doi: 10.1038/nature08365

Table S1. Autosomal F_{ST} among 25 Indian groups (no inbreeding correction)

			Kashmiri Pandit	Vaish	Srivastava	Sahariya	Lodi	Satnami	Bhil	Tharu	Meghawal	Vysya	Naidu	Velama	Madiga	Mala	Kamsali	Chenchu	Kurumba	Hallaki	Santhal	Kharia	Nyshi	Ao Naga	Siddi	Onge	Great Andamanese
Kashmiri Pandit	5	Kashmir		0.0005	0.0058	0.0228	0.0085	0.0134	0.0113	0.0049	0.0039	0.0190	0.0112	0.0116	0.0147	0.0150	0.0141	0.0599	0.0122	0.0136	0.0236	0.0324	0.0824	0.0860	0.0772	0.1335	0.0522
Vaish	4	Uttar Pradesh	0.0007		0.0035	0.0166	0.0050	0.0083	0.0061	0.0009	0.0034	0.0141	0.0091	0.0078	0.0103	0.0108	0.0096	0.0563	0.0064	0.0101	0.0163	0.0251	0.0801	0.0836	0.0746	0.1277	0.0484
Srivastava	2	Uttar Pradesh	0.0012	0.0013		0.0142	0.0054	0.0062	0.0059	0.0029	0.0058	0.0120	0.0076	0.0088	0.0068	0.0103	0.0091	0.0554	0.0069	0.0094	0.0168	0.0233	0.0786	0.0835	0.0774	0.1292	0.0477
Sahariya	4	Uttar Pradesh	0.0010	0.0010	0.0015		0.0130	0.0089	0.0113	0.0095	0.0187	0.0203	0.0146	0.0202	0.0119	0.0113	0.0141	0.0636	0.0113	0.0176	0.0111	0.0138	0.0667	0.0692	0.0838	0.1205	0.0443
Lodi	5	Uttar Pradesh	0.0007	0.0008	0.0012	0.0009		0.0069	0.0058	0.0029	0.0071	0.0116	0.0065	0.0092	0.0065	0.0062	0.0079	0.0563	0.0064	0.0081	0.0146	0.0213	0.0794	0.0824	0.0788	0.1281	0.0488
Satnami	4	Madhya Pradesh	0.0009	0.0010	0.0013	0.0010	0.0009		0.0057	0.0038	0.0099	0.0140	0.0087	0.0118	0.0053	0.0062	0.0069	0.0566	0.0059	0.0104	0.0078	0.0125	0.0665	0.0695	0.0789	0.1204	0.0419
Bhil	7	Gujarat	0.0007	0.0008	0.0011	0.0007	0.0006	0.0007		0.0022	0.0082	0.0129	0.0081	0.0092	0.0053	0.0052	0.0077	0.0560	0.0036	0.0102	0.0094	0.0170	0.0765	0.0796	0.0805	0.1235	0.0447
Tharu	9	Uttarkhand	0.0006	0.0006	0.0011	0.0007	0.0005	0.0007	0.0004		0.0049	0.0108	0.0052	0.0078	0.0047	0.0049	0.0055	0.0524	0.0021	0.0072	0.0080	0.0150	0.0701	0.0740	0.0753	0.1204	0.0409
Meghawal	5	Rajasthan	0.0007	0.0008	0.0012	0.0009	0.0007	0.0008	0.0006	0.0005		0.0158	0.0090	0.0107	0.0103	0.0110	0.0108	0.0592	0.0096	0.0117	0.0192	0.0279	0.0818	0.0858	0.0777	0.1300	0.0509
Vysya	5	Andhra Pradesh	0.0008	0.0009	0.0013	0.0010	0.0008	0.0009	0.0007	0.0006	0.0007		0.0147	0.0155	0.0135	0.0121	0.0119	0.0646	0.0128	0.0164	0.0211	0.0280	0.0858	0.0893	0.0877	0.1354	0.0563
Naidu	4	Andhra Pradesh	0.0009	0.0010	0.0014	0.0010	0.0009	0.0010	0.0008	0.0007	0.0008	0.0009		0.0083	0.0077	0.0073	0.0082	0.0599	0.0082	0.0109	0.0166	0.0230	0.0809	0.0843	0.0812	0.1297	0.0515
Velama	4	Andhra Pradesh	0.0008	0.0009	0.0014	0.0010	0.0009	0.0010	0.0007	0.0007	0.0008	0.0009	0.0009		0.0097	0.0103	0.0105	0.0626	0.0100	0.0114	0.0200	0.0288	0.0834	0.0871	0.0802	0.1318	0.0537
Madiga	4	Andhra Pradesh	0.0009	0.0009	0.0013	0.0009	0.0008	0.0009	0.0007	0.0006	0.0008	0.0009	0.0009	0.0009		0.0038	0.0063	0.0576	0.0046	0.0094	0.0115	0.0185	0.0775	0.0798	0.0803	0.1239	0.0469
Mala	3	Andhra Pradesh	0.0010	0.0011	0.0015	0.0012	0.0010	0.0011	0.0009	0.0008	0.0010	0.0010	0.0011	0.0011	0.0010		0.0060	0.0582	0.0040	0.0089	0.0105	0.0169	0.0780	0.0805	0.0830	0.1250	0.0475
Kamsali	4	Andhra Pradesh	0.0009	0.0009	0.0014	0.0010	0.0008	0.0009	0.0007	0.0006	0.0008	0.0009	0.0009	0.0009	0.0009	0.0011		0.0090	0.0068	0.0101	0.0139	0.0210	0.0790	0.0824	0.0824	0.1269	0.0492
Chenchu	6	Andhra Pradesh	0.0013	0.0014	0.0017	0.0014	0.0013	0.0014	0.0013	0.0012	0.0013	0.0014	0.0014	0.0014	0.0014	0.0015	0.0015		0.0543	0.0612	0.0577	0.0655	0.1205	0.1233	0.1245	0.1711	0.0918
Kurumba	9	Kerala	0.0006	0.0007	0.0011	0.0007	0.0006	0.0007	0.0004	0.0004	0.0006	0.0006	0.0007	0.0007	0.0006	0.0009	0.0007	0.0012		0.0095	0.0070	0.0136	0.0747	0.0767	0.0795	0.1185	0.0409
Hallaki	7	Karnataka	0.0007	0.0007	0.0012	0.0008	0.0007	0.0008	0.0005	0.0005	0.0007	0.0007	0.0008	0.0008	0.0007	0.0009	0.0007	0.0013	0.0005		0.0170	0.0243	0.0816	0.0848	0.0828	0.1295	0.0508
Santhal	7	Jharkhand	0.0008	0.0008	0.0012	0.0007	0.0007	0.0007	0.0006	0.0005	0.0007	0.0008	0.0008	0.0009	0.0007	0.0009	0.0008	0.0013	0.0005	0.0006		0.0045	0.0638	0.0674	0.0865	0.1152	0.0386
Kharia	6	Madhya Pradesh	0.0008	0.0009	0.0013	0.0008	0.0008	0.0008	0.0007	0.0006	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	0.0008	0.0014	0.0006	0.0007	0.0006		0.0616	0.0633	0.0907	0.1197	0.0411
Nyshi	4	Arunachal Pradesh	0.0013	0.0013	0.0017	0.0013	0.0013	0.0013	0.0011	0.0011	0.0013	0.0013	0.0014	0.0014	0.0013	0.0015	0.0013	0.0018	0.0011	0.0012	0.0011	0.0011		0.0215	0.1315	0.1559	0.0729
Ao Naga	4	Nagaland	0.0014	0.0014	0.0018	0.0013	0.0013	0.0013	0.0012	0.0012	0.0013	0.0014	0.0015	0.0014	0.0014	0.0016	0.0014	0.0018	0.0012	0.0013	0.0012	0.0012	0.0011		0.1338	0.1584	0.0752
Siddi	4	Karnataka	0.0017	0.0016	0.0020	0.0018	0.0017	0.0017	0.0017	0.0016	0.0016	0.0018	0.0017	0.0017	0.0018	0.0018	0.0018	0.0019	0.0016	0.0016	0.0016	0.0016	0.0018	0.0018		0.1748	0.1079
Onge	9	Andaman & Nicobar	0.0013	0.0014	0.0017	0.0014	0.0013	0.0013	0.0012	0.0012	0.0013	0.0013	0.0014	0.0015	0.0013	0.0016	0.0014	0.0018	0.0011	0.0012	0.0012	0.0013	0.0016	0.0016	0.0017		0.0905
Great Andamanese	7	Andaman & Nicobar	0.0010	0.0010	0.0014	0.0010	0.0009	0.0010	0.0008	0.0008	0.0010	0.0010	0.0010	0.0010	0.0010	0.0012	0.0010	0.0014	0.0008	0.0008	0.0008	0.0008	0.0012	0.0013	0.0017	0.0012	

Note: F_{ST} values are presented in the top right of the matrix, and standard errors are presented in the bottom left.

Table S2. Autosomal F_{ST} among 25 Indian groups (inbreeding correction)

			Kashmiri Pandit	Vaish	Srivastava	Sahariya	Lodi	Satnami	Bhil	Tharu	Meghawal	Vysya	Naidu	Velama	Madiga	Mala	Kamsali	Chenchu	Kurumba	Hallaki	Santhal	Kharia	Nyshi	Ao Naga	Siddi	Onge	Great Andamanese
Kashmiri Pandit	5	Kashmir		0.0023	0.0059	0.0234	0.0091	0.0142	0.0125	0.0059	0.0052	0.0194	0.0089	0.0111	0.0150	0.0158	0.0122	0.0618	0.0124	0.0119	0.0252	0.0337	0.0826	0.0860	0.0783	0.1349	0.0550
Vaish	4	Uttar Pradesh	0.0008		0.0036	0.0173	0.0057	0.0092	0.0074	0.0020	0.0048	0.0147	0.0070	0.0074	0.0107	0.0117	0.0078	0.0583	0.0067	0.0084	0.0180	0.0265	0.0803	0.0837	0.0758	0.1293	0.0513
Srivastava	2	Uttar Pradesh	0.0014	0.0015		0.0133	0.0045	0.0054	0.0055	0.0023	0.0056	0.0108	0.0037	0.0068	0.0054	0.0096	0.0055	0.0558	0.0055	0.0060	0.0168	0.0230	0.0772	0.0819	0.0771	0.1291	0.0490
Sahariya	4	Uttar Pradesh	0.0010	0.0010	0.0016		0.0125	0.0087	0.0114	0.0094	0.0189	0.0197	0.0112	0.0187	0.0111	0.0110	0.0110	0.0644	0.0104	0.0147	0.0117	0.0140	0.0658	0.0682	0.0839	0.1209	0.0461
Lodi	5	Uttar Pradesh	0.0007	0.0008	0.0015	0.0009		0.0066	0.0059	0.0028	0.0073	0.0110	0.0030	0.0076	0.0057	0.0058	0.0048	0.0571	0.0055	0.0051	0.0151	0.0215	0.0784	0.0814	0.0788	0.1284	0.0506
Satnami	4	Madhya Pradesh	0.0009	0.0010	0.0015	0.0010	0.0009		0.0061	0.0039	0.0103	0.0135	0.0055	0.0104	0.0046	0.0061	0.0040	0.0577	0.0053	0.0077	0.0085	0.0129	0.0657	0.0686	0.0792	0.1210	0.0438
Bhil	7	Gujarat	0.0007	0.0008	0.0013	0.0008	0.0006	0.0007		0.0027	0.0091	0.0129	0.0054	0.0083	0.0051	0.0055	0.0053	0.0574	0.0033	0.0079	0.0105	0.0178	0.0762	0.0791	0.0811	0.1245	0.0471
Tharu	9	Uttarkhand	0.0006	0.0006	0.0013	0.0008	0.0005	0.0007	0.0004		0.0056	0.0106	0.0022	0.0066	0.0043	0.0050	0.0028	0.0536	0.0017	0.0047	0.0089	0.0156	0.0695	0.0733	0.0757	0.1212	0.0430
Meghawal	5	Rajasthan	0.0007	0.0008	0.0014	0.0010	0.0007	0.0008	0.0006	0.0006		0.0159	0.0064	0.0099	0.0102	0.0114	0.0084	0.0608	0.0094	0.0096	0.0204	0.0288	0.0816	0.0854	0.0784	0.1311	0.0534
Vysya	5	Andhra Pradesh	0.0008	0.0009	0.0016	0.0010	0.0008	0.0009	0.0007	0.0006	0.0008		0.0111	0.0138	0.0125	0.0116	0.0087	0.0653	0.0118	0.0133	0.0215	0.0281	0.0847	0.0881	0.0877	0.1357	0.0579
Naidu	4	Andhra Pradesh	0.0009	0.0010	0.0016	0.0010	0.0009	0.0010	0.0008	0.0007	0.0008	0.0009		0.0038	0.0040	0.0041	0.0022	0.0579	0.0044	0.0051	0.0142	0.0204	0.0772	0.0804	0.0787	0.1272	0.0504
Velama	4	Andhra Pradesh	0.0009	0.0009	0.0015	0.0010	0.0009	0.0010	0.0008	0.0007	0.0008	0.0009	0.0009		0.0078	0.0090	0.0063	0.0624	0.0081	0.0074	0.0194	0.0280	0.0814	0.0850	0.0793	0.1312	0.0544
Madiga	4	Andhra Pradesh	0.0009	0.0009	0.0014	0.0010	0.0008	0.0009	0.0007	0.0006	0.0008	0.0009	0.0009	0.0009		0.0031	0.0028	0.0581	0.0034	0.0062	0.0117	0.0184	0.0762	0.0784	0.0801	0.1239	0.0483
Mala	3	Andhra Pradesh	0.0011	0.0011	0.0017	0.0012	0.0010	0.0011	0.0009	0.0009	0.0010	0.0010	0.0012	0.0012	0.0011		0.0030	0.0592	0.0033	0.0061	0.0112	0.0173	0.0772	0.0795	0.0833	0.1256	0.0494
Kamsali	4	Andhra Pradesh	0.0009	0.0009	0.0016	0.0010	0.0008	0.0009	0.0007	0.0006	0.0008	0.0009	0.0009	0.0009	0.0008	0.0011		0.0090	0.0033	0.0045	0.0118	0.0186	0.0755	0.0788	0.0801	0.1247	0.0484
Chenchu	6	Andhra Pradesh	0.0013	0.0014	0.0019	0.0014	0.0013	0.0014	0.0013	0.0012	0.0013	0.0014	0.0014	0.0014	0.0014	0.0015	0.0015		0.0547	0.0596	0.0595	0.0670	0.1209	0.1235	0.1257	0.1727	0.0948
Kurumba	9	Kerala	0.0006	0.0007	0.0013	0.0007	0.0006	0.0007	0.0004	0.0004	0.0006	0.0006	0.0007	0.0007	0.0006	0.0009	0.0006	0.0012		0.0062	0.0071	0.0134	0.0734	0.0753	0.0792	0.1185	0.0423
Hallaki	7	Karnataka	0.0007	0.0007	0.0014	0.0008	0.0007	0.0008	0.0005	0.0005	0.0007	0.0007	0.0008	0.0008	0.0007	0.0010	0.0007	0.0013	0.0005		0.0151	0.0220	0.0783	0.0814	0.0807	0.1275	0.0501
Santhal	7	Jharkhand	0.0008	0.0008	0.0014	0.0008	0.0007	0.0007	0.0006	0.0005	0.0007	0.0008	0.0008	0.0009	0.0008	0.0010	0.0008	0.0013	0.0005	0.0006		0.0057	0.0638	0.0673	0.0874	0.1166	0.0414
Kharia	6	Madhya Pradesh	0.0009	0.0009	0.0015	0.0009	0.0008	0.0008	0.0007	0.0006	0.0008	0.0009	0.0009	0.0009	0.0009	0.0011	0.0009	0.0014	0.0006	0.0007	0.0006		0.0613	0.0629	0.0914	0.1208	0.0436
Nyshi	4	Arunachal Pradesh	0.0013	0.0013	0.0019	0.0013	0.0013	0.0013	0.0012	0.0011	0.0013	0.0013	0.0014	0.0014	0.0013	0.0015	0.0013	0.0018	0.0011	0.0012	0.0011	0.0011		0.0198	0.1311	0.1557	0.0742
Ao Naga	4	Nagaland	0.0014	0.0014	0.0020	0.0014	0.0013	0.0013	0.0013	0.0012	0.0013	0.0014	0.0015	0.0015	0.0014	0.0016	0.0014	0.0018	0.0012	0.0013	0.0012	0.0013	0.0012		0.1334	0.1581	0.0764
Siddi	4	Karnataka	0.0017	0.0017	0.0021	0.0019	0.0017	0.0017	0.0017	0.0016	0.0017	0.0018	0.0018	0.0017	0.0018	0.0018	0.0018	0.0020	0.0016	0.0017	0.0016	0.0017	0.0019	0.0019		0.1756	0.1099
Onge	9	Andaman & Nicobar	0.0013	0.0014	0.0018	0.0014	0.0013	0.0013	0.0012	0.0012	0.0013	0.0013	0.0014	0.0015	0.0013	0.0016	0.0014	0.0018	0.0011	0.0012	0.0012	0.0013	0.0016	0.0017	0.0018		0.0934
Great Andamanese	7	Andaman & Nicobar	0.0010	0.0010	0.0016	0.0010	0.0009	0.0010	0.0008	0.0008	0.0010	0.0010	0.0010	0.0010	0.0010	0.0012	0.0010	0.0015	0.0008	0.0008	0.0008	0.0008	0.0012	0.0013	0.0017	0.0012	

Note: F_{ST} values are presented in the top right of the matrix, and standard errors are presented in the bottom left.

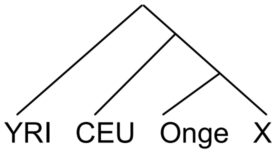
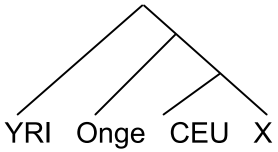
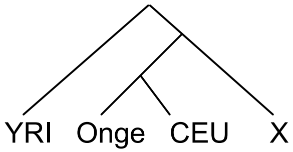
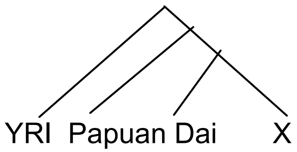
Table S3. Pairwise F_{ST} for combinations of Indian groups

Category of comparison	Details of comparison	No. of groups	Average F_{ST}	Average F_{ST} correcting for inbreeding
All India[†]	All pairs	19	0.0109	0.0100
	Comparing matched groups (both Uttar Pradesh or both Andhra Pradesh and both traditionally upper caste or both traditionally lower or middle caste)	9 pairs	0.0087	0.0069
Restricting to language	Indo-European speaking pairs	9	0.0076	0.0080
	Dravidian speaking pairs	8	0.0096	0.0067
Restricting to caste level	Traditionally upper caste pairs	5	0.0074	0.0061
	Traditionally lower and middle caste pairs	6	0.0010	0.0093
Restricting to a state	Uttar Pradesh pairs	4	0.0096	0.0095
	Andhra Pradesh pairs	6	0.0097	0.0069

* We exclude 6 outlier groups: the Onge, Great Andamanese, Ao Naga, Nyshi, Siddi and Chenchu. Individual pairwise F_{ST} values for all possible pairs of 25 groups are presented in Tables S1 and S2.

[†] The inbreeding corrected average F_{ST} between all pairs of 19 Indian groups (0.0100) is higher than the average F_{ST} between all pairs of 23 European groups in ref. 1 (0.0033). This phenomenon persists when we restrict to pairs of Indian groups of the same traditional caste level that are matched by geographic region (0.0069), and compare this to pairs of European groups that are matched by geographic region (0.0018). For performing a regional analysis of the European data in ref. 1, we defined five European "regions": Scandinavia (Helsinki, Førde, and Uppsala), Northern Europe (Kopenhagen, Rotterdam, Dublin, London and Kiel), Central Europe (Budapest, Lausanne, Augsburg, Innsbruck and Lyon), Eastern Europe (Prague, Belgrade, Bucharest and Warsaw), and Southern Europe (Rome, Lisbon, Madrid, Greece, Ancona and Barcelona).

Table S4. Formal tests for mixture on the Indian Cline (expansion of Table 2 in the main text)

Group (ordered from most ASI-related to most ANI)	No. samples after pruning	Z-score for *3 Population Test* $(P_X-P_{CEU})(P_X-P_{Santhal})$ (negative values indicate violation)	Z-score for *4 Pop Test* $(P_{YRI}-P_{CEU})(P_{Onge}-P_X)$  	Z-score for *4 Pop Test* $(P_{YRI}-P_{Onge})(P_{CEU}-P_X)$  	Z-score for *4 Pop Test* $(P_{YRI}-P_X)(P_{CEU}-P_{Onge})$  	Z-score for *4 Pop Test* $(P_{YRI}-P_{Papuan})(P_{Dai}-P_X)$[††] 
Onge	9	77.3 (not significant)	n/a	n/a	n/a	1.7 (not significant)
Mala	3	-2.5	13.8	20.4	7.1	-9.7
Madiga	4	-2.7	15.1	21.6	6.8	-11.2
Chenchu	6	31.3 (not significant)	16.9	21.2	5.6	-9.7
Kurumba	6	-12.6	17.1	24.5	6.0	-11.8
Bhil	7	-10.6	18.1	23.9	5.0	-13.0
Kamsali	3	-6.5	17.1	20.7	3.5	-10.9
Satnami	3	-5.6	16.7	19.0	3.4	-10.5
Vysya	5	5.4 (not significant)	18.1	21.1	1.8 (not significant[†])	-11.5
Naidu	4	-3.3	18.4	20.2	-0.3 (not significant[†])	-12.8
Lodi	5	-8.9	21.9	20.8	-1.1 (not significant[†])	-12.9
Tharu	5	-20.6	20.6	21.5	-1.4 (not significant[†])	-14.3
Velama	4	-3.2	19.4	17.2	-2.7	-14.4
Srivastava	2	-7.5	19.8	14.1	-5.5	-11.9
Meghawal	5	-13.3	24.8	18.0	-8.1	-15.6
Vaish	4	-22.0	25.7	18.0	-10.1	-15.6
Kashmiri Pandit	5	-20.6	30.7	17.0	-15.7	-17.1
Sindhi [*]	10	-26.3	27.8	13.0	-18.3	-20.7
Pathan [*]	15	-34.3	30.8	14.3	-21.2	-20.0

[*] Tests using HGDP samples use the reduced set of 119,744 autosomal SNPs, while all other tests use 560,123 autosomal SNPs.

[†] Four groups in the middle of the Indian Cline (from the Vysya to the Tharu) give non-significant Z-scores for the *4 Population Test* for the third tree topology ((YRI,X),(CEU,Onge)), which we hypothesize reflects the fact that two other topologies are both present (due to ancient mixture) and balance in their contribution to the *4 Population Test* statistic. However, we can show by another argument that this topology is not consistent with the data in the absence of mixture. Fitting this topology to the data and using a Weighted Block Jackknife to obtain a standard error, we estimate that the internal branches have negative length with high statistical significance (normally distributed Z-scores of -34 (Vysya), -34 (Naidu), -39 (Lodi) and -38 (Tharu) (Note S3)). Since the internal branch length is proportional to genetic drift under the null hypothesis of a correct topology, the topology cannot be correct.

[††] The Onge are the only ASI-related group with no evidence at all of ANI-related mixture, as assessed by a *4 Population Test* of the topology ((YRI,Papuan),(Dai,X)) in the last column. The f_4 statistic is extremely significantly different from 0 (Z-score << -9 standard deviations) for all Indian Cline groups, but is consistent with 0 (Z = 1.7) for the Onge. Thus, all the Indian Cline group have a component of mixture that the Onge do not.

Table S5. ANI ancestry estimates based on three alternative methods

Group	***f_3* Ancestry Estimation**					***f_4* Ancestry Estimation ***					***Regression Ancestry Estimation* †**
	Auto-somes	**Stand. error**	**X chrom.**	**Stand. error**	**P-value for X-autosome difference**	**Auto-somes**	**Stand. error**	**X chrom.**	**Stand. error**	**P-value for X-autosome difference**	
Mala	38.8%	1.2%	38%	9%	0.46	38.2%	1.7%	40%	13%	0.54	41%
Madiga	40.6%	1.2%	35%	14%	0.33	40.6%	1.7%	49%	13%	0.73	41%
Chenchu	40.7%	1.3%	31%	11%	0.18	42.1%	1.7%	23%	9%	0.021	42%
Bhil	42.9%	1.1%	42%	10%	0.45	42.5%	1.4%	37%	10%	0.30	44%
Satnami	43.0%	1.3%	33%	15%	0.26	43.6%	1.8%	39%	11%	0.35	46%
Kurumba	43.2%	1.1%	28%	10%	0.06	42.3%	1.5%	36%	10%	0.25	43%
Kamsali	44.5%	1.3%	44%	10%	0.50	43.8%	1.7%	49%	18%	0.62	45%
Vysya	46.2%	1.2%	40%	11%	0.29	44.7%	1.7%	44%	10%	0.48	49%
Lodi	49.9%	1.1%	43%	10%	0.25	47.7%	1.6%	47%	8%	0.48	52%
Naidu	50.1%	1.2%	54%	12%	0.62	48.6%	1.6%	54%	11%	0.69	52%
Tharu	51.0%	1.2%	34%	9%	0.03	50.9%	1.5%	35%	9%	0.04	53%
Velama	54.7%	1.3%	53%	11%	0.43	52.4%	1.7%	44%	13%	0.26	57%
Srivastava	56.4%	1.5%	43%	11%	0.11	55.0%	1.9%	47%	15%	0.30	60%
Meghawal	60.3%	1.2%	67%	13%	0.69	57.1%	1.4%	58%	11%	0.53	61%
Vaish	62.6%	1.2%	55%	13%	0.26	60.3%	1.5%	51%	12%	0.23	64%
Kashmiri Pandit	70.6%	1.2%	64%	11%	0.28	69.3%	1.3%	52%	7%	0.004	72%
Sindhi	73.7%	1.1%	81%	12%	0.71	70.7%	1.0%	65%	6%	0.17	78%
Pathan	76.9%	1.1%	83%	11%	0.71	74.2%	0.9%	73%	6%	0.40	81%

* For *f_4 Ancestry Estimation*, we use the statistic f_4(Adygei,Papuan; India,Onge)/f_4(Adygei,Papuan; CEU,Onge) to estimate ANI ancestry proportion, and obtain a standard error for each group by a Block Jackknife. This calculation only analyzes one Indian Cline group at a time, and hence the estimates are not expected to be biased by the outlier-removal procedure we used to eliminate specific groups from the Indian Cline (i.e. Kharia, Santhal, Sahariya and Hallaki).

† For *Regression Ancestry Estimation*, we plot f_4(YRI,Adygei; Onge,India$_k$), a number proportional to ANI ancestry, against f_4(YRI,Onge; Adygei,India$_k$), a number proportional to ASI ancestry. We then use regression analysis over all 18 groups to extrapolate the x-intercept and y-intercept, and interpolate the ANI ancestry proportion for each group (Note S5).

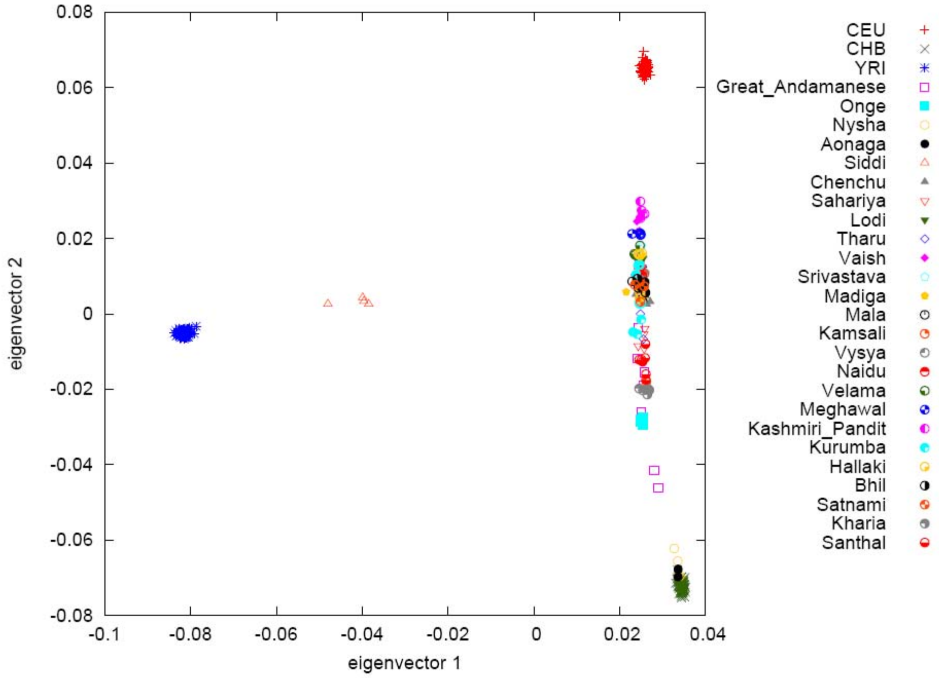
Table S6. mtDNA and Y chromosome data

		Mala	Madiga	Chenchu	Kurumba	Bhil	Kamsali	Satnami	Vysya	Naidu	Lodhi	Tharu	Velama	Srivastava	Meghawal	Vaish	Kash. Pandit	Ao Naga	Kharia	Santhal	Sahariya	Nyshi	Siddi	Hallaki	Onge	Gr. Andaman
mtDNA																										
Samples		37	20	75	68	147	43	28	58	60	49	104	16	35	30	9	61	60	101	110	23	45	94	29	33	9
ASI %		38%	90%	48%	56%	52%	67%	39%	31%	20%	51%	8%	25%	46%	50%	89%	31%	12%	47%	50%	57%	18%	4%	45%	0%	0%
ANI %		0%	10%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	13%	0%	0%	0%	73%	89%
M18	ASI	1	4	0	0	0	0	1	0	4	0	0	0	2	3	0	0	0	0	6	0	0	0	0	0	0
M2	ASI	0	3	30	15	10	1	3	1	3	0	0	0	0	0	0	0	1	17	11	2	3	0	0	0	0
M25	ASI	5	0	0	0	10	1	0	2	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
M2a	ASI	0	1	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	2	0	0	0	9	0	0
M2b	ASI	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	3	0	4	0	1	0	0
M3	ASI	0	1	0	4	9	3	0	5	3	9	0	2	2	3	7	3	1	9	1	3	1	0	1	0	0
M3a	ASI	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M4	ASI	2	1	5	5	30	0	1	3	1	4	2	0	0	1	1	5	0	6	5	2	0	0	0	0	0
M4a	ASI	0	0	0	0	0	3	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0
M5	ASI	2	2	1	10	5	9	2	4	0	5	1	0	0	1	0	6	5	13	19	0	0	0	0	0	0
M5?	ASI	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M5a	ASI	0	2	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	2	0	0	0	0	0
M6	ASI	0	2	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
R5	ASI	4	0	0	4	10	3	0	3	0	0	2	0	0	0	0	2	0	0	0	1	0	0	2	0	0
R6	ASI	0	0	0	0	3	0	4	0	0	0	0	1	0	2	0	3	0	0	5	1	0	0	0	0	0
R7	ASI	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
U2	ASI	0	1	0	0	0	0	0	0	0	7	0	0	4	0	0	0	0	0	0	0	0	4	0	0	0
U2c	ASI	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M40	ASI	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
M31	ASI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
M31a	ASI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	8
M32	ASI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	1
M35	ASI	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
I	ASI	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
M*	ASI	11	0	37	23	34	12	6	22	18	21	23	5	13	8	0	20	31	30	29	6	27	52	11	0	0
U3	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
U8	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
U7	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0
M30	ANI	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
M39	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
B4	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
B5a	ANI	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	ANI	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F1	ANI	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	5	0	0	0	6	0	0	0	0
F1a	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
F1c	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	2	0	0	0	0
R	ANI	7	0	1	5	25	2	5	10	24	3	66	2	5	3	0	13	11	17	20	0	1	10	3	0	0
R1	ANI	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
T	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0
U	ANI	2	0	1	2	11	0	6	5	5	0	1	0	0	4	0	4	0	6	1	0	0	0	1	0	0
U10	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
U11	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
U9	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Ua	ANI	2	0	0	0	0	0	0	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ub	ANI	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0
W	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L0	unclassified	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
L2	unclassified	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0
L3	unclassified	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0
L3?	unclassified	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Y chromosome																										
Samples		17	21	41	21	27	42	30	9	17	74	68	15	38	10	19	19	57	30	249	23	45	65	27	10	10
ASI %		47%	67%	59%	38%	89%	43%	70%	67%	65%	26%	65%	80%	21%	10%	74%	32%	61%	80%	71%	65%	82%	26%	89%	0%	40%
ANI %		53%	33%	41%	62%	11%	57%	30%	33%	35%	74%	35%	20%	79%	90%	26%	68%	39%	20%	29%	35%	18%	11%	11%	0%	60%
H	ASI	1	6	15	7	10	15	9	0	2	3	22	1	8	0	2	0	12	2	56	0	2	0	13	0	0
H1	ASI	1	0	0	0	0	0	0	0	3	0	0	0	0	0	3	1	0	9	0	1	0	9	0	0	0
H2	ASI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
L	ASI	4	1	6	0	0	3	0	0	0	4	0	9	0	0	0	1	5	0	0	0	0	0	0	0	0
O	ASI	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	10	0	0	1	0	1
O2	ASI	0	0	0	0	14	0	2	0	0	12	14	0	0	0	0	0	0	0	117	0	0	0	9	0	1
O3	ASI	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	35	0	0	0	2
R2	ASI	0	4	3	1	0	0	10	6	2	0	0	2	0	0	6	4	18	0	3	4	0	0	1	0	0
F	ASI	2	2	0	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	5	0	0	0
R	ASI	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
J	ANI	1	1	3	0	0	5	0	0	0	40	0	1	8	6	3	0	0	0	0	1	0	2	0	0	0
R1	ANI	8	6	12	3	3	19	9	2	5	7	13	2	22	1	2	11	21	1	18	7	8	2	1	0	0
C	ANI	0	0	2	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0
K	ANI	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K*	ANI	0	0	0	4	0	0	0	1	0	6	11	0	0	0	0	0	1	3	47	0	0	3	1	0	2
P	ANI	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	1	0	2	0	0	0	0	1	0	4
G	ANI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
B	unclassified	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	unclassified	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
E2	unclassified	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
E3a	unclassified	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28	0	0	0
B2	unclassified	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0
D*	unclassified	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0

Note: Haplogroups were designated as typical of Ancient South Indians (ASI) or Ancient North Indians (ANI) based on the judgement of an expert on mtDNA and Y chromosome variation (KT) who was blinded to ancestry estimates from the autosomes.

Figure S1

(a)



(b)

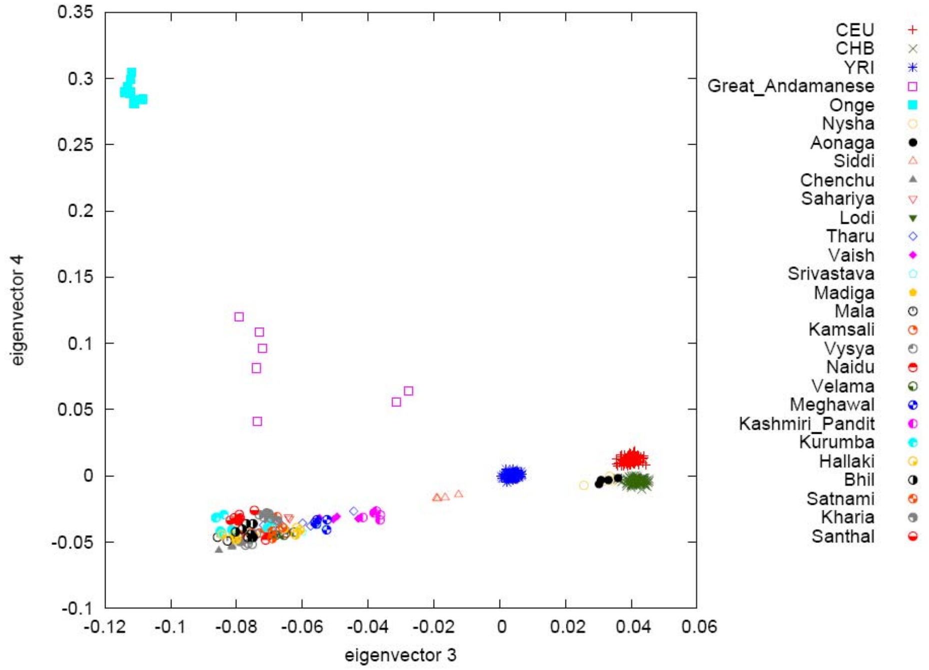


Figure S1 Legend: Principal components analysis of the 25 groups, together with CEU, CHB and YRI from HapMap. **(a)** The top two PCs show that the Siddi are an outlying group with ancestry that is related to West Africans (YRI), consistent with the known origin of this group in the Arab slave trade. They also show that the Nyshi and Ao Naga are closely related to East Asians (CHB), as expected from the fact that these groups speak a Tibeto-Burman language. **(b)** The third and fourth PCs distinguish the Andaman Island groups, and show that the Great Andamanese do not cluster in the plot. This is a signature of recent gene flow from the mainland in the last handful of generations (Note S1).

Figure S2

[see next page for the figure]

Figure S2 Legend: Decay of allele sharing provides evidence for ancient founder effects, which in many Indian Cline groups appear to have occurred at least 30 generations ago. For each of the groups that we genotyped (except for the Srivastava with just two individuals), we examined all pairs of samples, and recorded whether 0, 1 or 2 alleles were shared at each SNP (we scored SNPs that were heterozygous in both individuals as sharing 1 allele to account for phase ambiguity). Founder events are expected to cause segments of the genome to be identical by state (IBS) for at least one allele over a stretch of sequence due to their descent from a shared founder, with the extent of the shared segment providing information about the age of the event. To correct for allele sharing inherited from the ancestral population, we subtracted the curve obtained by comparisons across different Indian Cline groups, picking the closest match among the groups with 65% ± 5% ANI ancestry (Meghawal, Vaish and Kashmiri Pandit), 58% ± 5% ANI ancestry (Velama, Srivastava, Meghawal and Vaish), 53% ± 5% ANI ancestry (Lodi, Naidu, Tharu, Velama and Srivastava), 47% ± 5% ANI ancestry (Bhil, Satnami, Kurumba, Kamsali, Vysya, Lodi, Naidu and Tharu), and 42% ± 5% ANI ancestry (Mala, Madiga, Chenchu, Bhil, Satnami, Kurumba, Kamsali and Vysya). We performed a least-squares fit of $y = a + be^{-2Dt}$ to the data from each group where a, b and *t* are constants, D is the distance in Morgans between SNPs, and the factor of 2 corresponds to the fact that a recombination can occur on either haplotype that is being compared. Computer simulations reported in Figure S3 show that this procedure can infer the age *t* of founder events with reasonable accuracy under the assumption of a single founder event. As an example, in the Vysya, allele sharing decreases with an exponential decay of 0.461 cM, suggesting a founder event roughly 100/(2*0.461) = 108 generations ago (see also Figure 2). There are 6 Indo-European and Dravidian speaking groups with estimated founder events of >30 generations ago: Bhil (40), Hallaki (32), Meghawal (59), Sahariya (108), Vysya (108) and Velama (88).

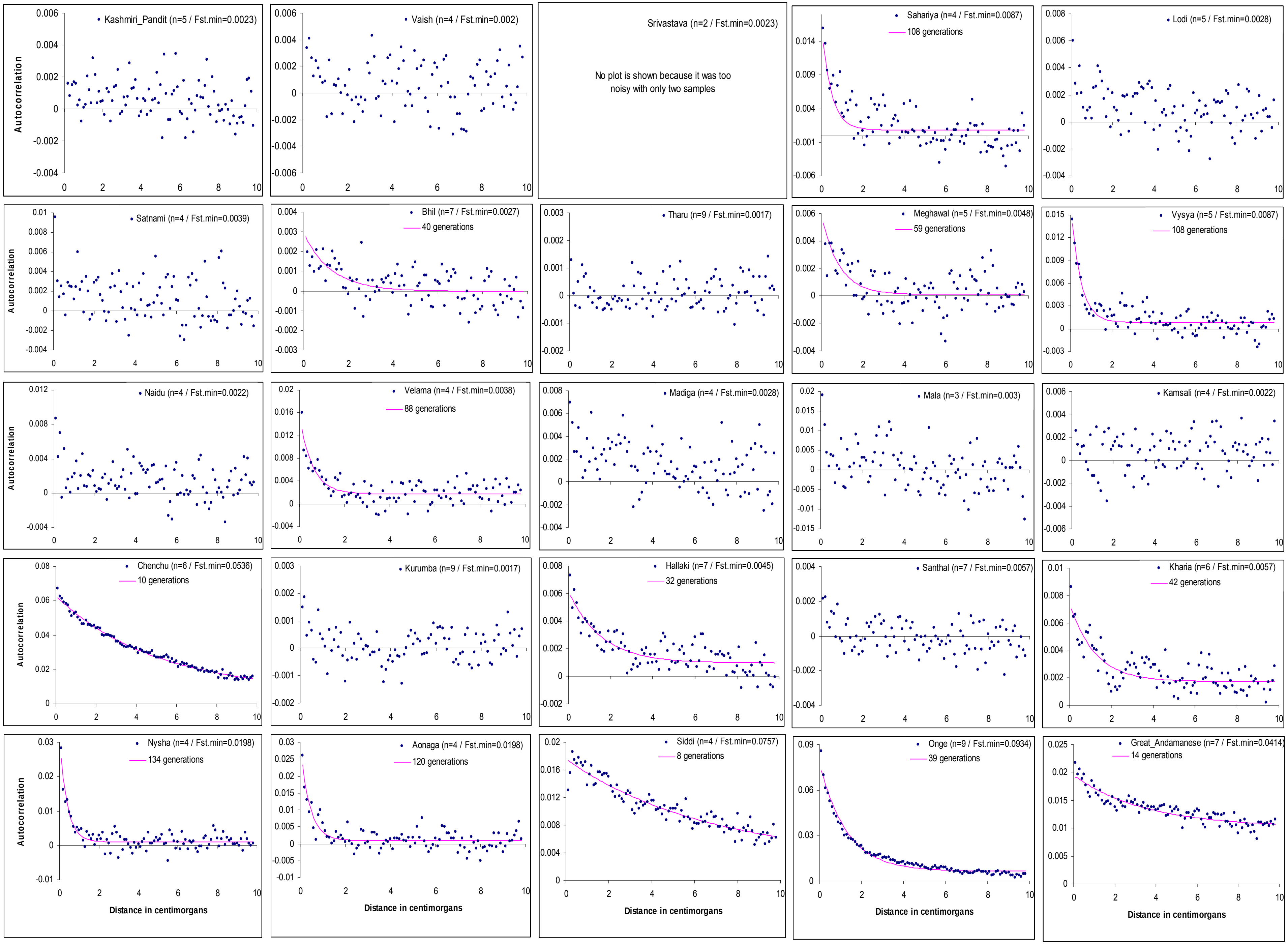
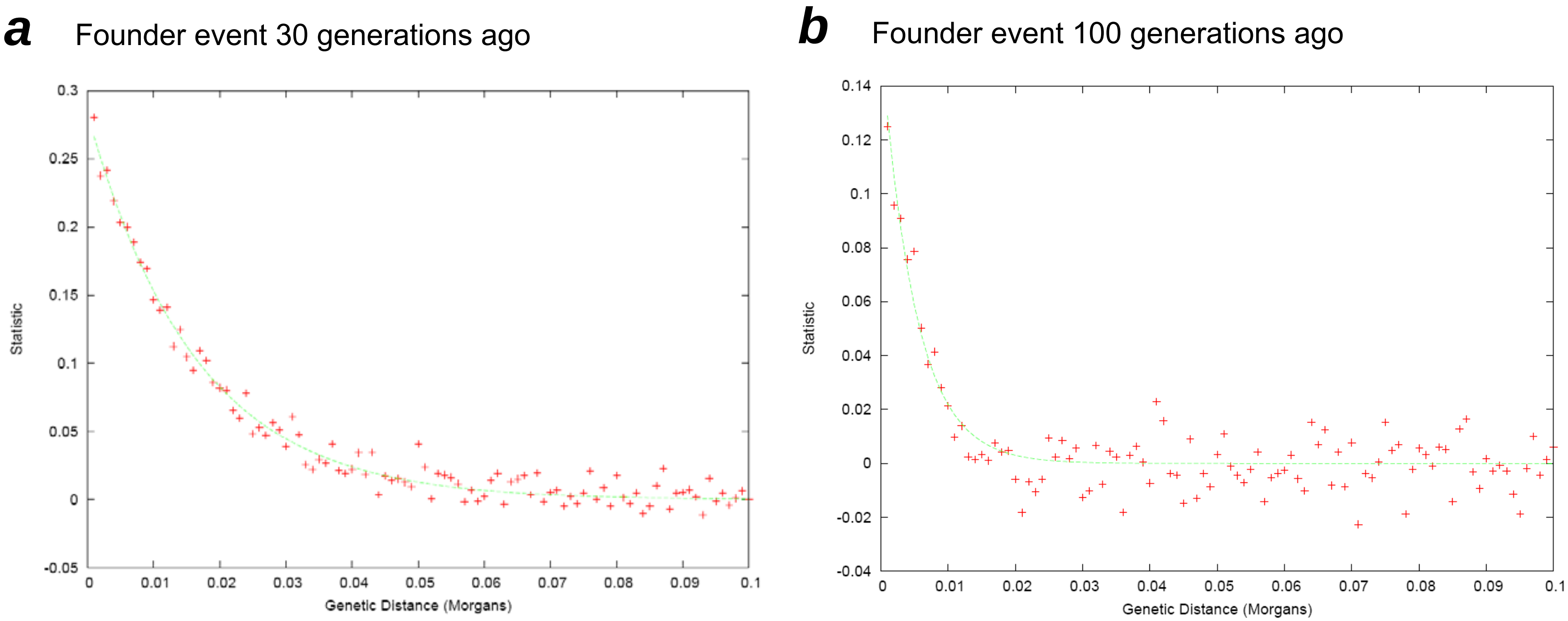

Kashmiri_Pandit (n=5 / Fst.min=0.0023)
Vaish (n=4 / Fst.min=0.002)
Srivastava (n=2 / Fst.min=0.0023)
No plot is shown because it was too noisy with only two samples
Sahariya (n=4 / Fst.min=0.0087)
108 generations
Lodi (n=5 / Fst.min=0.0028)
Satnami (n=4 / Fst.min=0.0039)
Bhil (n=7 / Fst.min=0.0027)
40 generations
Tharu (n=9 / Fst.min=0.0017)
Meghawal (n=5 / Fst.min=0.0048)
59 generations
Vysya (n=5 / Fst.min=0.0087)
108 generations
Naidu (n=4 / Fst.min=0.0022)
Velama (n=4 / Fst.min=0.0038)
88 generations
Madiga (n=4 / Fst.min=0.0028)
Mala (n=3 / Fst.min=0.003)
Kamsali (n=4 / Fst.min=0.0022)
Chenchu (n=6 / Fst.min=0.0536)
10 generations
Kurumba (n=9 / Fst.min=0.0017)
Hallaki (n=7 / Fst.min=0.0045)
32 generations
Santhal (n=7 / Fst.min=0.0057)
Kharia (n=6 / Fst.min=0.0057)
42 generations
Nysha (n=4 / Fst.min=0.0198)
134 generations
Aonaga (n=4 / Fst.min=0.0198)
120 generations
Siddi (n=4 / Fst.min=0.0757)
8 generations
Onge (n=9 / Fst.min=0.0934)
39 generations
Great_Andamanese (n=7 / Fst.min=0.0414)
14 generations
Autocorrelation
Distance in centimorgans
0
2
4
6
8
10

Figure S3



a Founder event 30 generations ago

b Founder event 100 generations ago

Figure S3 Legend: Simulations suggest that the decay of the autocorrelation of allele sharing calculated as in Figure S2 can be used to infer the age of a founder event. We simulated histories with a constant diploid size of 10,000 at all times except during the founder events. We sampled 5 individuals from each of two groups that experienced founder events (a) 30 and (b) 100 generations ago in which there was a contraction to 5 individuals for one generation. The two groups had the following simulated history: (i) Divergence from a common ancestral population 150 generations ago; (ii) Origin of this ancestral population by mixture of ANI-like (40%) and ASI-like (60%) populations 160 generations ago; and (iii) Splitting of the ANI-like and ASI-like populations 500 generations ago. We ascertained SNPs as heterozygotes in a single individual of entirely ANI-related ancestry, and generated data for 100,000 linked pairs of SNPs with a range of recombination distances. The plots are based on computing the autocorrelation of allele sharing within groups, and then subtracting the across-population autocorrelation to remove the effects of ancestral allele sharing (Methods). The fitted exponential function $y = a + be^{-2Dt}$ is shown in green, and the fitted value of t corresponds to (a) 31 generations for one population and (b) 99 generations for the other, roughly matching the input values used in the simulation.

Figure S4

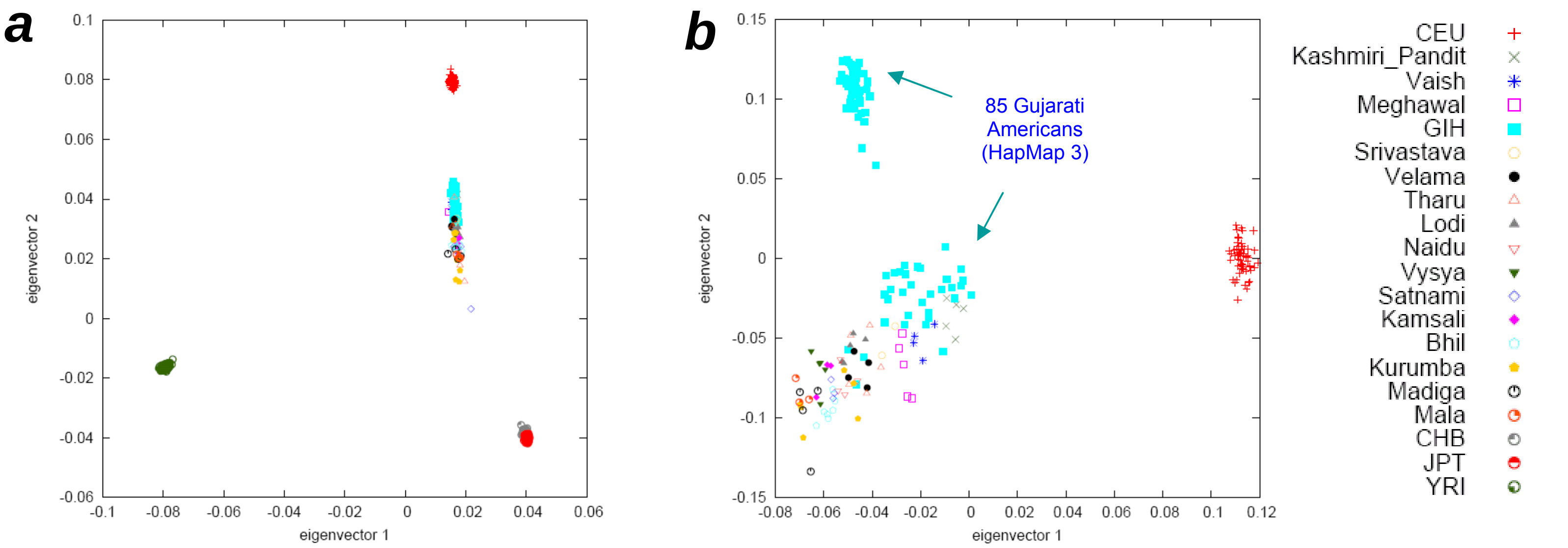


Figure S4 Legend: Genetic relationship of Gujarati Americans from HapMap Phase 3 (GIH) to other groups in India and worldwide. (a) We carried out a PCA of HapMap samples (YRI, CEU, CHB and JPT), and projected selected Indian groups onto the axes of variation defined by HapMap. The GIH (blue squares) fall along the main gradient of variation of Indian populations without unusual relatedness to West Africans (YRI) or East Asians (CHB or JPT). (b) A PCA of the same Indian groups together with the CEU and GIH shows that the GIH fall into at least two discrete clusters that are substantially differentiated (F_{ST}=0.005), confirming that defining an Indian Americans group based on its state-of-origin can mask substantial substructure, which presumably reflects the fact that Indian American groups from a single state are often derived from multiple effectively endogamous groups. Interestingly, one of the GIH subgroups fall outside the main gradient of Indian groups, suggesting that they harbor substantial ancestry that is not a simple mixture of ASI and ANI. A speculative hypothesis that some Gujarati groups descend from the founders of the “Gurjara Pratihara” empire, which is thought to have been founded by Central Asian invaders in the 7th century A.D. and to have ruled parts of northwest India from the 7-12th centuries. I. Karve noted that endogamous groups with names like “Gurjar” are now distributed throughout the northwest of the subcontinent, and hypothesized that that they likely trace their names to this invading group[2].

Figure S5

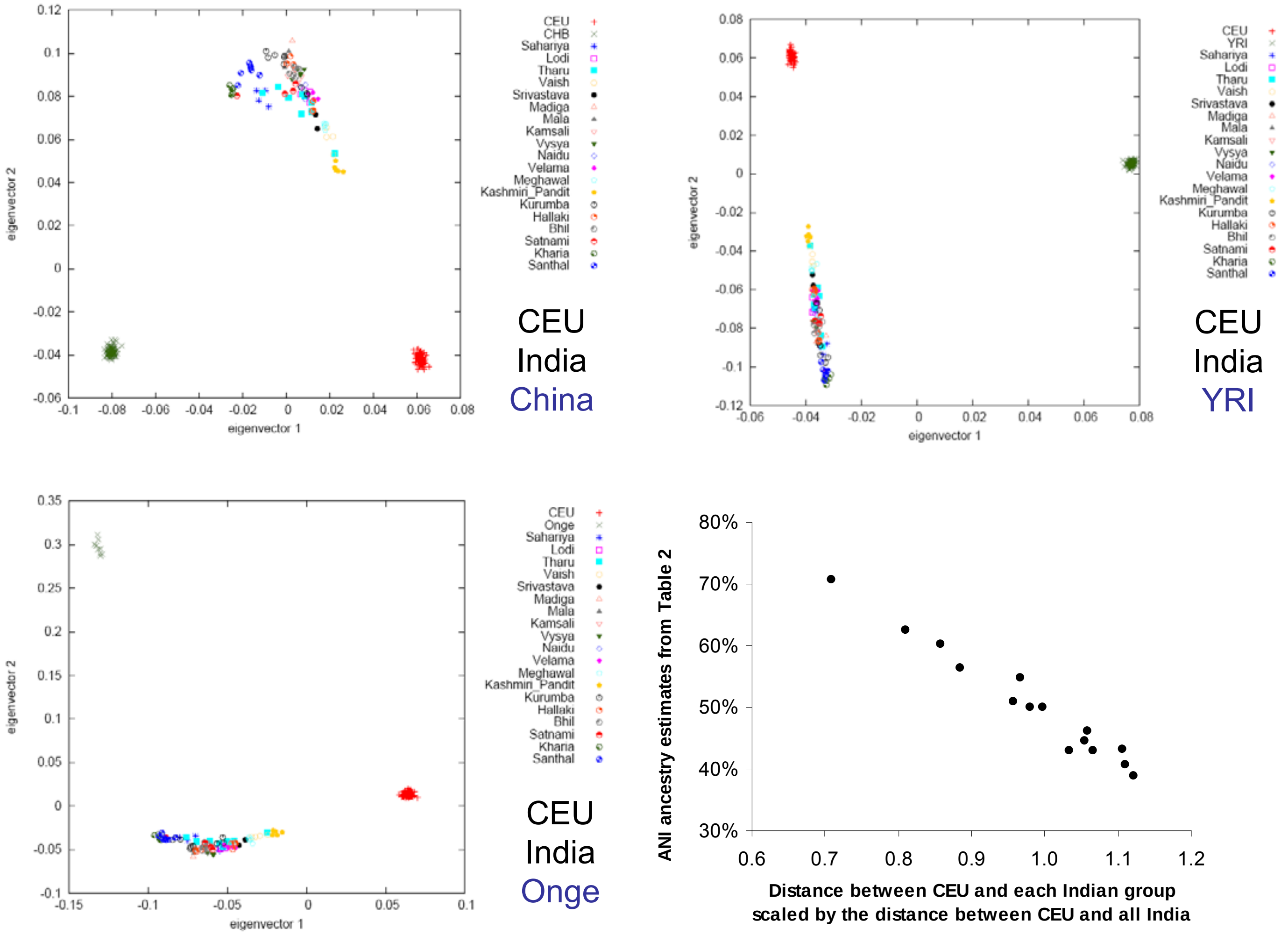
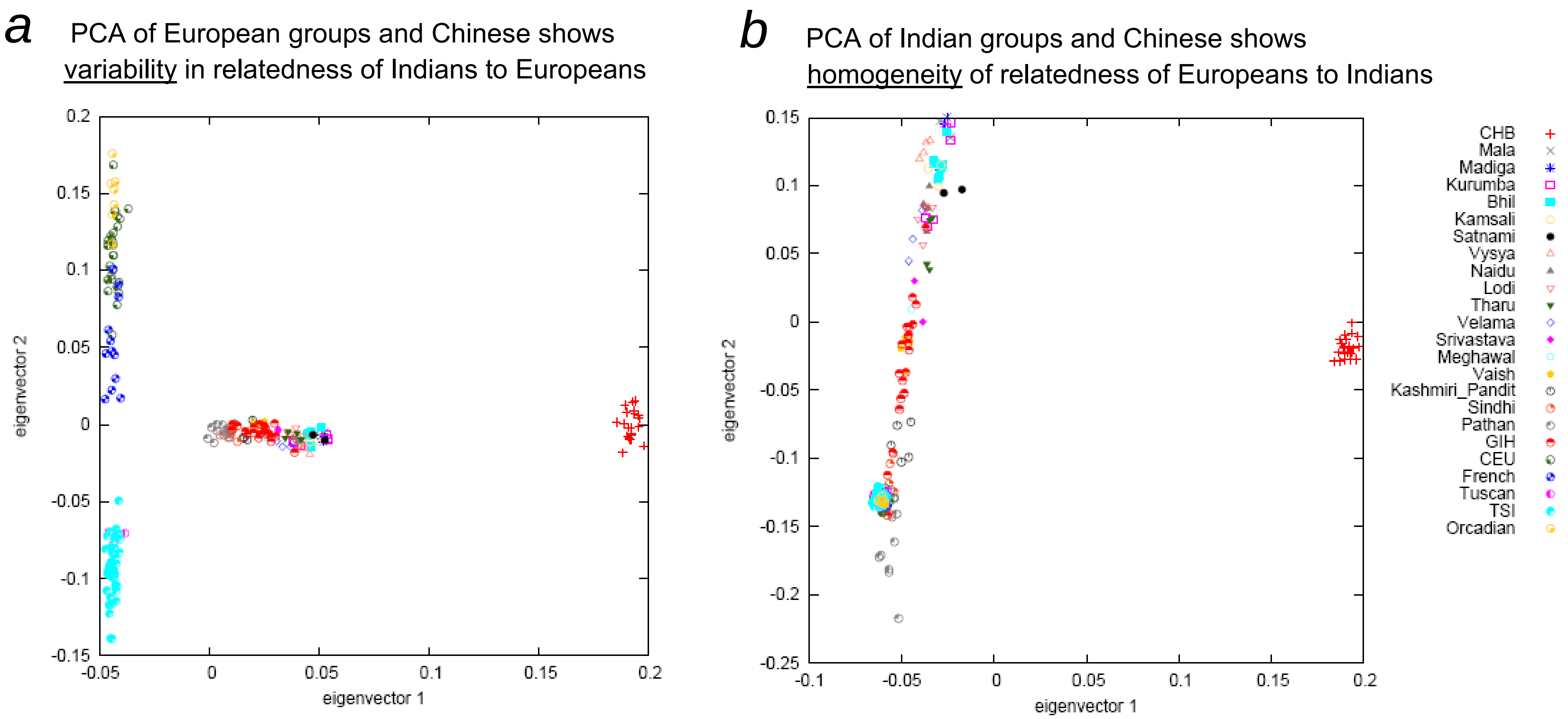


Figure S5 Legend: After controlling for relatedness to West Eurasian groups, genetic differences among non-Indians have little correlation to differences within India. We carried out PCA of 19 Indian groups with different pairs of non-Indian groups, after excluding 6 groups identified in Figure S1 and in the text as being outliers in ancestry (Onge, Great Andamanese, Siddi, Nyshi, Aonaga, and Chenchu). We find that the 19 Indian groups are largely distributed along a one-dimensional gradient including CEU and the centroid of the Indian groups. The only exceptions to this are the Kharia, Santhal and Sahariya who are "off cline" suggesting a more complex mixture history (consistent with the Kharia and Santhal speaking Austro-Asiatic languages). The ordering and relative distance from CEU are preserved whether we choose the non-Indian-subcontinent groups to be **(a)** CEU and CHB, **(b)** CEU and YRI, **(c)** or CEU and Onge. **(d)** We used the distance from CEU in the PCA to estimate a quantity that we hypothesized was linearly related to the proportion of West Eurasian-related mixture in each Indian group, which we confirmed by comparing the quantity to the model-based estimate of ANI-like ancestry in Table 2 for groups that overlapped between the two analyses.

Figure S6



a PCA of European groups and Chinese shows variability in relatedness of Indians to Europeans

b PCA of Indian groups and Chinese shows homogeneity of relatedness of Europeans to Indians

Figure S6 Legend: Indian groups show a gradient of relatedness to Europeans, but Europeans show no analogous gradient of relatedness to Indians. **(a)** We carried out a PCA of groups of European ancestry from HapMap and the HGDP (CEU, TSI, French, Tuscan and Orcadian) along with Chinese (CHB). Using the SNP weights for PC1 and PC2 that emerge from this analysis[3], we projected the Indian groups onto the pattern of variation defined in groups outside India, and replicate the Indian Cline found in Figure 3. These results support the hypothesis that different Indian groups have different proportions of ancestry from a hypothetical ANI ancestral group. **(b)** To test for evidence of an analogous "European Cline" of relatedness to India, we carried out a PCA of groups on the Indian Cline with HapMap Phase 3 Gujarati Americans (GIH) and CHB, and projected five groups of European ancestry (CEU, TSI, French, Tuscan and Orcadian) onto the PCA. We observe no variability among Europeans in their proximity to Indians (they all pile up at the same position on the PCA). This is consistent with these groups having all received about the same proportion of ASI-related ancestry.

Figure S7

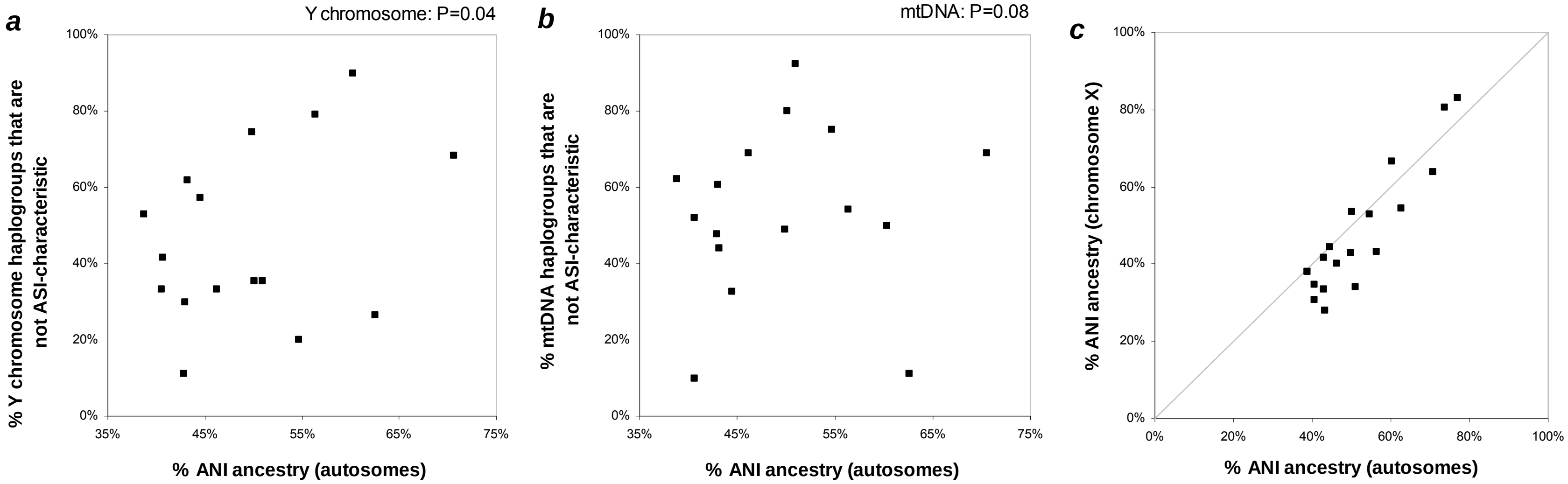


Figure S7 Legend: ANI-related ancestry in India measured in four different parts of the genome with different inheritance patterns. **(a)** For all 16 Indian groups in Table 2, we plot our autosomal estimate of ANI ancestry against the proportion of haplogroups that are not characteristic of ASI ancestry (Table S6). This analysis suggests that the Y chromosome estimates of ancestry are positively correlated to the autosomal ones, consistent with previous reports of a gradient of male relatedness to West Eurasians among Indian groups (P=0.04 by a 1-sided test from a weighted least squares regression that takes into account the variable precision of the estimates of haplotype frequencies in Table S6). **(b)** Further supporting the view that the gradient of relatedness to West Eurasians in India is primarily associated with male ancestry, the same analysis on mtDNA data shows weaker evidence of correlation (P=0.08). **(c)** We also compared estimates of ANI ancestry in Indian groups on the autosomes and chromosome X (Table S5). While our autosomal estimate of ANI ancestry is higher than the X chromosome estimate of ANI ancestry by about 7.4%, this pattern is not statistically significant (Z=1.2 standard deviations) given the large errors in our X chromosome ANI ancestry estimates. Standard errors are ±1.2% on average for the autosomes and ±11% on average for the X chromosome (Table S5).

Note S1:
Genetic structure of the Great Andamanese and Onge

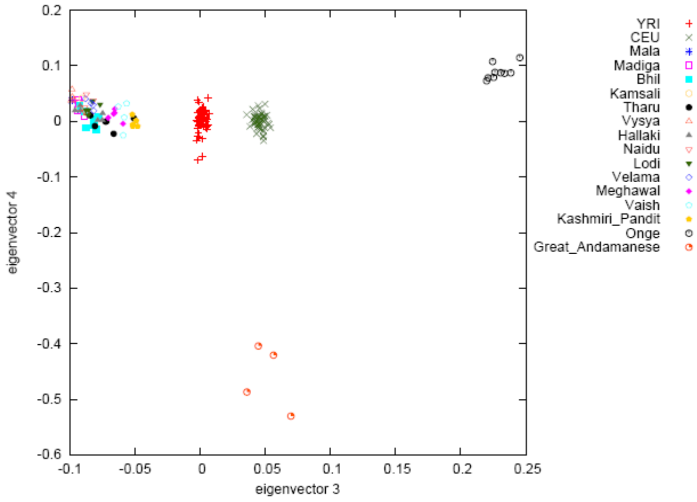
The SNP array data provide more information about the genetic structure of the Great Andamanese and Onge than has been available from studies of mtDNA and the Y chromosome.

Evidence for recent mixture in the history of the Great Andamanese

The PCA plot of Figure S1b, which is based on our autosomal SNP array data, shows that the 9 Onge fall into a tight cluster while the 7 Great Andamanese are dispersed into at least three clusters. The tight clustering of the Onge suggests that they have not received recent gene flow from the mainland, as such gene flow is expected to have a differential effect on different members of a group. By contrast, the Great Andamanese are very dispersed in the PCA, which is a signature of recent mixture.

The lack of evidence for recent mixture in the Onge is consistent with previously reports based on mtDNA and Y chromosome data from an overlapping set of the same samples. These reports suggested that the Onge share no common ancestry with non-Andamanese groups for the last few tens of thousands of years[4,5,6]. While the same holds true for the Great Andamanese on mtDNA, on the Y chromosome this latter group's ancestry appears to be almost entirely from the mainland[4].

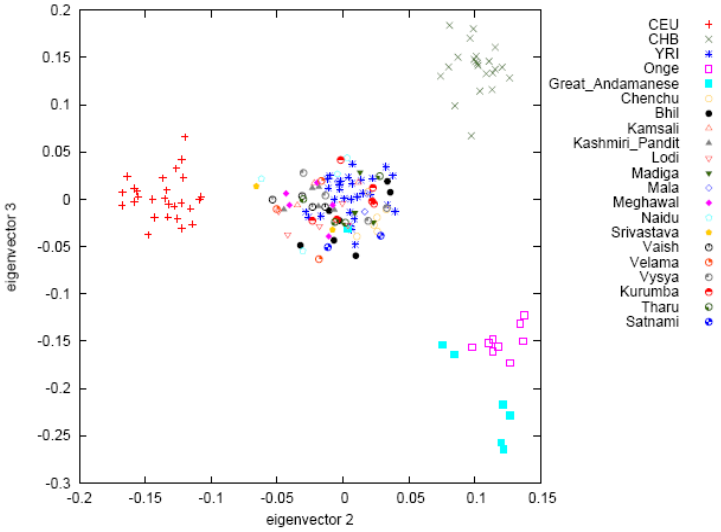
To further elucidate the population structure in the Great Andamanese, we carried out additional PCA of 4 samples that appeared in Figure S1b as if they might come from a homogeneous group. Note S1 Figure 1 shows a PCA of these four samples along with the Onge, YRI, CEU and some mainland Indian groups. The fourth principal component corresponds to genetic drift that appears to reflect the specific ancestry of the Great Andamanese that is not present in the Onge.



Note S1 Figure 1: Focusing on the four Great Andamanese that appear as if they might be homogeneous (from the top left of Figure S1b), we carried out a PCA limited to the Onge, Great Andamanese YRI, CEU, and some groups from the "Indian Cline" (Note S2). The first and second PCs are not relevant to Andaman Island genetics, but the fourth shows genetic drift specific to the Great Andamanese.

The Great Andamanese have less mixture on the X chromosome than on the autosomes

We next carried out PCA of the Great Andamanese and the Onge on the X chromosome. All samples used in this analysis are male, and hence the X chromosome analyses use haploid rather than diploid data. We found that 6 of the 7 Great Andamanese samples are as distant as the Onge from mainland Indians, suggesting that they may be unmixed on the X chromosome (Note S1 Figure 2).To formally test for mixture on chromosome X, we carried out a *4 Population Test* on the CEU, YRI, Onge, and the 6 Great Andamanese that fell into an approximate cluster on the X chromosome (Note S1 Figure 2). The 6 samples are consistent with being unmixed and falling into a clade with the Onge (Z=0.9). The other two topologies are rejected (Z=8.9 and Z=6.1).



Note S1 Figure 2: PCA of male Great Andamanese, Onge, CEU, CHB, and some Indian Cline groups on chromosome X shows that the Great Andamanese are as distinct from the other groups as the Onge, in contrast to the autosomal analyses of Fig. S1b. This suggests that on chromosome X, the Great Andamanese are mostly unmixed, potentially because their mothers are of unmixed Great Andamanese ancestry. The one exception is a male who in the X chromosome analysis falls within the main cluster of Indian variation, consistent with their father being Great Andamanese and their mother being of mainland Indian ancestry. On the autosomes, this individual's ancestry is identical to the main cluster of 4 Great Andamanese in Fig. S1b.

The most surprising difference between the X chromosome and autosomal analyses of the Great Andamanese is that one of the 4 Great Andamanese that fall into the largest cluster in Figure S1b on the autosomes is an outlier on the X chromosome. A speculative explanation is that the autosomal cluster of 4 Great Andamanese represents first generation admixed individuals with 50% Great Andamanese and 50% mainland Indian ancestry. We hypothesize that 3 individuals have a Great Andamanese mother, and 1 has a mother of mainland Indian ancestry. Men receive their X chromosome entirely from their mother, and this would explain why 3 of the individual appear as unmixed as the Onge on their X chromosome, while 1 individual appears to be entirely of mainland ancestry. Some of the individuals could also be second generation mixes.

We use the Onge to represent the genetic relationship of the Andaman Islands to other groups in the main study, since it is easier to analyze data from groups without a recent history of mixture.

Note S2:
Identifying a core set of 96 samples to represent the "Indian Cline"

Many of the analyses in this study are based on modeling the history of Indo-European and Dravidian speaking groups of the Indian subcontinent in terms of a two-way historical mixture of an "Ancestral North Indian" (ANI) population that is genetically close to Central Asians, Middle Easterners, and Europeans, and an "Ancestral South Indian" (ASI) population that is not close to any large modern group outside the Indian subcontinent.

The idea of an ancient mixture event in India has been previously suggested based on the presence of both Indo-European and Dravidian languages in India today, and by genetic data showing differences in Y chromosome haplotype frequencies that are associated with caste, language and geography[7,8,9,10]. In our data, the hypothesis of mixture emerges naturally from PCA (Figure 3), which shows that nearly all the Indo-European and Dravidian speaking groups spread out on a one dimensional gradient in a plot of the first versus the second PC.

Modeling the history of many Indian groups as a mixture of two ancestral populations is an oversimplification. In reality, even if ancient mixture did occur, it is likely to have been between substructured populations instead of homogeneous populations, and it is likely to have occurred at multiple times and at multiple geographic locations. However, approximating the history of many Indian groups as a simple mixture of two homogeneous ancestral populations provides a good fit to the summary statistics of allele frequency differentiation, and we believe that in this sense it is a useful starting point for future analyses that can detect more subtle events.

Note S2 Table 1 – Outlier samples removed during the filtering process

Pop.	No.	Sample IDs
Kamsali	1	Kamsali_192_R2
Satnami	1	Satnami_206_R2
Kurumba	3	Kurumba_41_R1, Kurumba_42_R1, Kurumba_48_R1
Tharu	3	Tharu_101_R1, Tharu_102_R1, Tharu_103_R1, Tharu_104_R1
Pathan	7	224, 234, 243, 251, 258, 259, 262
Sindhi	14	163, 165, 169, 171, 173, 175, 177, 179, 181, 191, 192, 199, 206, 208

Choosing samples for the Indian Cline

To define a set of samples to model the Indian Cline, we used three principles.

(i) We restricted analysis to groups that fell visually along a one dimensional gradient in the PCA of Figure 3, leading us to the hypothesis that we could model them as a simple mixture.

This caused us to remove three tribal groups (Sahariya, Kharia and Santhal) that were visually "off-cline" in the direction of being more closely related to East Asians (CHB). The fact that the "off-cline" groups include both of the Austro-Asiatic speaking groups (Kharia and Santhal), makes it likely that the PCA pattern genuinely reflects complex mixture in these groups—possibly gene flow from groups that are (distantly) related to East Asians—and is not a mathematical artifact of PCA that can arise due to isolation-by-distance[11].

(ii) We restricted analysis to samples that were homogeneous with their own group in PCA

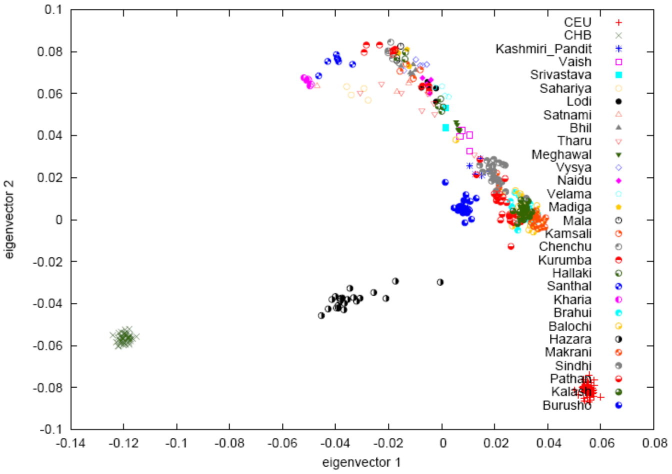
If the samples from a group are not homogeneous in a PCA, this comprises evidence that the group experienced mixture from a range of ancestries in the last handful of generations. In practice, we found that the majority of groups showed clear clusters in the PCA (with only a few outliers), justifying our removal of 9 samples that had evidence of inhomogeneity (Note S2 Table 1). We also removed an entire group based on the criterion of homogeneity: the Hallaki. While the Hallaki were all on the Indian Cline, they were so dispersed in the PCA (suggesting recent mixture with other groups in the Indian Cline) that we could not identify a main cluster.

(iii) We extended the Indian Cline by merging with 2 Pakistani groups

We also jointly analyzed the 25 Indian groups with 8 Pakistani groups from the Human Genome Diversity Panel (HGDP) that had been genotyped on an Illumina 650Y array[12]. We used PCA on these data to explore which of the 8 Pakistani groups are consistent with the Indian Cline. We began by removing samples that appeared to have outlying ancestry compared with other samples from the same groups (suggesting gene flow in the last handful of generations), or evidence of African gene flow (related to YRI), which is present in many of the HGDP samples from Pakistan as previously reported[12].

We found that 6 Pakistani groups (the Hazara, Kalash, Burusho, Makrani, Balochi and Brahui) were difficult to model as part off the Indian cline, since when we added samples from them into the PCA, they all generated new PCs that correlated to genetic differences among non-Indian groups (CHB, CEU, YRI and Adygei) suggesting a more complex history than a simple mixture of two ancestral groups. The Hazara and Burusho, in particular, show clear evidence of East Asian related mixture in the PCA (Note S2 Figure 1).

We identified 2 Pakistani groups (Pathan and Sindhi) as fully consistent with the Indian cline within the limits of our resolution. After removing 7 Pathan and 14 Sindhi samples with evidence of outlying ancestry (mostly West African related) that appears to be due to mixture in the last handful of generations, we added these 2 groups to the 16 Indian groups. This provided us with a set of 18 groups that we could use for modeling of the Indian Cline. The 2 Pakistani groups have more CEU-related ancestry than the Indian groups, allowing us to extend the Indian Cline in a way that increased power for analysis. A version of Figure 3 that is restricted to the 18 groups we used to represent the Indian Cline for modeling is shown in Note S2 Figure 1.

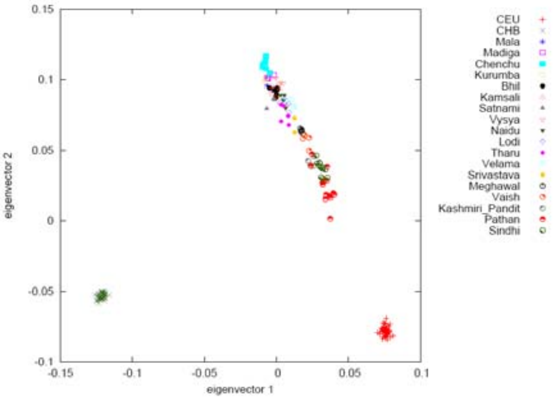


Note S2 Figure 1: PCA of 20 groups from India, together with CEU and CHB and 8 Pakistani groups from the HGDP. The Pakistani groups generally fall on the Indian Cline, but with more relatedness to CEU than any groups in India. The Hazara and Burusho are clear outliers with substantial amounts of East Asian-related ancestry. (This plot is similar to Figure 3 except that we have added Pakistani groups.)

Tabulation of samples that remain in the data set after defining the Indian Cline

After applying these filters to the merged data from 18 Indian Cline groups, there were only four statistically significant PCs ($P<0.05$ by the Tracy-Widom test of population structure[3]) which each had a clear qualitative interpretation: 1 = the difference between West Eurasians and East

Asians, 2 = “Indian Cline”, 3 = Separates Chenchu from all other samples, and 4 = Separates Vysya from all other samples.



Note S2 Figure 2: PCA of the 96 samples in 18 groups that we used to represent the Indian Cline for modeling analyses, along with CEU and CHB. To generate this plot, we removed 6 groups identified as having very different ancestry (Nyshi, Ao Naga, Kharia, Santhal, Sahariya, and Hallaki), and 9 outlier samples. We also added in the Pathan and Sindhi, two Pakistani groups with greater genetic relatedness to the CEUs, providing more statistical power to analyze variation in ancestry on the Indian Cline.

As a resource for subsequent work with these data, Note S2 Table 2 presents the groups and total number of samples (n=96) that remained after applying the filters.

Note S2 Table 2 – Filtering of samples to identify 96 on the Indian cline

Group	Source	Traditional caste or social designation	Before filtering	After filtering
Chenchu	This study	Tribal	6	6
Mala	This study	Lower	3	3
Madiga	This study	Lower	4	4
Bhil	This study	Tribal	7	7
Kurumba	This study	Tribal	9	6
Kamsali	This study	Lower	4	3
Vysya	This study	Middle	5	5
Satnami	This study	Lower	4	3
Naidu	This study	Upper	4	4
Lodi	This study	Lower	5	5
Velama	This study	Upper	4	4
Tharu	This study	Tribal	9	5
Srivastava	This study	Upper	2	2
Meghawal	This study	Lower	5	5
Vaish	This study	Upper	4	4
Kashmiri Pandit	This study	Upper	5	5
Sindhi	HGDP	Pakistan	24	10
Pathan	HGDP	Pakistan	22	15
Onge	This study	Hunter gatherer	9	dropped
Santhal	This study	Tribal	7	dropped
Kharia	This study	Tribal	6	dropped
Sahariya	This study	Lower	4	dropped
Siddi	This study	Tribal	4	dropped
Hallaki	This study	Tribal	7	dropped
Aonaga	This study	Tribal	4	dropped
Nysha	This study	Tribal	4	dropped
Great Andamanese	This study	Hunter gatherer	7	dropped
Burusho	HGDP	Pakistan	25	dropped
Brahui	HGDP	Pakistan	25	dropped
Hazara	HGDP	Pakistan	22	dropped
Makrani	HGDP	Pakistan	25	dropped
Balochi	HGDP	Pakistan	24	dropped
Kalash	HGDP	Pakistan	23	dropped

Note S3:
A framework for learning about history using genetic drift, and evidence that all Indian Cline groups are of mixed ancestry

We develop a novel series of methods for learning about history that are based on the idea of measuring "genetic drift", defined as the variance in allele frequencies that has occurred on any lineage of a phylogenetic tree. Cavalli-Sforza and Edwards first had the idea of fitting genetic drift parameters to a phylogenetic tree[13], and here we extend this framework in three ways.

(1) We present updated methods for fitting a phylogenetic tree to the measured drifts. We use a new formulation of f-statistics that is designed to be proportional to the genetic drift that occurred on any lineage. Our f-statistics contrast with F_{ST}, which is normalized differently in a way that makes it less proportional to genetic drift (Appendix).

(2) We extend the framework of Cavalli-Sforza and Edwards to model population mixture.

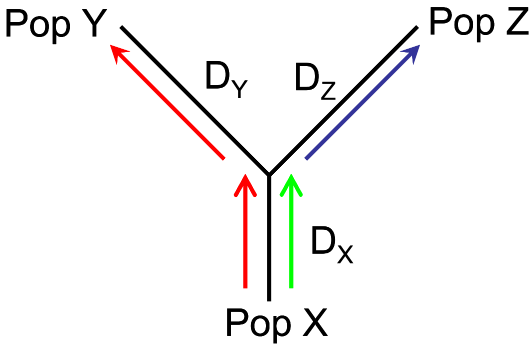
(3) We provide tools for rigorously testing whether a proposed tree is consistent with the data.

The *3 Population Test*

We applied two distinct methods based on measurement of genetic drift to formally test for a history of mixture on the Indian Cline. The first is a novel *3 Population Test*, which provides a direct test for whether a group has inherited a mixture of ancestries while making minimal assumptions about demography. The second is a *4 Population Test*[14,15], which is more sensitive, but is also more model-based so that a positive signal is more difficult to interpret.

The *3 Population Test* compares a tested population X to two reference populations Y and W, and calculates an f_3 statistic $f_3(X;Y,W)$ that we define as the product of the frequency difference between population X and Y, and the frequency difference between population X and W, normalized as described in the Appendix and averaged over all SNPs (Note S3 Figure 1).

In practice, we normalize by the frequency of the population X that appears twice in the f_3 statistic. The form of the normalization reflects the fact that the binomial variance in frequency of an allele as it is sampled from generation is expected to be proportion to p(1-p).[16]



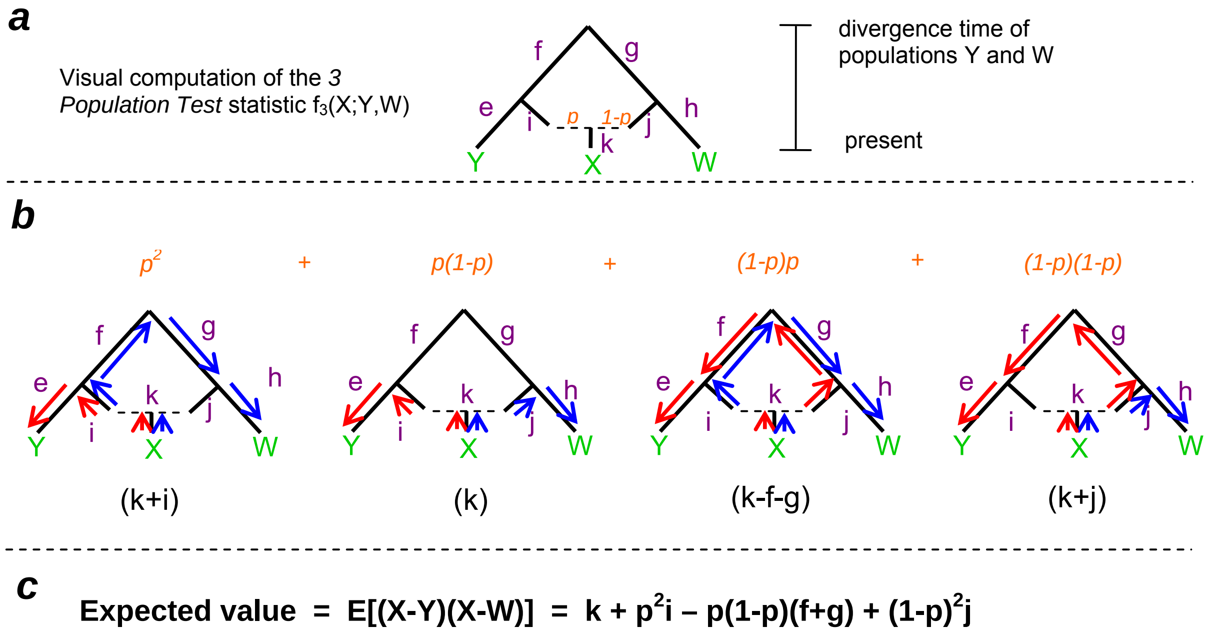
Note S3 Figure 1: The expected value of the *3 Population Test* statistic can be calculated visually. In the case that populations X, Y and Z are unmixed and can be related by an unrooted tree with drifts of D_X, D_y, and D_Z on each lineage, the product of the frequency difference between populations X and Y, and X and Z, suitably normalized and averaged over SNPs, is just proportional to the genetic drift D_X on the shared drift path. (The genetic drifts D_Y and D_Z are uncorrelated with respect to the *3 Population Test* statistic, and do not contribute to the expected value of the statistic).

Expected value of the *3 Population Test* statistic

The expected value of the *3 Population Test* statistic can be calculated visually.

In the case of no mixture, the expected value of the *3 Population Test* statistic is positive
If groups X, Y and W are related by a simple unrooted tree, the value of the *3 Population Test* statistic is expected to be proportional to the correlation in allele frequency difference between groups X and Y, and X and W. In the absence of mixture, this is proportional to the genetic drift D_X that is specific to the lineage leading to population X since its divergence from the node in the unrooted tree joining groups Y and W (Note S3 Figure 1). Genetic drift D_X is expected to be at least 0, and thus the expected value of the *3 Population Test* statistic is also positive.

In the case of mixture, the expected value of the *3 Population Test* statistic can be negative
If population X has a history of mixture with a proportion *p* from a population related to Y, and the rest of its ancestry from a population more related to W, we can calculate the expected value by tracing drift paths through the graph (Equation S3.1). Since the quantity in Equation S3.1 is quadratic, there are four terms, each of whose values can be calculated by following the path of frequency differences through the tree (Note S3 Figure 2). In the Appendix, we show that it is mathematically appropriate to calculate the expectation of f-statistics by tracing drift paths through an admixture graph, and in particular we show why it is appropriate to decompose a phylogenetic tree with admixture into its component parts to calculate expectations. Simulations in Note S5 confirm that this procedure works robustly for the application of estimating mixture.


a
Visual computation of the *3 Population Test* statistic $f_3(X;Y,W)$
f
g
e
h
i
p
1-p
j
k
Y
X
W
divergence time of populations Y and W
present
b
p^2
+
$p(1\text{-}p)$
+
$(1\text{-}p)p$
+
$(1\text{-}p)(1\text{-}p)$
(k+i)
(k)
(k-f-g)
(k+j)
c
Expected value = $E[(X\text{-}Y)(X\text{-}W)] = k + p^2 i - p(1\text{-}p)(f+g) + (1\text{-}p)^2 j$

Note S3 Figure 2: Calculation of the expected value of the *3 Population Test* statistic if population X is mixed but Y and W are not. (a) We show a generalized topology indicating that group X has inherited a proportion *p* of ancestry from a group related to Y, and a proportion *(1-p)* of ancestry from a group more closely related to W. The genetic drifts (variances in allele frequencies) are specified by lower case letters. (b) To compute the expected value of the *3 Population Test* statistic, we can break the graph into its four quadratic components with weights p^2, p(1-p), (1-p)p and $(1-p)^2$. The expected contribution that each of the four trees makes to the sum can be obtained by adding the shared drift between the first and second terms, where the red and blue arrows overlap. The sign is determined by whether the edge is traversed in the same or opposite direction by the frequency differences (X-Y) and (X-W). (c) Adding the results from the four trees with the appropriate weights, we note that one tree contributes a negative term p(1-p)(f+g), reflecting the fact that the drift paths move in opposite directions. We note that f+g is a substantial quantity. For India and the statistic f_3(India;CEU,Sathal), we believe that it is proportional to the genetic drift that occurred between ANI and ASI since their ancient divergence, which we estimate is about 0.092 in units comparable to F_{ST} (Figure 4). Thus, if there is mixture, the statistic can be negative.

Three of the terms contribute positively to the expected value, but one can contribute negatively because the drift takes opposite paths through some edges of the tree (Note S3 Figure 2):

$$\begin{aligned} E[f_3(X;Y,W)] &= \text{shared } (X\rightarrow Y) \text{ and } (X\rightarrow W) \text{ drift} \\ &= p^2(k+i) + p(1-p)k + (1-p)p(k-f-g) + (1-p)^2(k+j) \\ &= k + p^2 i - (1-p)p(f+g) + (1-p)^2 j \qquad \text{(S3.1)} \end{aligned}$$

Empirically, we find that when we calculate the statistic f_3(India;CEU,Santhal), 16 out of 18 Indian groups give highly negative values (Table 2). To understand why this occurs, we consider the genetic drift values (in units scale to be comparable to F_{ST}) from the model of history we fit in Figure 4 (derived in Note S4). We use CEU as an unmixed surrogate for the Ancestral North Indian (ANI) population and for the sake of argument, we consider Santhal to be an unmixed surrogate for the Ancestral South Indian (ASI) population:

$i = 0.0030$ = genetic drift in the ANI lineage since its divergence from CEU (Figure 4).

j = genetic drift in the ASI lineage since divergence from Santhal (we assume it is small).

k = genetic drift in each Indian Cline groups since mixture. This can be large in groups with histories of very strong founder effects (like the Chenchu or Vysya that are the only groups in Table 2 without significantly negative values) but is less than 0.006 for the others in Table 2.

$f+g = 0.092$ = genetic drift between ancestors of ANI and ASI after dispersion out of Africa.

The term (f+g) is much larger than i, j and k. Thus, in an Indian Cline group with substantial mixture, the term p(1-p)(f+g) may be large enough to exceed the magnitude of the three positive terms and to cause the value of the *3 Population Test* statistic to be negative.

For the argument above, we made a simplification in assuming that the Santhal (and CEU) were unmixed themselves. However, the sign of the *3 Population Test* statistic can not be affected by mixture in groups Y or W. If groups Y or W are mixed, the same patterns are expected as if they are unmixed. The reason is that we can split the trees algebraically into the unmixed components, in which case the expected value can be calculated as in Note 3 Figure 2. The observation of a significantly negative value for $f_3(X;Y,W)$ means unambiguously that the ancestors of group X experienced a history of mixture subsequent to their divergence from Y and Z.

Robustness of standard error calculation

To test for a reduction of the *3 Population Test* statistic below zero (providing evidence of mixture in the history of population X), and to test for significant deviations of other statistics from expectation, the simplest approach would be to treat all SNPs as independent, and then to assess the significance of tests of mixture. However, this is not appropriate, because not all SNPs are independent due to linkage disequilibrium (LD).

To address the problem of non-independence of SNPs, we assessed the variability of each test statistic (the f_2, f_3, f_4 and F_{ST} statistics described in the Appendix) using a Block Jackknife[17,18]. We divided the entire data set into 5 cM chunks (approximately 700 across the genome), choosing the span to be much larger than the typical extent of linkage disequilibrium among markers (typically tens of kilobases). We then dropped each chunk in turn and measured the variance of the test statistic weighting each chunk by its number of SNPs. This quantity could then be converted into a standard error by a standard formula[17,18].

In carrying out a Block Jackknife, it is important to assess whether the blocks are sufficiently large to correct for non-independence among SNPs. To assess this, we computed standard errors for different block sizes, for all pairwise calculations of F_{ST} for the four HapMap groups[19] (CEU, YRI, CHB and JPT). Compared with 1 cM blocks, standard errors increase by on average 8%, 10%, 14% and 15% respectively for blocks of 2 cM, 3 cM, 4 cM and 5 cM. The standard error is approaching an asymptote for blocks as large as 5 cM, and hence we conclude that blocks of 5 cM are sufficient to effectively correct for non-independence of SNPs.

We also explored whether the standard errors could be tightened by pruning the SNPs in our data to remove ones in LD. We implemented a greedy algorithm that removed SNPs in the data set based on the pattern of LD in West Africans, until all had a pairwise r^2 with neighboring SNPs less than a specified threshold (we explored $r^2 < 0.1$, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9). We found that standard errors could be reduced by 1%-9% by pruning SNPs, depending on the comparison. This slight increase in precision is promising for future analyses; however, we judged that it is not a major advantage over our strategy of simply using all SNPs.

Application of the *3 Population Test* suggests that all Indian Cline groups are mixed

We applied the *3 Population Test* to each of the 18 Indian Cline groups as X, using Y=CEU and W=Santhal as the reference groups. We chose the Santhal because the PCA of Figure 3 indicates that they have a relatively high proportion of ancestry from ASI compared with most of the Indian Cline groups. Although they deviate in the PCA in the direction of East Asian ancestry, a mixture history does not make a group less useful for the *3 Population Test* (see above).

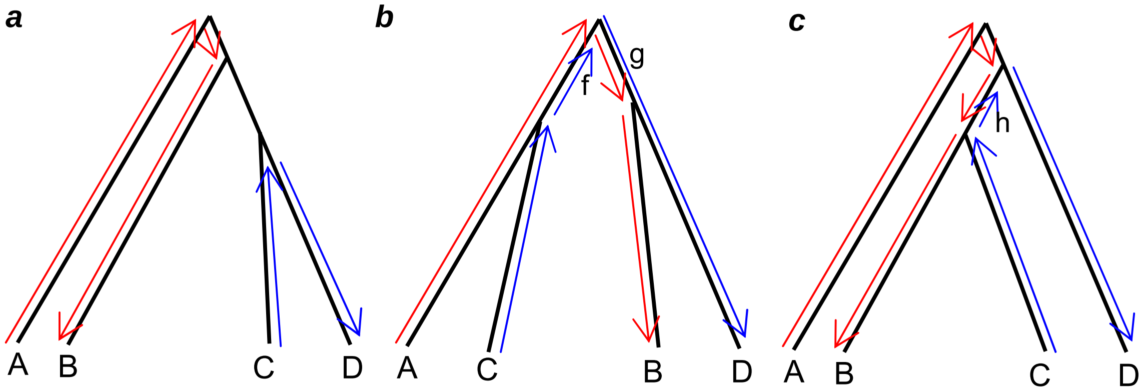
Results are presented in Table 2, and show significant evidence of mixture ($Z << -2$) for 16 of the 18 Indian Cline groups. The only Indian Cline groups that do not give a significant signal of mixture according to the *3 Population Test* are the Chenchu and Vysya. The lack of a significant result by the *3 Population Test* does not mean that these groups are not mixed. Instead, it is likely to reflect the fact that there has been substantial genetic drift in these two groups since mixture (*k* is large in the notation of Note S3 Figure 2), a fact that is also supported by the observation that the Chechu have a minimum F_{ST} of 0.052 with all other Indian groups and that the Vysya have a minimum F_{ST} of 0.011 (Table 1). Even though we do not obtain formal evidence of mixture by the *3 Population Test* for the Chenchu and Vysya, there is evidence that they are mixed based on the fact that the groups at the extremes of the Cline (Pathan and Mala) show evidence of mixture.

The *3 Population Test* is not prone to produce false-positives due to ascertainment bias

To generate a negative value of the *3 Population Test* statistic in the absence of population mixture, ascertainment bias would have to generate an *anti-correlation* in allele frequency differences between the Santhal and an Indian group, and between CEU and an Indian group. We can not see how such an artifact could occur based on the schemes used to select SNPs for medical genetics arrays. To produce this artifact, those studies must have tended to choose SNPs that were unusually different in frequency between some pairs of groups, and unusually similar in others. Indian groups were not used in SNP ascertainment, and hence this bias seems unlikely.

The *4 Population Test*

We also implemented a *4 Population Test*[14,15] that is again based on measuring genetic drift along lineages quantitatively (in the sense of variance in allele frequencies). The *4 Population Test* is more model-based, and hence provides more information about the nature of inferred population mixture that the *3 Population Test.*



Prediction of ((A,B),(C,D)) phylogeny
No overlap of paths, so E[f(A,B;C,D]=0

Prediction of ((A,C),(B,D)) phylogeny
Overlap of paths on edge f in same direction, so E[f(A,B;C,D] = f+g

Prediction of ((A,D) ,(B,C)) phylogeny
Overlap of paths on edge f in opposite directions, so E[f(A,B;C,D)] = -h

Note S3 Figure 3: Calculation of the expected value of the *4 Population Test* statistic f_4(A,B; C,D) for each of 3 possible phylogenetic trees consistent with unmixed populations. (a) If the phylogenetic tree is ((A,B),(C,D)), the expected value of the test statistic is 0, as the frequency differences between the first and second terms in the f_4 statistic take different and thus uncorrelated paths through the tree. (b) If the phylogenetic tree is ((A,C),(B,D)), the first and second terms are positively correlated and the expected value is positive, as the frequency differences share paths through the tree along the edges f and g. (c) If the phylogenetic tree is ((A,D),(B,C)), the frequency differences take paths in opposite directions along edge h, and the expected value is negative. We note that if the populations are mixed—so that no single phylogenetic tree can describe the data—the test statistic is not expected to be consistent with 0, unless the contributions from trees (b) and (c) balance. For example, if we write the populations as a linear combination of the three trees with weights p_A, p_B, and p_C respectively ($p_A+p_B+p_C=1$), and $p_B(f+g)=p_C(h)$, the two non-zero trees are expected to cancel and the *4 Population Test* will not detect mixture. We hypothesize that this is what is happening for the test statistic f_4(YRI,India; CEU,Onge) in the second-to-last column of Table S4 for some Indian groups. This test statistic is consistent with 0 for the Indian Cline groups from the Vysya through the Tharu, even though the position of these groups in the middle of the Cline indicates that they are mixed.

The idea of the *4 Population Test*, is that for each set of populations A, B, C and D, there are three possible unrooted trees that describe the relationships in the absence of admixture: ((A,B),(C,D)), ((A,C),(B,D)), and ((A,D),(B,C)). To test whether any one topology is consistent with the data, we can compute the frequency differences at each SNP—for example (p_A-p_B) and (p_C-p_D)—which should be uncorrelated to each other if the topology is ((A,B),(C,D)) and these two pairs form clades. To test the topology ((A,B),(C,D)) we calculate the statistic f_4(A,B; C,D) over all populations, and test for consistency with 0 by calculating a normally distributed Z-score (using a Block Jackknife to obtain a standard error correcting for linkage disequilibrium among SNPs as described for the *3 Population Test*). The expected value of the f_4 statistic can be calculated visually as shown in Note S3 Figure 3, and is only expected to be consistent with 0 if the topology is ((A,B),(C,D)). For all *4 Population Test* applications, we in practice use the outgroup West African (YRI) allele frequency in the denominator; this seems like a reasonable choice of outgroup since African history has little to do with Indian history (Appendix).

Application of *4 Population Test* suggests that all Indian Cline groups are admixed

To test the groups on the Indian Cline for consistency with all ASI-related ancestry, we used the *4 Population Test* to compare each group to Onge, CEU and YRI. For each Indian Cline group, we tested whether the statistics f_4(YRI,CEU; Onge,India) and f_4(YRI,Onge; CEU,India) are consistent with 0. If they are consistent, the population is consistent with having entirely ASI or ANI ancestry. However, even the most ASI-related group (Mala) gives a significant violation of the first topology (Z=13.8 standard deviations), and even the most ANI-related group (Pathan) gives a significant violation of the second topology (Z=14.3) (Table S4). Thus, all 18 groups on the Indian Cline have inherited a mixture of both ANI and ASI ancestry.

There is a third topology, ((YRI,India),(Onge,CEU)), that superficially appears to be consistent with the *4 Population Test* for four groups in the middle of the Indian Cline (Vysya, Naidu, Lodi and Tharu). However, the *4 Population Test* can artifactually give a passing Z-score if a group is mixed but has mixture proportions that cancel in just the right way (Note S3 Figure 3 legend). To further study this, we computed an extension of the *4 Population Test* that we call the "Negative Internal Branch Test". If the topology ((YRI,India),(Onge,CEU)) is correct, the inferred drift on the edge of the graph connecting the two clades should be positive, as reflected in the f_4 statistic f_4(YRI,Onge; India,CEU). However, the inferred drift is negative with high significance (Z<<-34 for all these groups), ruling out this topology.

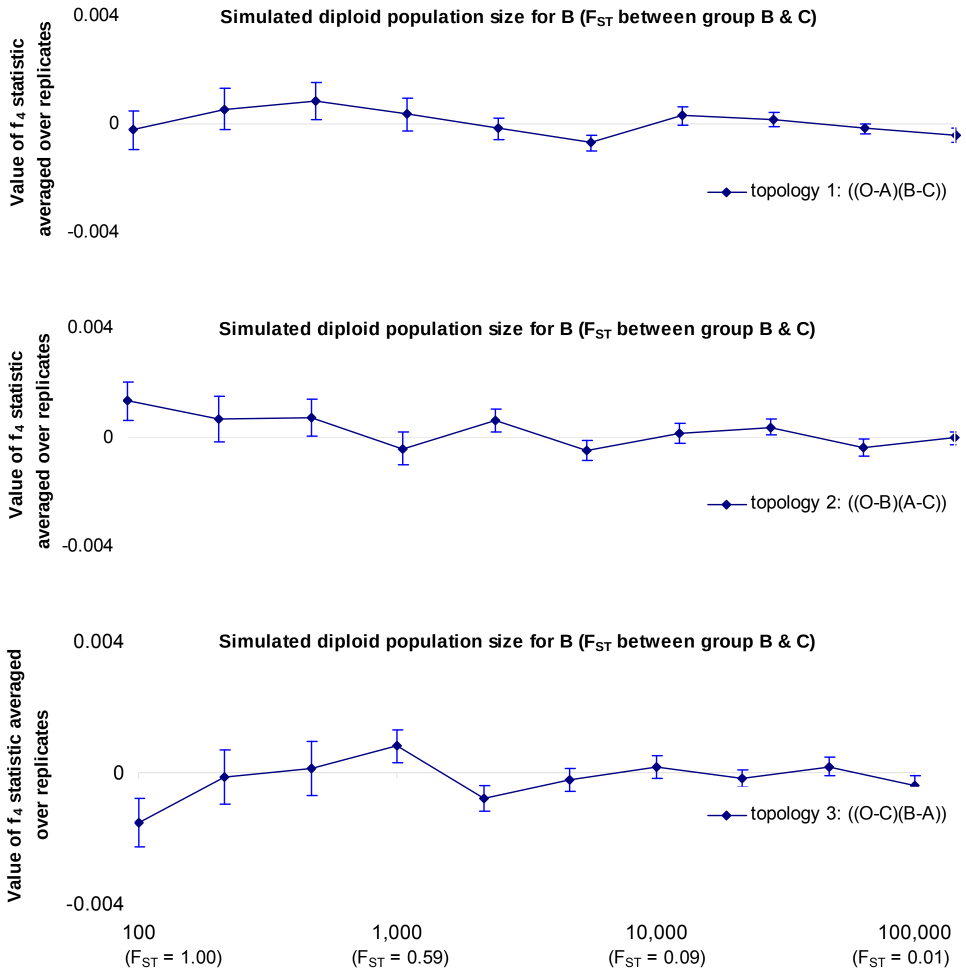
Minimal effect of SNP ascertainment bias on the *4 Population Test*

We determined that the statistical evidence of mixture that we detected in Indian Cline groups by the *4 Population Test* is not likely to be an artifact of "ascertainment bias" on the medical genetics arrays. While ascertainment bias is known to be a serious concern in the context of using genetic data to learn about history[20], and while in particular the site frequency spectrum is known to be very affected by ascertainment bias, f-statistic-based tests of gene flow (which use allele frequency differentiation information) are only weakly affected by ascertainment bias.

The most important empirical evidence that ascertainment bias is not causing false-positive signals of mixture using the *4 Population Test* is the fact that there are some sets of widely diverged groups in our data set that do satisfy the test. For example, in Table S4, we calculated the *4 Population Test* statistic f_4(YRI,Papuan;Dai,X), which tests whether the genealogy ((YRI,Papuan)(Dai,X)) is consistent with the data. We obtain a reasonable fit when X = Onge (Z-score of 1.7), but a very poor fit when X is any of the 18 Indian Cline groups (Z<<-9). These results are important in themselves, as they indicate that the Onge do not have a history of mixture from a West Eurasian-related population (ANI), even while all Indian Cline groups (including tribal groups like Chenchu and Kurumba) do have substantial ANI related mixture.

No evidence for "long branch attraction": high drift does not bias inference of topologies

A common concern in inferences of phylogenies is "long branch attraction", whereby highly diverged lineages tend to cluster together even though the truth is that the lineages are not most closely related. The most common context in which long branch attraction arises is in phylogenetic analysis to discover the relationship among species[21]. However, it has been suggested that it may also be relevant to reconstructing human population relationships. For example, an analysis of copy number variation in the HGDP found that a tree-building algorithm clustered Melanesian groups (Papuans and Bougainville islanders) with a highly drifted Pakistani population (the Kalash). The authors hypothesized that a reason why these populations might have erroneously formed a cluster in this analysis was long branch attraction[22].



Note S3 Figure 4: No evidence for a bias in the *4 Population Test* due to long branch attraction. We present the results of coalescent simulations comparing an outgroup O and three groups A, B and C that diverged simultaneously in a trifurcation 1,700 generations ago. We simulated all effective population sizes to be 10,000, except for groups B and C whose sizes we varied in simulation between 100 and 100,000. Standard errors were obtained from 50 replicates of a 60,000 locus simulation for each set of parameters. If the *4 Population Test* is robust to long branch attraction, the statistics f_4(O,A; B,C), f_4(O,B; A,C) and f_4(O,C; B,C) are expected to be consistent with 0, and in fact this is observed even when simulated genetic drift on lineages B and C is high (F_{ST} = 1.00) Thus, the f_4 statistics provide no support for the topology ((O,A),(B,C)) expected from long branch attraction.

To assess whether long branch attraction can bias our f_4 statistics, we carried out a series of coalescent computer simulations[23] of four populations that had a star-like phylogeny consistent with all three simple topologies. The outgroup "O" was simulated to have diverged from the other groups 4,000 generations ago, while groups "A", "B" and "C" were simulated to have diverged in a trifurcation 1,700 generations ago, so that no phylogeny was more correct than the others. The diploid population size was simulated to be 10,000 on all lineages, except for populations B and C, whose sizes we varied between 100 and 100,000 to explore the effect of long branch attraction. We simulated 10 chromosomes for each population. We included all

SNPs that were generated by the simulation software except ones that were monomorphic in the outgroup population (the frequency of the SNP in the outgroup population appears in the denominator of Equation 2, and hence it is required to be polymorphic). Each locus was simulated to be 1,000 base pairs in size and to have a mutation rate of $2x10^{-8}$ per generation.

Note S3 Figure 4 shows the results of the simulations. For each parameter combination, we simulated 60,000 loci (which in practice we found translated to about 130,000 SNPs), and performed 20 replicates of the simulation in order to obtain a mean and standard error. As expected for a phylogeny that is consistent with all three topologies, all f_4 statistics—f_4(O,A; B,C), f_4(O,B; A,C) and f_4(O,C; B,C)—are consistent with 0. We find no evidence of systematic bias in the statistics when the effective population sizes on lineages B and C become small (high genetic drift on these lineages). Thus, there is no evidence of a bias in our methods for inferring tree topologies.

Our tests of mixture are robust to the fact that f-statistics are non-linear for large drift.

A potential concern for our *3 Population Test* and for our *4 Population Test* is that f-statistics and genetic drift are only linearly related for small times on a diffusion time scale.

To understand why the f-statistics can be non-linear for large time scales, it is important to think about the denominator in the f-statistics. Suppose we normalize using YRI, as we do for the *4 Population Test*. In this case we calculate f_2, f_3 and f_4 using a weight $1/p^{YRI}(1-p^{YRI})$ where p^{YRI} is an estimate of the allele frequency in Yoruba. If we are interested in the genetic drift values in a part of the graph highly diverged from Yoruba (say East Asia), then a ‘better’ normalizer would in principle be $1/p^{CHB}(1-p^{CHB})$ where p^{CHB} is the allele frequency in Han Chinese. Using a normalizing population that is highly diverged from the populations of interest is expected to cause a shrinkage of genetic drift estimates in the part of the graph that is of interest, since genetic drift tends to drive allele frequencies to fixation and on average p(1-p) will decrease. In practice, however, this does not cause ‘local’ distortion in our graph, since all genetic drifts are shrunk by a similar magnitude. An analogy is representing a region of the globe on a flat map. It is easy to obtain local linearity but large-scale distortion is inevitable.

While the admixture graphs that we fit (Figure 4) can be somewhat distorted by the non-linearity of f-statistics and our choice of a normalizing population, this is not expected to generate false-positive evidence of mixture by the *3 Population Test* and *4 Population Test*. The reason for this is that these tests are based on detecting correlations of allele frequency. Non-linearity may affect the power of the tests, but will not generate false-positive correlations. Confirming this, the simulations that we report in Note S3 Figure 4 show that the *4 Population Test* remains robust even for extraordinarily high genetic drift.

We conclude that non-linearity affects f_2, f_3 and f_4 statistics, but that it in no way affects the validity of our statistical tests of population mixture (such as the *4 Population Test*). For this reason, we also do not believe that it affects inferences of the topology of the graph.

Note S4:
Relationship of Indian Cline groups to other groups worldwide

Based on the technology we developed to measure genetic drift on different lineages using f_2, f_3 and f_4 statistics (Appendix), we carried out a series of analyses that had the goal of finding a model of population divergence and mixture that was consistent with the data from the Indian and non-Indian groups.

(1) The first step was to study f_4 statistics among different sets of four populations, and to use the results to work out topologies relating YRI, CEU, ANI, ASI, Onge, Adygei, Papuans and Dai that were plausibly consistent with the data.

(2) The second step was to carry out a formal test of fit of the model, taking into account the uncertainty in the frequency differentiation measurements as assessed by a jackknife analysis.

For all the analyses in this section, we use data from 119,744 autosomal SNPs that overlap between our data (Affymetrix 6.0 array), the HGDP (Illumina 650Y array), and HapMap.

Sets of non-Indian groups that can be related by simple phylogenetic trees

Since all the Indian Cline groups have evidence of historical mixture (Table 2 and Note S3), our strategy for understanding the history of mixture in India was to go outside of the Indian mainland, and to identify sets of groups that could be related to each other in a simple way with requiring a history of mixture. We then attempted to relate these groups to the putative ancestral populations of the Indian Cline: ANI and ASI.

We began by applying a *4 Population Test* to determine that 4 non-mainland Indian groups—YRI, Papuans, Dai and Onge—could be related via a simple topology. (The Dai are an ethnic group from southern China who speak a Thai-related language, and we used them in preference to Han Chinese, since aspects of the data suggested a simpler mixture history.) We found that the topology ((YRI, Papuan),(Dai, Onge)) was consistent with the data (Z=1.7), but the topologies ((YRI, Dai), (Papuan, Onge)) and ((YRI, Onge),(Papuan, Dai)) were not (Z=5.9 and Z=4.2).

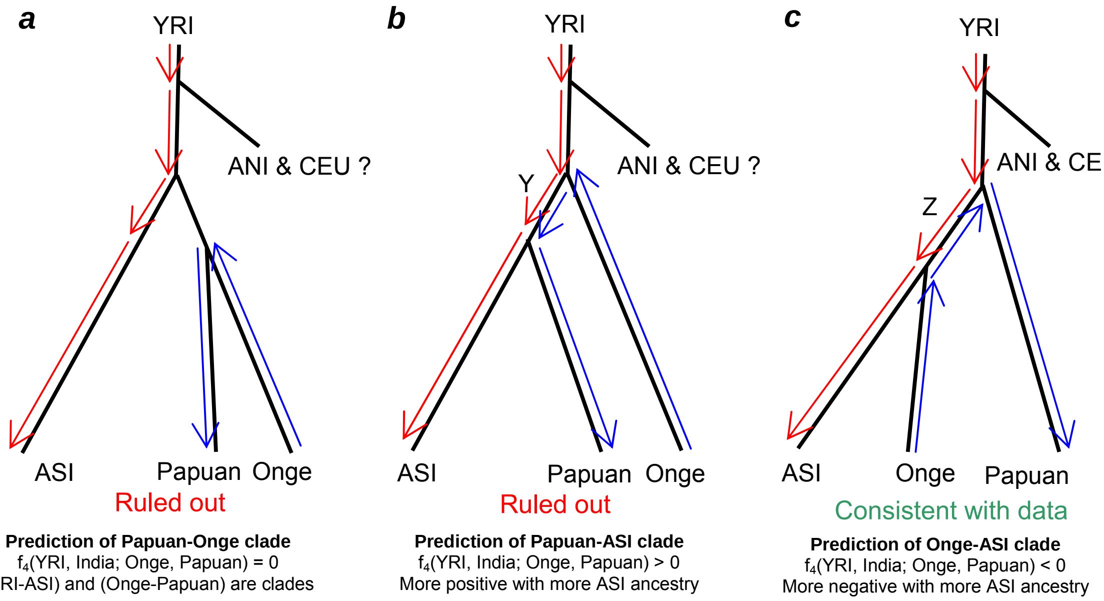
We also applied the *4 Population Test* to assess the relationship among the groups YRI, Onge, Adygei and CEU (the Adygei are a West Eurasian group from the Caucasus that we found to be useful for a number of analyses). We found that the topology ((YRI, Onge),(Adygei,CEU)) (Z=-2.4) is much more strongly supported than the alternatives (Z=37.4 and Z=38.0). Even though the best topology is not a perfect fit given our high resolution data (for example, some slight gene-flow from an East Asian ancestral group into the Caucasian Adygei could explain the Z=-2.4 observation), we proceed in what follows by assuming that this topology is correct.

Evidence that the Onge and ASI are a clade

We found that the Onge and the proposed ancestral population of South India (ASI) are consistent with forming a clade relative to the CEU.

We began by noting that the Onge and ASI are consistent with forming a clade relative to the Papuans by considering the *4 Population Test* statistic f_4(YRI,India; Onge,Papuan) and its trend as the Indian groups become increasingly distant from CEU in the PCA of Figure 3 (more ASI ancestry).

The expected value E[f_4(YRI,India; Onge,Papuan)] in terms of an f_4 statistic involving ASI can be calculated by writing each Indian Cline group as q(ASI)+p(ANI). The second term, which we can write as pf_4(YRI,ANI;Onge,Papuan), is equal to pf_4(YRI,CEU;Onge,Papuan) since ANI and CEU form a clade as discussed above. We empirically find that this term is very small relative to the first, and ignore it in practice. Thus, by analyzing the behavior of the statistic f_4(YRI,India; Onge,Papuan) for Indian Cline groups with decreasing proximity to CEU in the PCA of Figure 3, we can learn about the value of the quantity qf_4(YRI,ASI; Onge,Papuan) (Note S4 Figure 1).


a
YRI
ANI & CEU ?
ASI
Papuan
Onge
Ruled out
Prediction of Papuan-Onge clade
f_4(YRI, India; Onge, Papuan) = 0
(YRI-ASI) and (Onge-Papuan) are clades
b
YRI
ANI & CEU ?
Y
ASI
Papuan
Onge
Ruled out
Prediction of Papuan-ASI clade
f_4(YRI, India; Onge, Papuan) > 0
More positive with more ASI ancestry
c
YRI
ANI & CEU ?
Z
ASI
Onge
Papuan
Consistent with data
Prediction of Onge-ASI clade
f_4(YRI, India; Onge, Papuan) < 0
More negative with more ASI ancestry

Note S4 Figure 1: We tested 3 topologies relating ASI, Papuans and Onge, and found that only topology “c” is consistent with the data. To compute the inner product f_4(YRI,India;Onge,Papuan) pictorially, we examine the difference in frequency between YRI-India (red arrows) and intersect it with the difference Onge-Papuan (blue arrows). The expected value is the shared genetic drift between the two paths through the topology, and the sign is determined by whether the paths traverse the same direction. Since all Indian Cline groups are hypothesized to be a mixture of ASI and ANI, we can write India = q(ASI)+p(ANI), with the second term contributing nearly 0 to the f_4 statistic since the frequency difference between Onge and Papuan is independent of that between YRI and ANI (see above). Thus, the f_4 statistic is expected to equal q[f_4(YRI,ANI;Onge,Papuan)]. (a) If Papuans and Onge form a clade, then $f_4 = 0$ as the two frequency differences traverse independent parts of the tree. (b) If ANI and Papuans form a clade, they overlap on a branch with genetic drift “Y”, and the quantity is positive since the frequency differences move in the same way. (c) If ANI and Onge form a clade, the frequency comparison overlap on a branch with genetic drift “Z” and the frequency differences are in opposite directions so that a negative value is expected. The data support scenario “c”, since the f_4 statistics for the extended Indian Cline groups are all negative, and become more negative for groups with larger q and thus more ASI ancestry (Pathan -0.004; Mala -0.006).

Calculating f_4(YRI,India; Onge,Papuan) for a range of Indian Cline groups, we find that the statistic has a negative value, and becomes of larger magnitude for groups that have more ASI-related ancestry as assessed by proximity to CEU in the PCA of Figure 3: Pathan (-0.004), Kashmiri Pandit (-0.005), Bhil (-0.006) and Mala (-0.006).

To interpret the negative value of f_4(YRI,India; Onge,Papuan), we note that the expected value of f_4(YRI,ASI; Onge,Papuan) has a different expectation for the three possible simple topologies that could in theory relate these groups (Note S4 Figure 1):

(a) If Papuan and Onge form a clade, the expected value is 0 because there is no overlapping genetic drift history between these two groups and between YRI and ASI.
(b) If Papuan and ASI form a clade, then the expected value is non-zero because the (YRI-India) and (Onge-Papuan) frequency differences are expected to be correlated, in proportion to the length of the drift branch “Y” in Note S4 Figure 1. This quantity is expected to be positive because the frequency comparisons are in the same direction (the arrows flow the same way).
(c) If Onge and ASI form a clade, the expected value is negative because (YRI-India) and (Onge-Papuan) share genetic drift, marked by the length of the branch “Z” in Note S4 Figure 1. This quantity is expected to be negative because the frequency comparisons are in opposite directions (the arrows flow in different ways). Groups at the ASI end of the Indian Cline are expected to have larger absolute values because ASI is more heavily represented.

The observation that f_4(YRI,India; Onge,Papuan) is negative is thus consistent with the topology of Note S4 Figure 1c, and inconsistent with the alternatives. These results suggest that Papuans were the first Asian group to branch, and that Onge and ASI form a clade relative to Papuans.

Similar analyses place the ASI and Onge as a clade relative to CEU and Dai. In particular, we observe that the Z-scores for the *4 Population Test* for the topology ((YRI,CEU),(Onge,India)) become smaller as Indian Cline groups became increasingly distant from the CEU in the PCA of Figure 3 (Table S4). This is consistent with the hypothesis that for a group with all ASI ancestry, the topology would be ((YRI,CEU),(Onge,ASI)).

The Onge are the only ASI-related group without evidence of ANI mixture

We have demonstrated that the Onge are more closely related to the ASI than to any other of the HapMap or HGPD groups. However, we have not formally tested whether they have evidence of some ANI mixture (as do ASI-descended groups in mainland India).

To formally test whether the Onge have a history of ANI-related mixture (as do all the other groups in the Indian Cline; Table S4), we examined the *4 Population Test* statistic corresponding to the topology ((YRI,Papuan),(Dai,X)), which is f_4(YRI,Papuan; Dai,X). We studied this for each Indian Cline group separately and also for the Onge, who are ASI-related as describe above.

When X is an Indian Cline group, the Z-score for a deviation from zero is always extremely significant ($Z << -9$). From inspection of Table S4, we find that the Z-score is highly correlated ($r^2=0.84$) with the Z-score for the *4 Population Test* statistic corresponding to the topology ((YRI,CEU),(Onge,X)), indicating that they are both proportional to ANI ancestry.

When X=Onge, the *4 Population Test* statistic is not significantly different from zero (Z=1.7), in contrast with all the other Indian Cline groups where it is highly significantly below zero. Thus, there is no evidence for ANI-related mixture in the Onge. We conclude that the Onge are the only ASI-related group in this data set without the f_4-based evidence of ANI mixture.

Support for CEU and ANI forming a clade relative to the Adygei

We next attempted to discern the phylogenetic tree relating the three groups Adygei, CEU and ANI. To carry out this analysis, we constructed an argument analogous to that used to define the relationship of the Papuans, Onge and ASI in Note S4 Figure 1.

Writing each Indian Cline group as q(ASI)+p(ANI), and writing Y as an African or Asian group that has no evidence of West Eurasian-related gene flow (e.g. YRI, Onge or Papuan), we see that the expected value of the statistic E[f_4(CEU, Adygei; India, Y)] is about equal to pE[f_4(CEU, Adygei; ANI,Y)], since the term involving ASI is approximately 0.

The expected value E[f_4(CEU, Adygei; ANI,Y)] has three qualitatively different expectations depending on the topology relating the four groups (just as in the analogous Note S4 Figure 1):
(a) If CEU and Adygei form a clade, the expected value is = 0.
(b) If CEU and ANI form a clade, the expected value is > 0.
(c) If Adygei and ANI form a clade, the expected value is < 0.

Note S4 Table 1 shows that f_4(CEU, Adygei; ANI,Y) is positive, with Z-scores as high as 4.6 (for the topology (CEU,Adygei)(Pathan,YRI). Thus, the topology ((YRI,Adygei),(CEU,ANI)) is the only one consistent with our data.

We caution that the fact that the tree in which CEU and ANI are a clade is consistent with our data does not mean that the tree is accurate. More complex histories with multiple gene flow events are certainly possible, and at some level even likely, given the multiple historically documented waves of migration into India. Thus, one should not conclude that the above tree is “true”, but only that it fits the data better than other simple topologies.

Note S4 Table 1: f_4 analysis consistent with ANI and CEU forming a clade (Adygei outgroup)

Clade 1	Clade 2	Z-score for f_4 statistic
(CEU-Adygei)	(Pathan-YRI)	4.6
(CEU-Adygei)	(Pathan-Onge)	4.9
(CEU-Adygei)	(Pathan-Papuan)	3.0
(CEU-Adygei)	(Kashmiri Pandit-YRI)	2.9
(CEU-Adygei)	(Kashmiri Pandit-Onge)	2.9
(CEU-Adygei)	(Kashmiri Pandit-Papuan)	1.2

***f_3 Ancestry Estimation* suggests that the topology (YRI,(Adygei,(CEU,ANI),(ASI,Onge))) fits the data, and also provides estimates of mixture proportions along the Indian Cline**

The preceding results suggests that to at least a good degree of approximation:
- Onge and ASI form a clade with respect to all other groups we examined
- CEU and ANI form a clade with respect to all other groups we examined

Under the assumption that these clades are true, we developed a method of estimating mixture proportions along the Indian Cline. While it makes strong demographic assumptions, it also includes a goodness of fit test, which encourages us in the use of the analyses. The ancestry estimates are also consistent with the *Regression Ancestry Estimation* and *f_4 Ancestry Estimation* methods that we describe in Note S5, whose robustness we validate by computer simulation.

For *f_3 Ancestry Estimation*, we write each Indian Cline group as a linear combination of ANI and ASI: m_k(ANI)+(1-m_k)ASI. This should be interpreted as a group K having inherited ancestry with proportion m_k from ANI and (1-m_k) from ASI, followed by group-specific genetic drift.

We use three outgroups YRI, Papuan, Dai in our f_3 *Ancestry Estimation*. These are chosen to represent 3 groups, highly diverged from each other and from ANI, ASI and Onge. Since Onge and ASI are a clade, our demography implies that:

$$\begin{aligned} f_3(\text{Adygei;Outgroup,K}) &= m_k f_3(\text{Adygei;Outgroup,ANI})+(1\text{-}m_k)f_3(\text{Adygei;Outgroup,ASI}) \quad \text{(S4.1)} \\ &= m_k f_3(\text{Adygei;Outgroup,ANI})+(1\text{-}m_k)f_3(\text{Adygei;Outgroup,Onge}) \end{aligned}$$

We further note that f_3(Adygei;Outgroup,ANI) = z is small (because of the small genetic drift that appears to have occurred in the Adygei lineage since its split from ANI), and is independent of the choice of the outgroup.

We let K be an Indian Cline group and m_k be the corresponding ANI ancestry proportion, and then obtain a set of 54 = 3x18 equations:

$$f_3(\text{Adygei;Outgroup,K}) = (1\text{-}m_k)[f_3(\text{Adygei;Outgroup,Onge})] + (m_k)[z] \qquad \text{(S4.2)}$$

We can solve this set of equations by non-linear least squares, using each of the outgroups YRI, Papuan and Dai, and simultaneously fitting the m_k and z. We also tried allowing the coefficient z to depend on the outgroup, which would imply that the phylogeny was incorrect. However, this did not materially change the coefficients m_k or produce a significantly better fit, and hence we required z to be the same for all outgroups.

We fit the autosomal and X chromosome data separately. For the autosomal data we estimated z = 0.002, reflecting small genetic drift specific to the Adygei. On the X chromosome our estimates are noisy, and our best fit has substantial errors. Encouragingly, however, the autosomal estimates have a correlation coefficient around 0.9 with the chromosome X estimates. Ancestry estimates for each of the 18 Indian Cline groups are given in Table 2 and Table S4.

We computed the mean square error for f_3(Adygei;Outgroup,K) averaged across all Indian Cline groups and our 3 outgroup populations. The value was 4×10^{-7}, corresponding to a standard deviation of 0.0006. The standard error for our f_3 estimate is about 0.0004, and of course there is also a measurement error on the 'independent' variable f_3(Adygei;Outgroup,Onge). Thus, our fit seems satisfactory and a powerful check that our assumed demography is essentially correct. In particular, we can be confident that ASI and Onge are well described as a clade within the limits of our resolution, at least with respect to the outgroups Dai, Papuan, YRI and CEU.

ADMIXTUREGRAPH software for testing the fit of the model to the data

Having developed a model of population divergence and admixture that is consistent with multiple features of the data, we implemented a model fitting procedure that obtained a best estimate of the genetic drift on each branch of the tree (in units comparable to F_{ST}), as well as the mixture proportions in each Indian Cline group.

To implement this idea, we developed new software, ADMIXTUREGRAPH, which takes as input a proposed topology for the relationship among a set of groups, including population splits and mixture events but no specification of population sizes (an example is our simple model of history in Figure 4). In addition, the program takes as input a matrix of f_2 statistics between all pairs of groups, as well as the standard errors on the f_2 statistics (which are obtained by a Block Jackknife), and a covariance matrix that relates the errors on the f_2 statistics. The f_3 and f_4

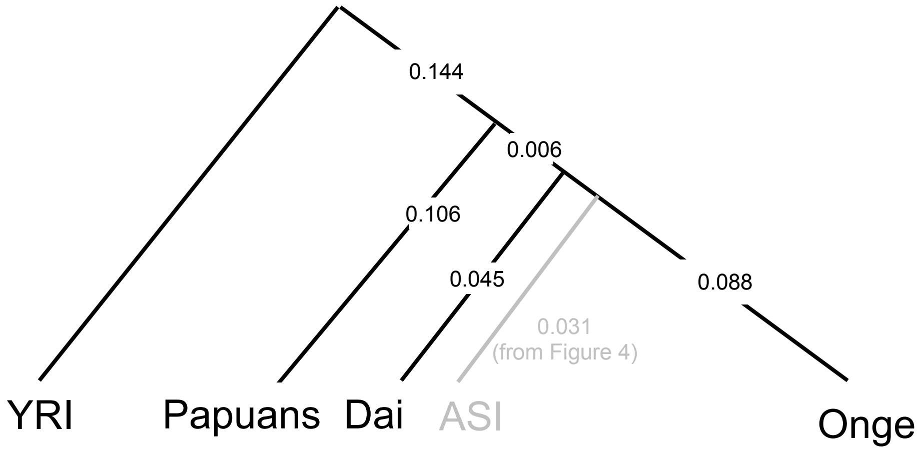
statistics can be calculated algebraically from the f_2 statistics, and hence the f_2 statistics provide all relevant information (Appendix).

With these inputs, ADMIXTUREGRAPH fits the f_2 statistics to the proposed model relating the groups, and thus obtains estimates of genetic drift on each edge of the topology, and mixture proportions for the mixed groups. ADMIXTUREGRAPH is currently a straightforward program that fits parameters to a model by minimizing the f_2 statistics. In the future, we hope to further develop it to screen through a space of possible topologies for those that match the data.

We first ran ADMIXTUREGRAPH on 7 groups—YRI, CEU, Onge, Pathan, Vaish, Meghawal and Bhil—and obtained the best fitting model in terms of genetic drift values and mixture proportions (Figure 4). To assess the adequacy of the fit, we computed 231 f-statistics with standard errors[17,18], and found that none was more than 3 standard deviations from expectation:

21	f_2 statistics:	(7x6)/2
105	f_3 statistics:	3(7x6x5)/(3x2)
105	f_4 statistics:	3(7x6x5x4)/(4x3x2)

A cautionary note is that the f_2, f_3 and f_4 statistics can all be computed from each other (Appendix), and hence there are no more than 21 degrees of freedom (the number of f_2 statistics). This is only modestly more than the 15 parameters (11 genetic drift values and 4 mixture proportions) that we estimate in the data. Thus, the stringency of our test is not as high as it might seem from the number of f-statistics we analyzed. We have nevertheless reported tests of all possible f_2, f_3 and f_4 statistics, as each probes a different linear combination of drift space and thus provides more opportunity to detect a discrepancy between the model and data.



Note S4 Figure 2: Application of the *4 Population Test* to the four groups YRI, Papuans, Dai and Onge shows that they are consistent with being related by a simple phylogenetic tree. We use the ADMIXTUREGRAPH software to estimate the genetic drift on each lineage in units comparable to F_{ST} based on the fit to this model. Based on the arguments in this note, the Onge and ASI are likely to form a clade relative to the Dai, Papuans and YRI (gray). The fact that the Onge and ASI are consistent with a clade, even thought all modern ASI-related populations now have a mixture of ANI-related ancestry, suggests that the ancestors of the Onge arrived in the Andaman Islands when there were still ASI-related populations without any ANI ancestry. This is also consistent with mtDNA analyses showing that some Andaman Islands and mainland ASI exchanged genes ~24,000 years ago[6].

A valuable feature of the ADMIXTUREGRAPH software is that it estimates the genetic drift that has occurred in each Indian Cline group since mixture. For example, ANCESTRYMAP estimates that the Vysya have experienced genetic drift of 0.0083 in units comparable to F_{ST}

since they arose as a mixture of the ancestral ANI and ASI. Table 2 reports this estimate for 18 Indian Cline groups, using f-statistics that are calculated in a way that is robust to the posibiliyt of inbreeding in the last few generations, which we verified occurs in some of the samples in our data set and which is known to be especially common in southern India[24]. These estimates provide the best quantification that we have for the intensity of founder events in Indian groups.

We finally used ADMIXTUREGRAPH to estimate the genetic drift that occurred on the lineages relating YRI, Papuans, Dai and Onge (Note S4 Figure 2). These groups represent a set of 4 highly diverged groups that have no substantial evidence for a history of mixture and that we use as a reference set for some of our analyses.

Limitations of the models

We have demonstrated that the topology of Figure 4 (YRI,((CEU,ANI),(ASI,Onge))) is consistent with the genetic data and that none of the simple alternative topologies is remotely consistent. However, the model is (intentionally) general, and makes no inferences about important features of history such as the timings of population splits, population expansions and contractions, or mixture events that do not disturb the topology. For example:

(a) We cannot rule out gene flow that might have occurred between African groups and West Eurasian groups since the out-of-African dispersal. Although such gene flow if it occurred would have been historically profound, it would not change the topology relating the groups, and hence would not have provided a positive score by our tests.

(b) We cannot rule out gene flow between Andaman Islanders and ASI since their divergence, even though mtDNA analysis has shown that Andaman Islanders and tribal East Indians share mtDNA ancestry in the last ~24,000 years[6]. This type of gene flow, if it occurred, would be very interesting. However, since the ASI and Onge form a clade in Figure 4, such gene flow would not violate the topology, and would not be detectable by our approach.

(c) We cannot distinguish between a history in which modern Indians descend from a single mixture between ANI and ASI, or multiple mixture events.

It is important to recognize that our methods may also not have statistical power to detect some violations of the topology, and thus Figures 3 is only meant to represent the simplest model that we could identify that is consistent with the data. For example, we used ADMIXTUREGRAPH to explore whether the data are consistent with a model in which there was some amount of gene flow from the ancestral ASI (after initial divergence from the Onge) into the ancestral population of Adygei, CEU and ANI. While we cannot currently distinguish between this history and the simpler one depicted in Figure 4, encouragingly our estimates of ANI ancestry proportion in Indian groups are expected to be reliable even in the presence of such a history, as indicated by our simulations including Asian gene flow in the ancestry of Adygei, CEU and ANI (Note S5).

We conclude that while our methods are robust for inferring tree topologies that are consistent with the data, the topology by no means provides a complete description of history. In future studies, it may be value to use methods like ours to infer the tree shape, and then make inferences about demography along each lineage by using additional information from sequence divergence data, allele frequency spectrum data, and linkage disequilibrium data.

Note S5:
Estimates of ancestry proportion on the Indian Cline

We have provided strong evidence that most groups in India are well described by an “Indian Cline”, whereby the groups have different proportions of ancestry inherited from an Ancestral North Indian population (“ANI”) and an Ancestral South Indian population (“ASI”).

Estimating ancestry in the absence of groups that are good surrogates for the ancestral populations is a difficult and unsolved problem in population genetics, and we are not aware of any published methods that are able to obtain estimates of ancestry proportion in this context in a way that has been validated by computer simulations over a range of scenarios. In the text we showed that the Adygei from the Caucasus and North Europeans have allele frequencies that are relatively close to those of the ANI, so that they can be used as reasonable surrogates. However, we found no modern group that provides a good proxy for the allele frequencies in the ASI. In the text, when we modeled 20 Indian groups (excluding the Siddi, Nyshi, Ao Naga, Onge and Great Andamanese) as the best fitting linear combination of Han Chinese (CHB) and North Europeans (CEU)[25,26], we obtained a poor fit of F_{ST}=0.026.

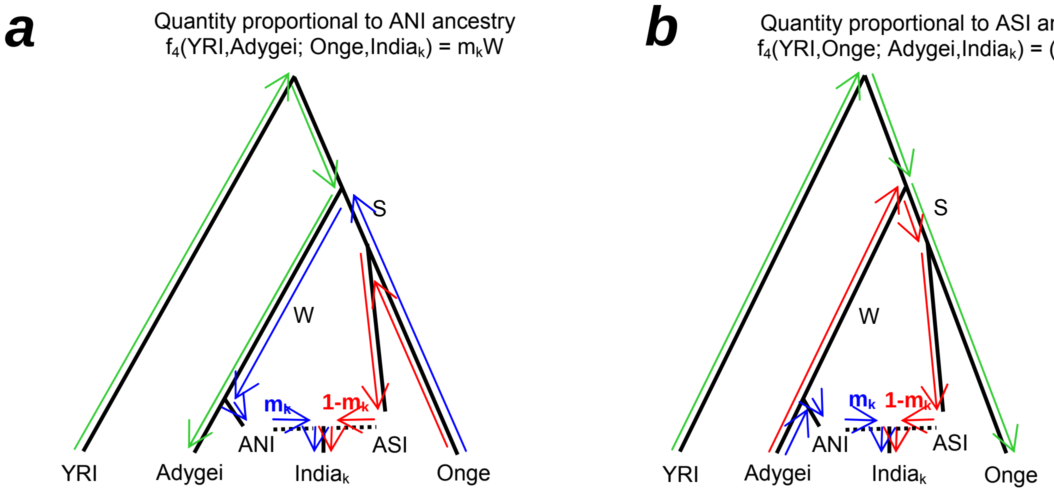
The challenge of estimating ancestry proportions in Indian Cline groups is one of obtaining absolute, unbiased estimates of ancestry. Relative estimates of ancestry are easy to obtain. For example, by inspecting the PCA of Figure 3, we can see that the groups closest to CEU have the most relatedness to ANI, while the groups furthest away have the least. The relative spacings of the groups in the PCA are linearly related to the differences in ancestry proportion (Figure S2d). However, PCA provides no information about absolute proportions of ancestry. The PCA results could in principle be consistent with very different ranges of ANI ancestry; for example, 5-20%, 35-80%, or 80-95%.

In what follows we introduce three novel procedures for inferring the proportion of ancestry in groups with a history of mixture for cases where we do not have access to good surrogates for the ancestral populations: *Regression Ancestry Estimation, f_4 Ancestry Estimation, and f_3 Ancestry Estimation*. All three methods are based on fitting f_3 and f_4-statistics (Appendix) to the data under the assumption that the phylogenetic tree relating the groups is known. Because these methods are based on f_3 and f_4 statistics that are not much affected by ascertainment bias, these methods for estimating ancestry are robust to SNP ascertainment, as we show via computer simulations.

Regression Ancestry Estimation and application to the Indian Cline

We introduce a novel idea, *Regression Ancestry Estimation*, which obtains unbiased estimates of ancestry proportions without requiring modern groups that are good proxies for the ancestral populations. The procedure only requires that: (i) The analyzed groups have a range of ancestry proportions, and (ii) data are available from 3 groups that are not mixed relative to each other: we use Adygei (or CEU), Onge and YRI.

We assume that modern groups on the Indian Cline are mixtures of two different ancestries, with a proportion m_k of ancestry from the ANI and a proportion $1-m_k$ from the ASI. For our analysis we assume the topology given in Note S5 Figure 1, where the ASI and Onge form a clade, and where the ancestors of ANI and West Eurasians diverged from this clade at the African root. This topology is based on that of Figure 4 in the main text and the discussions of Note S4.


a
Quantity proportional to ANI ancestry
f_4(YRI,Adygei; Onge,India$_k$) = m_kW
S
W
m_k
1-m_k
ANI
ASI
YRI
Adygei
India$_k$
Onge
b
Quantity proportional to ASI ancestry
f_4(YRI,Onge; Adygei,India$_k$) = (1-m_k)S
S
W
m_k
1-m_k
ANI
ASI
YRI
Adygei
India$_k$
Onge

Note S5 Figure 1: The idea of *Regression Ancestry Estimation* is based on our working model of ancient Indian history in Figure 4. The genetic drift specific to each branch of the tree is designated with a capital letter. The ANI ancestry proportion in each Indian Cline group is designated as m_k. (a) The expected value of the statistic f_4(YRI, Adygei; Onge, India$_k$) can be computed visually by overlapping the path taken by the first frequency comparison YRI→Adygei (indicated in green), with the second comparison Onge→India$_k$ (indicated in blue and red). Since there is no overlap between the two terms for the proportion of an Indian group's ancestry that derives from ASI (the green and red paths do not overlap), the expected value of this quantity is entirely determined by the overlap between the green and blue paths weighted by the ANI mixture proportion: m_kW. Thus, the f_4 statistic is proportional to the ANI proportion m_k in each Indian Cline group, up to an arbitrary unknown constant. (b) The expected value of the statistic f_4(YRI, Onge; Adygei, India$_k$) can be computed visually in a similar way, and is expected to equal (1-m_k)S, proportional to the ASI proportion in each Indian Cline group.

For each Indian Cline group, we compute two inner products over all SNPs *i*, using the frequencies of the SNP in each group and Equations S5.1 and S5.2. These inner products, or f_4 statistics, use normalizations that cause each SNP to contribute approximately the same amount of information to the measurement (Appendix). The normalization weights each SNP by a quantity that is proportional to its expected genetic drift in the ancestral groups $p_i(1\text{-}p_i)$ based on the binomial variance characteristic of genetic drift[27]. We use an outgroup YRI, to estimate the frequency, and thus only use SNPs that are polymorphic in YRI.

$$f_4(YRI, Adygei; Onge, India_k) = \frac{\sum_{i=1}^{n}\left(p^{YRI}{}_i - p^{Adygei}{}_i\right)\left(p^{Onge}{}_i - p^{India}{}_i\right)}{\sum_{i=1}^{n} p^{YRI}{}_i\left(1 - p^{YRI}{}_i\right)} \qquad \text{(S5.1)}$$

$$f_4(YRI, Onge; Adygei, India_k) = \frac{\sum_{i=1}^{n}\left(p^{YRI}{}_i - p^{Onge}{}_i\right)\left(p^{Adygei}{}_i - p^{India}{}_i\right)}{\sum_{i=1}^{n} p^{YRI}{}_i\left(1 - p^{YRI}{}_i\right)} \qquad \text{(S5.2)}$$

It is simple to compute the expected values of these quantities. The genetic drift is just the visual intersection of the genetic drift between the two populations in the first term and the two in the second term of the numerators of expressions (1) and (2), as diagrammed in Note S5 Figure 1.

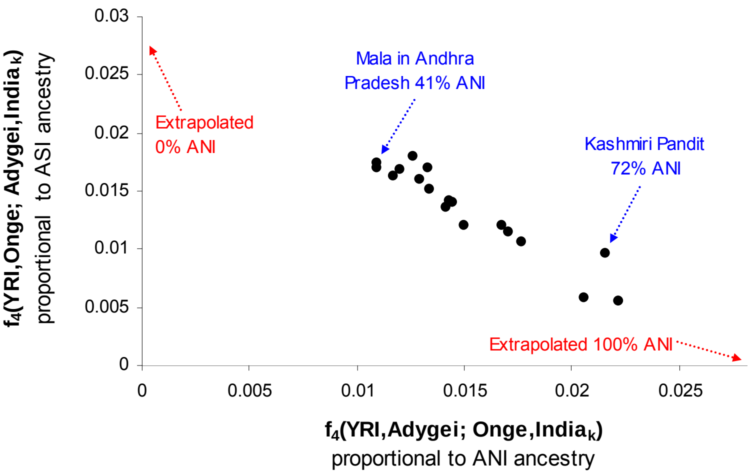
$$E[f_4(\text{YRI,Adygei; Onge,India}_k)] = m_kW \quad (3)$$

$$= m_k \times (\text{YRI to Adygei drift}) \text{ intersected with } (\text{Onge to ANI drift + post-mixture drift})$$
$$+ (1\text{-}m_k) \times (\text{YRI to Adygei drift}) \text{ intersected with } (\text{Onge to ASI drift + post-mixture drift})$$

$$E[f_4(\text{YRI,Onge; Adygei, India}_k)] = (1\text{-}m_k)S \quad (4)$$

$$= m_k \times (\text{YRI to Onge drift}) \text{ intersected with } (\text{Adygei to ANI drift + post-mixture drift})$$
$$+ (1\text{-}m_k) \times (\text{YRI to Onge drift}) \text{ intersected with } (\text{Adygei to ASI drift + post-mixture drift})$$

Here, W is the genetic drift on the ANI side of the tree, and S is the genetic drift on the ASI side. These are expected to be the same for all populations, since under the Indian Cline hypothesis all are derived from the same mix of two ancestral populations. Thus, when we plot f_4(YRI,Adygei; Onge,India) on the x-axis and f_4(YRI,Onge; Adygei, India) on the y-axis, they should fall along a line with a negative slope. We in fact observe this in real data (Note S5 Figure 2). If we have data from at least two groups with different proportions of ancestry, we can fit a trend-line to the data, and extrapolate where they intersect x-axis and y-axis to estimate the values of W and S, which in turn allow us to interpolate the ANI mixture proportion of any group.



Note S5 Figure 2: *Regression Ancestry Estimation* along the Indian Cline. We use f_4 statistics to estimate a statistic f_4(YRI, Adygei; Onge, India$_k$) that is expected to equal m_kW for each Indian group India$_k$, where W is the genetic drift that occurred ancestral to the divergence of Adygei and ANI. This value should be proportional to the ANI ancestry in each Indian Cline group. We similarly calculate f_4(YRI, Onge; Adygei, India$_k$), which we expect to equal $(1\text{-}m_k)S$ and should be proportional to the ASI drift in each Indian Cline group. By carrying out a least-squares fit to the 18 groups, we extrapolate the x- and y-intercepts, which correspond to the values expected for groups with entirely ANI and entirely ASI ancestry. We then interpolate the mixture proportions.

An important feature of *Regression Ancestry Estimation* is that its estimates are not expected to be affected by the fact that the YRI, Adygei and Onge to which we compare the Indian groups are poor surrogates for the ancestral populations, or by the fact that there may have been genetic drift in Indian groups after mixture. Because our estimates are dependent only on the W and S quantities reflecting genetic drift that occurred deep in the phylogenetic tree after the out-of-Africa dispersal, the ancestry estimate are independent of:

- The genetic drift specific to the YRI.
- The genetic drift specific to the Adygei.

- The genetic drift specific to the Onge.
- The genetic drift that occurred in each Indian Cline group *k* since mixture.

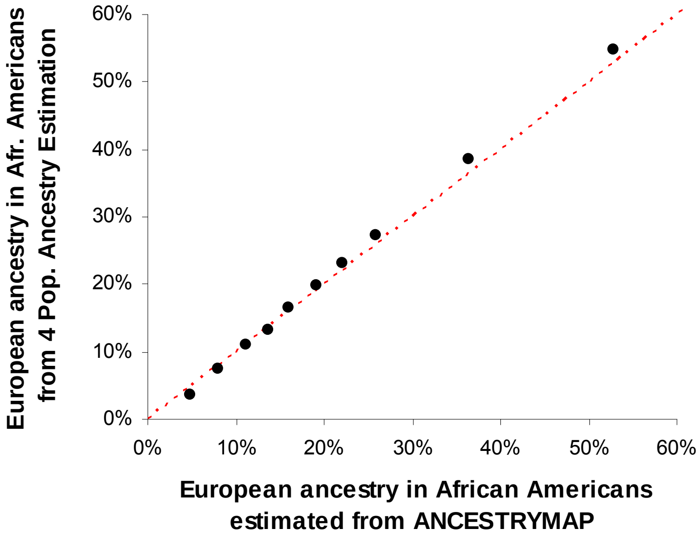
Application of *Regression Ancestry Estimation* to our data (Notes S5 Figure 2 suggests that the 18 groups in the Indian Cline range from 41% European-related ancestry (Mala) to 81% European-related ancestry (Pathan). To explore the robustness of our ancestry estimates, we repeated the entirely analysis substituting Papuans for Onge, and encouragingly obtained similar estimates (not shown). We did not obtain standard errors on the ancestry estimates using the *Regression Ancestry Estimation* procedure.

Validating *Regression Ancestry Estimation* by empirical comparisons and simulation

Empirical validation of *Regression Ancestry Estimation* in African Americans

To demonstrate empirically that *Regression Ancestry Estimation* can produce accurate estimates of ancestry, we first applied it to a group with a well understood history of recent mixture: African Americans. For African Americans, Nigerians (YRI) and European Americans (CEU) have been shown to be excellent surrogates for the ancestral African and European populations (F_{ST} of <0.001 between African Americans and the optimal mixture of these two groups)[28]. Thus, we can compare the results of *Regression Ancestry Estimation* to a "gold standard" estimate of ancestry in African Americans by treating the group as a mix of YRI and CEU, and estimating ancestry proportion using standard methods[28].

We implemented this analysis using real data from 89 African Americans who were genotyped on an Affymetrix 6.0 array and that we had previously analyzed by PCA to remove individuals who had some ancestry other than African or European[29]. We rank-ordered these samples into deciles based on estimates from the ANCESTRYMAP software[28], and treated these 10 deciles as groups on an 'African American Cline'. To estimate a quantity proportion to European ancestry 'W' and African ancestry 'S', we carried out the same procedure as for the Indian Cline, substituting YRI in place of Onge, and substituting San (Bushmen from HGDP) in place of YRI. The ANCESTRYMAP and *Regression Ancestry Estimates* are very similar (Note S5 Figure 3), providing confidence in the new method.



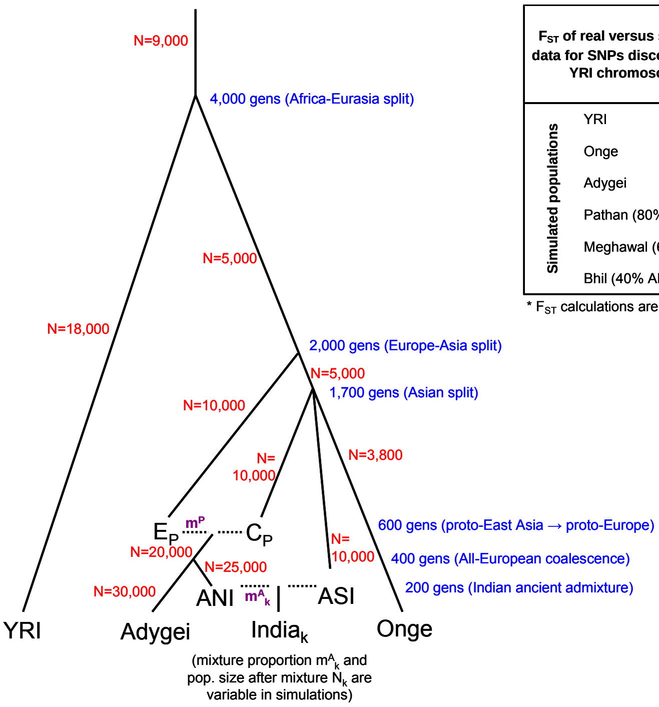
Note S5 Figure 3: Validation of *Regression Ancestry Estimation* by comparison to conventional estimates of mixture proportion in African Americans where we have good surrogates for the ancestral populations. We focused on 89 African Americans for which we had dense genotyping data from an Affymetrix 6.0 array. We rank-ordered these samples by their ANCESTRYMAP European ancestry estimates into 10 deciles, and used these as inputs into the *Regression Ancestry Estimation* (using YRI, San and Adygei in place of Onge, YRI and Adygei respectively in the Indian analysis). The *Regression Ancestry Estimates* closely match the results from ANCESTRYMAP when YRI and CEU are treated as ancestral populations.

Coalescent computer simulations used to validate of *Regression Ancestry Estimation*

We further tested the robustness of *Regression Ancestry Estimation* by carrying out coalescent computer simulations[30] of data from 15 groups with histories simulated to be similar to those of the groups used in our study. We simulated 4 non-Indian Cline groups with sample sizes chosen to match that in our data and to have crudely similar population histories; we called these simulated groups YRI (n=56), Onge (n=9), Adygei (n=17) and CEU (n=55)[15]. We also simulated 11 Indian Cline groups with ANI mixture proportions of 0-100% ancestry in 10% increments (5 from each group). We simulated 10,000,000 trees for each parameter set. To generate SNPs, we assumed loci of 1,000 base pairs without recombination, and distributed SNPs on the genealogy using a Poisson process assuming a mutation rate of 2×10^{-8} per base per generation.

The demographic parameters were chosen to roughly mimic parameters that emerged from previous studies of human historical expansions and contractions[15]. The demographic parameters that we used (split times and effective population sizes between splits) are presented below along with the assumed constant population sizes during each epoch (Note S5 Figure 4). For the purpose of validating the *Regression Ancestry Estimation* procedure, it was not important that the parameters exactly matched the truth, but we did adjust the parameters so that the pairwise F_{ST} matrix was a qualitative match to real data (Note S5 Figure 4, top right).

4,000 gens ago	Split of West African and Eurasian ancestors
2,000 gens ago	Split of ANI and ASI ancestors
1,700 gens ago	Split of Asian populations ('proto-East Asia', ASI, and Onge)
600 gens ago	Gene flow from 'proto-East Asia' into the ancestral population of ANI and West Eurasians, so that the proto-West Eurasian/ANI mixture proportion is m^P. Most of our simulations assume m^P=100% (no gene flow), but we vary this parameter to test the robustness of our procedure if the ancestors of ANI and West Eurasians were mixed.
400 gens ago	Split of CEU and Adygei
200 gens ago	Age of the ancient mixture event that formed the Indian Cline.

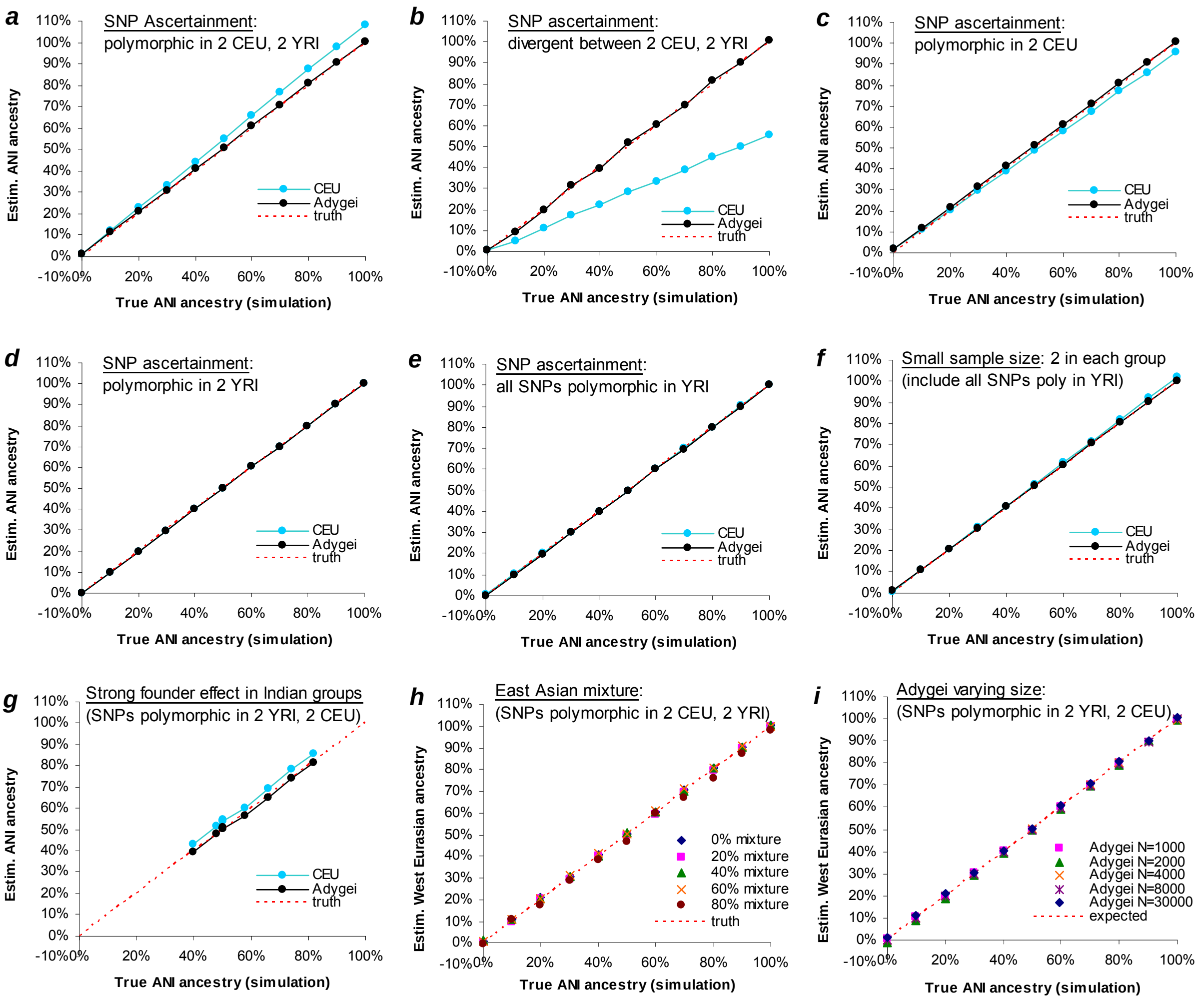

N=9,000
4,000 gens (Africa-Eurasia split)
N=5,000
N=18,000
2,000 gens (Europe-Asia split)
N=5,000
1,700 gens (Asian split)
N=10,000
N=
10,000
N=3,800
E_P
m^P
C_P
600 gens (proto-East Asia → proto-Europe)
N=20,000
N=25,000
N=
10,000
400 gens (All-European coalescence)
N=30,000
ANI
m^A_k
ASI
200 gens (Indian ancient admixture)
YRI
Adygei
$India_k$
Onge
(mixture proportion m^A_k and
pop. size after mixture N_k are
variable in simulations)

F_{ST} of real versus simulated data for SNPs discovered in 2 YRI chromosomes		Real populations*					
		YRI	Onge	Adygei	Pathan	Meghawal	Bhil
Simulated populations	YRI		0.228	0.145	0.136	0.139	0.141
	Onge	0.193		0.156	0.138	0.129	0.123
	Adygei	0.146	0.160		0.010	0.025	0.039
	Pathan (80% ANI)	0.132	0.135	0.019		0.008	0.017
	Meghawal (60% ANI)	0.129	0.125	0.029	0.007		0.009
	Bhil (40% ANI)	0.130	0.122	0.046	0.018	0.007	

* F_{ST} calculations are based on the merged data set of 119,744 autosomal SNPs.

Note S5 Figure 4: (Left) Model of population split times and sizes used in simulations. We simulated 15 populations meant to be similar to 4 non-Indians (YRI, Onge, Adygei and CEU), 11 Indian Cline groups with mixture proportions of 0-100%, and population sizes after mixture of 30,000 unless otherwise specified. (Top right) We show the observed F_{ST} (above diagonal) vs. the simulated F_{ST} (below diagonal) under our model, using groups with 80%, 60% and 50% European-related ancestry to approximately represent the Pathan, Meghawal and Bhil respectively.

For our simulations, we computed the *Regression Ancestry Estimates* using the simulated YRI, Onge, and either Adygei or CEU as non-Indian groups, and a subset of 5 of the Indian Cline groups with ANI mixture proportions of 40%, 50%, 60%, 70% and 80%. This range of mixture proportions was chosen to match what we think is true for the real Indian Cline.



Note S5 Figure 5: Robustness of *Regression Ancestry Estimation* as assessed by simulation. (a-e) For five different SNP ascertainments, the estimates are nearly unbiased relative to the truth as long as we use Adygei as the ANI-related ancestral population. (f) The estimates continue to be unbiased even with tiny sample sizes (2 for each simulated Indian group matching our smallest sample sizes and also for the simulated CEU, Adygei, YRI and Onge). Focusing on Adygei as the ANI-related population, the estimates continue to be robust even if (g) the effective population sizes of the Indian Cline groups vary from 1,000-30,000 after mixture, (h) the ANI-related ancestral population received up to 80% East Asian related ancient mixture, or (i) the Adygei are a poor surrogate for ANI (effective population size of 1,000-30,000 after their separation for ANI).

Regression Ancestry Estimates are robust to ascertainment bias

We carried out a series of simulations to explore how SNP ascertainment affects the inferences of ancestry. This is important to consider, as the ascertainment of SNPs in the Affymetrix 6.0 and Illumina 650Y arrays was influenced by the CEU and YRI groups from HapMap[19]. We examined 5 ascertainment procedures meant to mimic an extreme range of SNP ascertainments:

(i) Discovery as polymorphic in 2 CEU chromosomes

(ii) Discovery as polymorphic in 2 YRI chromosomes
(iii) Discovery as polymorphic in a mixture of 2 CEU and 2 YRI chromosomes
(iv) Discovery as completely different in frequency in 2 CEU and 2 YRI chromosomes
(v) No ascertainment except for the requirement of polymorphism in YRI

Note S5 Figure 5 (panels a-e) shows that when we use Adygei as the ANI-related group, we obtain ancestry estimates in the Indian Cline groups that are unbiased. However, using CEU as the ANI-related ancestral population, the estimates can be somewhat biased except for scenario (i). This suggests that SNP ascertainment in CEU induces artifactual correlations in the tree.

Regression Ancestry Estimates are robust to small sample sizes

The Indian Cline groups had sample sizes as low as 2 (for the Srivastava). To assess whether *Regression Ancestry Estimation* provided unbiased estimates of ancestry even for such small sample sizes, we carried out a computer simulation of low samples sizes. We simulated 2 samples (4 chromosomes) for the CEU, YRI, Adygei, Onge and all Indian Cline groups, and restricted analysis to SNPs that were polymorphic in the YRI. *Regression Ancestry Estimation* continues to provide unbiased estimates of ancestry proportion (Note S5 Figure 5f).

Regression Ancestry Estimation is robust to variable Indian population sizes since mixture

Most of our simulations were carried out by assuming that the Indian Cline groups after mixture were of the same size, which we picked to be sufficiently large (N = 30,000) that there was expected to be little drift in the simulated 200 generations since admixture. To assess the robustness of our procedure to variable effective population sizes (simulating, for example, what is seen in the Chenchu), we simulated 5 groups with mixture coefficients of 0.5 and sizes of 16,000, 8,000, 4,000, 2,000, 1000. There is no evidence that the effective population size of an Indian Cline group after mixture affects inferences when we use the Adygei as the ANI-related group (Note S5 Figure 5g).

Regression Ancestry Estimates are robust to mixture in the ancestors of ANI

We explored whether the assumption that the ancestral ANI and West Eurasian populations are unmixed since the dispersal from Africa might be problematic. If there has been very ancient mixture in the ancestral population—for example, due to an isolation-by-distance phenomenon involving gene flow from an ancestral population related to East Asians[31]—it is possible that this could bias our inferences. We simulated scenarios in which the ANI ancestral population derives 0%, 20%, 40%, 60% and 80% from a proto-Asian population. The ANI-related ancestry estimates are accurate when we use Adygei as the ANI-related group (Note S5 Figure 5h).

Regression Ancestry Estimates are robust to using an inaccurate modern surrogate for ANI

We also carried out simulations in which the effective population size of the Adygei (used as the surrogate ancestral population for the ANI-related ancestors of India) was much less than the relatively large size of 30,000 that we chose to use in our simulations (which was motivated by the low group-specific genetic drift in the Adygei inferred in Note S4). We find that even a substantial amount of genetic drift in the Adygei does not bias the ancestry estimates (Note S5 Figure 5i), confirming that *Regression Ancestry Estimation* does not require the availability of accurate ancestral populations for ANI. In practice, the situation is much better than this. We estimate that the F_{ST} between Adygei and ANI is in fact modest (about 0.007, analysis not shown), so that the Adygei provide a reasonable proxy for the ANI ancestral population.

f_4 and f_3 Ancestry Estimation

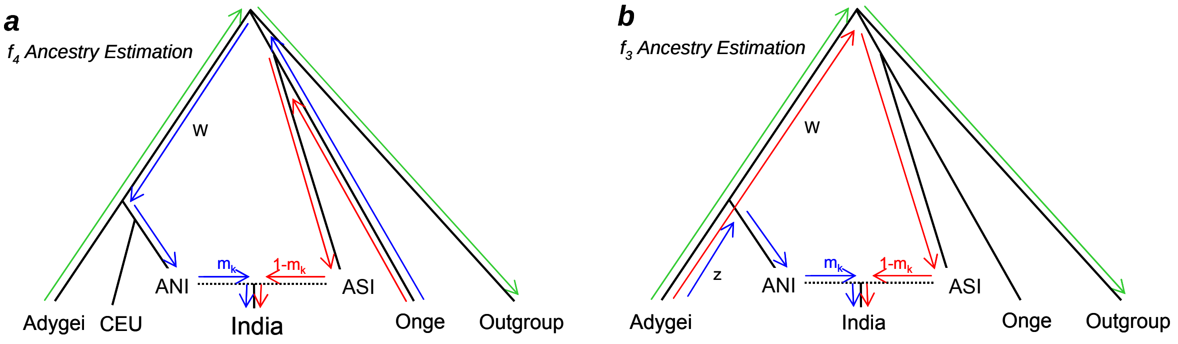
f_4 Ancestry Estimation

f_4 Ancestry Estimation is based on calculating the ratio of two f_4 statistics for each group:

$$\frac{f_4(Adygei, Outgroup; Onge, India_k)}{f_4(Adygei, Outgroup; Onge, CEU)} = \frac{\sum_{i=1}^{n}\left(p^{Adygei}{}_i - p^{Outgroup}{}_i\right)\left(p^{Onge}{}_i - p^{India}{}_i\right)}{\sum_{i=1}^{n}\left(p^{Adygei}{}_i - p^{Outgroup}{}_i\right)\left(p^{Onge}{}_i - p^{CEU}{}_i\right)} \qquad (5)$$

The value of this quantity can be estimated visually (Note S5 Figure 6a). The numerator, which should be proportional to the overlap between the drift paths Adygei→Outgroup and Onge→India$_k$, is expected to equal $-m_kW$, where m_k is the ANI ancestry proportion in an Indian Cline group. The denominator should be equal to -W, since m_k is effectively 1 for this group. By taking the ratio of the numerator and denominator, we can obtain an estimate of the ANI mixture proportion in an Indian Cline group m_k (assuming that the model of history in Figure 4 is correct). We can calculate a standard error by a Block Jackknife[17,18].

We implemented *f_4 Ancestry Estimation* using the three outgroups from Note S4 Figure 2 (Papuan, YRI, and Dai), and found that they gave consistent results. We quote results with Papuans as the outgroup in Table S4, because we found that this in practice gave the smallest standard errors (an average of 1.5% for each of the Indian Cline groups).



f_4(Adygei, Outgroup; Onge, India) is expected to be proportional to $-m_kW$: the overlap between Adygei→Outgroup (green) & Onge→India (blue and red) drift.

f_4(Adygei,Outgroup; Onge, CEU) is similarly expected to be proportional to -W. Thus, the ratio is expected to equal m_k, the ANI mixture in each group.

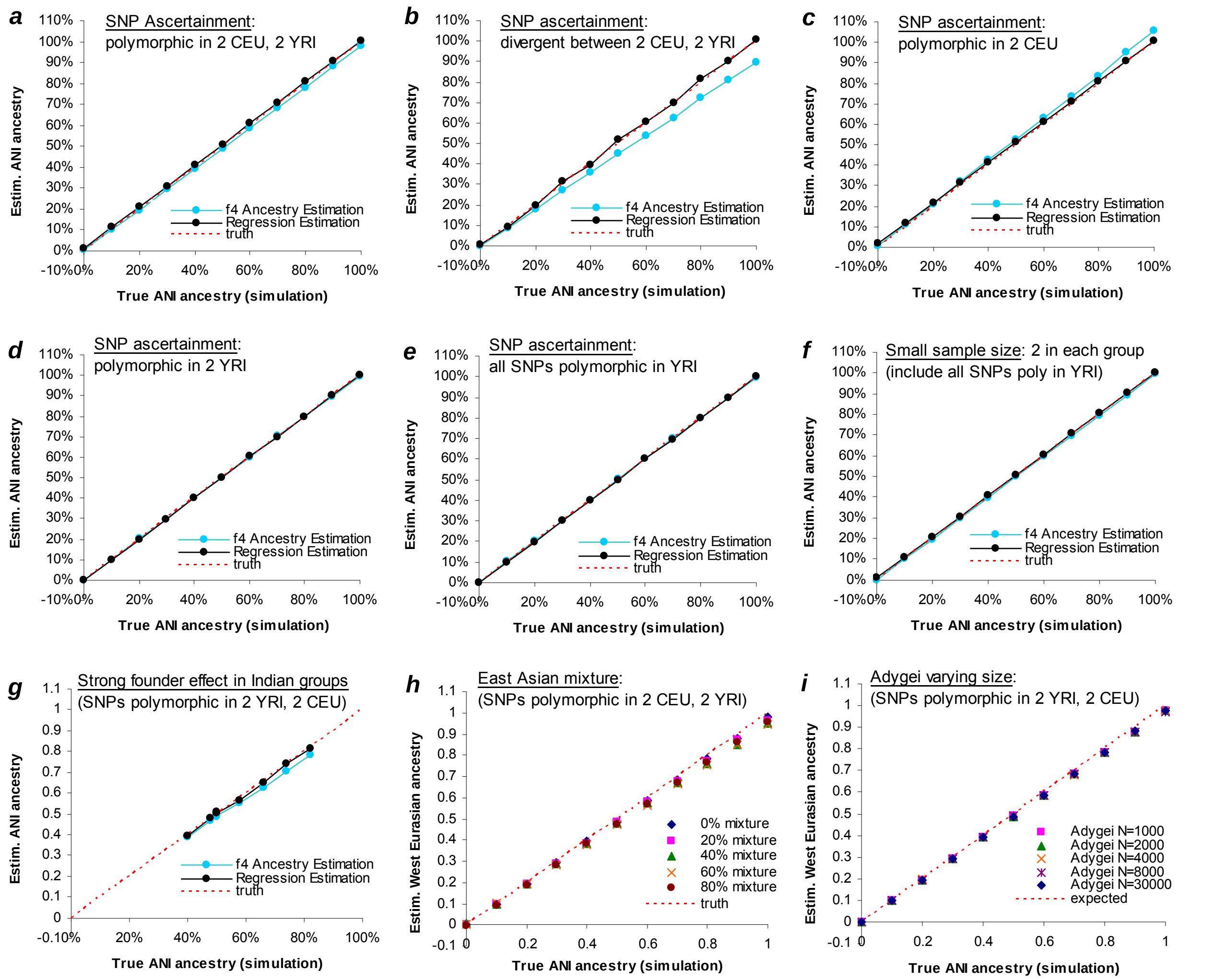
f_3(Adygei; Outgroup, India) is expected to be proportional to z + m_kW: the overlap of Adygei→Outgroup (green) and India→Onge (blue and red) drift.

f_3(Adygei; CEU, Onge) is similarly expected to be proportional to z + W. We fit z using three different populations as an outgroup, allowing us to solve for m_k.

Note S5 Figure 6: *f_4 and f_3 Ancestry Estimation*. (a) *f_4 Ancestry Estimation* is based on calculating a ratio of two f_4 statistics, which are expected to have the value of m_k. (b) *f_3 Ancestry Estimation* is based on a similar strategy with two f_3 statistics, which have expected values $z+m_kW$ and z+W. Here, z is the genetic drift specific to the Adygei since this group split from ANI, which is small and can be estimated from the data (Note S4). After obtaining z and measuring the other quantities, we can estimate the ANI mixture proportion m_k in each group.

Coalescent computer simulations establish the robustness of f_4 Ancestry Estimation

We explored the properties of *f_4 Ancestry Estimation* for the same simulation scenarios as we used to evaluate *Regression Ancestry Estimation* as discussed in detail above. Both methods are robust to a range of deviations from our assumptions, as shown in Note S5 Figure 7.



Note S5 Figure 7: Robustness of *f_4 Ancestry Estimation* as assessed by coalescent computer simulations for the same set of scenarios as in Note S5 Figure 5. (a-g) We compared the results for *Regression Ancestry Estimation* using Adygei as the ANI-related population, to the results of *f_4 Ancestry Estimation,* showing that both give robust results. The only exception is the extreme scenario where we only analyze SNPs discovered as differences between 2 CEU and 2 YRI chromosomes, in which case *f_4 Ancestry Estimation* produces slightly biased estimates. (h-i) We also show that f_4 Ancestry Estimation is robust for scenarios of East Asian mixture and varying population sizes in the Adygei outgroup.

f_3 Ancestry Estimation

The *f_3 Ancestry Estimation* procedure is described in the Methods and Note S4, and is summarized visually in Note S5 Figure 6b to highlight the parallels to *f_4 Ancestry Estimation*.

The ancestry estimates that we quote in the main text come from *f_3 Ancestry Estimation* in preference to the other two methods because the standard errors are lower. The smaller standard errors are likely due to a computational improvement, in which we did not require the use of a single outgroup as we did for *f_4 Ancestry Estimation.* Instead, our implementation allows us to use all three possible outgroups from Note S4 Figure 2 (Papuan, YRI and Dai), and find the best joint fit. We found that standard errors averaged 1.2% in practice (Table S5).

Consistency of different ancestry estimation methods

The autosomal estimates of ancestry obtained by *Regression Ancestry Estimation* are very similar to those from f_4 *Ancestry Estimation* and f_3 *Ancestry Estimation* (Table S5). This provides confidence in all three approaches. For example, the ANI estimates for Mala are 41%, 38% and 39% respectively, and for Kashmiri Pandit they are 72%, 69% and 71%.

Comparison of ancestry estimates on the autosomes and the X chromosome

An interesting question is whether more ANI ancestry has been inherited on the male lineage in India than on the female lineage, as might be expected from previous analysis of Y chromosome and mtDNA. We tested for this in our genome-wide SNP data by comparing estimates of ANI ancestry on the autosomes and the X chromosome. This analysis took advantage of the fact that the X chromosome is inherited in two thirds of instances through the female lineage compared to the autosomes which are inherited in only one half of instances through the female lineage.

We found that for *Regression Ancestry Estimation*, standard errors were so large (because of the limited data) that they were unusable. We chose f_3 *Ancestry Estimation* in preference to f_4 *Ancestry Estimation* because we had computationally implemented a method to integrate information from all three outgroups (Papuans, Dai and YRI), which reduced errors (Table S5).

Using f_3 *Ancestry Estimation*, we tested if the mean ANI was significantly different on the autosomes and X chromosome. These two values are plotted against each other in Figure S7c. The mean estimated ANI ancestry is lower by 0.074, but our X chromosome estimates are so noisy that this corresponds to a Z-score of only -1.2. More data are necessary to detect evidence of gender bias by a comparison of autosome and X chromosome ANI estimates in India.

References

[1] Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann HE, Rüther A, Schreiber S, Becker C, Nürnberg P, Nelson MR, Krawczak M, Kayser M (2008) Correlation between genetic and geographic structure in Europe. Curr Biol. 18, 1241-1248.

[2] Karve I (1968) Hindu Society—An interpretation. S. R. Deshmukh.

[3] Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet. 2, e190.

[4] Thangaraj K, Singh L, Reddy AG, Rao VR, Sehgal SC, Underhill PA, Pierson M, Frame IG, Hagelberg E (2003) Genetic affinities of the Andaman Islanders, a vanishing human population. Curr Biol. 13, 86-93

[5] K Thangaraj et al. (2005) Reconstructing the origin of Andaman Islanders. Science 308, 996.

[6] Barik SS, Sahani R, Prasad BV, Endicott P, Metspalu M, Sarkar BN, Bhattacharya S, Annapoorna PC, Sreenath J, Sun D, Sanchez JJ, Ho SY, Chandrasekar A, Rao VR (2008) Detailed mtDNA genotypes permit a reassessment of the settlement and population structure of the Andaman Islands. Am J Phys Anthropol. 136, 19-27.

[7] Bamshad M, Kivisild T, Scott Watkins W, Dixon ME, Ricker CE, Rao BB, Mastan Naidu J, Ravi Prasad BV, Govinda Reddy P, Rasanayagam A, Papiha SS, Villems R, Redd AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB (2001) Genetic evidence on the origins of Indian caste populations. Genome Res 11, 994–1004.

[8] Cordaux R, Saha N, Bentely GR, Aunger R, Sirajuddin SM, Stoneking M (2003) Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. Eur J Hum Genet 11:253–264.

[9] Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, Jin L, Su B, Pitchappan R, Shanmugalakshmi S, Balakrishnan K, Read M, Pearson NM, Zerjal T, Webster MT, Zholoshvili I, Jamarjashvili E, Gambarov S, Nikbin B, Dostiev A, Aknazarov O, Zalloua P, Tsoy I, Kitaev M, Mirrakhimov M, Chariev A, Bodmer WF (2001) The Eurasian heartland: a continental perspective on Y-chromosome diversity. Proc Natl Acad Sci USA 98, 10244-10249.

[10] Thanseem I, Thangaraj K, Chaubey G, Singh VK, Bhaskar LV, Reddy BM, Reddy AG, Singh L (2006) Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. BMC Genet. 7, 42.

[11] Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 40, 646-649.

[12] Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100-1104.

[13] Cavalli-Sforza LL, Edwards AW (1967) Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet. 19, 233-257.

[14] Becquet C, Patterson N, Stone AC, Przeworski M, Reich D (2007) Genetic structure of chimpanzee populations. PLoS Genet. 3, e66.

[15] Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet. 39, 1251-1255.

[16] Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identify and paternity. Genetica 96, 3-12.

[17] Busing FMTA, Meijer E, and van der Leeden R (1999) Delete-m jackknife for unequal m. Statistics and Computing, 9, 3-8.

[18] Künsch HR (1989) The jackknife and the bootstrap for general stationary observations. Ann. Statist., 17, 1217-1241.

[19] International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature. 449, 851-861.

[20] Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Research 15, 1496-1502.

[21] Bergsten J (2005) A review of long-branch attraction. Cladistics 21, 163-193.

[22] Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) G

[23] Hudson RR (1990) Oxford Surveys in Evolutionary Biology (eds Futuyma, D. J. & Antonovics, J.) 1-44 (Oxford Univ. Press, Oxford).

[24] Dronamraju KR (1964) Mating systems of the Andhra Pradesh people. Cold Spring Harb Symp Quant Biol. 29, 81-84.

[25] Pemberton TJ, Jakobsson M, Conrad DF, Coop G, Wall JD, Pritchard JK, Patel PI, Rosenberg NA (2008) Using population mixtures to optimize the utility of genomic databases: linkage disequilibrium and association study design in India. Ann Hum Genet. 72, 535-546.

[26] Rosenberg NA, Mahajan S, Gonzalez-Quevedo C, Blum MG, Nino-Rosales L, Ninis V, Das P, Hegde M, Molinari L, Zapata G, Weber JL, Belmont JW, Patel PI (2006) Low levels of genetic divergence across geographically and linguistically diverse populations from India. PLoS Genet. 2, e215.

[27] Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96, 3-12.

[28] Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly M, Reich D (2004) Methods for high-density admixture mapping of disease genes. Am J Hum Genet. 74, 1001-1013.

[29] Price AL, Patterson N, Hancks DC, Myers S, Reich D, Cheung VG, Spielman RS (2008) Effects of cis and trans genetic ancestry on gene expression in African Americans. PLoS Genet. 4, e1000294.

[30] Hudson RR (1990) Gene genealogies and the coalescent process. Oxf Surv Evol Biol 7, 1–44

[31] Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA 102, 15942-15947.