A Shift towards Endogamy: Genetic Evidence of Population Mixture in India in the last 4000 years.

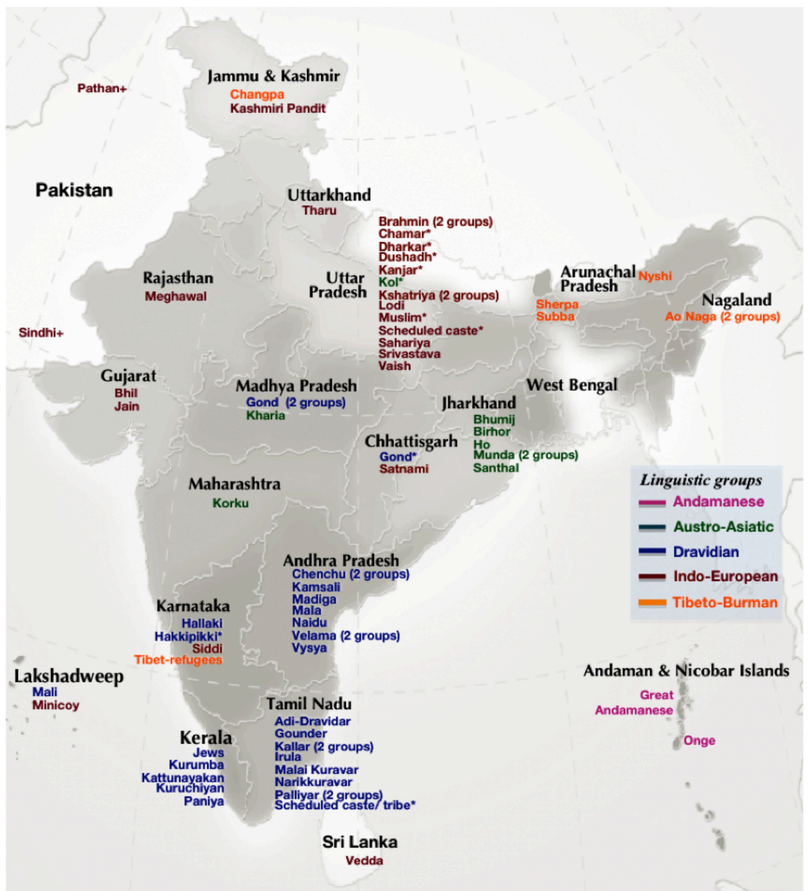
Priya Moorjani[1,2,6+], Kumarasamy Thangaraj[3,+], Nick Patterson[2], Mark Lipson[4], Po-Ru Loh[4], Periasamy Govindaraj[3], Bonnie Berger[4], David Reich[1,2*], Lalji Singh[3,5*]

[1]Department of Genetics, Harvard Medical School, USA, [2]Broad Institute, USA, [3]Centre for Cellular and Molecular Biology, India, [4]Department of Mathematics and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA, [5]Genome Foundation, India, [6]Columbia University, USA, [+] These authors contributed equally, *These authors co-mentored the project

Introduction

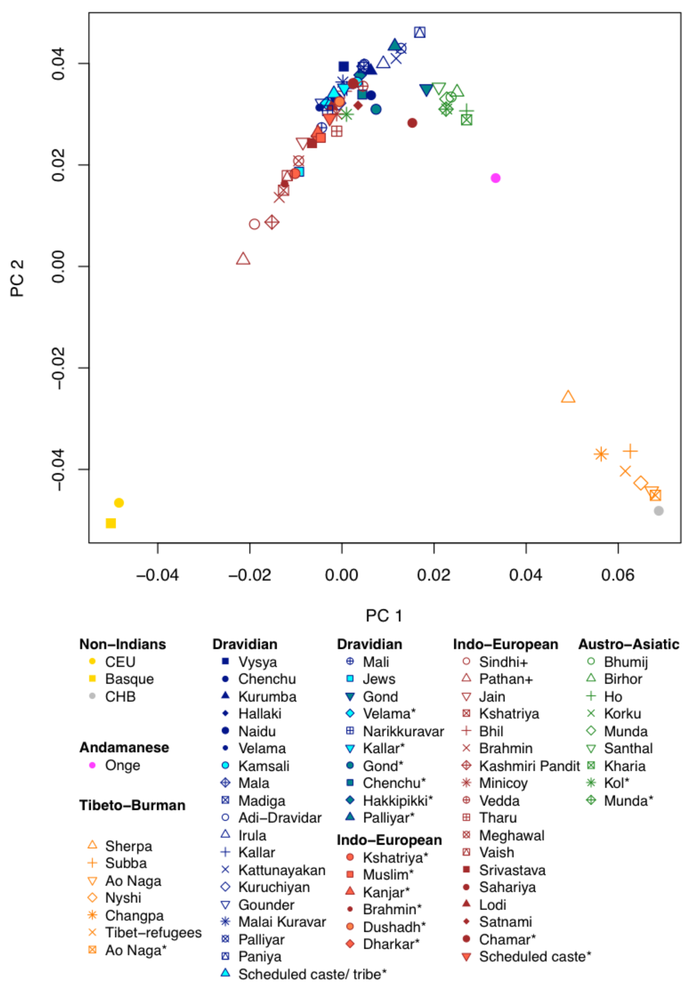
All Indian groups descend from a mixture of two highly divergent populations: Ancestral North Indians (ANI) related to Central Asians, Middle Easterners and Europeans, and Ancestral South Indians (ASI) not closely related to groups outside the subcontinent. The date of mixture is unknown but is central for understanding Indian history. We report genome-wide data from 73 groups from the Indian subcontinent and analyze linkage disequilibrium to estimate ANI-ASI mixture dates of 1,900-4,200 years ago. In at least a subset of groups 100% of the mixture is consistent with having occurred during this period. These results show that India experienced a demographic and cultural transformation several thousand years ago, from a region in which major population mixture was common, to one in which mixture even between closely related groups became rare because of a shift to endogamy.

Sampling locations

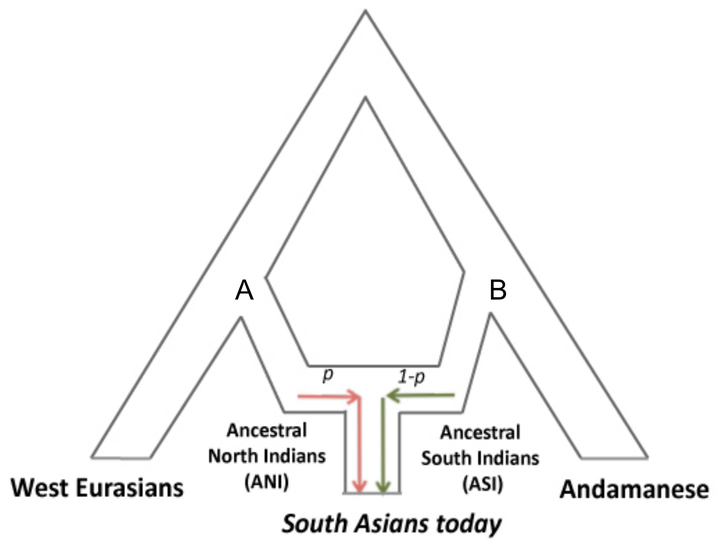


Map of Sampling locations showing the state of origin of the 70 groups of samples included in the study.

Population structure in South Asia



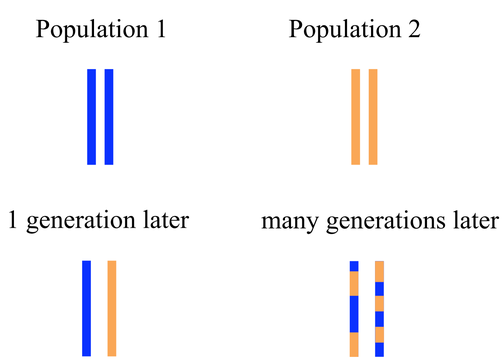
All South Asians are admixed



Model of admixture between Ancestral North Indians (ANI) related to Europeans and Middle Easterners and Ancestral South Indian (ASI) related to Onge (from Andaman and Nicobar islands) provides the best fit to South Asian data.

Estimating dates using weighted LD

Population mixtures create *mosaic* chromosomes



After n generations: chromosomal segments are $\sim 1/n$ cM

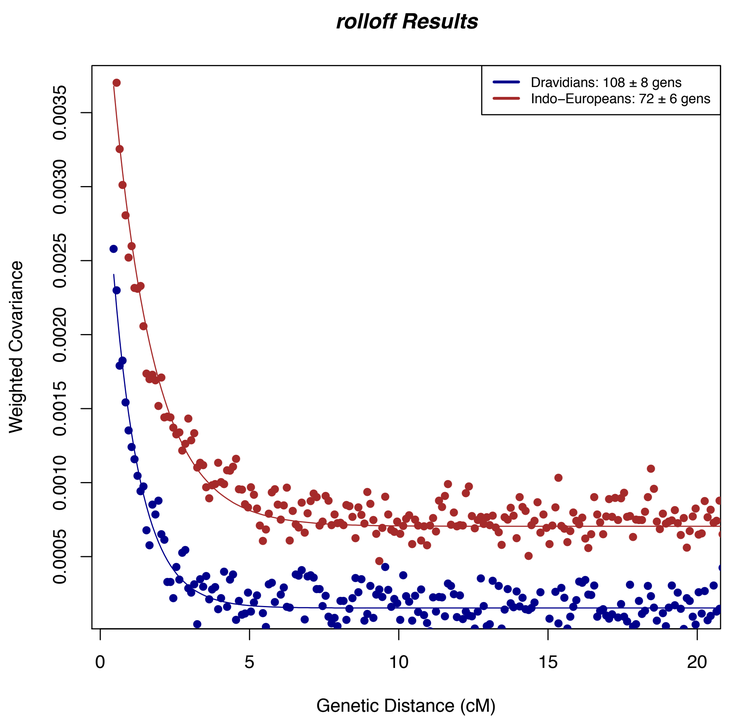
To estimate the date of ANI-ASI mixture, we applied our new method called *ROLLOFF/ALDER* that studies ***admixture related linkage disequilibrium*** (LD) to infer the date of admixture.

Specifically, we compute the statistic -

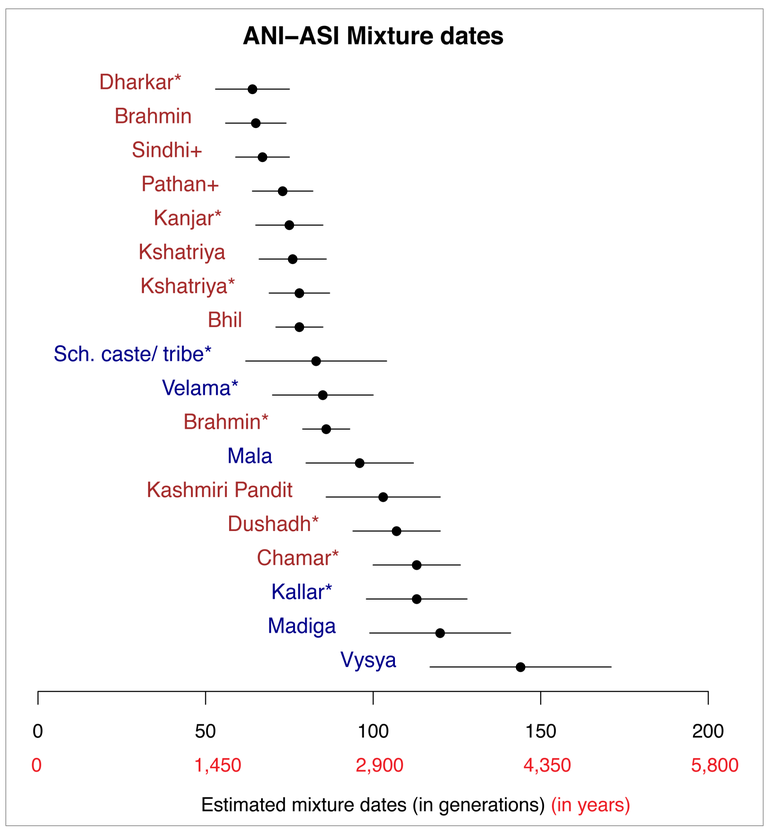
$$R(d) = \frac{\sum_{|x-y|\approx d} z(x,y)w(x,y)}{\sum_{|x-y|\approx d} w(x,y)^2}$$

where $z(x,y)$ = correlation/ covariance between SNPs x, y and $w(x,y)$ = weight of SNPs x and y that can be (a) the allele frequency difference between two groups we use as surrogates for the ancestors; (b) the allele frequency difference between a tested Indian group and one reference; (c) the PCA-based SNP loadings for SNPs (x, y). We plot the weighted covariance with distance and obtain a date by fitting an exponential function: $y = e^{-nd} + c$, where d is the distance in Morgans and we interpret n as the number of generations since admixture.

When did ANI- ASI admixture occur?



Estimated dates of admixture



The number of pulses of gene flow

We used our newly developed method called *qpwave*[5], which estimates the number of pulses of gene flow by computing the rank of the f_4 *relationship matrix.* We set up the matrix as follows –

$$Y(m,n) = f_4(India_{base}, India_{other}; NonIndia_{base}, NonIndia_{other})$$

where $India_{base}$ = fixed Indian group; $India_{other}$ = set of m Indian populations; $NonIndia_{base}$ = Yoruba; $NonIndia_{other}$ = set of n Eurasian populations including Onge, Middle East, Europe, Dai, Papuan and Karitiana.

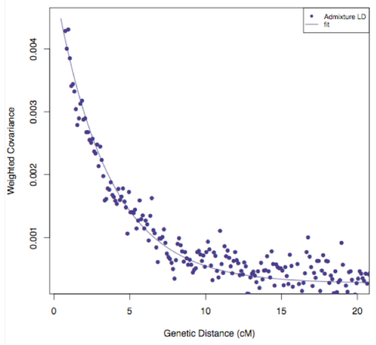
To score the rank of the matrix, we fit $Y(m,n) = A \times B$

where $A = m \times k$ and $B = k \times n$

Testing rank k+1 versus rank k is a standard Likelihood Ratio test, leading to a χ^2 statistic under the null hypothesis that the F_4 matrix has rank k. Coalescent simulations show that this method produces accurate estimates of the number of gene flow events, even when the ancestral populations are closely related to each other and the rank is not affected by drift.

Formal test for single mixture

Admixture LD statistic

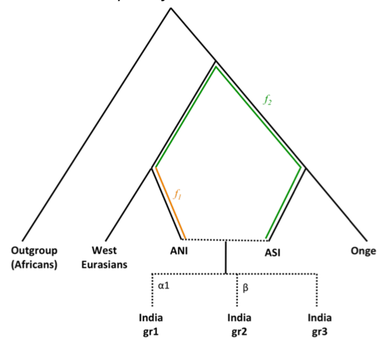


Run *ALDER*, Fit: $y = Ae^{-nd} + c$

↓

Observed amplitude ($a_o = A + c/2$)

Frequency statistic



Run *admixture graph*

↓

Expected amplitude ($a_E = 2\alpha\beta(\alpha f_1 - \beta f_2)^2$)

Group	Observed Amplitude	Expected Amplitude	*Z* score
Simulated from null model with $\alpha = 0.3$	3.3	3.18	0.5
Simulated from alternate model with $\alpha = 0.3$	1.76	3.16	-3.7
Indo-Europeans rank 1	0.6 ± 0.1	0.7 ± 0.2	-0.35
Dravidians rank 1	0.8 ± 0.1	1.1 ± 0.2	-1.06

* Reject null model of single wave if $|Z| > 3$

References

1. Reich, D. *et al.* Reconstructing Indian population history. *Nature* 461, 489-494 (2009)
2. Patterson *et al.* Population Structure and Eigenanalysis. *PLoS Genet* 2:e190 (2006)
3. Moorjani et al. History of African Gene Flow into Southern Europeans, Levantines and Jews. *PLoS Genet* 7:e1001373 (2010)
4. Fenner. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropoly* 128:415 (2005)
5. Reich, D. *et al.* Reconstructing Native American population history, *Nature* (2009)
6. Loh, PR. *et al.* Inferring admixture histories of human populations using linkage disequilibrium, *Genetics* 193, 1233-1254.