The promise of disease gene discovery in South Asia

Nathan Nakatsuka[1,2], Priya Moorjani[1,3], Niraj Rai[4], Biswanath Sarkar[5], Arti Tandon[1,6], Nick Patterson[6], Lalji Singh[4], David Reich[1,6,7,*], 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]Anthrological Survey of India, Kolkata, West Bengal 700016, India

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

[7]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA

It is tempting to think of the more than 1.5 billion people who live in South Asia as one large ethnic group, but in fact, South Asia is better viewed as comprised of very many small endogamous groups that usually marry within their own group (caste or tribe). To perform a high resolution assessment of South Asian demography, we assembled genome-wide data from over 2,000 individuals from over 250 distinct South Asian groups, more than tripling the number of diverse India groups for which such data are available, and including tribe and caste groups sampled from every state in India. We document shared ancestry across groups that correlates with geography, language, and caste affiliation, and characterize the strength of the founder events that gave rise to many of these groups. Over a third of the groups—including eighteen with census sizes of more than a million—descend from founder events stronger than those in Ashkenazi Jews and Finns, both of which have high rates of recessive disease due to their histories of strong founder events. These results highlight a major and unappreciated opportunity for reducing the disease burden among South Asians through the discovery of and genetic testing for recessive disease genes.

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 intermarriage[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 proximal neighbors[2-4], indicating that they have experienced strong founder events, whereby a small number of ancestors gave rise to the many descendants that exist today[4]. The evidence that a substantial fraction of groups in South Asia might descend from founder events represents a major opportunity for improving health. Detailed studies of founder populations of European ancestry, including Ashkenazi Jews, Finns, Amish, Hutterites, Sardinians, and French Canadians, have resulted in the discovery of dozens of rare recessive diseases in each group, allowing genetic

counseling that has helped reduce disease burden in each of these communities[5]. Opportunities for improving health through founder event disease mapping in India are even greater due to more widespread endogamy.

To characterize the medically relevant founder events in India, we carried out new genotyping of 890 samples from 206 endogamous groups in India on the Affymetrix Human Origins single nucleotide polymorphism (SNP) array[6]. Based on power calculations to determine the number of samples needed to confidently detect a founder event at least as strong as that in Ashkenazi Jews or Finns (Supplementary Figure 1), we aimed in most cases to genotype up to five individuals per group. Previous studies that sampled the genetic diversity of South Asia focused to a disproportionate extent on tribal groups and castes with small census sizes in order to capture the largest possible amount of anthropological diversity[3,4,7-9]. In this study, our sampling included many groups with large census sizes to investigate the prospects for future disease gene mapping. 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 1,192 individuals from 231 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[7]. We merged 1000 Genomes Phase 3 data (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 those that were outliers in Principal Component Analysis (PCA). To remove close relatives, we also removed one individual from each pair that were outliers in their group for Identity-by-Descent (IBD) genomic segments, and we removed all IBD segments that were larger than 20 centimorgans (cM).

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)[4,11,12]. We next removed YRI, Siddi and indigenous Andamanese (another known outlier) from the datasets and repeated PCA (Figure 1b, Supplementary Figure 3). The PCA documents three broad clusters[4,8,7]. First, almost all Indian groups speaking Indo-European and Dravidian languages lie along the "Indian Cline," reflecting the fact that they are admixed, with different proportions of Ancestral Northern Indian (ANI) ancestry related to Europeans, Central Asians, and Near Easterners; and Ancestral Southern 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 Austro-Asiatic speaking neighbors or 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-Burmese speaking groups and other groups with high proportions of East Asian ancestry such as the Austroasiatic speaking Khasi and Tharu form a gradient of ancestry relating them to East Asian groups such as Han Chinese. These three clusters are also evident in a neighbor-joining tree based on F_{ST} (Supplementary Figure 4). We confirmed the East Asian related admixture in some groups using the statistic f_3(Test; Mala, Chinese); significantly negative values of this statistic provide unambiguous evidence that the *Test* population is admixed of populations related (perhaps distantly) to Mala (an Indian Cline group with high ASI ancestry) and Chinese[6] (Supplementary Table 1).

For each pair of individuals, we used GERMLINE to detect segments of the genome where the individuals are likely to share a common ancestor within the last few dozen generations, that is, where they are IBD[13]. We used HaploScore to filter out segments that were likely to be false-positives[14]. After normalizing for sample size, we estimated the distribution of IBD genome-wide within each group with at least two samples (Figure 2, Online Data Table 1). We found systematic differences in the

inferred IBD on different platforms; however, by normalizing by the average IBD in each group by that detected in European Americans (CEU) (present in all three datasets), we were able to meaningfully compare groups across platforms (Supplementary Figure 5). We confirmed the accuracy of this method for detecting founder events by using two other methods that we found gave highly correlated results (correlation r=0.86-0.98): first, we computed F_{ST} between each group and other groups with similar ancestry sources, and second, we fit a formal model of history using *qpGraph*[6] and measured the founder event as the population-specific genetic drift post-admixture (Supplementary Figure 6)(Online Data Table 1). These analyses suggest that over a third of the Indian groups we analyzed (111 in total) have stronger founder effects than those that occurred in both Finns and Ashkenazi Jews (Figure 3). These groups are geographically and anthropologically varied, include diverse tribe, caste, and different religious groups, and also include eighteen groups with census sizes of over a million (Figure 3; Table 1). However, the groups with smaller census sizes are also medically important, as the per-individual rate of recessive disease is expected to be higher in proportion to their IBD score. Study of groups with small population size such as Amish, Hutterites, and the people of the Saguenay Lac-St. Jean region have proven to be powerful, leading to the discovery of dozens of novel disease variants that are specific to each group.

To better understand the history of the groups in which we detected founder events, we computed IBD for all pairs of individuals across groups. After applying a cutoff on the IBD score corresponding to ~1/3 of the founder event size of Ashkenazi Jews, most of the groups either have no matches or only match other individuals in their group. However, some groups also share IBD across groups, typically following a religious affiliation (e.g. Catholic Brahmins) or a distinctive linguistic affiliation (particularly Austroasiatic speakers) (Supplementary Table 3). These results point to recent gene flows among some of these pairs of groups.

The strong founder events offer major opportunities for improving health in South Asia. The first opportunity lies in targeted discovery of new genetic risk factors for disease. It is already known that one group we identified as having a strong founder event, the Vysya, has over a 100-fold higher rate of butyrylcholinesterase deficiency than other Indian groups, so that in India, Vysya ancestry is a known counter-indication for the use of common muscle relaxants such as succinylcholine or mivacurium that are given prior to anasthesia[15]. The Agarwal community in North India, though not present in our study, is also known to have founder mutations causing higher rates of Hereditary Fructose Intolerance[16] and Megalencephalic Leukoencephalopathy[17]. Systematic studies in the Vysya and other founder event groups—involving collaboration with clinical geneticists, local pediatricians, obstetricians, midwives, and social workers to identify congenital syndromes that are common in these communities—would discover many more examples. Identification of pathogenic mutations responsible for such syndromes is straightforward with present technology. 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. While rare recessive diseases would be a prime target for gene mapping, the founder groups we have identified may also be of substantial importance for disease gene mapping studies of common disease, as rare variant association analyses are known to have enhanced power in such groups[18,19].

Once group specific founder event disease mutations are discovered, they can be tested for prenatally, and indeed, much of the improvement in human health that has come from founder event disease gene mapping studies is due to prenatal testing. Another way that discovery of rare recessive disease genes is likely to be important, especially in India, is through pre-marriage counseling in traditional communities where arranged marriages are common. An example of the power of this approach is *Dor Yeshorim*, a community genetic testing program among orthodox Ashkenazi Jews in both the United States and Israel[20]. Matchmaking is the norm among the hundreds of thousands of traditional religious orthodox Ashkenazi

Jews in both the United States and Israel, and *Dor Yeshorim* has taken the approach of visiting schools, genetically screening students for common recessive disease causing mutations known to affect Ashkenazi Jews using an inexpensive test, and entering the results into a confidential database. Match-makers query the *Dor Yeshorim* database prior to making their 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. The program is successful in both the United States and Israel, such that ~95% of community members whose marriages are arranged participate; as a result, recessive diseases like Tay Sachs have virtually disappeared from these communities. A similar approach should work as well in Indian communities where arranged marriages are common, and where there is already recognition of the power of clinical screening to affect birth outcomes. Given the potential for saving lives and ultimately financial and medical resources, this or similar kinds of research could serve as an important investment for future generations[21].

This study of more than 250 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 for any region in the world. Despite the breadth of this data, the groups surveyed here represent only about 5% of the anthropologically well-defined groups in India. Extensions of the survey to all well-defined anthropological groups would make it possible to identify large numbers of additional founder groups susceptible to recessive diseases and to assess the extent to which the founder events we have already detected are localized to the specific regions from which our samples were drawn, or are shared across people of the same ethnic group across different regions in India. An important priority for future work is also to carry out pilot studies to find real disease genes in our groups, thereby proving by example the power of this approach for directing future disease mapping studies.

Supplementary Data:

Supplementary Data include an excel spreadsheet detailing all groups and their scores on the IBD, Fst, and Population-specific drift analyses. Also included are 6 supplementary figures and 3 supplementary tables.

Acknowledgements:

We are grateful to the many Indian, Pakistani, Bangladeshi, and Nepalese communities and individuals who contributed the DNA samples analyzed here. We thank Raj Rajkumar (deceased) for his assistance in assembling the sample collection. We would also like to acknowledge Analabha Basu and Patha Majumdar for helping to share their data for analysis in this study. Funding for this project was provided by the NIGMS (T32GM007753) to NN. This work was supported by a Translational Seed Fund grant from the Dean's Office of Harvard Medical School to DR, who is also a member of the Howard Hughes Medical Institute. KT is supported by CSIR network project GENESIS (BSC0121). PM was supported by the National Institutes of Health (NIH) under a Ruth L. Kirschstein National Research Service Award F32 GM115006-01. 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. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-93 (2012).
7. 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).
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. Narang, A. *et al.* Recent admixture in an Indian population of African ancestry. *Am J Hum Genet* **89**, 111-20 (2011).
12. Shah, A.M. *et al.* Indian Siddis: African descendants with Indian admixture. *Am J Hum Genet* **89**, 154-61 (2011).
13. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res* **19**, 318-26 (2009).

14. 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).
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. Bijarnia-Mahay, S. *et al.* Molecular Diagnosis of Hereditary Fructose Intolerance: Founder Mutation in a Community from India. *JIMD Rep* **19**, 85-93 (2015).
17. Shukla, P. *et al.* Molecular genetic studies in Indian patients with megalencephalic leukoencephalopathy. *Pediatr Neurol* **44**, 450-8 (2011).
18. Wang, S.R. *et al.* Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland. *Am J Hum Genet* **94**, 710-20 (2014).
19. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* **41**, 35-46 (2009).
20. Raz, A.E. Can population-based carrier screening be left to the community? *J Genet Couns* **18**, 114-8 (2009).
21. 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).
22. Sudmant, P.H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
23. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).
24. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).
25. Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166 (1989).

26. 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).

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

28. 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).

Population	Sample Size	IBD Rank	F_{ST} Rank	Drift Rank	Census Size	Location
Pattapu_Kapu	4	15	17	12	13,697,000	Andhra Pradesh
Kumhar	5	20	28	18	3,144,000	Uttar Pradesh
Rajbanshi	4	29	NA	NA	3,801,677	West Bengal
Kallar	5	38	37	29	2,426,929	Tamil Nadu
Arunthathiyar	5	40	48	34	1,084,162	Tamil Nadu
Yadav	13	41	66	44	20,000,000	Puducherry
Naga	4	51	NA	NA	1,667,712	Nagaland
Scheduled_Caste_Haryana	4	53	66	47	5,113,615	Haryana
Reddy	8	56	70	53	22,500,000	Andhra Pradesh
Baniyas	4	60	81	56	4,848,000	Uttar Pradesh
Dawoodi	2	62	41	40	4,592,854	Gujarat
Muslim_Bihar	4	64	62	44	13,722,048	Bihar
Vysya	10	73	30	21	3,200,000	Andhra Pradesh
Rajput	4	75	139	136	51,137,113	Haryana
Garasia	5	77	70	43	4,215,603	Gujarat
Satnami	5	85	96	65	4,200,000	Chhattisgarh
Shani	3	87	139	80	4,029,411	Bihar
Newar	6	95	NA	NA	1,523,000	Nepal

Table 1. Indian groups with strong IBD scores. Eighteen Indian groups with IBD scores higher than Ashkenazi Jews and census sizes over 1 million that are of particularly high interest for founder event disease gene mapping studies.

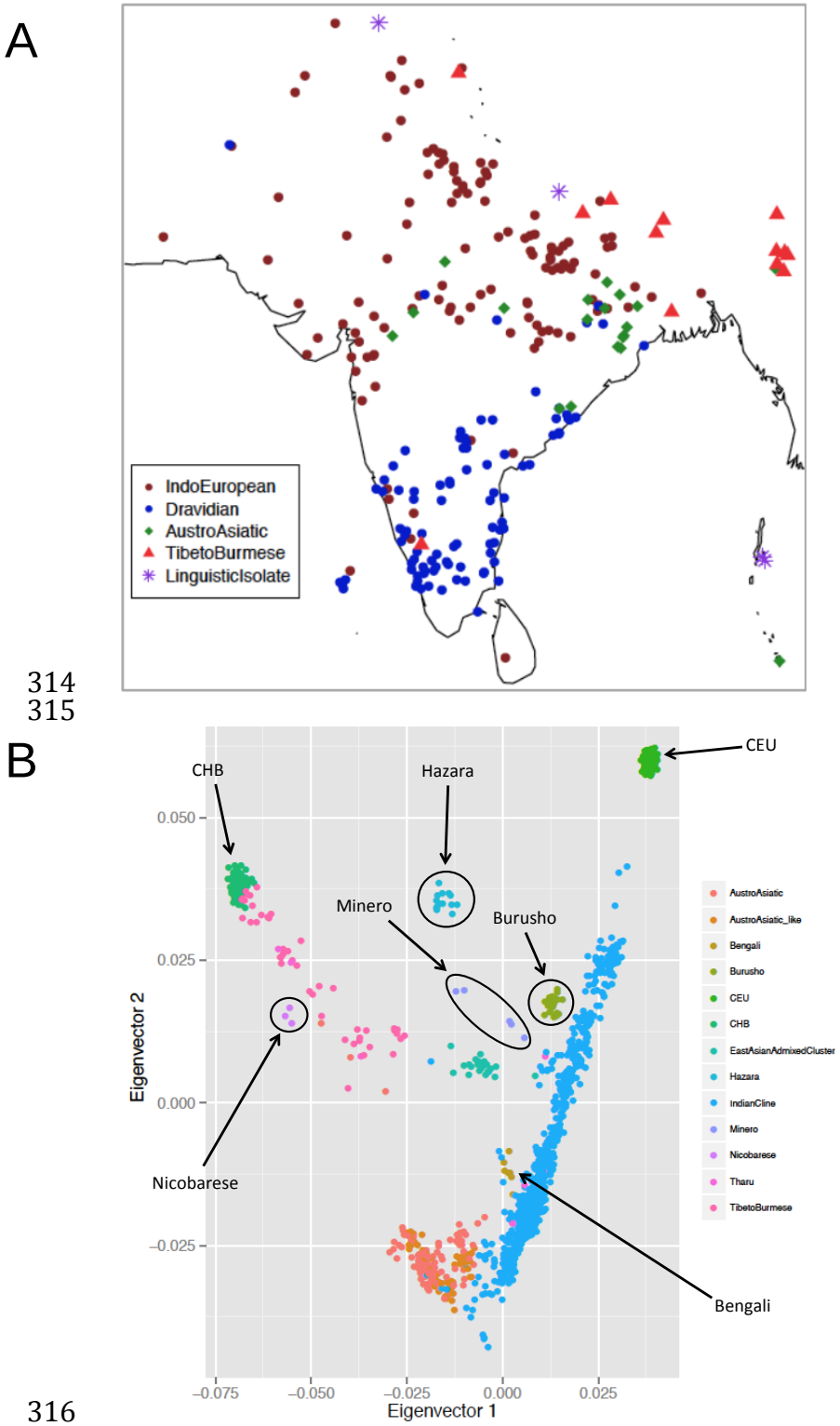
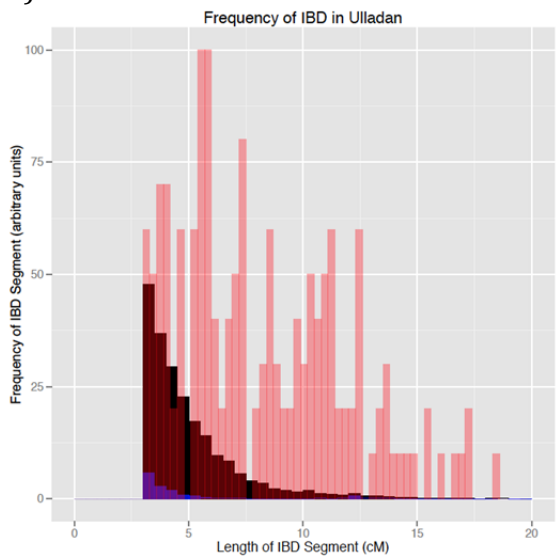
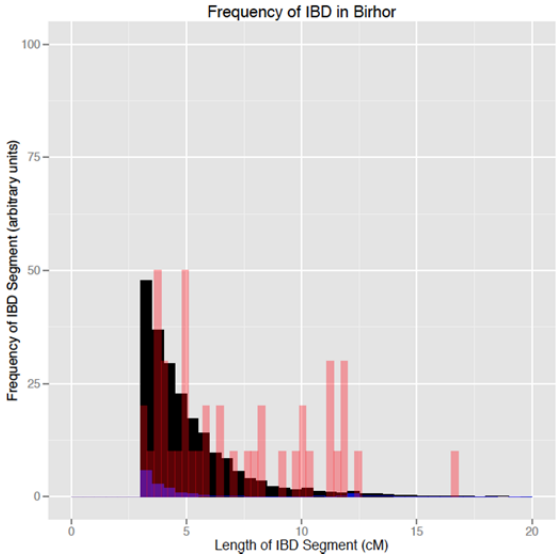


Figure 1. Dataset overview. (a) Sampling locations for all analyzed groups. Each point indicates a distinct group (random jitter added to help in visualization at locations where there are many groups). (b) PCA of Human Origins dataset along with European Americans (CEU) and Han Chinese (CHB).

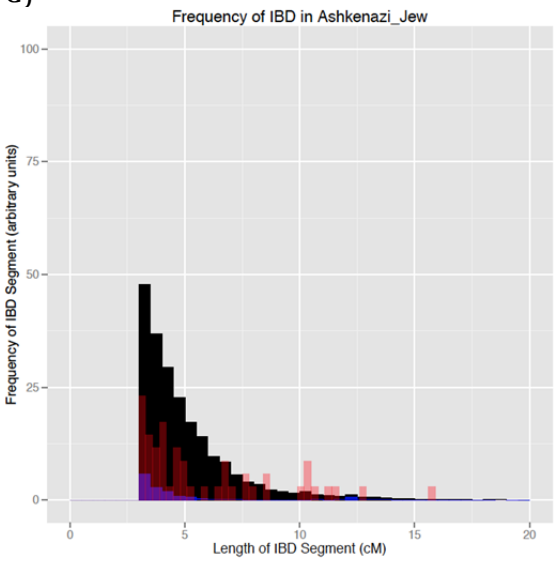
A)



B)



C)



D)

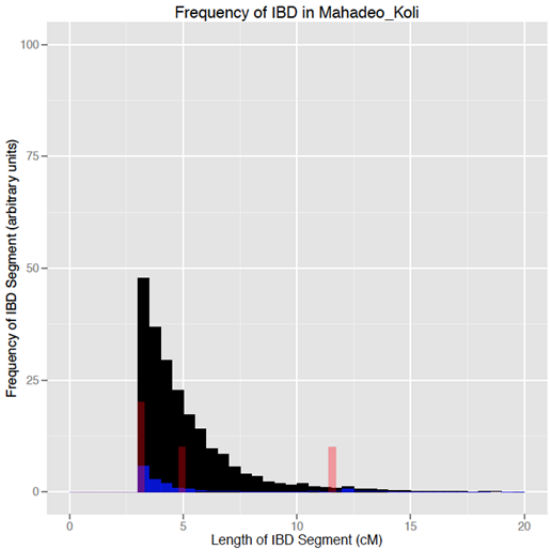


Figure 2. Histogram of IBD in groups with founder events of different magnitudes: (A) very large in Ulladan, (B) large in Birhor, (C) moderate in Ashkenazi Jews, and (D) small in Mahadeo_Koli. In each plot, we showed for comparison the histogram of IBD for European Americans (CEU) with a negligible founder event in blue, and that in Finns (FIN) with a large founder event in black.

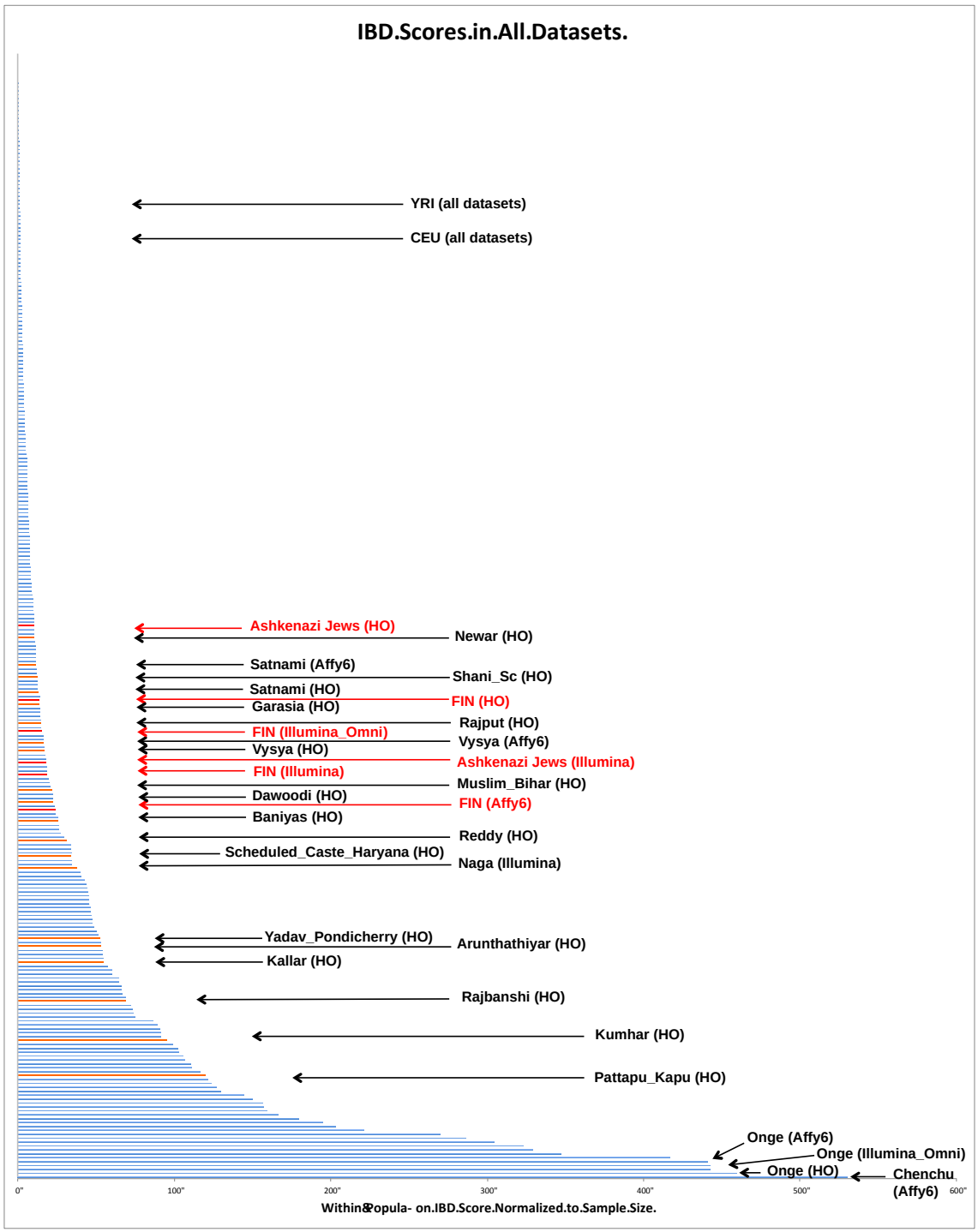


Figure 3. IBD scores normalized by that in European Americans (CEU). Histogram ordered by IBD score, which is roughly proportional to the per-individual risk for recessive disease due to the founder event. (These results are also given quantitatively for each group in Online Table 1.) We restrict to 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 stronger founder events and census sizes of more than a million in orange.

Online Methods:

Data Sets:

We used genotyping array data from multiple sources. We assembled a dataset of 1,182 individuals from 225 groups genotyped on the Affymetrix Human Origins array, of which data from 890 individuals from 206 groups is newly reported here (Figure 1a). We merged these data with a 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*.[7], which consisted of 367 individuals from 20 groups generated on Illumina Omni 1-Quad arrays. These groups came from India, Pakistan, Nepal, Sri Lanka, and Bangladesh, and we refer to all of them here as the "South Asia" dataset.

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

Our "Affymetrix 6.0" dataset consisted of 332 individuals genotyped on 329,261 SNPs, and our "Illumina_Omni" dataset consisted 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 consisting 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[22] to create a dataset with 1,225 individuals genotyped on 512,615 SNPs. We analyzed the four datasets separately due to the small intersection of SNPs between them and the possible systematic differences across genotyping platforms. We merged in the 1000 Genomes Phase 3 data[23] (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 on both the SNP and individual level. On the SNP level, we required at least 95% genotyping completeness for each SNP (across all individuals). On the individual level, we required at least 95% genotyping completeness for each individual (across all SNPs).

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

We used EIGENSOFT 5.0.1 smartpca[24] on each group. We also developed a procedure to distinguish recent relatedness from founder effects so that we could remove recently related individuals. We first identified all duplicates or obvious close relatives by using Plink "genome" and removed all individuals who had both a PI_HAT score greater than 0.45 and the presence of at least 1 IBD fragment greater than 30cM long. We used an iterative procedure of identifying 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. After each round we repeated the

process if the new IBD score was at least 30% lower than the prior IBD score. Due to their known very small census size (our sample consists of a substantial fraction of the entire population) and exceptional anthropological interest, we excluded Onge, a tribal population of the Andaman and Nicobar Islands, from this analysis. After data quality control and merging with the 1000 Genomes Project data, the Affymetrix 6.0 dataset included 3,215 individuals genotyped on 326,181 SNPs, the Illumina dataset included 2,789 individuals genotyped on 484,293 SNPs, the Illumina Omni dataset included 2,834 individuals genotyped on 750,919 SNPs, and the Human Origins dataset included 3,696 individuals genotyped at 500,648 SNPs.

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[25] with Yoruba as the outgroup. We used Itol[26] to display the tree.

Power Calculations:

We performed power calculations to determine the approximate number of samples required to detect founder events of a pre-specified strength. We used Beagle 3.3.2 FastIBD[27] to identify all shared IBD segments between individuals within a group with parameters *missing=0; lowmem=true; gprobs=false; verbose=true; fastIBD=true; ibdscale=scale* (where scale = sqrt(#samples/100)). We used the output of this program to calculate mean IBD sharing. We computed standard errors via jackknife resampling over individuals for each group. These analyses demonstrate that only 3-5 individuals are needed to assess accurately the size of founder effects in groups with strong founder events (Supplementary Figure 1). Weaker founder effects are more difficult to detect, but these groups are of less interest from the perspective of founder event disease mapping, so we aimed to sample ~5 individuals per group in the new genotyping effort based on the Affymetrix Human Origins platform.

Phasing and IBD Detection:

We phased all datasets using Beagle 3.3.2 with the settings *missing=0; lowmem=true; gprobs=false; verbose=true*[28]. We left all other settings at default. We determined IBD segments using the GERMLINE algorithm[13] 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[14]. 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 to ignore the effect of consanguinity and shorter than 3cM to minimize false positives and better differentiate groups with stronger founder effects from those with weaker effects. We treated all groups as subpopulations (e.g. Vysya and Ashkenazi Jews) and only retained IBD segments within the subpopulation (e.g. only IBD segments shared between Vysya individuals). We computed "founder effect size" as the total length of IBD segments between 3-20cM divided by ((2*(# of individuals in sample)) choose 2) to normalize for sample size. We also repeated these analyses with FastIBD[27] for the Affy 6.0 and Illumina datasets and observed that the results were highly correlated (r>0.96) (Supplementary Table 1). 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 several different Jewish groups and all twenty-six 1000 Genomes groups to improve phasing and calibration of the IBD scores, but of these groups we only included results for two founder groups (Ashkenazi Jews and Finns for comparison with Indian groups), and two outbred populations (CEU and YRI for normalization) in the final IBD score ranking.

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 (2*# of individuals in the 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 3, corresponding to approximately three times the founder effect size of CEU (and ~1/3 the effect size of Ashkenazi Jews).

f_3-statistics:

We used the f_3-statistic[6] $f_3(Test; Ref_1, Ref_2)$ 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 assessed 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[6] to model each Indian population on the cline as a mixture of ANI and ASI ancestry. Within the limits of our resolution, this model (YRI, (Indian population, (Georgians, ANI)), [(ASI, Onge])) proposed by Moorjani *et al*.[8] is a good fit to the data, in the sense that none of the *f*-statistics relating the groups are greater than three standard errors from expectation. 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 recent East Asian admixture.

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 population (Supplementary Figure 6). Specifically, we calculated the F_{ST} between each group and the rest of the

499 individuals in their respective cluster based on EIGENSOFT *smartpca*. For these
500 analyses we only included groups with Austroasiatic-related genetic patterns (i.e.
501 those groups clustering near Austroasiatic speakers on the PCA) and those on the
502 Indian cline; we excluded all groups with recent East Asian admixture. For
503 Ashkenazi Jews and Finns, we used the minimum F_{ST} to their closest European
504 neighbors.
505
506
507