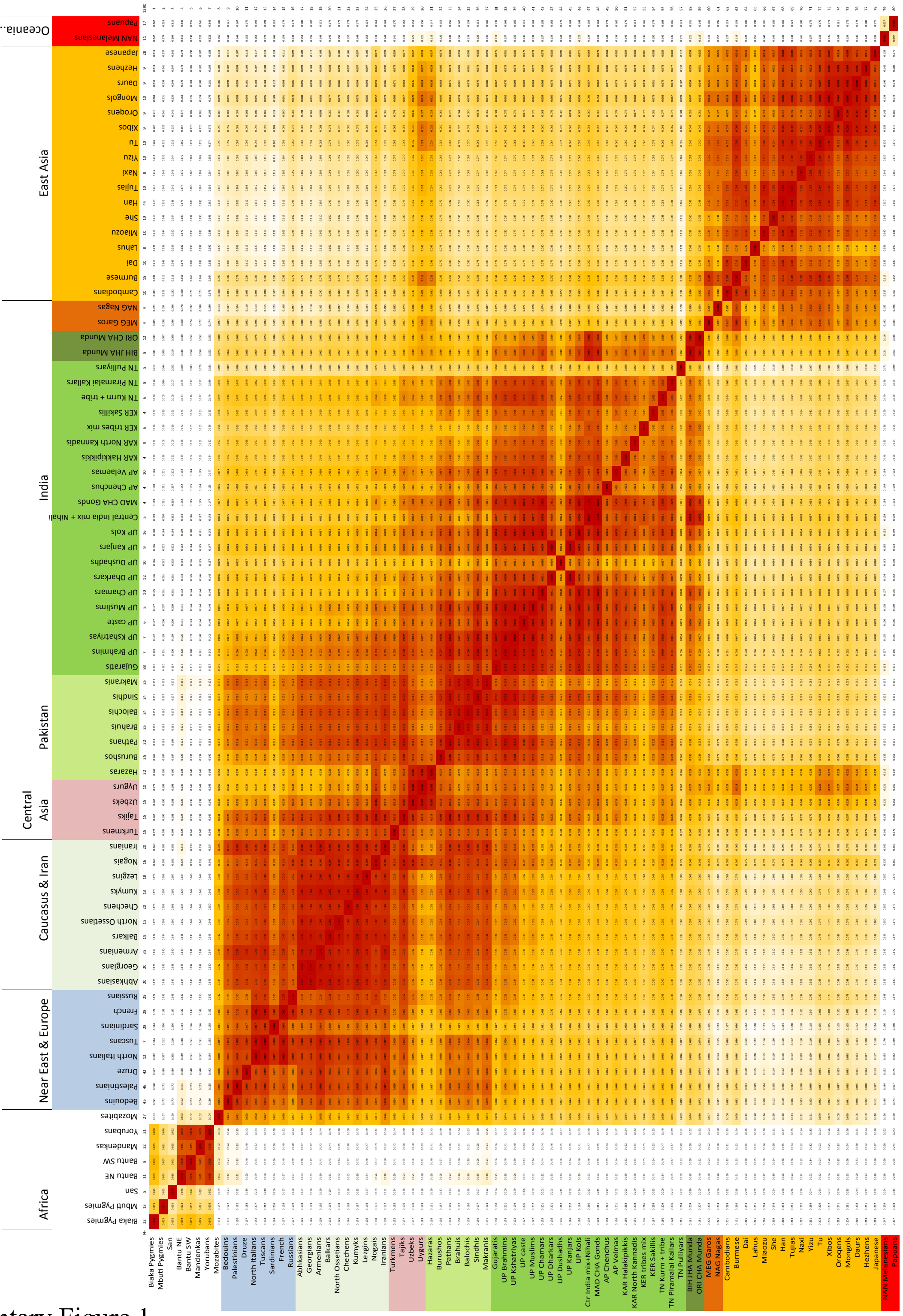
The American Journal of Human Genetics, Volume *89*
Supplemental Data

Shared and Unique Components of Human Population Structure and Genome-Wide Signals of Positive Selection in South Asia

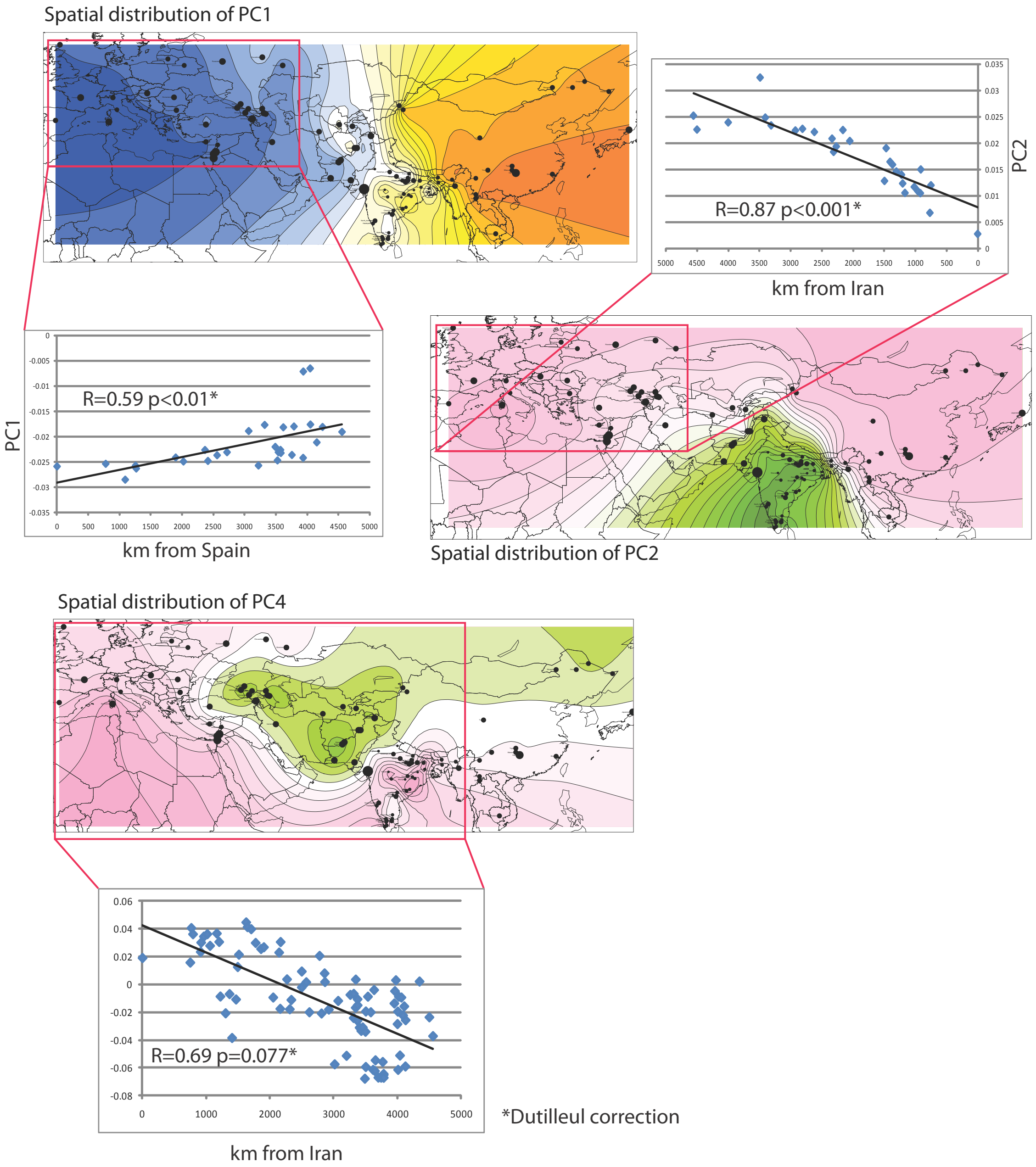
Mait Metspalu, Irene Gallego Romero, Bayazit Yunusbayev, Gyaneshwer Chaubey, Chandana Basu Mallick, Georgi Hudjashov, Mari Nelis, Reedik Mägi, Ene Metspalu, Maido Remm, Ramasamy Pitchappan, Lalji Singh, Kumarasamy Thangaraj, Richard Villems, and Toomas Kivisild



Supplementary Figure 1

Matrix of pairwise mean Fst values of the studied populations.

UP –Uttar Pradesh, MAD Madhya Pradesh, CHA – Chattisgarh, AP –Andhra Pradesh, KAR – Karnataka, KER – Kerala, TN – Tamil Nadu, BIH –Bihar, ORI –Orissa, MEG –Megalaya, NAG – Nagaland. Central Indiamix + Nihali and Gonds from Madhya Pradesh and Chattisgarh are composites of regional groupings of samples from different populations which makes the negative Fst uninformative.

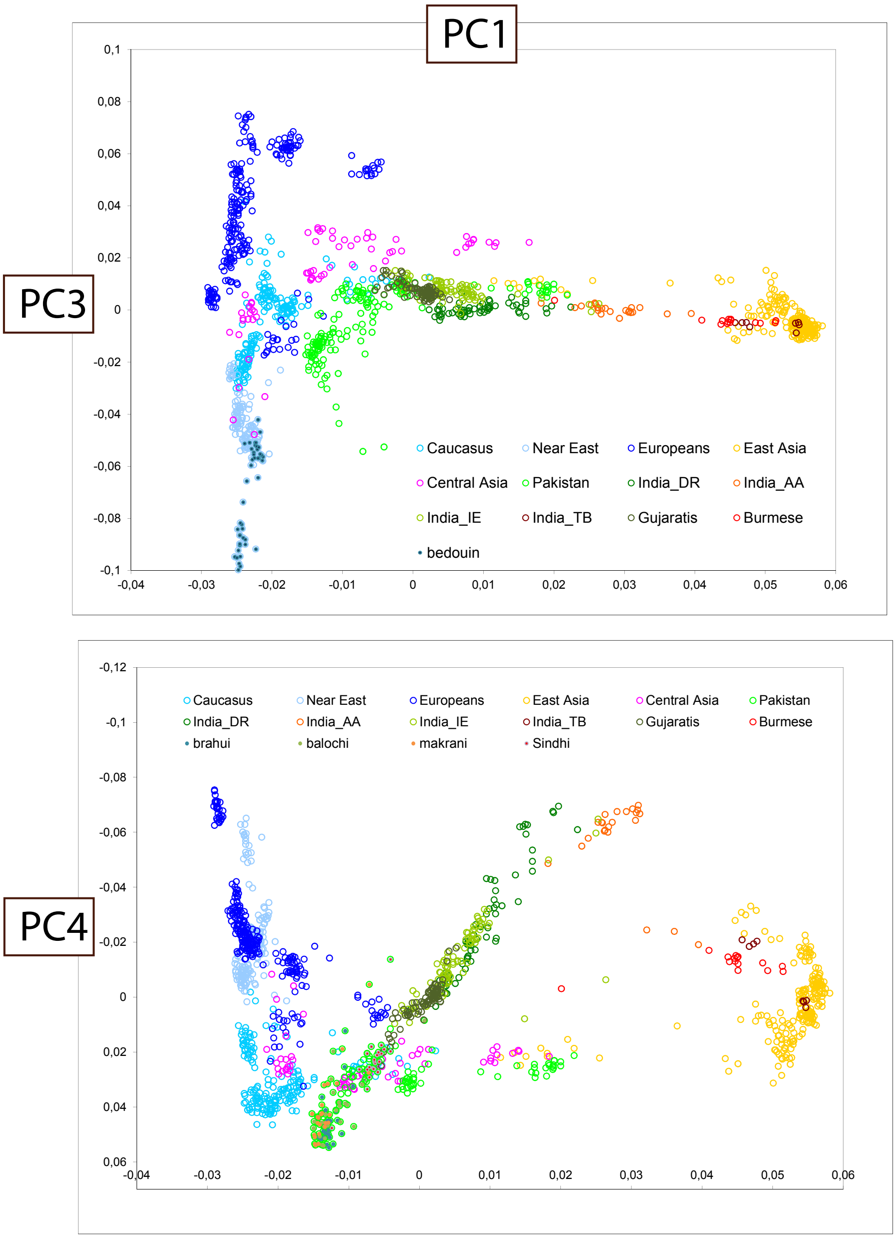


*Dutilleul correction

Supplementary Figure 2

Geographic representation of PC1, 2 and 4.

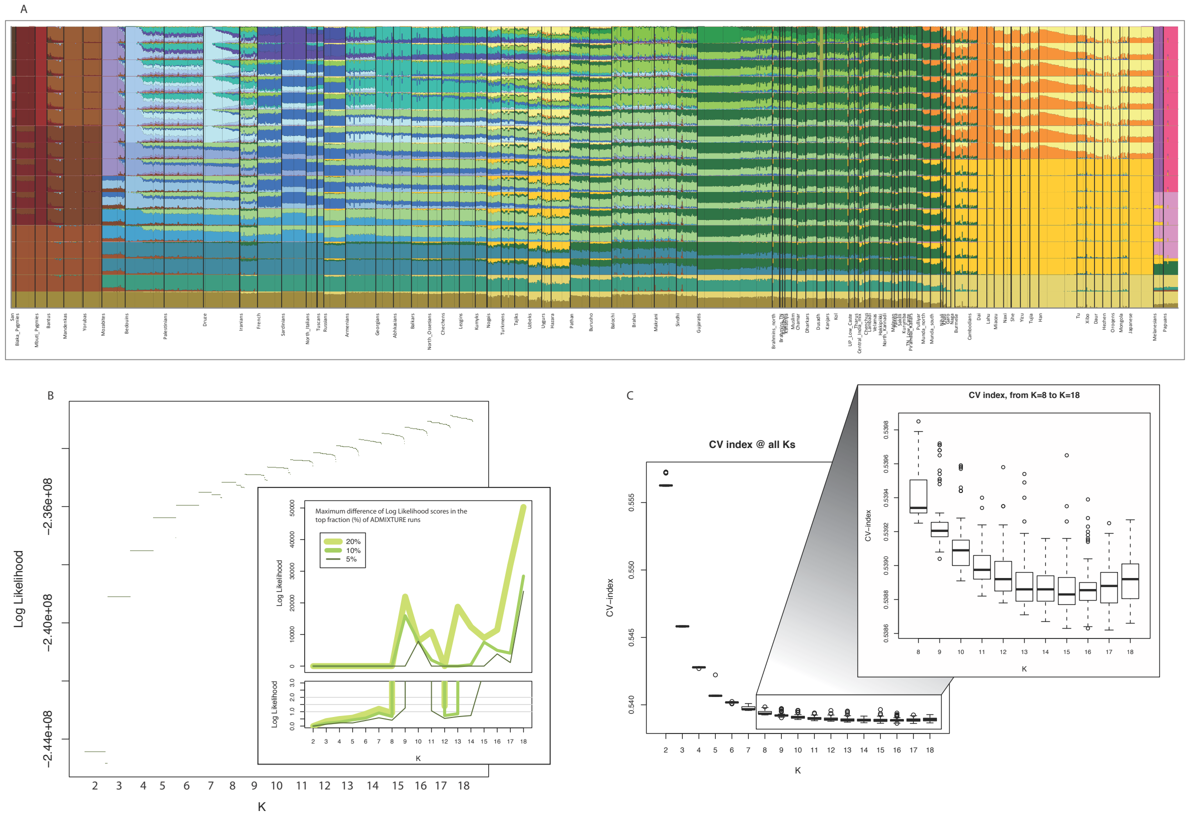
Black dots indicate the geographic locations of the populations their diameters represent sample sizes. Red rectangles show subsets of populations for which the spatial correlation (modified T test in Passage 2, see methods) tests are applied to. In case of PC4 the geographic patchiness of the data was explained by spatial autocorrelation (shown in correlogram next to the map) as the modified T test lost significance due to similar and highly dissimilar PC4 values in India and the Caucasus which are equidistant from Baluchistan.



Supplementary Figure 3

PC 3 and 4 plotted against PC1 in the Eurasian set of populations.

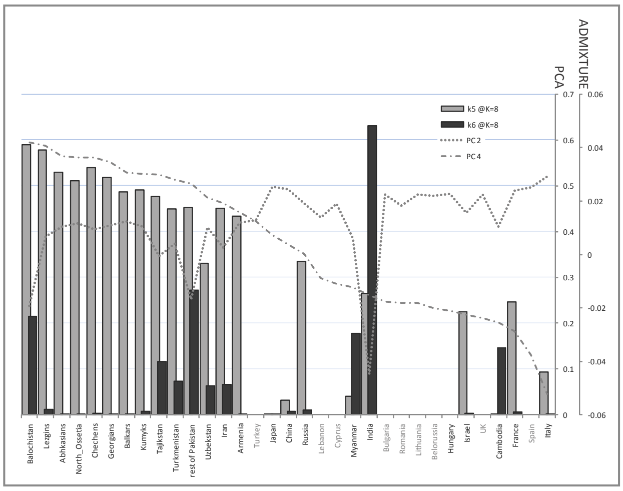
See text for explanation for differential population highlighting.



Supplementary Figure 4

ADMIXTURE analysis form K=2 to K=18

a) ADMIXTURE plots from K=2 to K=18. At each K the run (out of 100 runs) with the highest log-likelihood is plotted. Each vertical column represents one sample and represents its probability to have ancestry in the constructed ancestral populations differentiated by colors. b) log-likelihood scores (LLs) of all the 18 X 100 runs of ADMIXTURE. Note that while at low values of K all runs arrive at the same or very similar LLs, at high K-s the LLs vary. Inset shows the extent of this variation in the fractions (5%, 10%, 20%) of runs that reached the highest LLs. c) Box and whiskers plot of the cross validation indexes of all 1800 runs of ADMIXTURE.

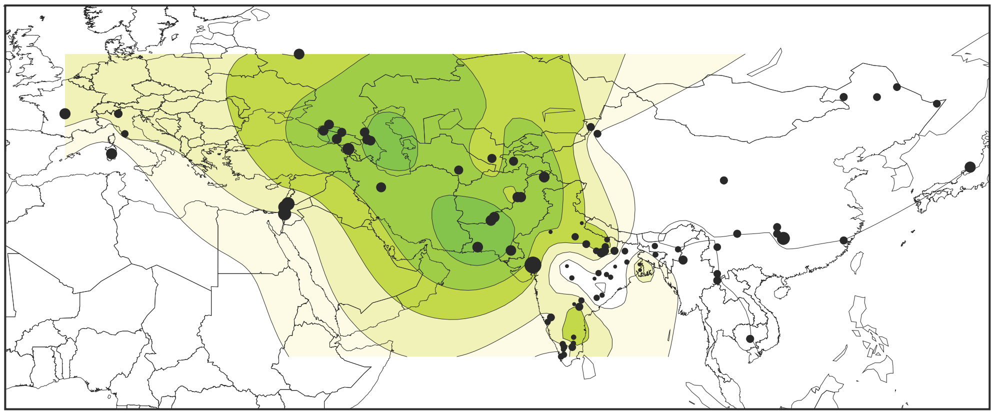


Supplementary Figure 5

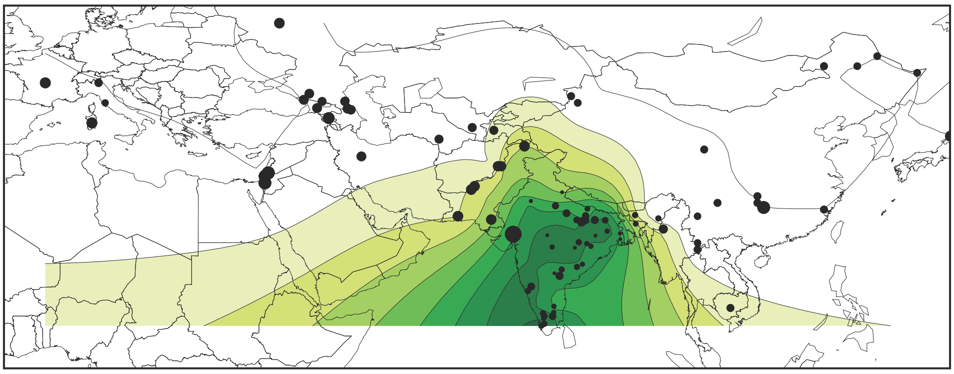
Average membership in k5 and k6 @ K=8 and population averages for PC2 and PC4 (rightmost scale).

Populations shown in grey were not included in the ADMIXTURE analysis.

Geographic spread of membership in k5 @ K=8



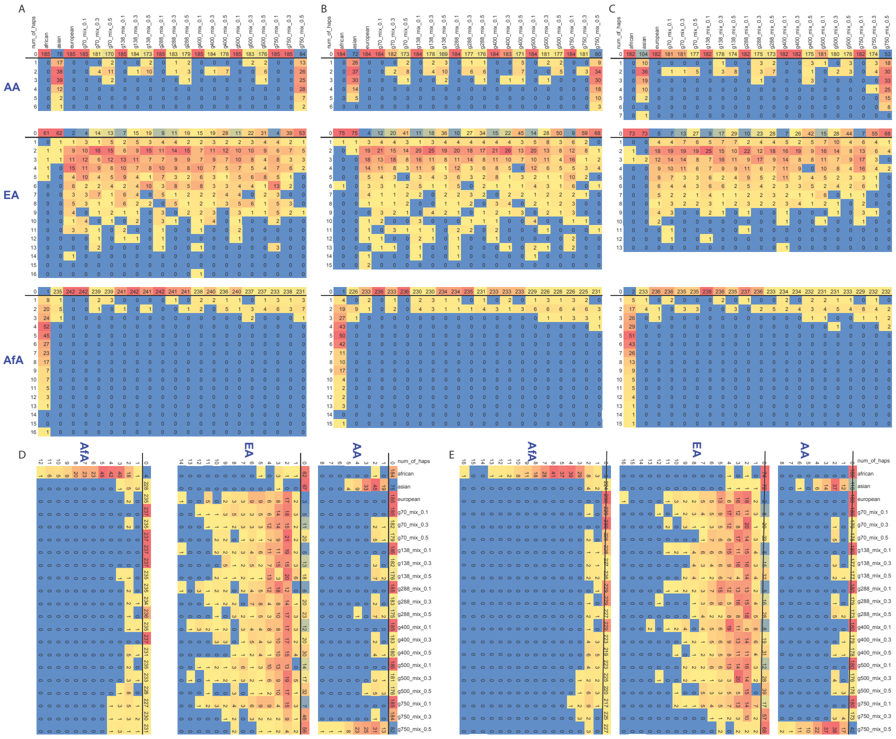
Geographic spread of membership in k6 @ K=8



Supplementary Figure 6

Geographic representation of membership in k5 and k6 @ K=8.

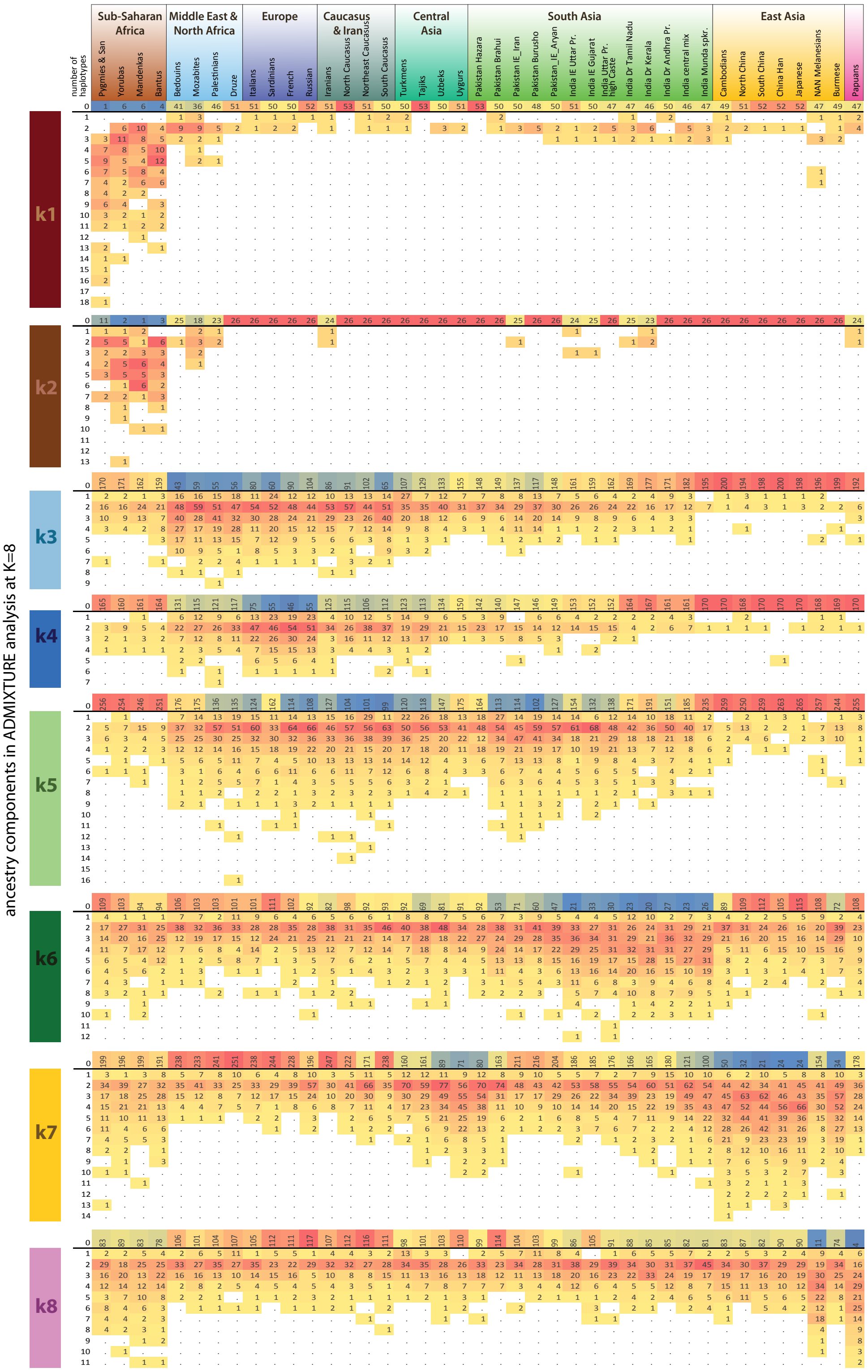
Black dots indicate the geographic locations of the populations their diameters represent sample sizes.



Supplementary Figure 7.

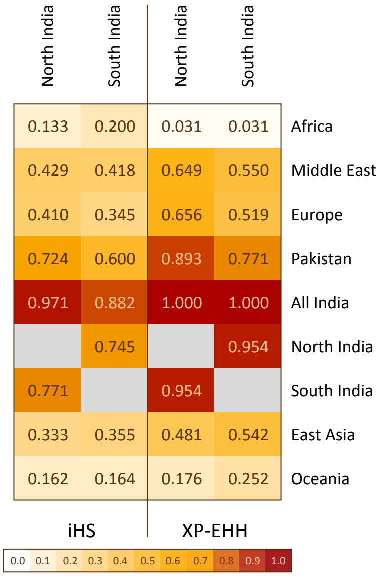
Haplotype diversity flanking ancestry associated SNPs in simulated populations.

For each ‘Asian’, ‘European’ and ‘African’ ancestry associated SNP, 260-kb genomic windows were defined and haplotypes bearing associated alleles were counted. Haplotypes bearing alternative alleles not associated with given ancestry component were omitted. Genomic windows were then binned by the number of observed haplotypes (numbers shown to the left of each row). Color filled cells show how many genomic windows with particular number of haplotypes were observed in a given population. Results for each simulated population (names are given on the top) are given in columns. Cell color changes from blue to white and red with increasing number of observed windows. Results for three ancestral populations (African, Asian and European) and 18 admixed populations are presented. Admixed populations were generated by simulating gene flow events from Asian to European population at different times in the past. Three different gene flow events replacing 10%, 30% and 50% of sequences were simulated. Parameters of gene flow event that produce given admixed population are given in population names shown on the top. For example, “g70_mix_0.3” means that gene flow event occurred 70 generations ago and Asian population contributed 30% of sequences, while remaining 70% of sequences are European. To show the extent of variation between different simulation runs, we present results for five independent runs started by different random seeds on separate panels (a, b, c, d, e). AA – “Asian” ancestry; EA – “European” ancestry; AfA – “African” ancestry



Supplementary Figure 8. **Haplotype diversity in African, West Eurasian, South Asian, Southeast Asian and Oceanian populations in genomic windows (0.26-cM) surrounding SNPs associated with eight ancestry components emerged from the ADMIXTURE analysis (K=8).**

Results for each ancestry component are organized in sections of rows and the eight components are indicated by color and number as in Figure 2b. In each genomic window only haplotypes with associated alleles were counted. Data is shown by populations (organized continentally in columns) and each color-coded cell of the matrix shows the number of genomic windows with particular number of haplotypes.

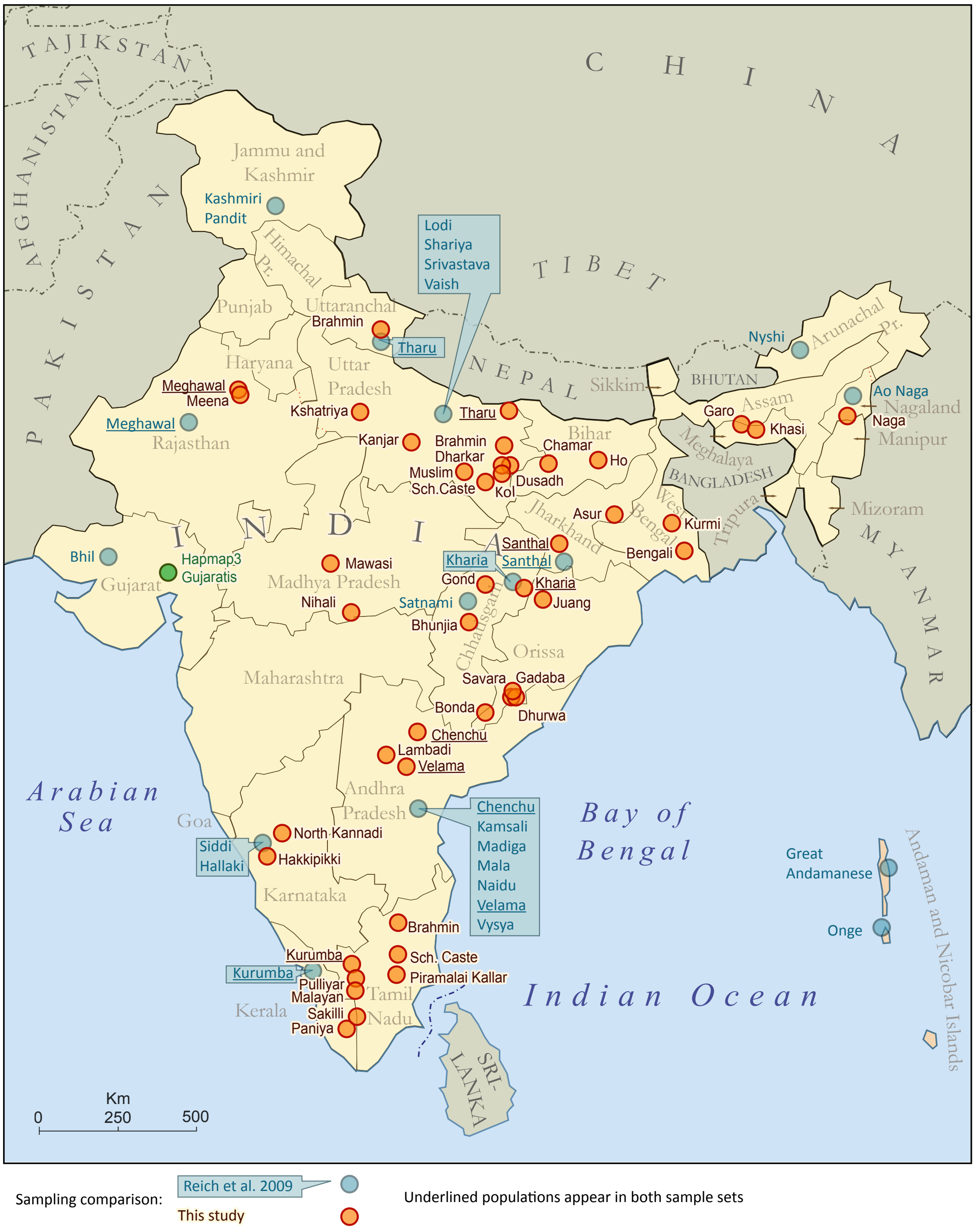


	iHS		XP-EHH	
	North India	South India	North India	South India
Africa	0.133	0.200	0.031	0.031
Middle East	0.429	0.418	0.649	0.550
Europe	0.410	0.345	0.656	0.519
Pakistan	0.724	0.600	0.893	0.771
All India	0.971	0.882	1.000	1.000
North India		0.745		0.954
South India	0.771		0.954	
East Asia	0.333	0.355	0.481	0.542
Oceania	0.162	0.164	0.176	0.252

Supplementary figure 9.

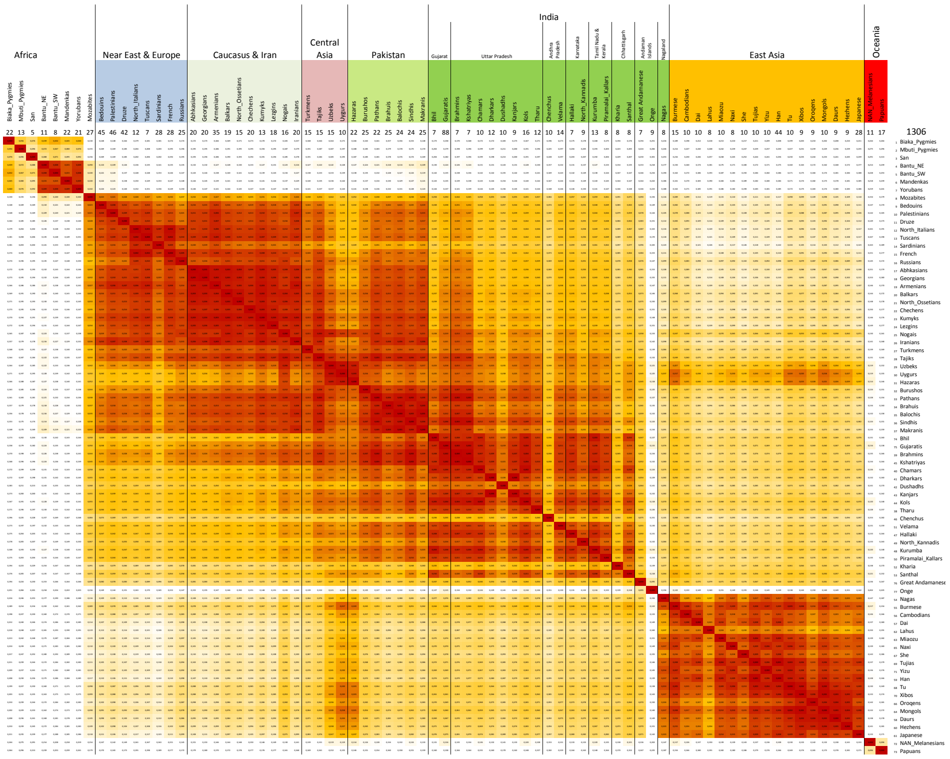
iHS and XP-EHH signal sharing between North and South India and continental populations.

The fraction of signals found in the top 1% of test scores in population *i* and the top 5% of population *j* is given in cell (*i*,*j*). Africa refers to Yoruba, Mandenka and Bantu individuals from the HGDP-CEPH panel.



Supplementary figure 10.

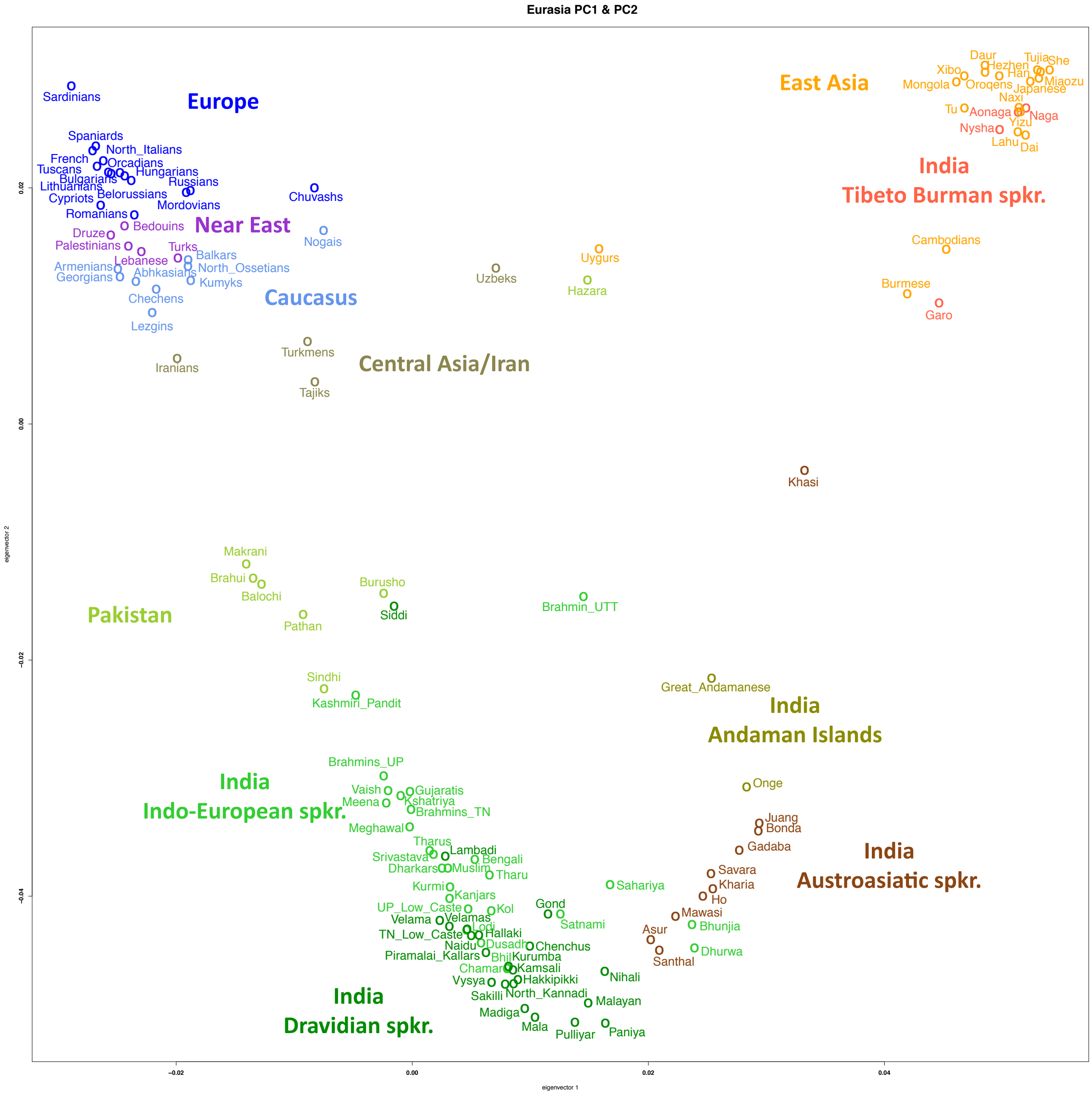
Comparison of sampling locations and populations in the current study and in the study published by Reich et al. (Reich et al. 2009)



Supplementary figure 11.

Matrix of pairwise mean Fst values of the studied populations analyzed together with genotyping data from Reich et al 2009.

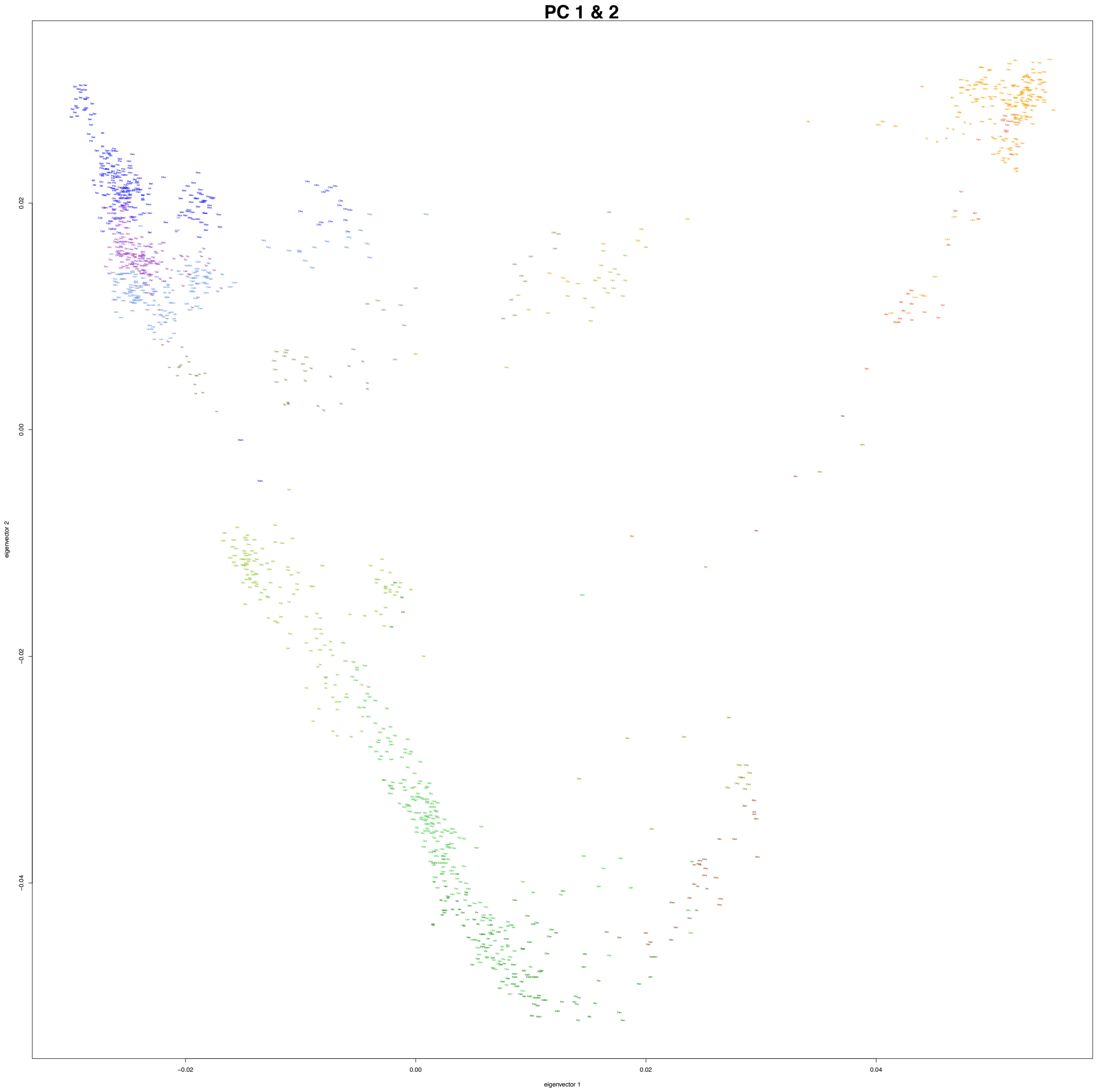
Only populations with N>7 are included. See Table S1 for sample details.



Supplementary figure 12.

The population means of the first and second principal components in the combined dataset including also data from Reich et al 2009.

See Table S1 for sample details.



Supplementary figure 13.

The first and second principal components in the combined dataset including also data from Reich et al 2009.

See Table S1 and Table S6 for sample details and population name aberrations. See Figure S12 for color code.