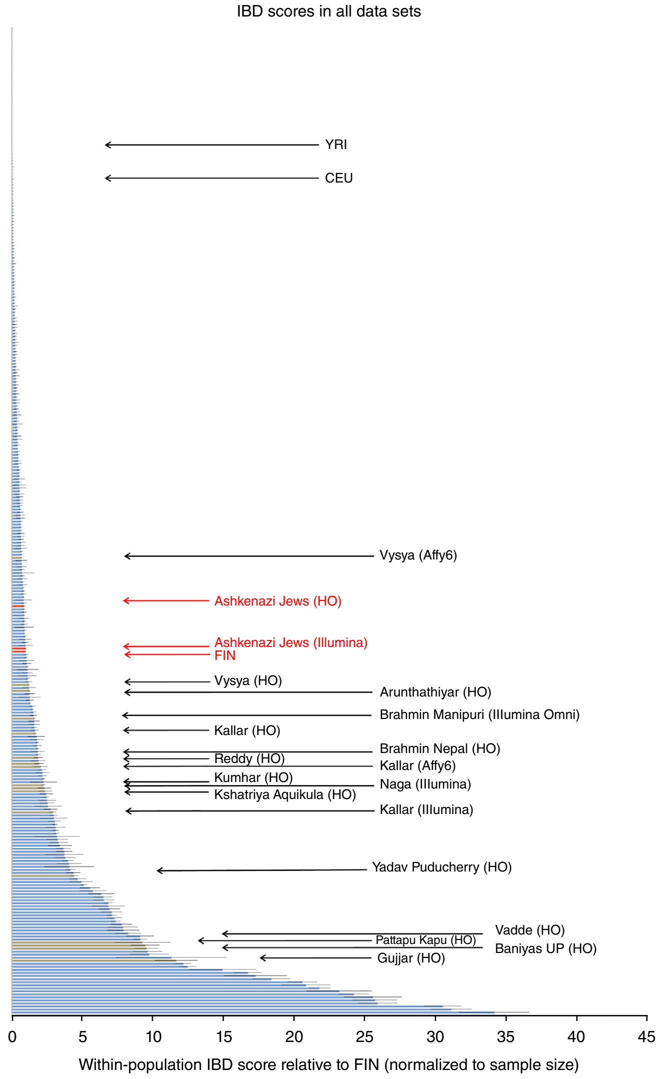

Figure 3 IBD scores relative to Finns (FIN). Histogram ordered by IBD score, which is approximately proportional to the per-individual risk for recessive disease due to the founder event. (These results are also given quantitatively for each group in **Supplementary Table 5**.) We restricted our analysis to groups with at least two samples, combining data from all four genotyping platforms into one plot. Data from Ashkenazi Jews and Finns are highlighted in red, and those from South Asian groups with significantly higher IBD scores than those of Finns and census sizes of more than 1 million are 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. Data sets are indicated in parentheses: Affy6, Affymetrix 6.0 data; HO, Human Origins data.

A fruitful way to approach gene mapping would be to proactively survey communities known to have strong founder events, searching for congenital diseases that occur at high rates in these communities. This approach was pioneered in the 1950s in studies of the Old Order Amish in the US, a founder population of approximately 100,000 individuals in whom many dozens of recessive diseases were mapped. That research program was crucial to founding modern medical genetics and provided extraordinary health benefits. Our results suggest that the potential for disease gene mapping in South Asia would be orders of magnitude greater.

Mapping of recessive diseases may be particularly important in communities practicing arranged marriages, which are common in South Asia. An example of the power of this approach can be found in Dor Yeshorim, a community genetic testing program among religious Ashkenazi Jews[17], which visits schools, screens students for common recessive disease-causing mutations previously identified to segregate at a higher frequency in the target group, and enters the results into a confidential database. Matchmakers query the database before 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. Because approximately 95% of community members whose marriages are arranged participate in this program, recessive diseases such as Tay–Sachs have virtually disappeared in these communities. A similar approach should work as well in South Asian communities. Given the potential for saving lives, this or similar types of research could be a valuable investment for future generations[18].

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are thankful to the many Indian, Pakistani, Bangladeshi, Sri Lankan, and Nepalese individuals who contributed the DNA samples analyzed here, including the patients with PPD and MPS IVA. We are grateful to A. Basu and P. P. Majumder (National Institute of Biomedical Genomics, Kalyani, India) for early sharing of data. Funding was provided by an NIGMS (GM007753) fellowship to N.N.; a Translational Seed Fund grant from the Dean's Office of Harvard Medical School, and grant HG006399 to D.R.; a Council of Scientific and Industrial Research, Government of India grant to K.T.; support from TIFAC-CORE to S.A.V. and K.S.; and NIGMS grant 115006 to P.M. The study of MPS IVA patients was funded by grants from the Department of Biotechnology, Government of India (BT/PR4224/MED/97/60/2011 to S.S. and S.M.J.) and the Department of Science and Technology, Government of India (SR/WOS-A/LS-83/2011 to S.S.). Funding for the mutation analysis of Indian patients with PPD was provided by the Indian Council of Medical Research (BMS 54/2/2013) to K.M.G. D.R. is supported as an Investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

N.N., P.M., D.R., and K.T. conceived the study. N.N., P.M., N.R., B.S., A.T., N.P., and D.R. performed analysis. G.S.B., K.M.G., M.S.M., S.S., A.K., S.A.V., S.M.J., K.S., L.S., and K.T. collected data. N.N., D.R., and K.T. wrote the manuscript with the help of all coauthors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Mastana, S.S. Unity in diversity: an overview of the genomic anthropology of India. *Ann. Hum. Biol.* **41**, 287–299 (2014).
2. Bamshad, M.J. *et al.* Female gene flow stratifies Hindu castes. *Nature* **395**, 651–652 (1998).
3. Basu, A. *et al.* Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.* **13**, 2277–2290 (2003).
4. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (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**, 233–247 (2002).

7. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
8. Moorjani, P. *et al.* Genetic evidence for recent population mixture in India. *Am. J. Hum. Genet.* **93**, 422–438 (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**, 731–744 (2011).
10. Behar, D.M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466**, 238–242 (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. USA* **113**, 1594–1599 (2016).
12. 1000 Genomes Project Consortium *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**, 461–468 (2006).
14. Shukla, A. *et al.* Homozygous p.(Glu87Lys) variant in ISCA1 is associated with a new multiple mitochondrial dysfunctions syndrome. *J. Hum. Genet.* **62**, 723–727 (2017).
15. Dalal, A. *et al.* Analysis of the WISP3 gene in Indian families with progressive pseudorheumatoid dysplasia. *Am. J. Med. Genet. A.* **158A**, 2820–2828 (2012).
16. Bhavani, G.S. *et al.* Novel and recurrent mutations in WISP3 and an atypical phenotype. *Am. J. Med. Genet. A.* **167A**, 2481–2484 (2015).
17. Raz, A.E. Can population-based carrier screening be left to the community? *J. Genet. Couns.* **18**, 114–118 (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).

ONLINE METHODS

Data sets. We assembled a data set 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 (**Fig. 1a**). We merged these data with the data set published in Moorjani *et al.*[8], which consisted of 332 individuals from 52 groups genotyped on the Affymetrix 6.0 array. We also merged the data with two additional data sets published in Metspalu *et al.*[9], consisting of 151 individuals from 21 groups genotyped on Illumina 650K arrays, as well as a data set published in Basu *et al.*[11], consisting of 367 individuals from 20 groups generated with Illumina Omni 1-Quad arrays. All the samples were collected with the approval of the Institutional Ethical Committees (IEC) of the CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India; Kasturba Hospital, Manipal, India; the Centre for Human Genetics, Bangalore, India; and the Fetal Care Research Foundation, Chennai, India.

We analyzed two different data sets of Ashkenazi Jews, 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’ data set consisted of 332 individuals genotyped for 329,261 SNPs, and our ‘Illumina Omni’ data set consisted of 367 individuals genotyped for 750,919 SNPs. We merged the data sets for South Asians and Ashkenazi Jews generated by the other Illumina arrays to create an ‘Illumina’ data set consisting of 172 individuals genotyped for 500,640 SNPs. We merged the data from the Affymetrix Human Origins arrays with the data for Ashkenazi Jews and data from the Simons Genome Diversity Project[19,20] to create a data set with 4,402 individuals genotyped for 512,615 SNPs. We analyzed the four data sets separately, owing to the small intersection of SNPs among them. We merged in the 1000 Genomes Phase 3 data[21] (comprising 2,504 individuals from 26 different groups and notably including 99 Finns) into all of the data sets. We used genome reference sequence coordinates (hg19) for analyses.

Quality control. We filtered the data at both the SNP and individual levels. 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 in which 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 the two groups *A* and *B* genotyped on the same plates,

$$\frac{1}{n}\Sigma_{i=1}^{n}\left(Diff_A^i\right)\left(Diff_B^i\right)$$

as well as a standard error from a weighted block jackknife across chromosomes. This quantity should be consistent with zero within several standard errors if there are no batch effects causing systematic differences across the plates, because allele-frequency differences between two samples of the same group should be random fluctuations that are unrelated to the array plates on which they are genotyped. This analysis revealed strong batch effects associated with one array plate, and we removed those samples from further analysis.

We used EIGENSOFT 5.0.1 smartpca[22] on each group to detect PCA outliers and removed 51 samples. 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[23] ‘genome’ and GERMLINE[24] to compute the 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 one IBD fragment greater than 30 cM. 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 (>20 cM) that were greater than 2.5 s.d. and 1 s.d., respectively, from the group mean. For sample sizes of 5 or below, we used modified *Z* scores of 0.6745 × (IBD score – median(score))/median absolute deviation, and identified all pairs with modified *Z* scores >3.5 for both total IBD and total long IBD, as suggested by Iglewicz and Hoaglin[25]. 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 data set by using this method (**Supplementary Table 1**). Together, these analyses removed 53 individuals from the Affymetrix 6.0 data set, 21 individuals from the Illumina data set, 43 individuals from the Illumina Omni data set, and 225 individuals from the Human Origins data set.

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

Simulations to test relatedness filtering and IBD analyses. We used ARGON[26] to simulate groups with different bottleneck strengths to test the IBD analyses and relatedness filtering. We used ARGON’s default settings, including a mutation rate of 1.65×10^{-8} per bp per generation and a recombination rate of 1×10^{-8} per bp per generation, and simulated 22 chromosomes of a size of 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 and formed a group with $N_e = 50{,}000$. At 72 generations ago, the group then split into three separate groups with bottlenecks leading to N_e of 400, 800, or 1,600. The three groups then exponentially expanded to the present size of $N_e = 50{,}000$. We designed these simulations to capture important features of demographic history typical of Indian groups[4,8]. We chose the bottleneck sizes because they represented founder events that were approximately the same strength (bottleneck to 800), twice as strong (bottleneck to 400) and half as strong (bottleneck to 1,600) as that of the Finns. We then performed the IBD analyses described below with 99 individuals from the group with a bottleneck strength similar to that of the Finns (such that 198 haploid individuals were simulated and merged to produce 99 diploid individuals) and different numbers of individuals from the other groups. These analyses demonstrated that with only four or five individuals, we were able to accurately assess the strength of founder events in groups with strong founder events (**Supplementary Fig. 2** and **Supplementary Table 2**). Weaker founder events were more difficult to assess, but these groups were of less interest for founder-event disease mapping, so we aimed to sample approximately five individuals per group.

We wrote custom R scripts to simulate first- and second-cousin pairs. We took individuals from the bottleneck of size 800 and performed ‘matings’ by taking two 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 new diploid offspring. The matings were performed to achieve first and second cousins. We placed these cousins back into the group with a group of size 800, then ran the relatedness filtering algorithms to evaluate whether they would identify these individuals.

Phasing, IBD detection, and IBD score algorithm. We phased all data sets with Beagle 3.3.2 with the settings missing = 0; lowmem = true; gprobs = false; verbose = true[27]. We left all other settings at default. We determined IBD segments by using GERMLINE[24] 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 among 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 (ref. 28). We chose a HaploScore threshold matrix on the basis of calculations from Durand *et al.*[28] for a ‘mean overlap’ of 0.8, which corresponds to a precision of approximately 0.9 for all genetic lengths from 2 to 10 cM. Measurement of IBD can sometimes be difficult in admixed populations, owing to differential proportions of the divergent ancestries among different individuals in the same group, but we found that in both the simulated and real data, we were able to detect IBD at the expected amounts.

In addition to performing the procedure that we developed to remove close relatives (described in ‘Quality control’ above), we also removed segments longer than 20 cM, because simulations showed that this procedure increased sensitivity of the analyses (**Supplementary Table 2**). To normalize for sample size, we computed the IBD score as the total length of IBD segments between 3 and 20 cM divided by

$$\left\{\binom{2n}{2}-n\right\}$$

where n is the number of individuals in each group. We then expressed each group's score as a ratio of their IBD score to that of the Finns and calculated standard errors for this score by using a weighted block jackknife over each chromosome, with 95% confidence intervals defined as IBD score ± 1.96 × s.e.

We repeated these analyses with FastIBD[29] for the Affymetrix 6.0 and Illumina data sets 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 data sets into different groups, because 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 26 of the 1000 Genomes groups to improve phasing, but among these groups, we included only results for Ashkenazi Jews and two outbred groups (CEU and YRI) in the final IBD-score ranking.

Analyses of patients with disease. We used Affymetrix Human Origins arrays to genotype 12 patients with PPD and six patients with MPS IVA, all of whom had previously determined disease-associated mutations[15,16,30] (three of the surveyed patients with MPS IVA are newly reported here). A total of six of the patients with PPD had p.Cys78Tyr substitutions, six had p.Cys337Tyr substitutions (all six of the patients with MPS IVA had p.Cys78Arg substitutions). 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 (twice 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 by using GERMLINE, as described above. We collapsed individuals into 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 approximately one-third the founder-event strength of that of the Ashkenazi Jews).

f_3 statistics. We used the f_3 statistic[7]f_3(Test; Ref_1, Ref_2) to determine whether there was evidence that the Test group was derived from an admixture of groups related to Ref_1 and Ref_2. A significantly negative statistic would provide unambiguous evidence of mixture in the Test group. We determined the significance of the f_3 statistic by using a block jackknife with a block size of 5 cM. We considered statistics over 3 s.e. below zero to be significant.

Computing group-specific drift. As a second method to measure the strength of founder events, we used qpGraph[7] to model each Indian group in the cline as a mixture of ANI and ASI ancestry, by using the model (YRI, (Indian group, (Georgians, ANI)), [(ASI, Onge])) proposed by Moorjani *et al.*[8] This approach provides estimates for postadmixture drift in each group (**Supplementary Fig. 5**), which reflects the strength of the founder event (such that high drift values indicate stronger founder events). We included only groups in the Indian cline in this analysis, and we removed all groups with evidence of East Asian related admixture (**Fig. 1b** and **Supplementary Table 6**) because this admixture is not accommodated in the above model.

PCA-normalized F_{ST} calculations. As a third method to measure the strength of founder events, we estimated the minimum F_{ST} between each South Asian group (**Supplementary Fig. 6**) and a broad cluster of individuals spanning many groups, defined as described below, that were closest to them in the PCA (**Supplementary Note**). For the Affymetrix 6.0, Illumina, and Illumina Omni data sets, 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 in **Fig. 1b**) into one cluster, thus yielding a total of three clusters (with all other groups ignored for this analysis). For the Human Origins data set, 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, thus yielding a total of four clusters. (Khasi and Nicobarese were ignored in this analysis, because they did 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, on the basis of 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.

F_{ST} calculations to determine overlapping groups. Overlapping groups between the data sets were initially determined on the basis of anthropological information (**Supplementary Table 1**). We further empirically tested for overlap by computing the F_{ST} between different groups across all data sets for groups with significantly stronger IBD scores than those of Finns. (We could not perform this analysis for groups with weaker founder events, because they would have low F_{ST} with respect to each other even if they were truly distinct groups.) We considered pairs with F_{ST} <0.004 to be overlapping. These included all groups known to be overlapping according to anthropological information as well as three additional pairs of groups that might be genetically similar because of recent mixing (for example, Kanjars and Dharkar, which are distinct nomadic groups that live near each other but intermarry, thus explaining the low F_{ST} between them). A **Life Sciences Reporting Summary** for this paper is available.

Code availability. Code for all calculations is available upon request.

Data availability. The informed consent and permits associated with the newly reported data do not allow for fully public release. Therefore, researchers who wish to analyze the data should send the corresponding authors a PDF of a signed letter containing the following language: '(a) We will not distribute the data outside our collaboration; (b) We will not post data publicly; (c) We will make no attempt to connect the genetic data to personal identifiers; and (d) We will not use the data for commercial purposes.'

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**, 318–326 (2009).
25. Hoaglin, D.C. & Iglewicz, B. *How to Detect and Handle Outliers* (ASQC Quality Press, 1993).
26. Palamara, P.F. ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process. *Bioinformatics* **32**, 3032–3034 (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**, 1084–1097 (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**, 2212–2222 (2014).
29. Browning, B.L. & Browning, S.R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
30. Bidchol, A.M. *et al.* GALNS mutations in Indian patients with mucopolysaccharidosis IVA. *Am. J. Med. Genet. A.* **164A**, 2793–2801 (2014).

nature research

Corresponding Author: David Reich

Date: 06/06/2017

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

▸ Experimental design

1. Sample size

Describe how sample size was determined.

We used simulations to determine the approximate number of individuals that would be needed to detect strong founder events (main text lines 457-490).

2. Data exclusions

Describe any data exclusions.

We used standard quality control metrics in the field to exclude SNPs and individuals based on genotyping completeness, batch effects, close relatives, or PCA outliers (main text lines 409-453).

3. Replication

Describe whether the experimental findings were reliably reproduced.

We increased our sample size for groups detected to have strong founder events and repeated the analyses with the larger sample sizes.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Individuals were grouped based on anthroplogical information.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was not relevant to the study, because it would not affect the results.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
□	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	□	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
□	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
□	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

We used PLINK version 1.7 (a genotype analysis tool for basic calculations of SNP data), smartpca (from EIGENSOFT to perform PCAs and Fst calculations), GERMLINE (version 1.5.1, to calculate IBD), ARGON (for simulations of population history), Beagle (version 3.3.2, to do phasing), and Haploscore (to find false positive IBD segments), qpgraph (from Eigensoft, to make a model of population history), ADMIXTOOLS (to do f3 statistics), and self-written code to do smaller data manipulations.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

The informed consents and permits associated with the newly reported data are not consistent with fully public release. Therefore, researchers who wish to analyze the data should send the corresponding authors a PDF of a signed letter containing the following language: "(a) We will not distribute the data outside my collaboration, (b) We will not post data publicly, (c) We will make no attempt to connect the genetic data to personal identifiers, (d) We will not use the data for commercial purposes."

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

N/A

b. Describe the method of cell line authentication used.

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used have been authenticated OR state that no eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination OR state that no eukaryotic cell lines were used.

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

Provide a rationale for the use of commonly misidentified cell lines OR state that no commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Individuals came from over 260 groups in India, Pakistan, Nepal, Sri Lanka, and Bangladesh. Age and gender were not provided. Patients with mucopolysaccharidosis type IVA and progressive pseudorheumatoid dysplasia were also studied. Their ethnic group, age and gender information were not provided or used for this study.