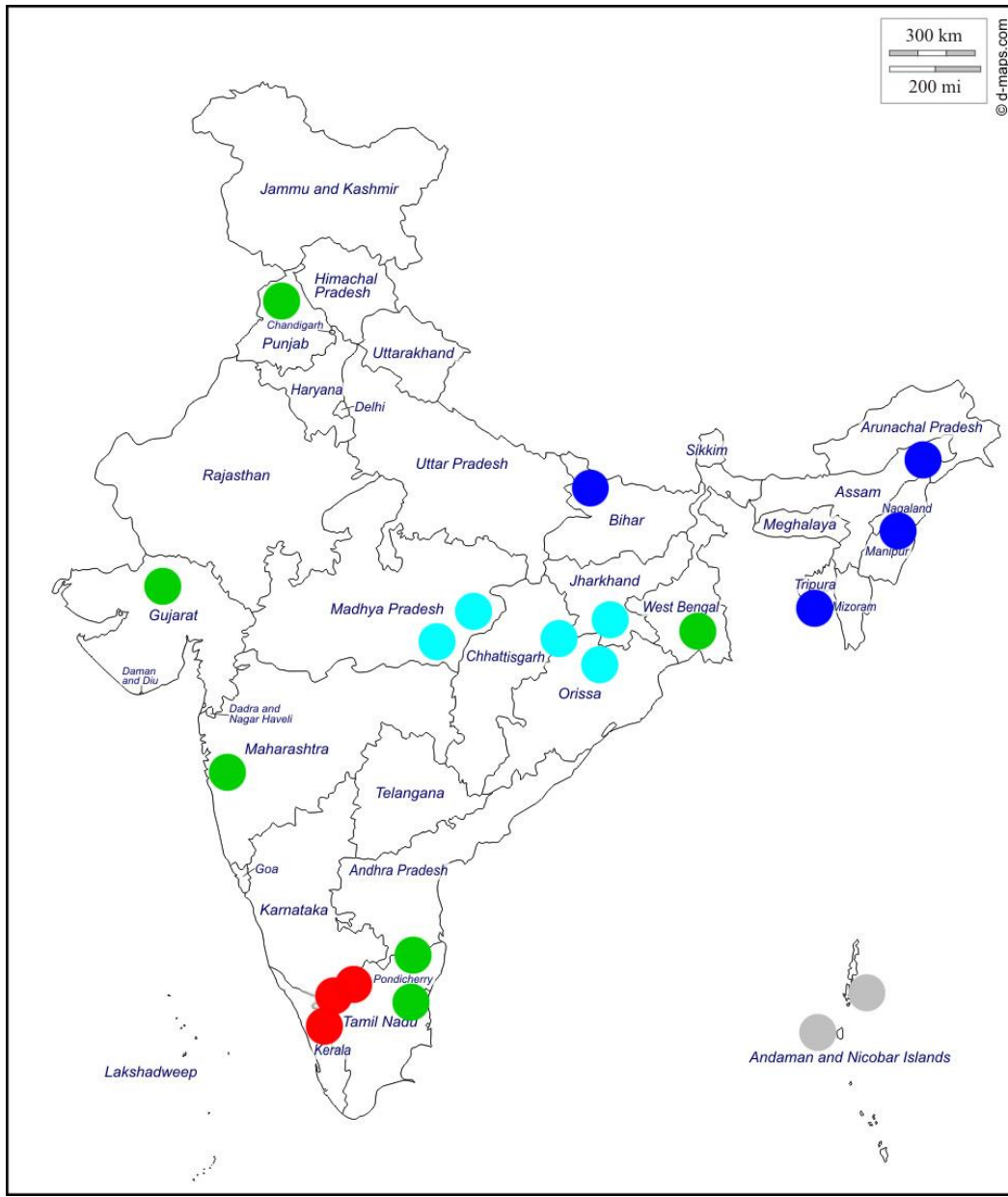


Supplementary Information

Fig S1: Map of India showing approximate locations of sampling of the populations included in this study. Populations shown in 'grey' are populations from the Andaman and Nicobar archipelago. Populations shown in 'red' are Dravidian speaking tribal populations from the Nilgiri Hills in Southern India. Populations shown in 'cyan' are Austro-Asiatic speaking tribal populations from the East and Central India. Populations shown in 'green' are caste populations primarily speaking the Indo-European language. Populations shown in 'blue' are Tibeto-Burman speaking populations of North-East India and are predominantly tribes except the Manipuri Brahmins. (More description in Table 1)



Supplementary Information 1 (SI-1)

Detailed results of ADMIXTURE analysis with *all* 20 populations

The conclusions that we wish to highlight in **SI-1** are:

- 1) The cross-validation (CV) error is minimized when $K=5$, irrespective of whether the entire dataset or the LD pruned dataset is used, or whether CV is taken to be 5% and 10%
- 2) At $K=2$, a small proportion (mean=0.06) of Jarwa and Onge ancestries are noted to be present in individuals drawn from mainland populations, (first panel in Fig. Supplement). However, this proportion decreases as we increase K . At $K=3$ it stands at 0.02 and at $K=4, 5$ it further reduces to 0.004). Therefore, it appears that these estimates are of statistical noise, rather than real admixture estimates.

Contents for this section

Fig. Supplement(i) Individual ancestry inferred with ADMIXTURE with $K = 2, 3$ and 4 are plotted. Each individual is represented by a vertical line partitioned into colored segments whose heights correspond to his/her ancestry coefficients for up to four inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; the labels were used to order the samples in plotting. (*SNPs were pruned only to include those for which the pairwise Linkage-Disequilibrium was less than 0.5*)

Fig. Supplement (ii) Individual ancestry inferred with ADMIXTURE with $K = 5$ for which the CV error is minimized. (*SNPs were pruned only to include those for which the pairwise Linkage-Disequilibrium was less than 0.5*)

Table S1. (A) (B) (C) and (D) Cross-Validation error for different choices of K ,

- (A) CV error calculated at 5% with all SNPs
- (B) CV error calculated at 5% after pruning SNPs in LD (*pairwise LD <0.5*)
- (C) CV error calculated at 10% with all SNPs
- (D) CV error calculated at 10% after pruning SNPs in LD (*pairwise LD <0.5*)

Fig. Supplement(iii).(A), (B), (C), (D): These graphs correspond to Table S1 (A) (B)(C) and (D) respectively

Table S1 E: The ADMIXTURE Estimates pertaining to $K=5$ for 20 populations(*SNPs were pruned only to include those for which the pairwise Linkage-Disequilibrium was less than 0.5*)

Fig. Supplement(i): Individual ancestry (*all 367 individuals from 20 populations*) inferred with ADMIXTURE with $K = 2, 3$ and 4. Each individual is represented by a vertical line partitioned into colored segments whose heights correspond to his/her ancestry coefficients in up to four inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; they were used to order the samples in plotting.

With $K=2$ The mainland Indian populations separate from the Jarawa (JRW) and Onge(ONG) or the hunter-gatherer tribal populations of Andaman and Nicobar Islands

With $K=3$ In addition to the Island and mainland separation, the Tibeto-Burman speaking populations from NE-India separate from the other mainland populations.

With $K=4$ The caste populations in India, primarily Indo-European speakers, separate from the tribal populations (i.e the Austro-Asiatic speaking tribes of East and Central India and the Dravidian speaking tribes of Nilgiri Hills)

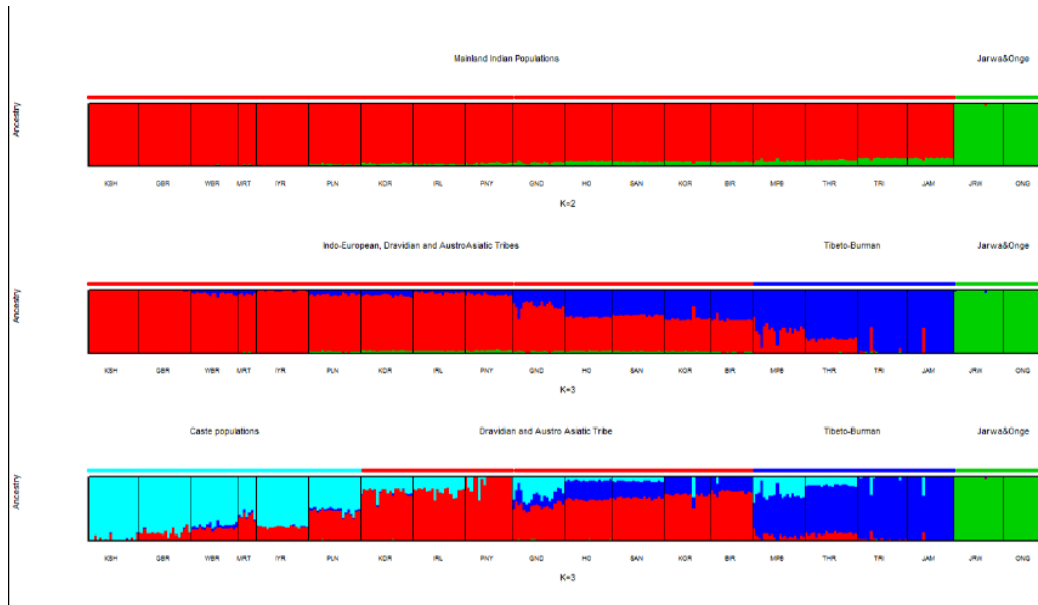


Fig. Supplement (ii): Individual ancestry (*all 367 individuals from 20 populations*) inferred with ADMIXTURE with $K = 5$ for which the CV error is minimized. The Austro-Asiatic speaking tribes of East and Central India separate from the Dravidian speaking tribes of Nilgiri Hills. However we see substantial evidence of admixture in the populations.

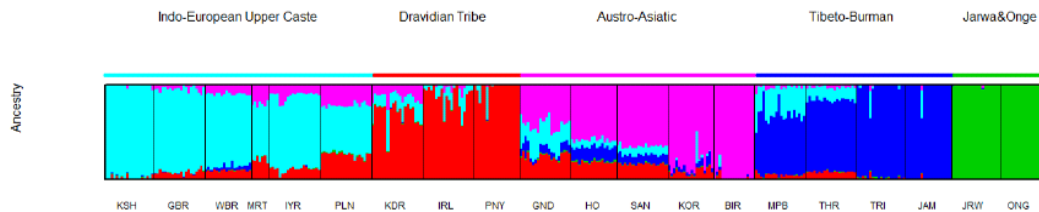


Table S1:

A. CV 5% all SNPs

K	CV Error
2	0.4416
3	0.43605
4	0.4327
5	0.4316
6	0.43255

B. CV 5% LD pruned SNPs

K	CV Error
2	0.51511
3	0.50892
4	0.5053
5	0.5039
6	0.50571

C. CV 10% all SNPs

K	CV Error
2	0.441
3	0.43522
4	0.43168
5	0.43022
6	0.43078

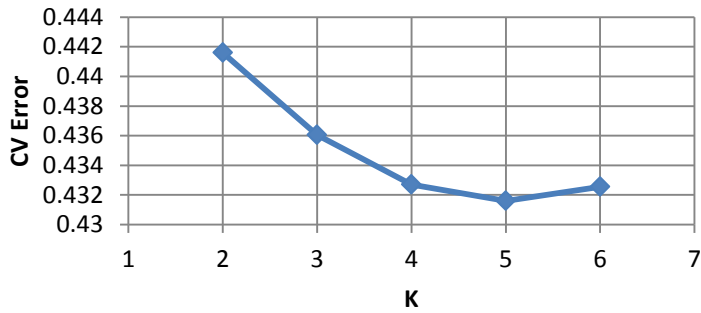
D. CV 10% LD pruned SNPs

K	CV Error
2	0.441
3	0.43522
4	0.43168
5	0.43022
6	0.43078

Fig. Supplement (iii) (A), (B), (C) and (D)

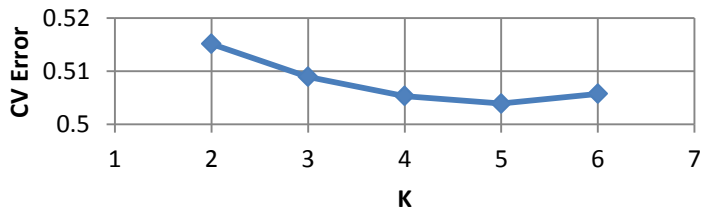
**Proportion of Cross-Validation(CV)
Error in ADMIXTURE run with
different values of K (CV=5%)**

367 individuals ALL Indian populations (all SNPs)



**Proportion of Cross-Validation(CV)
Error in ADMIXTURE run with
different values of K (CV=5%)**

**367 individuals ALL Indian populations (LD
removed)**



**Proportion of Cross-Validation(CV)
Error in ADMIXTURE run with
different values of K (CV=10%)**

**367 individuals ALL Indian populations (LD
removed)**

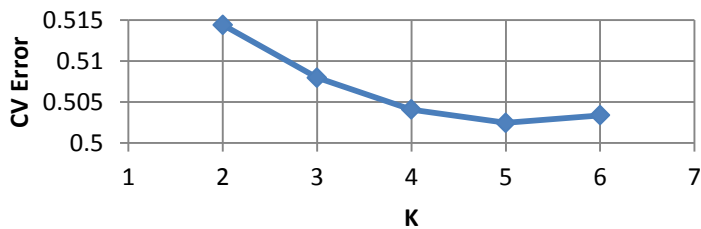


Table S1E: ADMIXTURE Estimates with K=5 for 20 populations

Population Name	ANI	ASI	AAA	ATB	Ancestral Andaman
KSH	0.97	0.014	0.007	0.007	0.002
GBR	0.875	0.07	0.045	0.006	0.003
WBR	0.769	0.086	0.097	0.042	0.005
MRT	0.59	0.199	0.207	0.001	0.002
IYR	0.785	0.106	0.105	0.001	0.003
PLN	0.523	0.249	0.219	0	0.009
KAD	0.171	0.682	0.136	0.003	0.008
IRL	0.132	0.835	0.032	0	0.002
PNY	0.037	0.959	0.002	0.002	0
GND	0.249	0.241	0.417	0.082	0.011
HO	0.072	0.174	0.591	0.15	0.013
SAN	0.087	0.157	0.656	0.093	0.007
KOR	0.027	0.079	0.823	0.066	0.005
BIR	0.008	0.013	0.972	0.006	0
MPB	0.292	0.038	0.032	0.634	0.005
THR	0.139	0.074	0.037	0.747	0.003
TRI	0.024	0.012	0.013	0.943	0.009
JAM	0.016	0.004	0.003	0.975	0.002
JRW	0	0	0.002	0.001	0.996
ONG	0	0	0	0	1

Supplementary Information 2 (SI-2)

In SI-2 we detail the results of our ADMIXTURE run with *the 18mainland Indian* populations

The observations of interest that we have emphasized in this **SI-2** are:

- 1) The cross-validation (CV) error is minimized when $K=4$, irrespective of whether we have used the entire dataset or the LD pruned dataset, or whether we have used CV at 5 fold or 10 fold. [Table S2 (A)–(F); Fig. S3 (A)–(F)]
- 2) ADMIXTURE, which explored a very high-dimensional likelihood space, was robust in detecting population structure and the inferences are stable in multiple runs of the program with a random initialization (Random seed as starting point). [Table S3 (A) – (D)]
- 3) Multiple programs which estimate ancestry and admixture proportions from genotype-data converge to similar inference about population structure and admixture in Indian populations. [Table S4, S5; Fig. S4]. Also elaborating on some findings using fineSTRUCTURE.
- 4) We explore the sex-bias in admixture proportions. [Fig. S5, S6]

Contents for this section

Fig. S2: Individual ancestry inferred with ADMIXTURE with $K = 2$ and 3. Each individual is represented by a vertical line partitioned into colored segments whose heights correspond to his/her ancestry coefficients in up to four inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; they were used to order the samples in plotting.

Table S2 (A)–(F): Cross-Validation error for different choice of K

- (A) CVE calculated at 5-fold when all the SNPs are included (no LD pruning)
- (B) CVE calculated at 10-fold when all the SNPs are included (no LD pruning)
- (C) CVE calculated at 5-fold when SNPs with pairwise LD <0.5 were included
- (D) CVE calculated at 10-fold when SNPs with pairwise LD <0.5 were included
- (E) CVE calculated at 5-fold when SNPs with pairwise LD <0.1 were included
- (F) CVE calculated at 10-fold when SNPs with pairwise LD <0.1 were included

Fig. S3 (A)–(F): The graphs corresponding to Table S2 (A)–(F) respectively

Table S3: Summary Table of cross-validation error (CVE) generated from multiple runs (10) of ADMIXTURE, using

- (A) CVE at 5 fold all SNPs with pairwise LD <0.5
- (B) CVE at 10 fold all SNPs with pairwise LD <0.5
- (C) CVE at 5 fold all SNPs with pairwise LD <0.1
- (D) CVE at 10 fold all SNPs with pairwise LD <0.1

Table S4: The *frappe* estimates with K=4 for 18 populations

Table S5: Details of the 69 populations as identified by fineSTRUCTURE

Fig. S4A and B: Relationships among the 69 populations identified by fineSTRUCTURE (Table S5)

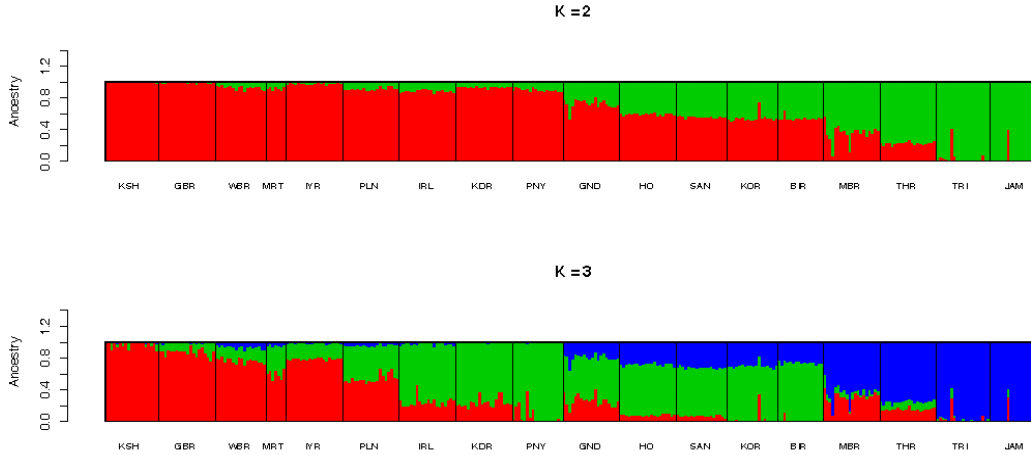
Fig. S5A: ADMIXTURE with K=4 on 107 females from 15 populations shows more ATB component and reduced ANI component in the X-Chromosome of individuals from KSH, GBR, MRT, IYR, PLN as well as GND, HO, SAN, KOR populations.

Fig. S5B: Q-Q Plot of the 107 females.

Fig. S6A: Dendrogram of the X-chromosome haplotypes show separate clades belonging to different populations (This is a large format figure and can be viewed clearly when magnified)

Fig. S6A: Dendrogram of the X-chromosome haplotypes show separate clades belonging to different populations (This is a large format figure and can be viewed clearly when magnified). The color codes are consistent with the colors used in previous figures.

Fig. S2: Individual ancestry inferred with ADMIXTURE with K = 2 and 3. Each individual is represented by a vertical line partitioned into colored segments whose heights correspond to his/her ancestry coefficients in up to four inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; they were used to order the samples in plotting.



With K=2, like before (SI-1), the close to 100% GREEN are the TB speakers from NE India and close to 100% RED are caste populations, primarily IE speakers of North India. We have identified these GREEN as the ATB component.

With K=3, like before (SI-1), the close to 100% BLUE are the TB speakers from NE India. The other population (RED with K=2) is split into RED and GREEN. We have identified the 'RED' component as the ANI ancestry. The GREEN is the combined (ASI+AAA), which separate at K=4.

Table S2: Cross-Validation error (CVE) for different choices of K clearly shows CVE to be minimum when K=4

- (A) CVE calculated at 5-fold when all the SNPs are included (no LD pruning)
- (B) CVE calculated at 10-fold when all the SNPs are included (no LD pruning)
- (C) CVE calculated at 5-fold when SNPs with pairwise LD <0.5 were included
- (D) CVE calculated at 10-fold when SNPs with pairwise LD <0.5 were included
- (E) CVE calculated at 5-fold when SNPs with pairwise LD <0.1 were included
- (F) CVE calculated at 10-fold when SNPs with pairwise LD <0.1 were included

Table S2A:

K	CV Error
2	0.54127
3	0.53672
4	0.53508
5	0.53678

Table S2B:

K	CV Error
2	0.54069
3	0.53585
4	0.53388
5	0.5347

Table S2C:

K	CV Error
2	0.50353
3	0.4998
4	0.4989
5	0.50011

Table S2D:

K	CV Error
2	0.50283
3	0.49872
4	0.49738
5	0.49859

Table S2E:

K	CV Error
2	0.50331
3	0.49983
4	0.49853
5	0.49953

Table S2F:

K	CV Error
2	0.50265
3	0.49882
4	0.49737
5	0.49766

Fig. S3 (A)–(F): The graphs corresponding to Table S2 (A)–(F) respectively

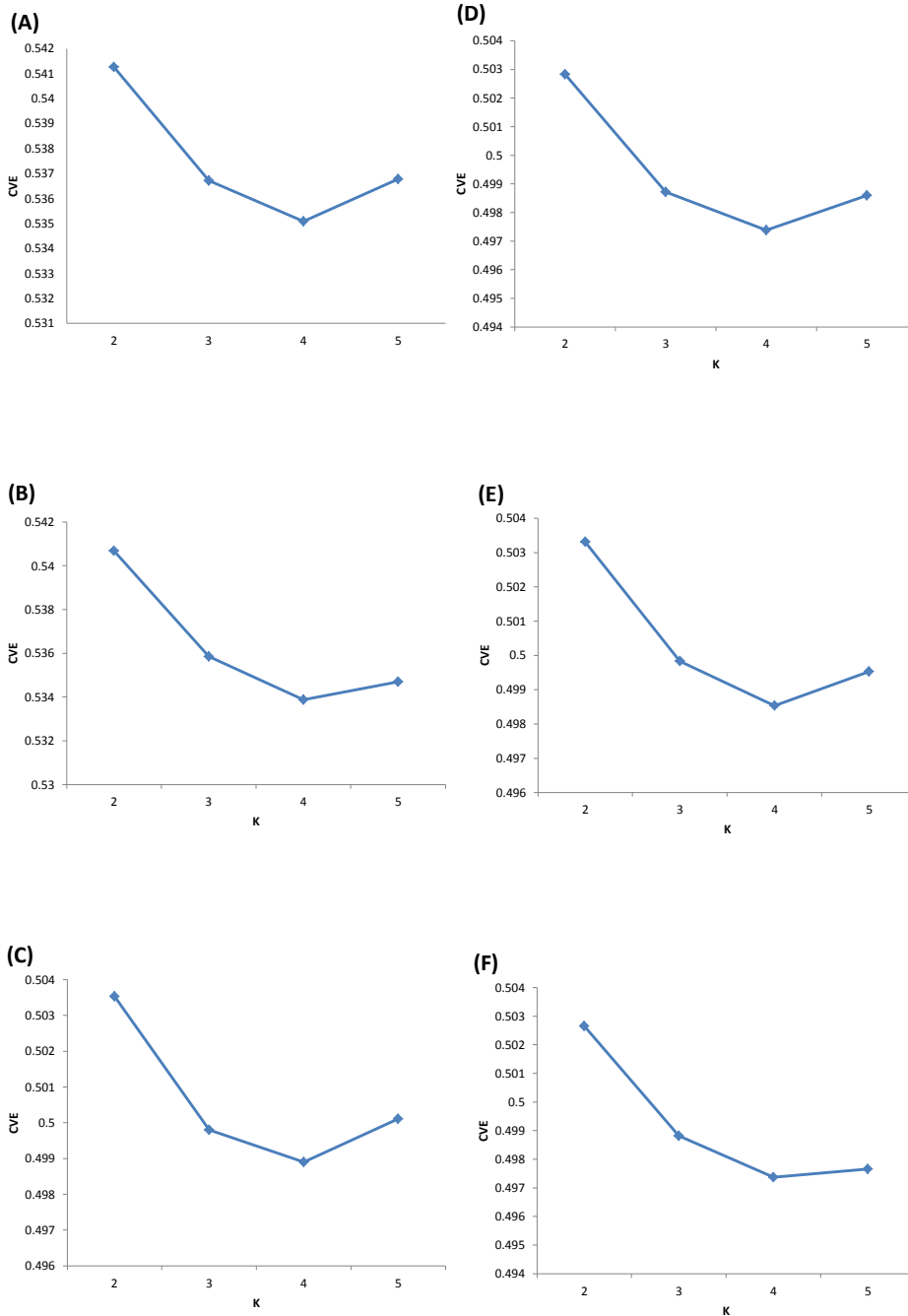


Table S3: Summary Table of cross-validation error (CVE) generated from multiple runs (10) of ADMIXTURE, using
(A) CVE at 5 fold all SNPs with pairwise LD <0.5

- (B) CVE at 10 fold all SNPs with pairwise LD <0.5
 (C) CVE at 5 fold all SNPs with pairwise LD <0.1
 (D) CVE at 10 fold all SNPs with pairwise LD <0.1

Table S3A

	K=2	K=3	K=4	K=5
Mean	0.5189967	0.5150022	0.5136544	0.5155000
Standard Deviation	3.08×10^{-05}	3.52×10^{-05}	6.48×10^{-05}	2.23×10^{-04}
Minimum	0.51896	0.51495	0.51358	0.51533
Maximum	0.51904	0.51508	0.51378	0.51608

Table S3B

	K=2	K=3	K=4	K=5
Mean	0.5183767	0.5140633	0.5123056	0.5133944
Standard Deviation	1.22×10^{-05}	2.34×10^{-05}	2.40×10^{-05}	2.59×10^{-04}
Minimum	0.51835	0.51402	0.51227	0.51317
Maximum	0.51839	0.51409	0.51234	0.51387

Table S3C

	K=2	K=3	K=4	K=5
Mean	0.5034133	0.4997322	0.4986922	0.4996167
Standard Deviation	9.92×10^{-05}	7.72×10^{-05}	1.88×10^{-04}	2.65×10^{-04}
Minimum	0.50331	0.49962	0.49845	0.49905
Maximum	0.50356	0.49983	0.49904	0.50001

Table S3D

	K=2	K=3	K=4	K=5
Mean	0.5026544	0.4987611	0.4972900	0.4974733
Standard Deviation	2.24×10^{-05}	4.28×10^{-05}	1.17×10^{-04}	2.08×10^{-04}
Minimum	0.50262	0.49866	0.49713	0.49741
Maximum	0.50268	0.49882	0.49741	0.49767

Table S4: Ancestry proportions of 18 mainland Indian populations as estimated by the best fit (K=4) model in *frappe*

Population Name	ANI ancestry	ASI ancestry	AAA ancestry	ATB ancestry
KSH	0.9793	0.0149	0.0045	0.0013
GBR	0.8823	0.0759	0.0412	6.00E-04
WBR	0.7663	0.0994	0.101	0.0332
MRT	0.5751	0.2141	0.2105	3.00E-04
IYR	0.8046	0.111	0.0837	7.00E-04
PLN	0.4902	0.2761	0.2331	6.00E-04
KAD	0.0895	0.7681	0.1414	0.0011
IRL	0.0532	0.9255	0.0213	0
PNY	0.0252	0.9696	0.0052	0
GND	0.3697	0.193	0.3756	0.0617
HO	0.0475	0.1705	0.7116	0.0704
SAN	0.0347	0.1933	0.6398	0.1321
KOR	0.0181	0.0471	0.9091	0.0257
BIR	0.0082	0.0054	0.9864	0
MPB	0.2635	0.0512	0.0351	0.6502
THR	0.0935	0.0951	0.0447	0.7667
TRI	0.0156	0.0084	0.0117	0.9643
JAM	0.0149	0.0044	0.0031	0.9776

Detailed result of fineSTRUCTURE analysis:

Table S5: The 69 subpopulations identified by fineSTRUCTURE

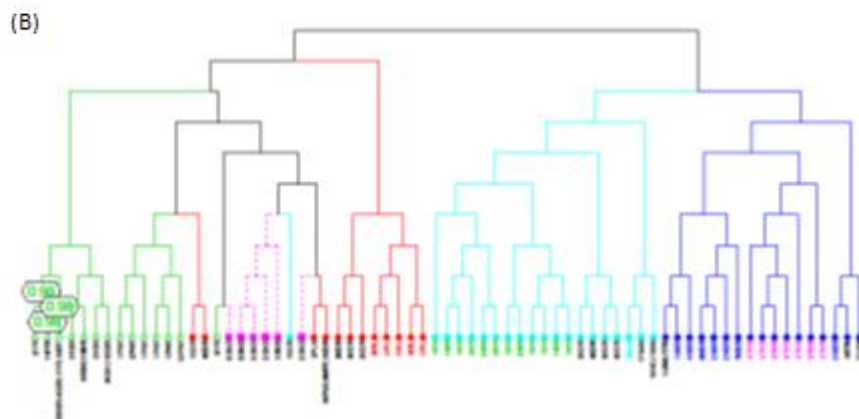
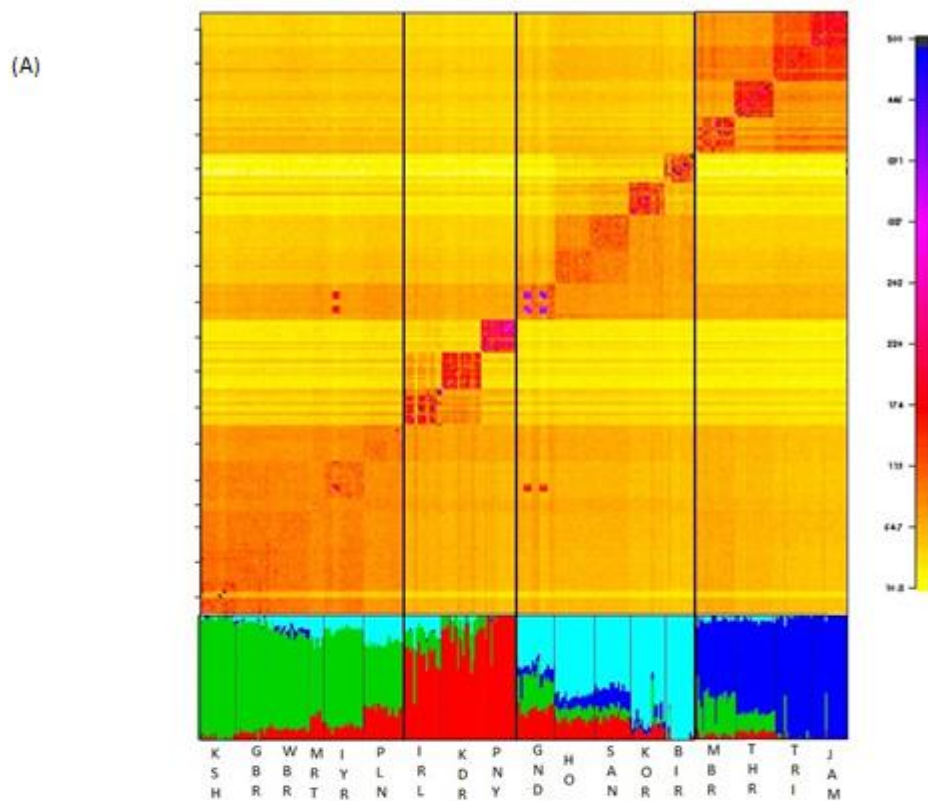
Sub-Population	Number of Individuals and Original Population Label
1	2 IYR
2	14 IYR
3	18 WBR; 4 GBR; 1 IYR; 1 MRT
4	2 KSH
5	6 KSH; 1 GBR
6	2 KSH
7	9 KSH; 15 GBR
8	1 PNY
9	2 PNY
10	1 PNY
11	1 PNY
12	1 PNY
13	12 PNY
14	1 KDR
15	4 KDR
16	3 IYR
17	8 GND

18	5GND
19	2GND
20	2GND
21	1GND
22	1KOR
23	2GND
24	2PLN
25	18PLN; 6MRT; 1KDR
26	3KDR
27	9KDR
28	2KDR
29	8IRL
30	2IRL
31	5IRL
32	3IRL
33	2IRL
34	1BIR
35	1BIR
36	1BIR
37	2BIR
38	2BIR
39	2BIR
40	1BIR
41	2BIR
42	1BIR
43	1BIR
44	1BIR
45	1BIR
46	2KOR
47	4KOR
48	5KOR
49	6KOR
50	2SAN
51	17SAN
52	18HO; 1SAN
53	1JAM; 1TRI
54	1MBR
55	4MBR
56	2MBR
57	2MBR
58	2MBR
59	7MBR
60	2THR
61	2THR
62	2THR
63	8THR
64	2THR
65	2THR
66	2THR

67	2MBR
68	18TRI
69	17JAM

Fig. S4A: Heat map of the ‘Coancestry Matrix’ of 331 individuals from 18 mainland Indian populations. The co-ancestry matrix broadly conforms to the inferences of the 4- ancestral components identified by ADMIXTURE.

Fig. S4B: Relationship between the 69 populations identified by fineSTRUCTURE (Table S5)



Supplementary text: Findings using fineSTRUCTURE

ADMIXTURE analysis indicated that the Gond (GND) is an extremely heterogeneous and admixed tribal population (Fig. 3B and Table 2). Both ADMIXTURE and fineSTRUCTURE have revealed that the upper caste Iyers (IYR), in spite of being Dravidian speakers and residing in south India, possess a high fraction of the ANI component (Fig. 3B, Table 2 and Fig. 3C). fineSTRUCTURE has also revealed the co-ancestry of the ANI component of IYR and GND, but no striking similarity of the ANI component with the other AA speaking Ho tribals living in the same geographical region (Fig. 3C). fineSTRUCTURE analysis has thus reestablished that some of the hunter-gatherer tribals of mainland India (Table 1) irrespective of their linguistic affiliation, have remained very isolated and demographically small after evolving from an ancestral population; these features have resulted in decreasing genomic similarities among them by genetic drift (Fig. 3C).

Sex-Bias in Admixture:

Fig. S5A: ADMIXTURE with $K=4$ on 107 females from 15 populations shows more ATB component and reduced ANI component in the X-Chromosome of individuals from KSH, GBR, MRT, IYR, PLN as well as GND, HO, SAN, KOR populations.

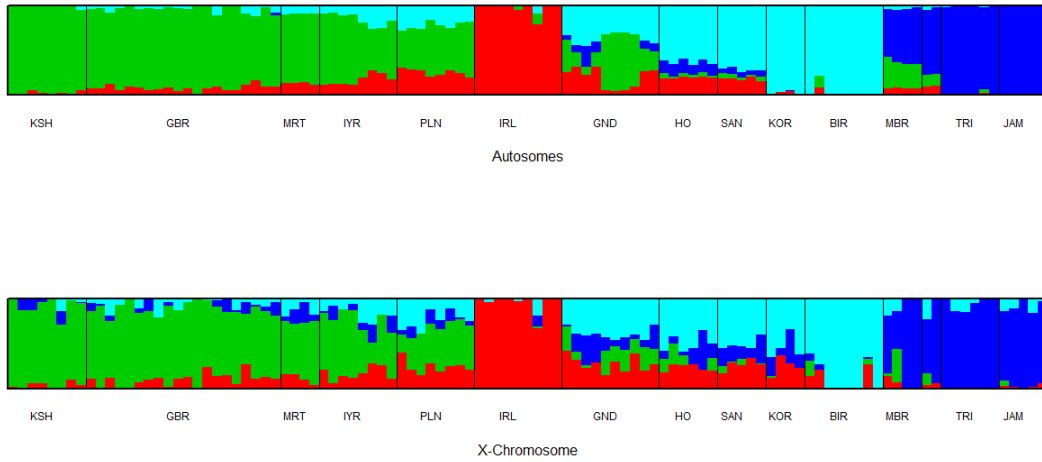


Fig. S5B: Q-Q Plot of the 107 females.

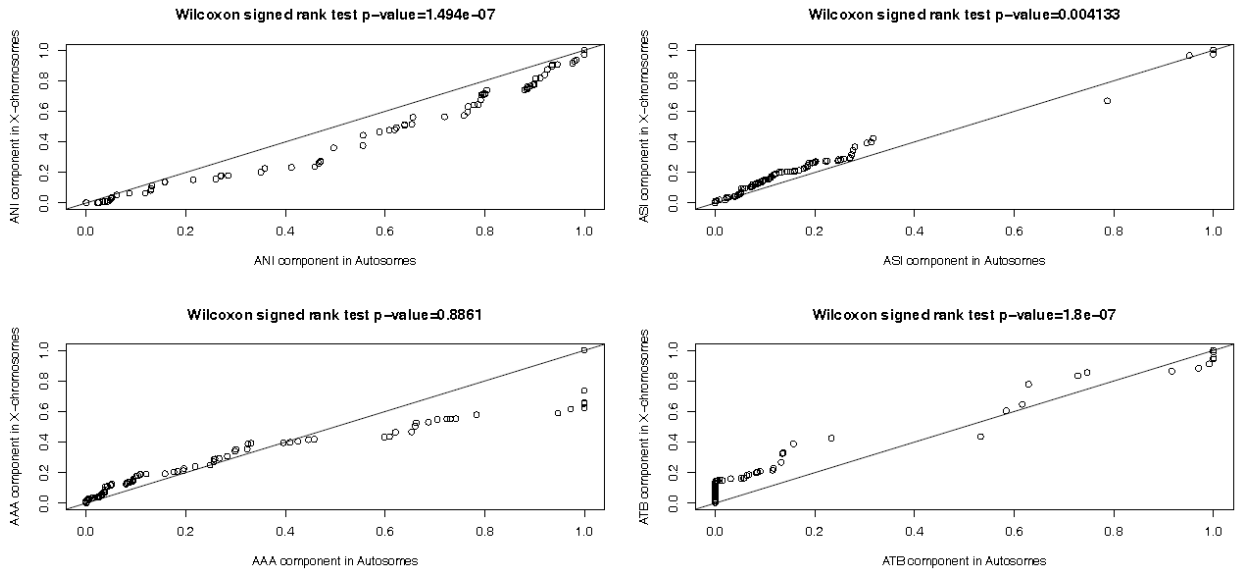


Fig. S6A: Dendrogram of the X-chromosome haplotypes show separate clades belonging to different populations (This is a large format figure and can be viewed clearly when magnified)

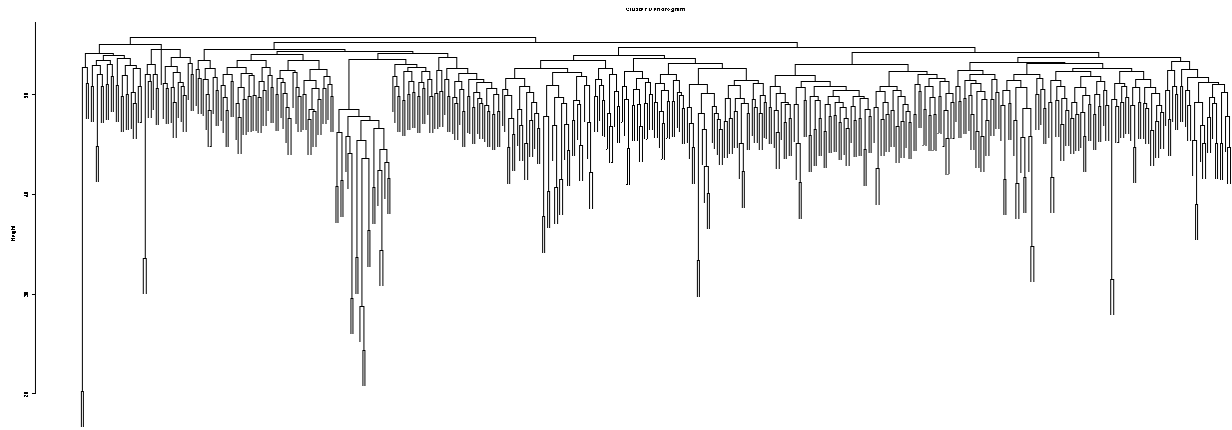


Fig. S6B: Dendrogram of the X-chromosome haplotypes show separate clades belonging to different populations (This is a large format figure and can be viewed clearly when magnified). The color codes are consistent with the colors used in previous figures.

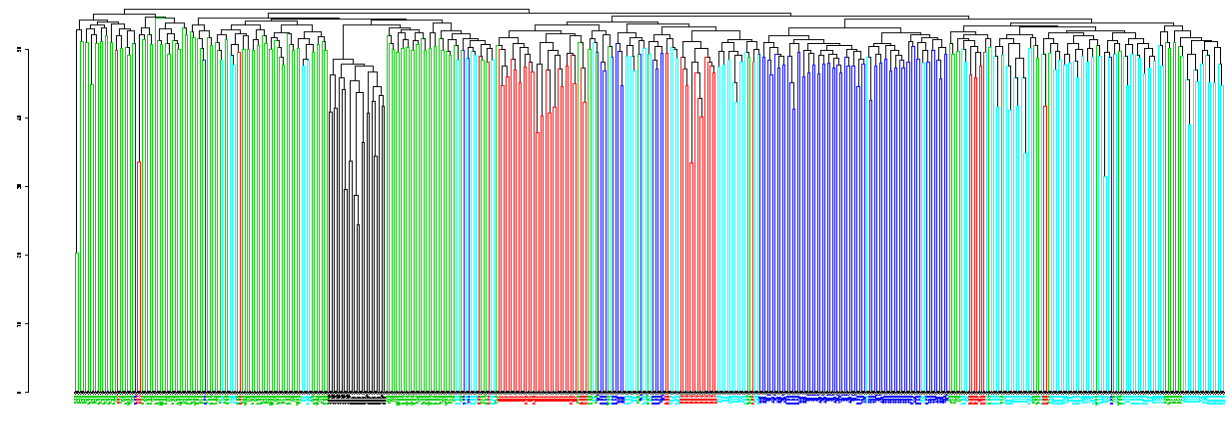
GREEN is used for haplotypes of individuals with a major ANI ancestry (i.e. KSH, GBR, IYR, MRT, PLN)

RED is used for haplotypes of individuals with a major ASI ancestry (IRL, KDR, PNY)

CYAN is used for haplotypes of individuals with a major AAA ancestry (GND, HO, SAN, KOR, BIR)

BLUE is used for haplotypes of individuals with a major ATB ancestry (MPB, THR, TRI, JAM)

BLACK is used for haplotypes of individuals from the JRW and ONG populations.



Supplementary Information 3 (SI 3)

The detail results exploring:

- 1) The genetic relationship of the ancestries present in mainland India with neighbouring populations. (page 26)

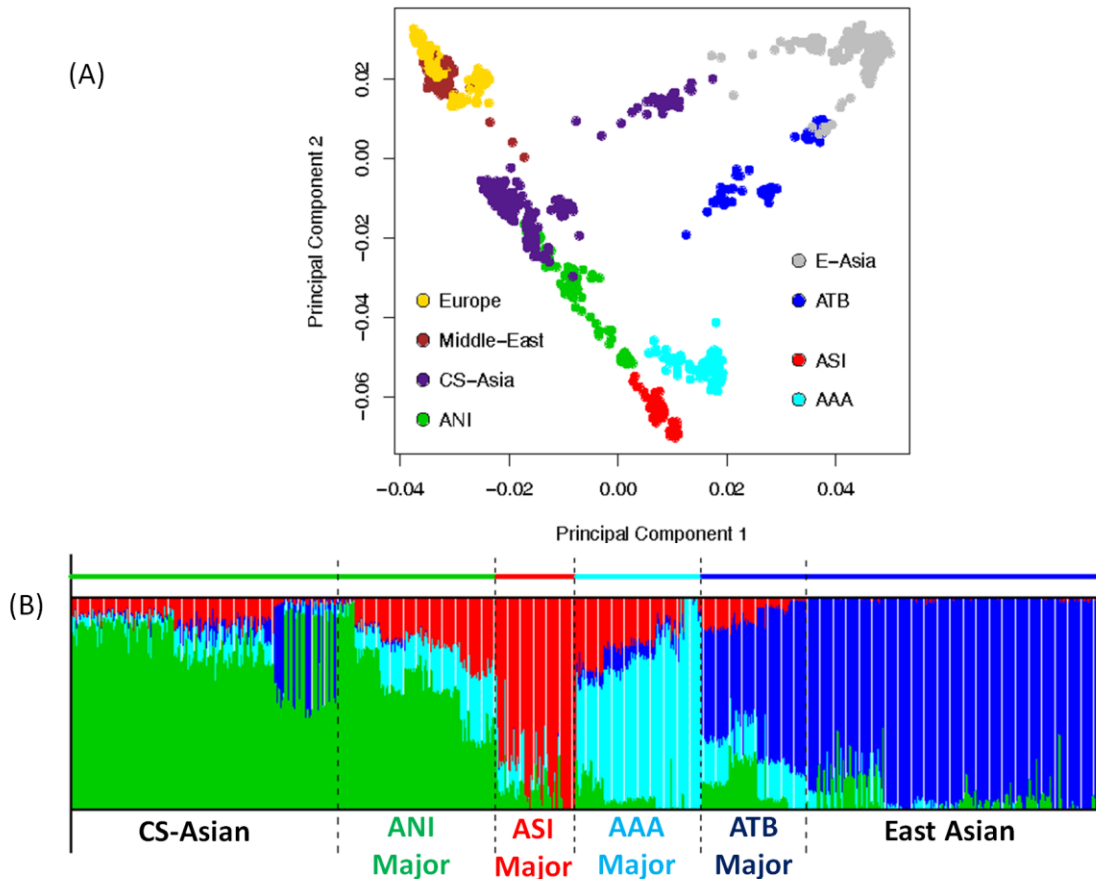


Fig. S7A: The PCA plot with Europeans, Middle-Easterners, Central-South Asians (CS-Asian), East-Asians (E-Asian) included in Human Genome Diversity Panel (HGDP) shows that the Europeans and Middle-Easterners cluster distinctly, in spite of being genetically close to the C-S Asians and populations which have high proportion of ANI ancestry

Fig. S7B: **Estimates of ancestral components of 331 individuals from 18 mainland Indian populations along with 207 CS Asian and 235 E-Asian individuals of HGDP.** A model with four ancestral components ($K=4$) was the most parsimonious to explain the variation and similarities of the genome-wide genotype data. The CS-Asians are similar to ANI-major populations and E-Asians are similar to ATB major population indicating common ancestry for the respective populations before subdividing into the population identities that we see today. It also clearly shows that the AAA and the ASI cannot be readily identified with any of these global population groups. Population labels were added only after each individual's ancestry had

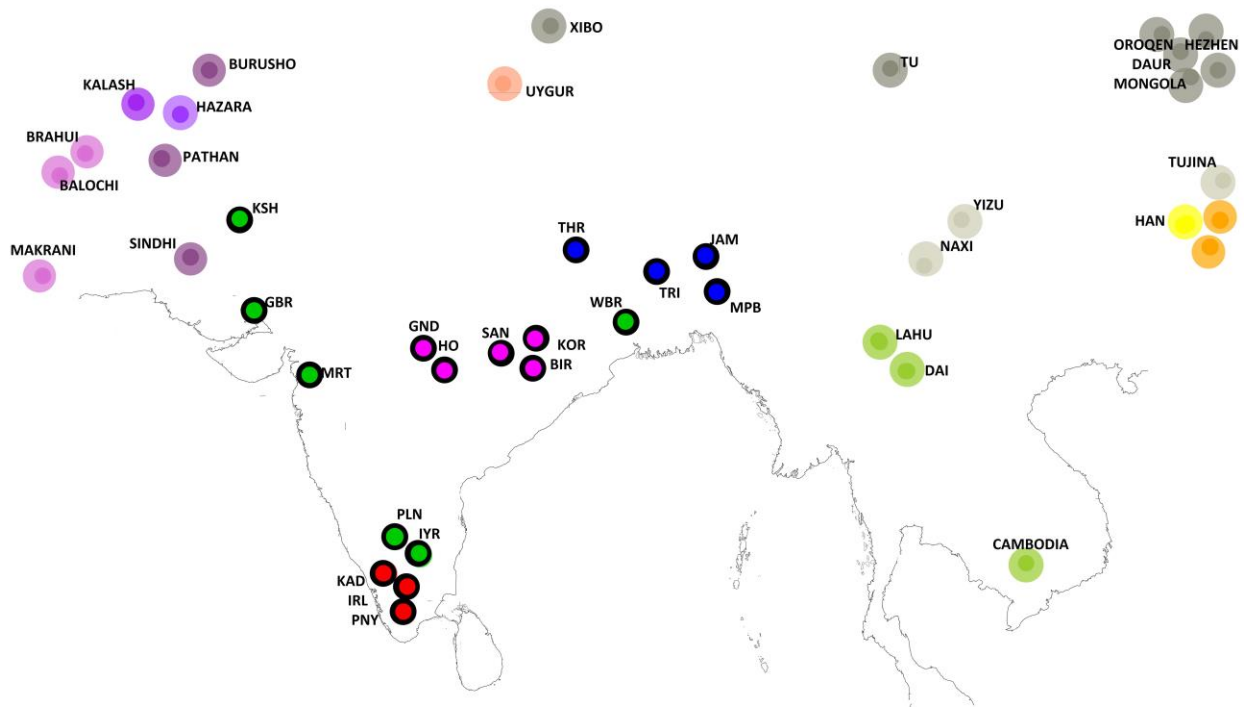
been estimated. The colors correspond to the colors used to encircle clusters of individuals in Fig. 2A. [also mainland Indians in Fig. S2, S3 and S7A]

Supplementary Text:

Analysis of 18 Mainland Indian Populations combined with the Central-South Asian and East-Asian samples of HGDP.

We combined our data set of 18 mainland populations with the Central-South Asians (CS-Asians) and East Asians (E-Asians) from HGDP. The CS-Asians populations included are: Brahui, Balochi, Hazara, Makrani, Sindhi, Pathan, Kalash and Burusho. While the E-Asians populations included are: Han, Tujia, Yizu(Yi), Miao zu, Oroqen, Daur, Mongola, Hezhen, Xibo, Dai, Lahu, She, Naxi, Tu, Yakut, Japanese, Combidian. Uygur who are admixed between CS-Asians and E-Asians are also included

Fig. Supplement : Approximate sampling location and population names



Definition of the Groups:

The Li et al paper(22) (Supplementary material 2.2 and figure S2 B), has subdivided the CS-Asians into 2 major groups: Group-1 consisting Brahui, Makrani, Balochi and Group-2 consisting Burusho, Pathan and Sindhi. They identified Hazara and Kalash as outlier populations.

The detailed analysis of the HGDP E-Asians (Supplementary material 2.2 and figure S2 B) shows that populations in E-Asia also have multiple subgroups. We have defined:

E-Asian-group 1: Populations with high ‘northern’ ancestry include Mongola, Oroqen, Hezhen, Daur, Tu, Xibo, and Japanese. These groups reside in high latitude areas and speak languages of the Altaic family.

E-Asian-group 2: In contrast to E-Asian-group-1, populations like Dai, Lahu and Cambodian, who live in or near southwestern China have the lowest northern ancestry.

The Han and northern Han Chinese can be distinguished, although the former is most likely a mixture of southern and more central individuals.

E-Asian group 3: The Naxi and Yi are from the Yunnan Province in Southwest China, also have high northern ancestry, possibly due to their shared ancestry with the nomadic Qiang, an ethnic group from the Tibetan plateau.

E-Asian group 4: Other southern populations to the east (She and Miao)

Yakut as a separate group because it is an admixed population.

Tracing the ANI and the ATB ancestries

We have followed the above definition. In Fig. 3 (main text) PC-1 represents the systematic variation broadly separating the CS-Asian ancestry from E-Asian ancestry (Fig Supplement above shows the approximate positions on the map from where the populations were sampled), whereas the PC-2 represents the systematic variation broadly between the AAA + ASI ancestry and others.

In Fig. 3 we have broadly recapitulated the findings of Li et al(22). The Hazara and the Kalash are isolated clustered populations in the scatter of PC-1 versus PC-2. There is a thin line of separation between CSA-group-1 and group-2, with group-2 slightly closer to E-Asians. The ANI-major populations of India, particularly KSH which has ~97% ANI ancestry is inseparable from the CS-Asian group-2. Similarly, the JAM and TRI who have more than 95% ATB ancestry are inseparable from E-Asian-group 2. This identifies the origin of the ANI and the ATB ancestries with other major ancestries of the world, thus emphasizing the possible migration corridors through NE and NW India.

The proportion of variation explained by PC-1 (5.16%) and PC-2 (3.62%) in Figure 3, are both large compared to population data with 630918 markers. This indicates that the systematic variation separating the (AAA + ASI) ancestry from others is large, and thus the origin of these ancestries remain not well understood.

Our inferences inform that (i) four ancestral populations arrived in India with the ANI major populations probably using the NW corridor and the ATB major populations using the NE corridor (ii) after their arrival there was considerable admixture among them (iii) endogamy was abruptly established about 1600 years ago, and (iv) the practice of endogamy has been strictly followed resulting in strong ethnic sub-structuring that is evident even to this day.

Supplementary Information 4 (SI-4)

We show the genetic similarity of the Andaman Island populations (Jarawa and Onge) with Papuan and Melanesian populations of HGDP.

Pages 29, 30

The joint analysis of the 20 Indian populations (18 mainland Indian + 2 Island population (JWA and ONG) along with the CS-Asians, E-Asians and Oceania population of HGDP reveal that the Island ancestry of JWA and ONG, which is clearly distinct from all ancestries found in mainland India is indeed also different from CS-Asians and E-Asians but is very similar to the Oceania ancestry.

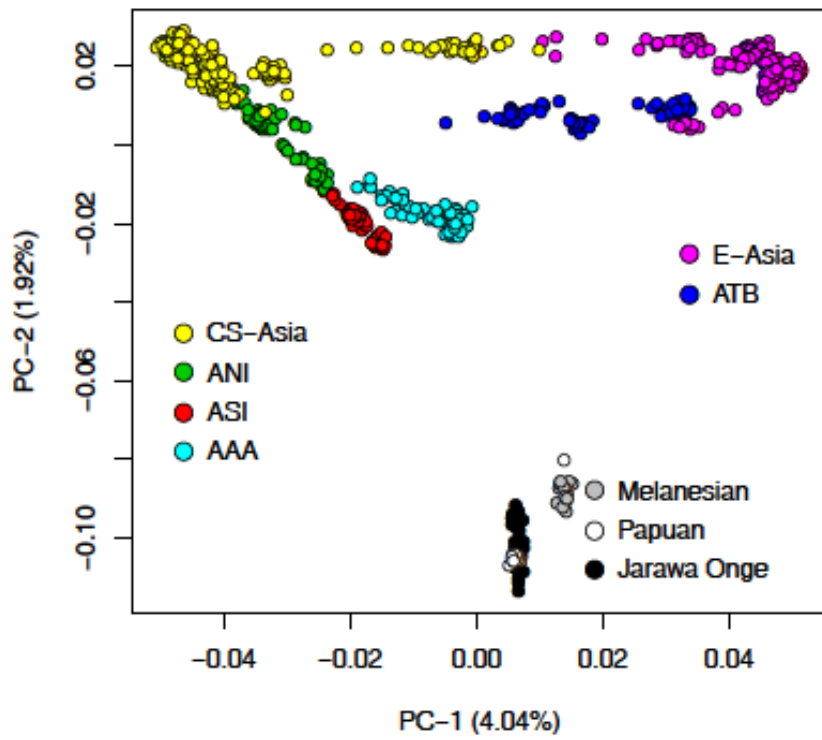
The PC-1 versus PC-2 scatter plot reveals that the Oceania populations of HGDP, especially the Papuans are close to the JWA and ONG (Supplementary Figure 4.1). The variation explained by PC-1 and PC-2 are both high. However, the JWA and ONG separate from the Oceania population along PC-3 (Supplementary Figure 4.2 and Supplementary Figure 4.3). This indicates that the genetic difference between island populations + Oceania population is large compared to the 4 mainland Indian population clusters as well as the CS-Asia and E-Asians **also** it establishes that the genetic difference between the Island populations and Oceania (archipelago ancestry) is small compared to that between the other ancestries and this archipelago.

Fig. S8 A: The PCA plot (PC-1 versus PC-2) of JWA and ONG along with mainland Indians and CS-Asians, E-Asians and Oceania populations of HGDP. It shows clustering of the JWA and ONG populations with Oceania population of HGDP.

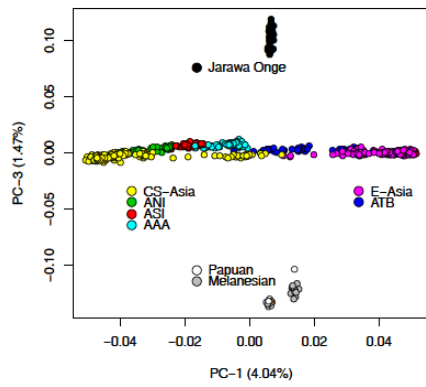
Fig. S8 B: The PC-1 versus PC-3 plot of JWA and ONG along with mainland Indians and CS-Asians, E-Asians and Oceania populations of HGDP. It shows separation of the JWA and ONG populations.

Fig. S8 C: The PC-2 versus PC-3 plot of JWA and ONG along with mainland Indians and CS-Asians, E-Asians and Oceania populations of HGDP. It shows separation of the JWA and ONG populations with Oceania population of HGDP

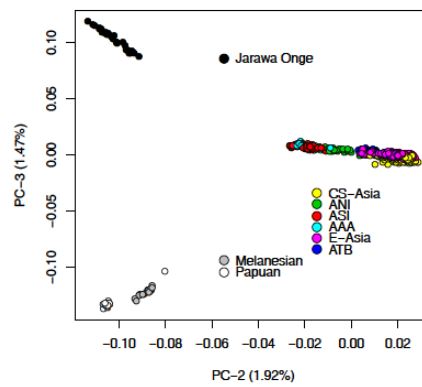
(A)



(B)



(C)



Supplementary Information 5 (SI-5)

In **SI-5** we show

(1) The Ancestral Chromosomal Block Length (ACBL) distribution fits to the theoretical exponential distribution.

Fig. Supplementary A: The distribution of ACSL pertaining to ASI, AAA and ATB, and the fitted exponential distribution among GBR, WBR and IYR population. (The Kolmogorov-Smirnov Test was performed to check the equality of the distribution of ACSL and the fitted exponential)

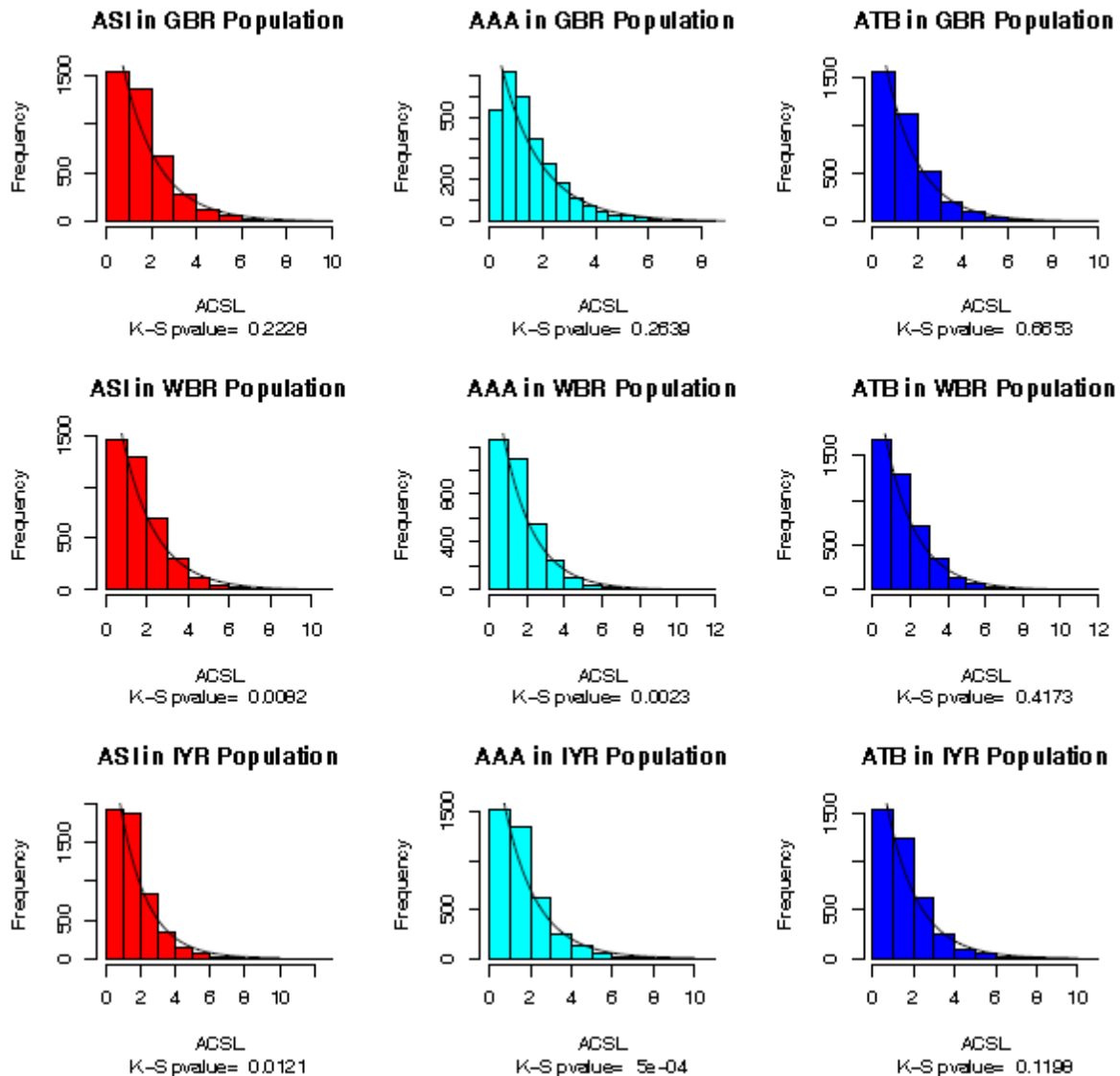


Fig. Supplementary B: The distribution of ACSL pertaining to ASI, AAA and ATB, and the fitted exponential distribution among MRT and PLN population. (The Kolmogorov-Smirnov Test was performed to check the equality of the distribution of ACSL and the fitted exponential)

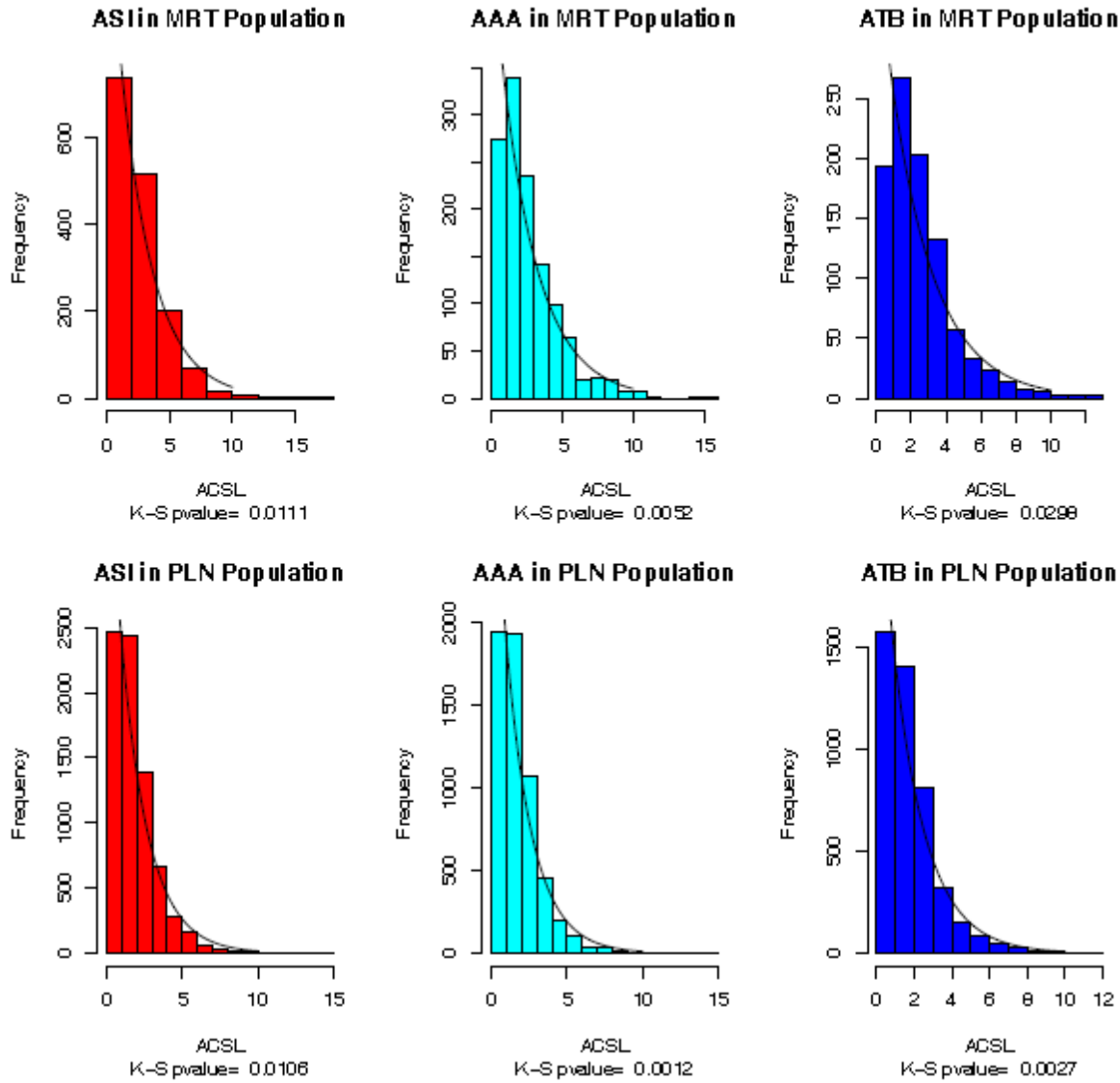


Fig. Supplementary C: The distribution of ACSL pertaining to ANI, AAA and ATB, and the fitted exponential distribution among KDR and IRL population. (The Kolmogorov-Smirnov Test was performed to check the equality of the distribution of ACSL and the fitted exponential)

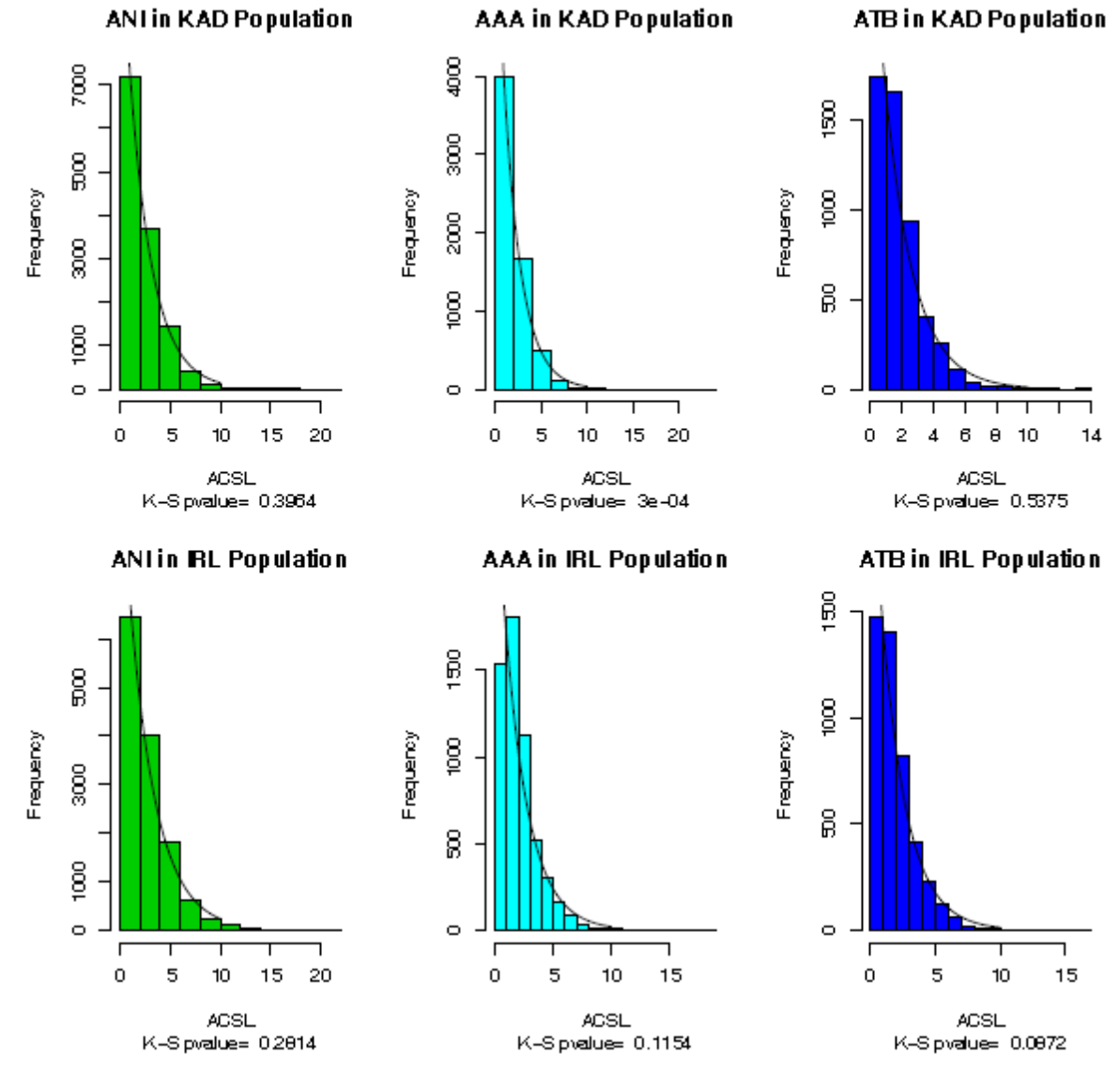


Fig. Supplementary D: The distribution of ACSL pertaining to ANI, ASI and ATB, and the fitted exponential distribution among GND, HO, SAN and KOR population. (The Kolmogorov-

Smirnov Test was performed to check the equality of the distribution of ACSL and the fitted exponential)

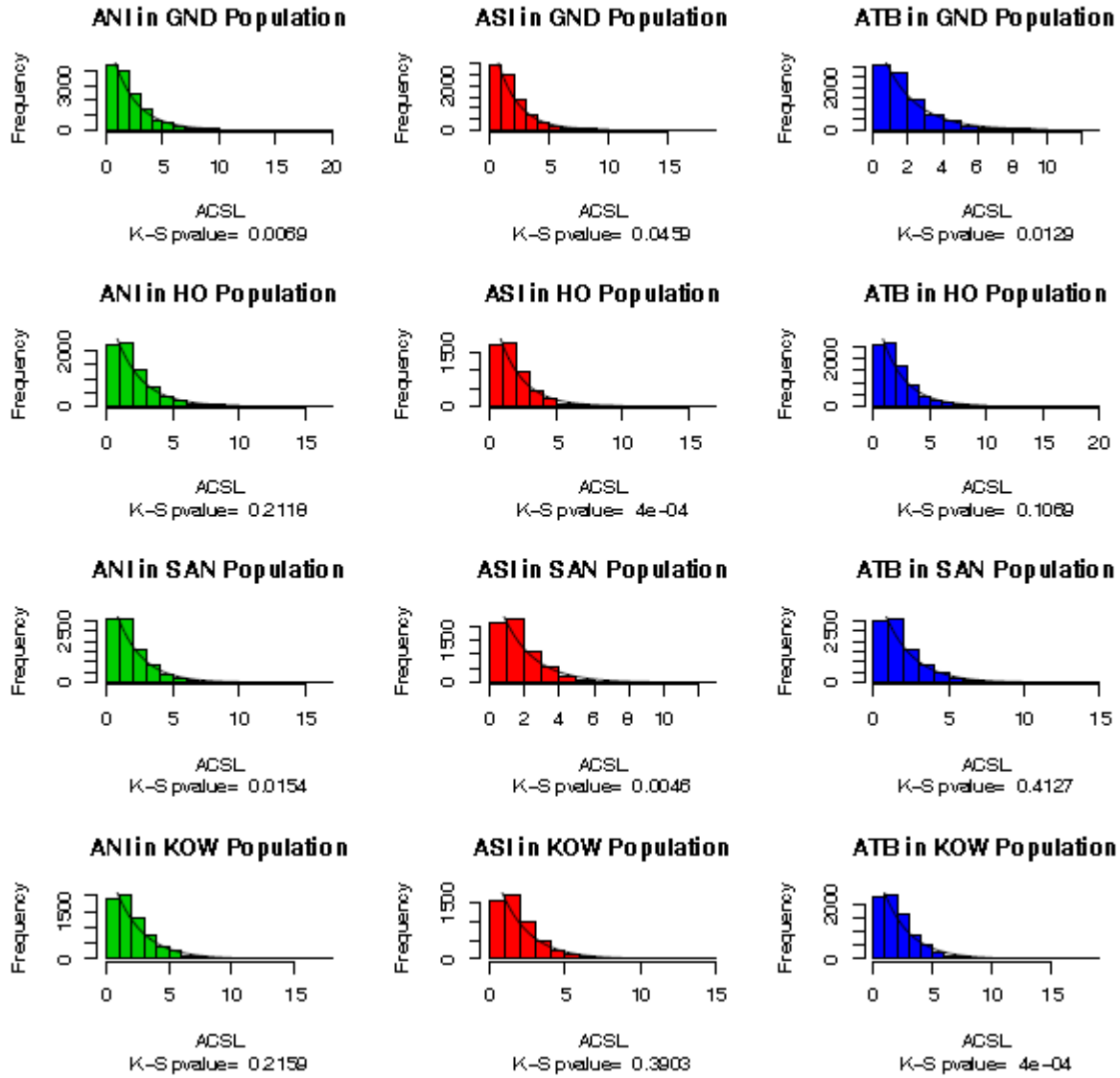
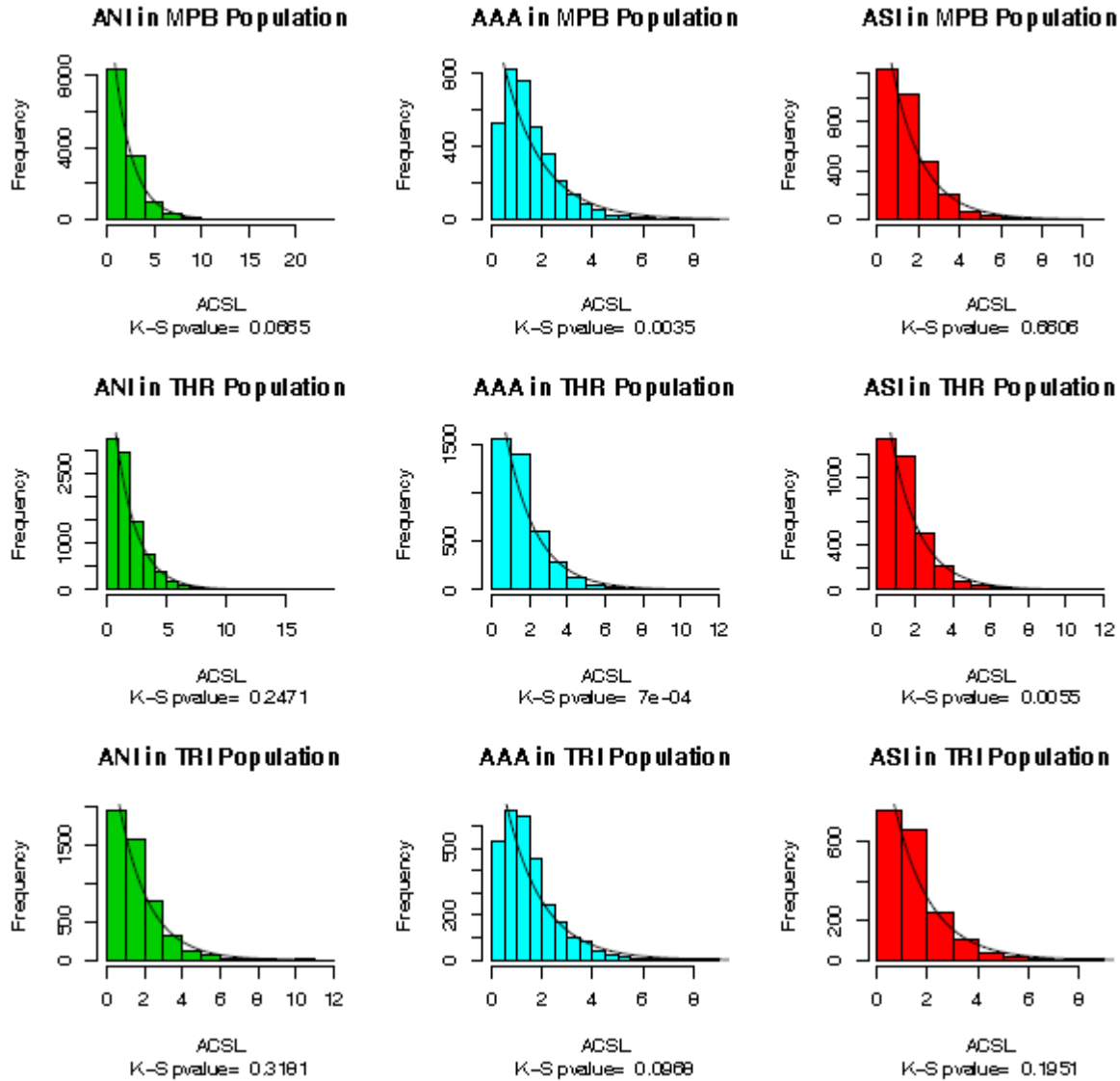


Fig. Supplementary E: The distribution of ACSL pertaining to ANI, ASI and AAA, and the fitted exponential distribution among MPB, THR and TRI population. (The Kolmogorov-Smirnov Test was performed to check the equality of the distribution of ACSL and the fitted exponential)



Supplementary Information 6 (SI-6)

SI-6 is the detailed methods section:

DNA microarray analysis and data curation:

The study was originally planned with 20 individuals from 20 populations. Individuals with genotype calls at <90% of markers were eliminated. Two individual from the Birhor (BIR) and one individual each from the populations Korwa (KOR), Onge (ONG) and Ho were excluded because of relatedness closer to second cousin, inferred by high IBD. While choosing between a relative pair thus identified, we have retained the individual with higher proportion of genotype calls. Markers with minor allele frequency <5% in one or more populations or those that deviated from HWE ($p < 0.001$) were excluded. The final data set comprised data on 367 individuals and 803570 markers.

X-Chromosome haplotyping:

As mentioned in the main text, females in the samples were identified using X-chromosome data. In order to infer the X-chromosome haplotypes for each female individual we used Shapeit2 (30,31).

Sex-Bias in admixture:

Sex bias in ancestry contributions was evaluated by selecting only females (to ensure we compare a diploid X chromosome to diploid autosomes), and running ADMIXTURE with $K = 4$ on the X chromosome and autosomes separately. The Wilcoxon signed rank test, a non-parametric version of the paired Student's t-test that does not require the normality assumption, was applied to assess the significance of the difference in X and autosomal ancestry proportions.

The distance matrix for the phylogenetic tree on the phased X-chromosome haplotype was generated using the ‘complete linkage’ method in hierarchical cluster analysis. The clustering and the dendrogram plotting was done using R 2.12.2 (<https://www.r-project.org/>).

Distribution of ancestral block lengths (ABL)

For all the 16 admixed mainland populations (except the Khatri (KSH), Paniya (PNY), Birhor (BIR) and Jamatia (JAM) which were used as reference, ancestral block segments were inferred for each individual haplotype. We calculated the mean and variance from the distribution of the observed ABLs belonging to each of the 3 ancestral components, except the major one, within a population. That was then compared with an exponential distribution with the same mean. We used the non-parametric Kolmogorov-Smirnov test to compare the distributions.