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 on the X chromosome with reliable genotyping across >95% of the samples, and used the Birdsuite software[50] to assign genotypes. We removed 10 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 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 artifacts, we could not find any PC that distinguished all the Indians from the HGDP samples.

Statistical methods for analyzing 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 (Appendix). We also calculated an inbreeding corrected F_{ST} that is asymptotically consistent in the presence of excess homozygosity (Appendix)[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 (Appendix).

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 implies 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 caution 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[51].

Inferring the age of founder events via 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, Kashmiri Pandit), "58 ± 5% ANI" (Velama, Srivastava, Meghawal, Vaish), "53% ± 5% ANI" (Lodi, Naidu, Tharu, Velama, Srivastava), "47 ± 5% ANI" (Bhil, Satnami, Kurumba, Kamsali, Vysya, Lodi, Naidu, Tharu) and "42 ± 5% ANI" (Mala, Madiga, Chenchu, Bhil, Satnami, Kurumba, Kamsali, Vysya). To convert the observed allele sharing decay to a date estimate, we perform a least squares fit of 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 Note S3 and the Appendix 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, since the genetic drift along the lineage leading from the ancestor to X must be positive. By 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 (Appendix). 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 4 groups cannot be related via 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* (Note S5), which produce consistent results on the Indian Cline groups as shown in Table S5. 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, implying that each has inherited a proportion m_k of ANI ancestry followed by genetic drift. The topology of Figure 4 implies 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 non-linear 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.

Additional References for Methods

[50] McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 40, 1166-1174.

[51] Thorburn D (1977) On the asymptotic normality of the jackknife. Scandinavian Journal of Statistics 4, 113-118.