ARTICLES

Reconstructing Indian population history

David Reich[1,2]*, Kumarasamy Thangaraj[3]*, Nick Patterson[2]*, Alkes L. Price[2,4]* & Lalji Singh[3]

India has been underrepresented in genome-wide surveys of human variation. We analyse 25 diverse groups in India to provide strong evidence for two ancient populations, genetically divergent, that are ancestral to most Indians today. One, the 'Ancestral North Indians' (ANI), is genetically close to Middle Easterners, Central Asians, and Europeans, whereas the other, the 'Ancestral South Indians' (ASI), is as distinct from ANI and East Asians as they are from each other. By introducing methods that can estimate ancestry without accurate ancestral populations, we show that ANI ancestry ranges from 39–71% in most Indian groups, and is higher in traditionally upper caste and Indo-European speakers. Groups with only ASI ancestry may no longer exist in mainland India. However, the indigenous Andaman Islanders are unique in being ASI-related groups without ANI ancestry. Allele frequency differences between groups in India are larger than in Europe, reflecting strong founder effects whose signatures have been maintained for thousands of years owing to endogamy. We therefore predict that there will be an excess of recessive diseases in India, which should be possible to screen and map genetically.

The first systematic surveys of human variation in India focused on anthropometric traits, and found that India is structured along lines of ethnicity as well as geography[1], a result that has since been confirmed by blood group, protein polymorphism[2,3] and genetic analysis[4]. Genetic studies have further documented differences in relatedness to west Eurasians[5–8], and mitochondrial DNA (mtDNA) studies have shown that India contains deep-rooted lineages that share no common ancestry with groups outside of South Asia for tens of thousands of years[9]. The most comprehensive survey of genetic variation in India so far analysed 405 single nucleotide polymorphisms (SNPs) in 55 groups and identified distinct clusters correlated to language and geography[10], while another study analysed 1,200 polymorphisms in 15 Indian American groups[11]. However, neither study analysed enough data to more finely discern patterns of genetic variation.

We genotyped 132 Indian samples from 25 groups. To survey a wide range of ancestries, we sampled 15 states and six language families (including two language families from the Andaman Islands[12]) (Table 1 and Fig. 1). To compare traditionally 'upper' and 'lower' castes after controlling for geography, we focused on castes from two states: Uttar Pradesh and Andhra Pradesh. We genotyped all samples on an Affymetrix 6.0 array, yielding data for 560,123 autosomal SNPs after filtering (Methods). Allele frequency differentiation between groups was estimated with high accuracy (F_{ST} had an average standard error of ± 0.0011; Supplementary Tables 1 and 2). For some analyses, we also merged our data with HapMap[13] and data from the Human Genome Diversity Panel (HGDP)[14,15] (Methods).

We analysed these data to address five questions. Does India contain more substructure than Europe? Has endogamy been long-standing in Indian groups? Do nearly all Indians descend from a mixture of populations? Is the ancestry of tribal groups systematically different from castes? What is the origin of the indigenous Andaman Islanders?

Extensive population structure in India

We applied principal components analysis (PCA)[16,17] to identify outlier groups (Supplementary Fig. 1). The first principal component shows that the Siddi have African ancestry, consistent with their origin involving the Arab slave trade[18]. The second shows that the Nyshi and Ao Naga cluster with the Chinese (CHB), consistent with them speaking Tibeto-Burman languages. The third and fourth show that the Great Andamanese do not cluster tightly, consistent with gene flow from the mainland in the last few generations[19]. However, the Onge cluster tightly, making them more useful for studying the relationship of the indigenous Andamanese to groups worldwide (Supplementary Note 1). We treat the Chenchu as a sixth outlier group because of their high minimum F_{ST} of 0.052 from all other groups (Supplementary Table 3).

The average pairwise F_{ST} of the remaining 19 groups is 0.0109. This is much larger than the 0.0033 in a recent study of 23 European groups[20], although a strict comparison is difficult, as European studies have focused on cosmopolitan samples[20,21], which could underestimate differentiation relative to our village-centred sampling. We considered the possibility that the high F_{ST} could be an artefact due to marriage between close relatives, which is known to be common in southern India[22], and which can exaggerate measurements of frequency differentiation. However, when we recalculated F_{ST} correcting for consanguinity[23] (see Appendix in Supplementary Information), the average differentiation decreased only marginally to 0.0100. We also determined that the high F_{ST} was not due to our strategy of sampling diverse groups. Restricting to the nine pairs of groups that were from the same state and traditional caste level, the average inbreeding-corrected F_{ST} was 0.0069; much higher than the analogous 0.0018 in Europe when comparing within regions (Supplementary Table 3).

We propose that the high F_{ST} among Indian groups could be explained if many groups were founded by a few individuals, followed by limited gene flow[8,24]. This hypothesis predicts that within groups, pairs of individuals will tend to have substantial stretches of the genome in which they share at least one allele at each SNP. We find signals of excess allele sharing in many groups (Supplementary Fig. 2), which as expected tend to occur in the groups that have the highest F_{ST} values from all others ($P = 0.002$ for a correlation). To estimate the age of founder events, we measured the genetic distance scale over which allele-sharing decays, and verified the robustness of our procedure by simulation (Supplementary Fig. 3). Six Indo-European- and Dravidian-speaking groups have evidence of founder events dating to more than 30 generations ago (Supplementary Fig. 2), including the Vysya at more than 100 generations ago (Fig. 2). Strong endogamy must have applied since then (average gene flow less than 1 in 30 per generation) to prevent

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [3]Centre for Cellular and Molecular Biology, Hyderabad 500 007, India. [4]Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

Table 1 | 25 groups sampled from 13 states of India

Group	Samples	Language family	Traditional caste or social designation	Sampling location				Min F_{ST} to others	
				State/territory	Nearest large town, city or island	Latitude/longitude	Census size†	Uncorrected	Inbreeding corrected
Kashmiri Pandit	5	Indo-European	Upper caste	Kashmir	Dras	34°22′N/75°50′E	7,000	0.0005	0.0023
Vaish	4	Indo-European	Upper caste	Uttar Pradesh	Jaunpur	25°46′N/82°44′E	25,000,000	0.0005	0.0020
Srivastava	2	Indo-European	Upper caste	Uttar Pradesh	Mirzapur	25°10′N/82°37′E	10,000,000	0.0029	0.0023
Sahariya	4	Indo-European	Lower caste	Uttar Pradesh	Allahabad	25°28′N/81°54′E	41,000[a]	0.0089	0.0087
Lodi	5	Indo-European	Lower caste	Uttar Pradesh	Jhansi	26°45′N/83°24′E	57,000	0.0029	0.0028
Satnami	4	Indo-European	Lower caste	Chhattisgarh	Raipur	20°29′N/85°58′E	4,200,000	0.0038	0.0039
Bhil	7	Indo-European	Tribal	Gujarat	Ahmedabad	23°02′N/72°40′E	7,400,000[a]	0.0022	0.0027
Tharu	9	Indo-European	Tribal	Uttarkhand	Nainital	29°23′N/79°30′E	96,000[a]	0.0009	0.0017
Meghawal	5	Indo-European	Lower caste	Rajasthan	Jodhpur	26°18′N/73°04′E	890,000	0.0034	0.0048
Vysya	5	Dravidian	Middle caste	Andhra Pradesh	Anantapur	14°41′N/77°39′E	3,200,000	0.0108	0.0087
Naidu	4	Dravidian	Upper caste	Andhra Pradesh	Chittoor	13°13′N/79°06′E	19,000,000	0.0052	0.0022
Velama	4	Dravidian	Upper caste	Andhra Pradesh	Mahboob Nagar	16°31′N/75°51′E	13,000,000	0.0078	0.0038
Madiga	4	Dravidian	Lower caste	Andhra Pradesh	Warangal	17°58′N/79°35′E	1,600,000[b]	0.0038	0.0028
Mala	3	Dravidian	Lower caste	Andhra Pradesh	Hyderabad	17°22′N/78°29′E	2,900,000[b]	0.0038	0.0030
Kamsali	4	Dravidian	Lower caste	Andhra Pradesh	Kurnool	15°49′N/78°02′E	5,100,000	0.0055	0.0022
Chenchu	6	Dravidian	Tribal	Andhra Pradesh	Anantapur	17°22′N/78°28′E	28,000[a]	0.0524	0.0536
Kurumba	9	Dravidian	Tribal	Kerala	Palakkad	10°54′N/76°27′E	1,300[a]	0.0021	0.0017
Hallaki	7	Dravidian	Tribal	Karnataka	Uttara Kannada	13°55′N/74°09′E	75,000	0.0072	0.0045
Santhal	7	Austro-Asiatic	Tribal	Jharkhand	Santhal Pargana	24°30′N/87°30′E	2,100,000[a]	0.0045	0.0057
Kharia	6	Austro-Asiatic	Tribal	Madhya Pradesh	Raigarh	23°08′N/73°07′E	6,900[a]	0.0045	0.0057
Nyshi	4	Tibeto-Burman	Tribal	Arunachal Pradesh	Papum Pare	26°55′N/92°40′E	56,000[a]	0.0215	0.0198
Ao Naga	4	Tibeto-Burman	Tribal	Nagaland	Kohima	25°40′N/94°08′E	105,000[a]	0.0215	0.0198
Siddi	4	Dravidian*	Tribal	Karnataka	Dharwand	15°27′N/75°05′E	25,000	0.0746	0.0757
Onge	9	Jarawa-Onge	Hunter gatherer	Andaman & Nicobar	Little Andaman	10°30′N/92°30′E	97[a]	0.0905	0.0934
Gr. Andamanese	7	Andamanese	Hunter gatherer	Andaman & Nicobar	Great Andaman	12°12′N/93°00′E	42[a]	0.0386	0.0414

* The language of the Siddi is Dravidian, but their ancestors spoke a Bantu language.
† Census estimates correspond to all of India. Numbers are based on: [a]ref. 50, and [b]ref. 51. For some groups (without a superscript) we obtained estimates from the Census of India 1991, Registrar General Office, Government of India.

the genetic signatures of founder events from being erased by gene flow. Some historians have argued that 'caste' in modern India is an 'invention' of colonialism[25] in the sense that it became more rigid under colonial rule[26]. However, our results indicate that many current distinctions among groups are ancient and that strong endogamy must have shaped marriage patterns in India for thousands of years[24,27].

Medical implications

The high frequency differentiation among Indian groups is medically significant as it shows that 'population stratification' (systematic ancestry differences between cases and controls that can lead to false-positive disease associations) may be a confounder in gene-mapping studies. This is superficially at odds with a recent report that in Indian Americans, allele frequency differentiation is lower than among Europeans[11]. A potential explanation for the discrepancy is that the previous study pooled samples by state of origin, which can mask substructure. For example, when we performed PCA on an independent set of 85 Gujarati Americans[28], we found that they separate into two distinct clusters with high differentiation ($F_{ST} = 0.005$) (Supplementary Fig. 4). Similarly, pairs of Uttar Pradesh and Andhra

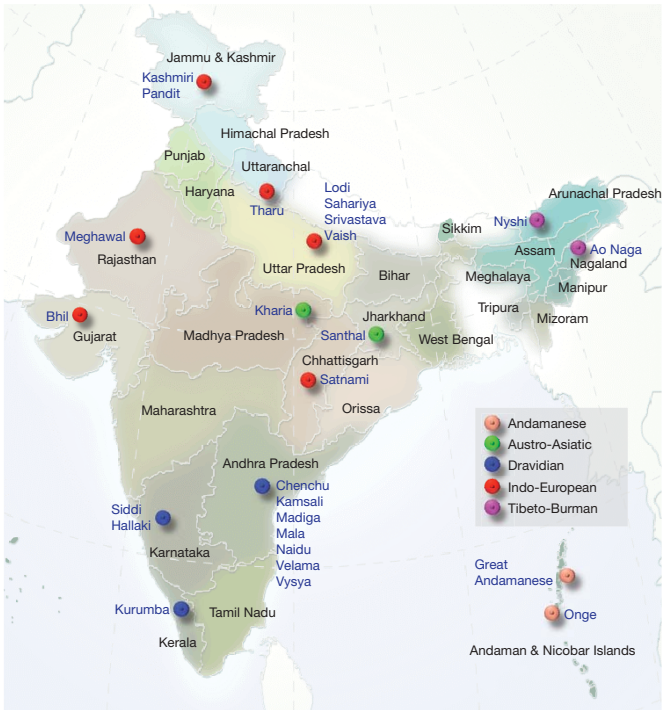

Figure 1 | Map of India. A map of India is shown with the state of origin of the 25 groups that we studied.

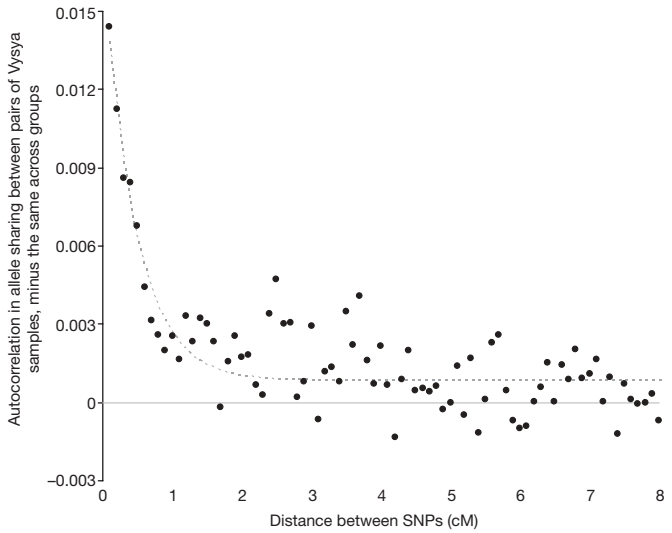

Figure 2 | Linkage disequilibrium based evidence for founder events in India. For each pair of samples, we calculate the autocorrelation of the number of shared alleles as a function of distance, recognizing that SNP genotypes should differ by at most one allele in regions of identity by descent. To correct for background allele sharing, we subtract the same quantity comparing across groups. Allele sharing in the Vysya decreases with an exponential decay of 0.461 cM as shown here, suggesting a founder event roughly $100/(2 \times 0.461) = 108$ generations ago. We present similar analyses for all Indian groups in Supplementary Fig. 2.

Pradesh groups in our data (excluding the outlying Chenchu) have an average F_{ST} of 0.0107, but their differentiation decreases to 0.0033 when we first pool by state. It was recently suggested that to correct for stratification in India, it may be adequate to adjust for membership in five broad genetic clusters[10]. However, our results show that many Indian groups have a degree of allele frequency differentiation from their neighbours that is at least as large as that between northern and southern Europeans, which is known to be sufficient to cause false-positive associations to disease if uncorrected[29].

The widespread history of founder events in India is also medically significant because it predicts a high rate of recessive disease. In Finland, there is a high rate of recessive diseases that has been shown to be due to a founder event, and that has resulted in a minimum F_{ST} of 0.005 with other European groups[20]. Our data show that many Indian groups have a minimum F_{ST} with all other groups that is at least as large (Table 1). Haldane wrote decades ago that ''if inter-caste marriages in India become common, various… recessive characters will become rarer''[30]. However, it has not been generally appreciated that this applies to groups throughout India, and not only to groups in which consanguinity is common[22]. We propose that founder effects are responsible for an even higher burden of recessive diseases in India than consanguinity. To test this hypothesis, we used our data to estimate the probability that two alleles from a group share a common ancestor more recently than that group's divergence from other Indians, and compared this to the probability that an individual's two alleles share an ancestor in the last few generations owing to consanguinity[23]. Nine of the 15 Indian groups for which we could make this assessment had a higher predicted rate of recessive disease owing to founder events than to consanguinity, including all the Indo-European-speaking groups (Table 2). These results highlight the value of systematically surveying Indian groups to identify those with the strongest founder effects, and prioritizing them for studies to identify recessive diseases and map genes.

A further reason why some diseases are expected to occur at increased frequencies in India is shared descent from a common Indian ancestral population[10]. An example is a 25-base-pair deletion in *MYBPC3* that increases heart failure risk by about sevenfold, and occurs at around 4% throughout India but is nearly absent elsewhere[31]. It has recently been shown that the power to discover disease risk variants can be increased by modelling Indian genetic variation using a reference panel of European and Chinese chromosomes[32]. However, the example of *MYBPC3* shows that this is an imperfect solution, because clinically significant alleles that are rare outside of India cannot be imputed by studying non-Indian genetic variation. It is important to specifically characterize Indian variation to permit full-powered gene mapping in India.

Population mixture in Indian history

To better understand the genetic ancestry that is only found in India, we carried out a PCA of Europeans (CEU) and Chinese (CHB) along with 22 Indian groups (Fig. 3). The first principal component distinguishes CEU from CHB, and the second reflects ancestry that is unique to India. The most remarkable feature of the PCA is a gradient of proximity to western Eurasians (Supplementary Fig. 5) (an analogous PCA in Europeans did not produce a gradient of proximity to India; Supplementary Fig. 6). We call this the 'Indian Cline', and propose that it reflects the fact that different Indian groups have inherited different proportions of ancestry from the 'Ancestral North Indians' (ANI) who are related to western Eurasians, and the 'Ancestral South Indians' (ASI). To model ANI–ASI mixture, we selected a subset of 18 groups that formed tight clusters along the Indian Cline, and included the Pathan and Sindhi from Pakistan[14] because they were consistent with the Indian Cline in the PCA but showed greater proximity to western Eurasians (Supplementary Note 2), providing more information about ANI–ASI mixture.

To test whether any of the 18 Indian Cline groups were consistent with all ANI or all ASI ancestry, we applied a new 3 Population Test (Methods). If group X is related to groups Y and W by a simple tree (through a history of divergence without subsequent mixture) then if we define the SNP allele frequencies as p_X, p_Y and p_Z, the quantity $(p_X - p_Y)(p_X - p_W)$ averaged over SNPs, should be proportional to the variance in allele frequency since group X split from Y and Z, and thus should be positive. However, this quantity can be negative if X descends from a mixture event (Supplementary Note 3 and Appendix in Supplementary Information). We applied this test to each of the 18 Indian Cline groups in turn using CEU = Y and Santhal = W, and obtained significantly negative scores for 16 groups (Table 2) as assessed by a Block Jackknife analysis that corrects for linkage

Table 2 | Detection and quantification of population mixture along the Indian Cline

Indian Cline group	Samples	*Z*-score from 3 Population Test for mixture	ANI ancestry (% ±1 s.e.)	Genetic drift *D* from the best fitting combination of ANI and ASI*	Wright's fixation index *F* (estimates inbreeding)†	Estimated fraction of recessive diseases due to founder events‡ (%)
Mala	3	−2.5	38.8 ± 1.2	0.0023	0	100
Madiga	4	−2.7	40.6 ± 1.2	0.0018	0.0061	23
Chenchu	6	31.3 (n.s.)	40.7 ± 1.3	0.0492	0	100
Bhil	7	−10.6	42.9 ± 1.1	0.0024	0	100
Satnami	3	−5.6	43.0 ± 1.3	0.0019	0	100
Kurumba	6	−12.6	43.2 ± 1.1	0.0001	0.0052	2
Kamsali	3	−6.5	44.5 ± 1.3	0.0016	0.0066	19
Vysya	5	5.4 (n.s.)	46.2 ± 1.2	0.0083	0.0071	54
Lodi	5	−8.9	49.9 ± 1.1	0.0027	0.0056	32
Naidu	4	−3.3	50.1 ± 1.2	0.0022	0.0435	5
Tharu	5	−20.6	51.0 ± 1.2	0.0000	0	NA
Velama	4	−3.2	54.7 ± 1.3	0.0044	0.0197	18
Srivastava	2	−7.5	56.4 ± 1.5	0.0023	0	100
Meghawal	5	−13.3	60.3 ± 1.2	0.0035	0	100
Vaish	4	−22.0	62.6 ± 1.2	0.0012	0	100
Kashmiri Pandit	5	−20.6	70.6 ± 1.2	0.0019	0	100
Sindhi	10	−26.3	73.7 ± 1.1	0.0008	0.0043	16
Pathan	15	−34.3	76.9 ± 1.1	0.0001	0.0039	3

NA, not applicable; n.s., not significant.

* Estimates of genetic drift (the variance in allele frequencies on any lineage) are based on a model in which each group is a simple mixture of ANI and ASI, followed by subsequent genetic drift specific to that group (corrected for inbreeding). To fit the model, we use the algorithm described in Supplementary Note 4, and fit f_2, f_3 and f_4 statistics that are calculated in a way that is unbiased by inbreeding (see Appendix in Supplementary Information).

† Wright's fixation index *F* is estimated as the excess rate at which the two copies of a chromosome within an individual from a group are identical by state, compared within across individuals from that group (see Appendix in Supplementary Information). We set negative values to 0; standard errors are typically ~0.003. Owing to the small sample sizes, these estimates are heavily influenced by the samples that happen to have been included in our analysis, and thus should be considered approximate.

‡ To estimate the proportion of recessive disease cases that are due to founder events, we consider the two alleles that a single individual carries at any locus. With probability *F* given by Wright's Fixation Index, they coalesce in the last few generations owing to consanguinity, and with probability $D(1-F)$, they coalesced since ANI–ASI mixture owing to founder events specific to that group. The fraction of recessive diseases due to founder events can thus be estimated as $D(1-F)/(F+D(1-F))$.

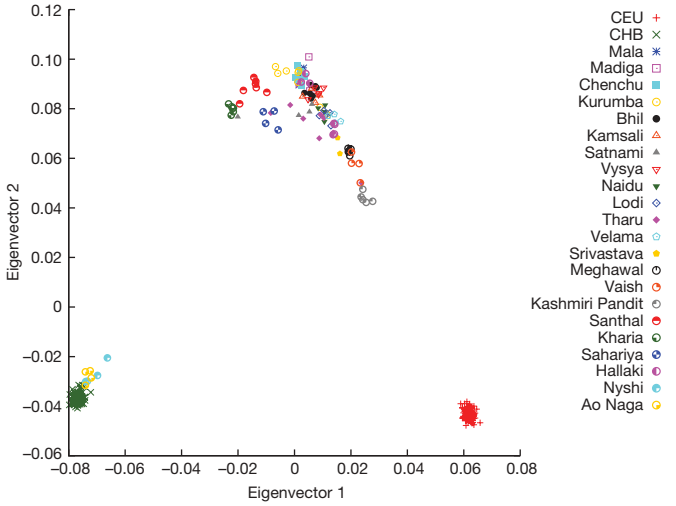

Figure 3 | PCA of 22 groups from the Indian subcontinent. Analysis of these groups along with Europeans (CEU) and Chinese (CHB) shows a gradient of relatedness to CEU that runs through most Indo-European and Dravidian groups, with the Kashmiri Pandit most related to CEU. Both the Austro-Asiatic speaking groups (Kharia and Santhal) and the tribal Sahariya are off-cline, whereas the two Tibeto-Burman speaking groups cluster with CHB. (Data from the outlying Siddi, Onge and Great Andamanese are not shown.)

disequilibrium among SNPs[33] (Methods). These results do not mean that the Indian groups descend from mixtures of European and Austro-Asiatic speakers, but only that they derive from at least two different groups that are (distantly) related to CEU and Santhal.

We verified the evidence of mixture by carrying out a 4 Population Test[34]. For any four groups there are three possible simple trees. If ((A,B),(C,D)) is correct, the allele frequency differences between A and B should be uncorrelated with those between C and D, which we can assess by averaging the quantity $(p_A - p_B)(p_C - p_D)$ across SNPs (see Appendix in Supplementary Information) and testing for consistency with 0 (Methods). No Indian Cline group could be related simply to CEU, Onge and West Africans (YRI) after testing all trees (Supplementary Table 4).

Relationship of Indians to non-Indians

We developed a model to study the historical relationship of Indian groups to those worldwide, on the basis of the hypothesis that most groups can be approximated as a mixture of two ancestral populations followed by group-specific drift. To fit the model to the data, we computed the squared allele frequency difference between all pairs of groups, and chose parameters by minimizing the difference between observation and expectation (Supplementary Note 4). The idea of fitting allele frequency differentiation to historical models was first explored by Cavalli-Sforza and Edwards[35], and here we extend it to trees with mixture. This approach contrasts with the STRUCTURE algorithm, which fits data without a tree[36], or a tree in which many groups split simultaneously from an ancestral population followed by mixture[37]. Although STRUCTURE is accurate for estimating individual mixture proportions in recently mixed groups, it is not clear whether its estimates of ancient mixture are biased because it does not model hierarchical relationships among groups, which could lead to inaccurate estimates of allele frequencies in ancestral populations. In contrast, we use a more realistic tree model, and provide a test of fit.

Applying our model-fitting procedure, we find that the tree (YRI,((CEU,ANI),(ASI, Onge))) provides an excellent fit to the data from Indian groups. In particular, when the Pathan, Vaish, Meghawal and Bhil are modelled as mixtures of ANI and ASI (Fig. 4), the observed allele frequency differentiation statistics are all consistent with the theoretical expectation within three standard deviations (Supplementary Note 4).

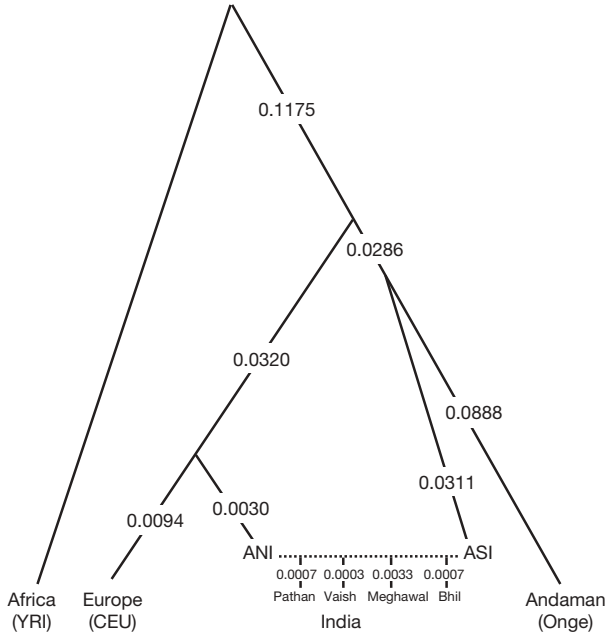

Figure 4 | A model relating the history of Indian and non-Indian groups. Modelling the Pathan, Vaish, Meghawal and Bhil as mixtures of ANI and ASI, and relating them to non-Indians by the phylogenetic tree (YRI,((CEU,ANI),(ASI, Onge))), provides an excellent fit to the data. Although the model is precise about tree topology and ordering of splits, it provides no information about population size changes or the timings of events. We estimate genetic drift on each lineage in the sense of variance in allele frequencies, which we rescale to be comparable to F_{ST} (standard errors are typically ± 0.001 but are not shown).

Two features of the inferred history are of special interest. First, the ANI and CEU form a clade, and further analysis shows that the Adygei, a Caucasian group, are an outgroup (Supplementary Note 4). Many Indian and European groups speak Indo-European languages, whereas the Adygei speak a Northwest Caucasian language. It is tempting to assume that the population ancestral to ANI and CEU spoke 'Proto-Indo-European', which has been reconstructed as ancestral to both Sanskrit and European languages[38], although we cannot be certain without a date for ANI–ASI mixture.

Second, our analysis shows that the Onge form a clade with the ASI (Supplementary Note 4), which we verified by running the 4 Population Test on ((YRI,Papuan)(Dai,X)), and finding that it is consistent when X = Onge ($Z = 1.7$) but inconsistent for all Indian Cline groups ($Z \ll -9$) (Supplementary Table 4). Previous mtDNA analyses suggested that the Onge do not share any maternal ancestry with groups outside India within the last ~48,000 years[19,39]. Although the Onge do share ancestry with some rare haplogroups in some Indian tribal populations within the last ~24,000 years[39,40], this observation is consistent with our inferred Onge–ASI clade, as long as the gene flow predated the ASI–ANI mixture that later occurred on the mainland.

We warn that 'models' in population genetics should be treated with caution. Although they provide an important framework for testing historical hypotheses, they are oversimplifications. For example, the true ancestral populations of India were probably not homogeneous as we assume in our model, but instead were probably formed by clusters of related groups that mixed at different times. However, modelling them as homogeneous fits the data and seems to capture meaningful features of history.

Estimates of mixture proportions in India

Estimating the proportions of ANI and ASI ancestry in India is challenging, because we are unaware of any published methods that produce unbiased estimates of mixture proportion in the absence

of accurate ancestral groups. We developed three methods for estimating ancestry, which we verified were accurate even in the face of SNP ascertainment bias and some inaccuracies in our phylogenetic model (Supplementary Note 5), and which we found provided consistent estimates (Supplementary Table 5). The 18 Indian Cline groups all have between 39% and 77% ANI ancestry on the basis of f_3 Ancestry Estimates (Methods), which we quote because it has the smallest standard errors (Table 2). ANI ancestry is significantly higher in Indo-European than Dravidian speakers ($P = 0.013$ by a one-sided test)[5–8,41], suggesting that the ancestral ASI may have spoken a Dravidian language before mixing with the ANI[42]. We also find significantly more ANI ancestry in traditionally upper than in lower or middle caste groups ($P = 0.0025$)[5–8,41], and find that traditional caste level is significantly correlated to ANI ancestry even after controlling for language ($P = 0.0048$), suggesting a relationship between the history of caste formation in India and ANI–ASI mixture.

We compared our autosomal estimates of ANI ancestry to Y chromosome and mtDNA haplogroup frequencies. Y chromosome analysis has shown that traditionally upper caste and Indo-European speaking groups have increased frequencies of alleles that are also common in western Eurasians[5,6]. However, mtDNA analysis has shown increased frequencies of haplogroups common in western Eurasians only in northwest India[7,8,43]. Comparing the autosomal estimates of ANI ancestry to the frequencies of haplogroups characteristic of western Eurasians, we find a significant correlation on the Y chromosome ($P = 0.04$) and a more marginal correlation in mtDNA ($P = 0.08$) (Supplementary Table 6 and Supplementary Fig. 7). The stronger gradient in males, replicating previous reports, could reflect either male gene flow from groups with more ANI relatedness into ones with less, or female gene flow in the reverse direction. The latter hypothesis is unlikely, because extensive female gene flow in India would be expected to homogenize ANI ancestry on the autosomes just as in mtDNA, which we do not observe. Supporting the view of little female ANI ancestry in India, it has been reported that mtDNA 'haplogroup U' splits into two deep clades[44]. 'U2i' accounts for 77% of copies in India but about 0% in Europe, and 'U2e' accounts for 0% of all copies in India but about 10% in Europe. The split is estimated to have occurred about 50,000 years ago, indicating low female gene flow between Europe and India since that time.

Discussion

We have documented a high level of population substructure in India, and have shown that the model of mixture between two ancestral populations, ASI and ANI, provides an excellent description of genetic variation in many Indian groups. A priority for future work should be to estimate a date for the mixture, which may be possible by studying the length of stretches of ANI ancestry in Indian samples[45,46], and will shed light on the process leading to the present structure of Indian groups. A second priority should be to discern the details of the history of the ANI and ASI before they mixed, including the date of their separation and their history of expansion and contraction. This may be possible by analysing allele frequency spectrum[47] and linkage disequilibrium data[45,48,49]. Our findings finally have medical implications. By showing that a large proportion of Indian groups descend from strong founder events, these results highlight the importance of identifying recessive diseases in these groups and mapping causal genes.

METHODS SUMMARY

Blood samples were collected with informed consent from volunteers. We designate groups by their anthropological name as well as their geographic location, as it has been shown that both are required to specify an effectively endogamous group in India[1]. All DNA samples were genotyped on Affymetrix 6.0 arrays. We restricted most analyses to samples that had no evidence for genetic relatedness and to 560,123 autosomal SNPs for which there were no signs of problematic genotyping and for which the data were relatively complete. For some analyses we also intersected our data with Illumina 650Y genotyping of the Human Genome Diversity Panel[14] and HapMap[13,28], which produced a merged data set of 119,744 autosomal SNPs[14]. We carried out PCA using the EIGENSOFT software[17], assessed allele frequency differentiation among groups using F_{ST}, assessed inbreeding in each group using Wright's Fixation Index F[23], and computed standard errors using a Block Jackknife[33]. To detect the signature of founder events in linkage disequilibrium data, we studied all possible pairs of samples for each group, and recorded whether they share 0, 1 or 2 alleles at each SNP (at SNPs in which both individuals were heterozygous, we recorded 1 allele to be shared to account for the ambiguity in the haplotype phase). Long stretches of allele sharing can reflect regions that are shared identical by descent from a common founder, and by measuring the exponential decay of allele sharing with distance, we inferred the age of the founder event (Supplementary Fig. 3). To test for a history of mixture, we applied 3 and 4 Population Tests (Supplementary Note 3). To infer the proportion of ancestry in each Indian Cline group in the absence of accurate ancestral populations, we used f_3 Ancestry Estimation (Supplementary Note 5).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 April; accepted 5 August 2009.

1. Majumdar, D. N. & Rao, C. R. *Race Elements in Bengal: a Quantitative Study* (Asia Publishing House, 1960).
2. Roychoudhury, A. K. & Nei, M. Genetic relationships between Indians and their neighboring populations. *Hum. Hered.* **35,** 201–206 (1985).
3. Das, B. M., Das, P. B., Das, R., Walter, H. & Danker-Hopfe, H. Anthropological studies in Assam, India. *Anthropol. Anz.* **44,** 239–248 (1986).
4. Zerjal, T. *et al.* Y-chromosomal insights into the genetic impact of the caste system in India. *Hum. Genet.* **121,** 137–144 (2007).
5. Bamshad, M. *et al.* Genetic evidence on the origins of Indian caste populations. *Genome Res.* **11,** 994–1004 (2001).
6. Wells, R. S. *et al.* The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc. Natl Acad. Sci. USA* **98,** 10244–10249 (2001).
7. Thanseem I. *et al.* Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. *BMC Genet.* **7,** 42 (2006).
8. Basu, A. *et al.* Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.* **13,** 2277–2290 (2003).
9. Thangaraj, K. *et al. In situ* origin of deep rooting lineages of mitochondrial Macrohaplogroup 'M' in India. *BMC Genomics* **7,** 151 (2006).
10. Indian Genome Variation Consortium. Genetic landscape of the people of India: a canvas for disease gene exploration. *J. Genet.* **87,** 3–20 (2008).
11. Rosenberg, N. A. *et al.* Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet.* **2,** e215 (2006).
12. Abbi, A. Is Great Andamanese genealogically and typologically distinct from Onge and Jarawa? *Language Sciences* doi:10.1016/j.langsci.2008.02.002 (22 April 2008).
13. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
14. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100–1104 (2008).
15. Jakobsson, M. *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451,** 998–1003 (2008).
16. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201,** 786–792 (1978).
17. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
18. Thangaraj, K., Ramana, G. V. & Singh, L. Y-chromosome and mitochondrial DNA polymorphisms in Indian populations. *Electrophoresis* **20,** 1743–1747 (1999).
19. Thangaraj, K. *et al.* Genetic affinities of the Andaman Islanders, a vanishing human population. *Curr. Biol.* **13,** 86–93 (2003).
20. Lao, O. *et al.* Correlation between genetic and geographic structure in Europe. *Curr. Biol.* **18,** 1241–1248 (2008).
21. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).
22. Dronamraju, K. R. Mating systems of the Andhra Pradesh people. *Cold Spring Harb. Symp. Quant. Biol.* **29,** 81–84 (1964).
23. Nei, M. & Chesser, R. K. Estimation of fixation indices and gene diversities. *Ann. Hum. Genet.* **47,** 253–259 (1983).
24. Karve, I. *Hindu Society—an Interpretation* (S. R. Deshmukh, 1968).
25. Boivin, N. in *The Evolution and History of Human Populations in South Asia* (eds Petraglia, M. D. & Allchin, B.) 341–362 (Springer, 2007).
26. Dirks, N. B. *Castes of Mind: Colonialism and the Making of Modern India* (Princeton Univ. Press, 2001).
27. Bhasin, M. K. & Walter, H. *Genetics of Castes and Tribes of India* (Kamla-Raj Enterprises, 2001).
28. Index of /genotypes/2008-07_phaseIII. ⟨http://ftp.hapmap.org/genotypes/2008-07_phaseIII/⟩.
29. Campbell, C. D. *et al.* Demonstrating stratification in a European American population. *Nature Genet.* **37,** 868–872 (2005).

30. Haldane, J. B. S. A defense of beanbag genetics. *Perspect. Biol. Med.* **7,** 343–359 (1964).
31. Dhandapany, P. S. *et al.* A common Cardiac Myosin Binding Protein C variant associated with cardiomyopathies in South Asia. *Nature Genet.* **41,** 187–191 (2009).
32. Pemberton, T. J. *et al.* Using population mixtures to optimize the utility of genomic databases: linkage disequilibrium and association study design in India. *Ann. Hum. Genet.* **72,** 535–546 (2008).
33. Künsch, H. R. The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* **17,** 1217–1241 (1989).
34. Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genet.* **39,** 1251–1255 (2007).
35. Cavalli-Sforza, L. L. & Edwards, A. W. Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.* **19,** 233–257 (1967).
36. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
37. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164,** 1567–1587 (2003).
38. Mallory, J. P. & Adams, D. O. *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World* (Oxford Univ. Press, 2006).
39. Barik, S. S. *et al.* Detailed mtDNA genotypes permit a reassessment of the settlement and population structure of the Andaman Islands. *Am. J. Phys. Anthropol.* **136,** 19–27 (2008).
40. Palanichamy, M. G. *et al.* Comment on ''Reconstructing the Origin of Andaman Islanders''. *Science* **311,** 470 (2006).
41. Watkins, W. S. *et al.* Genetic variation in South Indian castes: evidence from Y-chromosome, mitochondrial, and autosomal polymorphisms. *BMC Genet.* **9,** 86 (2008).
42. Southworth, F. C. *Linguistic archaeology of South Asia* (Routledge-Curzon, 2005).
43. Cordaux, R. *et al.* Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. *Eur. J. Hum. Genet.* **11,** 253–264 (2003).
44. Kivisild, T. *et al.* Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Curr. Biol.* **9,** 1331–1334 (1999).
45. Falush, D. *et al.* Traces of human migrations in *Helicobacter pylori* populations. *Science* **299,** 1582–1585 (2003).
46. Baird, S. J. E. Phylogenetics: Fisher's markers of admixture. *Heredity* **97,** 81–83 (2006).
47. Chikhi, L., Bruford, M. W. & Beaumont, M. A. Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158,** 1347–1362 (2001).
48. Hellenthal, G., Auton, A. & Falush, D. Inferring human colonization history using a copying model. *PLoS Genet.* **4,** e1000078 (2008).
49. Lohmueller, K. E., Bustamante, C. D. & Clark, A. G. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* **182,** 217–231 (2009).
50. Singh, K. S. *People of India, National Series, Volume III, Scheduled Tribes* (Oxford Univ. Press, 1994).
51. Singh, K. S. *People of India, National Series, Volume III, Scheduled Castes* (Oxford Univ. Press, 1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the volunteers from throughout India who donated DNA; A. G. Reddy, A. Shah and R. Tamang for generating the Y chromosome and mtDNA data; J. Neubauer for sample preparation; and A. Tandon for data curation. We thank B. N. Sarkar and A. G. Roy for helping with group census size estimates, and D. Falush, J. Novembre, A. Ruiz-Linares and S. Watkins for comments on the manuscript. D.R., N.P. and A.L.P. were supported by NIH grant HG004168, and D.R. was supported by a Burroughs Wellcome Career Development Award in the Biomedical Sciences. K.T. and L.S. were supported by grants from the Council of Scientific and Industrial Research of the Government of India, and K.T. was supported by a UKIERI Major Award (RG-4772).

Author Contributions K.T. and L.S. collected the DNA samples, D.R., K.T. and L.S. collected the genetic data, N.P. developed the mathematical theory for *f*-statistics, and D.R., K.T., N.P. and A.L.P. analysed the data. D.R. wrote the manuscript and Supplementary Information with input from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.R. (reich@genetics.med.harvard.edu) or L.S. (lalji@ccmb.res.in).

METHODS

Sample collection. Blood samples were collected from volunteers with the help of local administrators, and with informed consent and approval of an Institutional Ethical Committee. The names we use are the ones by which the groups are described anthropologically, but are not unique identifiers. We use 'traditionally upper caste' to designate Brahmin and Kshatriya, 'traditionally middle caste' to refer to Vysya, and 'traditionally lower caste' to refer to Shudra. We use 'tribal' and 'hunter gatherer' to refer to non-caste groups.

Genotyping and data curation. We genotyped samples on Affymetrix 6.0 arrays using standard protocols. We restricted analysis to 560,123 SNPs on the autosomes and 27,630 SNPs on the X chromosome with reliable genotyping across >95% of the samples, and used the Birdsuite software[52] to assign genotypes. We removed ten samples with unusually high relatedness to others as assessed by the rate of genome-wide allele sharing (we included one sample per kinship group). We also intersected our data with HGDP samples genotyped on an Illumina 650Y array[14] and HapMap samples, resulting in 119,744 SNPs on the autosomes and 5,551 SNPs on the X chromosome. As evidence for the usefulness of the merged data set, and the absence of substantial structure in the data related to experimental artefacts, we could not find any PCA that distinguished all the Indians from the HGDP samples.

Statistical methods for analysing population structure. PCA was performed using the EIGENSOFT software[17]. We estimated allele frequency differentiation using F_{ST}, which we computed using a formula that has asymptotically minimal variance (see Appendix in Supplementary Information). We also calculated an inbreeding corrected F_{ST} that is asymptotically consistent in the presence of excess homozygosity (see Appendix in Supplementary Information)[23]. To compute Wright's Fixation Index F[23], an estimate of the inbreeding coefficient for each group, we compared the probability of two alleles being shared identical by state within the same individual, to across individuals from the same group (see Appendix in Supplementary Information).

Block Jackknife procedure to estimate standard errors. To obtain a standard error on F_{ST} as well as the f_2, f_3 and f_4 statistics, we used a Block Jackknife procedure[33]. We divided the genome into contiguous 5 cM chunks and deleted each in turn to quantify the variability of the statistic, which produces a standard error for the value of any estimated quantity. When the null hypothesis indicates that an f-statistic has mean zero as in the 4 Population Test, the jackknife standard error can be converted to a Z-score, which has mean 0 and variance 1 under the null hypothesis. We warn that the normality assumption becomes imperfect for $|Z|>2$ (not shown). Thus, large Z-scores should be viewed as statistically significant but not simply convertible to P-values[53].

Inferring the age of founder events by correlation of allele sharing. For each pair of samples in our data set we record whether they share 0, 1 or 2 alleles at each SNP in the genome. When both individuals are heterozygous we record 1 allele shared (to account for uncertainty about haplotype phase). For each Indian group, we compute the autocorrelation of this allele sharing statistic as a function of distance across all sample pairs, searching for the signature of stretches of allele sharing due to descent from a common founder whose extent reflects the age of the founder event. To correct for background allele sharing inherited from the ancestral populations, we subtract the curve obtained by comparing pairs across groups of similar ANI proportion, choosing from '65 ± 5% ANI' (Meghawal, Vaish and Kashmiri Pandit), '58 ± 5% ANI' (Velama, Srivastava, Meghawal and Vaish), '53% ± 5% ANI' (Lodi, Naidu, Tharu, Velama and Srivastava), '47 ± 5% ANI' (Bhil, Satnami, Kurumba, Kamsali, Vysya, Lodi, Naidu and Tharu) and '42 ± 5% ANI' (Mala, Madiga, Chenchu, Bhil, Satnami, Kurumba, Kamsali and Vysya). To convert the observed allele sharing decay to a date estimate, we perform a least squares fit to an exponential distribution, $y = a + be^{-2Dt}$. Here, t is the inferred number of generations since the founder event under the assumption of a single strong event, D the genetic distance in Morgans between SNPs, and the factor of 2 reflects the fact that a stretch of allele sharing can be broken by recombination on either haplotype.

3 Population Test for mixture. The 3 Population Test is based on an 'f_3 statistic', a 3-population generalization of F_{ST}. This statistic is equal to the inner product of the frequency differences between a group X and two other groups A and B, which we show in Supplementary Note 3 and the Appendix in the Supplementary Information is proportional to the correlated genetic drift between groups A and X, and groups A and B. If X is related in a simple way (without mixture) to an ancestor, we expect this quantity to be positive, because the genetic drift along the lineage leading from the ancestor to X must be positive. In contrast, if group X has arisen from a mixture of groups related to A and B, it can be negative, and thus the observation of a significantly negative value of the f_3 statistic provides an unambiguous signal of mixture.

4 Population Test for mixture. To assess whether an unrooted phylogenetic tree, for example (YRI,Papuan)(Dai,Onge), is consistent with the SNP allele frequency data, we calculate an 'f_4 statistic', which is expected to be proportional to the correlation in allele frequency differences between pairs of groups (see Appendix in Supplementary Information). If the topology (A,B)(C,D) is correct, then the frequency differences between A and B should reflect genetic drift that is uncorrelated with that between C and D. Thus, the expected value of the product of frequency differences is zero. We compute the statistic f_4(A;B;C,D) with a jackknife standard error. We interpret significant deviations of the f_4 statistics from 0 for all three possible topologies as evidence that the four groups cannot be related by a simple phylogeny without mixture.

f_3 Ancestry Estimation. To obtain estimates of ANI ancestry for each Indian Cline group in the absence of accurate ancestral populations, we used f_3 Ancestry Estimation, f_4 Ancestry Estimation and Regression Ancestry Estimation (Supplementary Note 5), which produce consistent results on the Indian Cline groups as shown in Supplementary Table 5. Here we restrict our description to the f_3 Ancestry Estimates, which we use for Table 2 as this method provides the smallest standard errors. To implement f_3 Ancestry Estimation, we model each Indian Cline group as a linear mixture $K = m_k$(ANI) + $(1 - m_k)$ASI, indicating that each has inherited a proportion m_k of ANI ancestry followed by genetic drift. The topology of Fig. 4 suggests that Onge and ASI are a clade, and hence f_3(Adygei;Outgroup,K) = $m_k f_3$(Adygei;Outgroup,ANI) + $(1 - m_k) f_3$(Adygei;Outgroup,ASI) = $m_k f_3$(Adygei;Outgroup,ANI) + $(1 - m_k) f_3$(Adygei;Outgroup,Onge). We thus obtain equations: $y_{K,Outgroup} = (1 - m_k)x_{Outgroup} + (m_k)z$, where $x_{Outgroup} = f_3$ (Adygei;Outgroup,Onge) and $y_{K,Outgroup} = f_3$(Adygei;Outgroup,K), and solve them using nonlinear least squares, fitting the m_k and z for all three outgroups simultaneously (YRI, Papuan and Dai). We explored whether allowing the coefficient z to depend on $x_{Outgroup}$ improves the fit, as might be expected if the three outgroups do not all have the same position in the phylogeny. We found that this did not change the coefficients m_k or produce a significantly better fit, and hence we allow z to be the same for all three outgroups.

52. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genet.* **40**, 1166–1174 (2008).
53. Thorburn, D. On the asymptotic normality of the jackknife. *Scand. J. Stat.* **4**, 113–118 (1977).