The Equidistance Index of Population Structure

Yaron Granot*[1], Saharon Rosset[2], and Karl Skorecki[1]

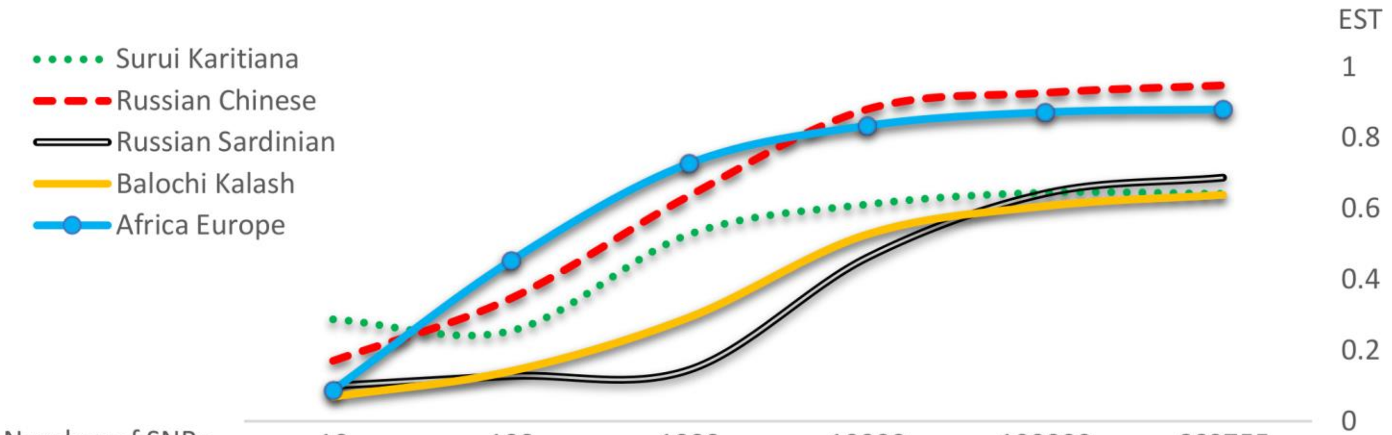
1 Rappaport Faculty of Medicine and Research Institute, Technion–Israel Institute of Technology, and Rambam Medical Center, Haifa, Israel, 2 School of Mathematical Sciences Tel Aviv University, Tel Aviv, Israel

* E-mail: yarongranot@hotmail.com

Supplemental Data

Table S1. E_{ST}min-max, F_{ST} and E_{BT} in 60 HGDP population pairs ranked by E_{BT}

Population Pair	E_{ST}min	E_{ST}mean	E_{ST}median	E_{ST}max	F_{ST}	E_{BT}
Surui vs. San	0.881	0.905	0.917	0.949	0.300	0.989
Surui Family vs. San	0.937	0.958	0.969	0.984	0.377	0.986
San vs. Han North	0.938	0.956	0.972	0.981	0.194	0.986
Japanese vs. Yoruba	0.969	0.974	0.978	0.984	0.143	0.980
Japanese vs. Mandenka	0.893	0.946	0.961	0.977	0.142	0.977
Han South vs. Yoruba	0.969	0.972	0.976	0.985	0.143	0.976
Italian vs. Yoruba	0.963	0.971	0.975	0.984	0.115	0.976
Russian vs. Yoruba	0.958	0.966	0.970	0.979	0.117	0.971
Surui vs. Tuscan	0.851	0.876	0.884	0.932	0.182	0.971
Yoruba vs. San	0.831	0.883	0.930	0.947	0.079	0.962
Surui vs. Mongola	0.765	0.803	0.815	0.893	0.147	0.954
Surui vs. Russian	0.775	0.810	0.821	0.892	0.170	0.953
French vs. Japanese	0.935	0.942	0.950	0.965	0.085	0.951
Han North vs. Bantu Kenya	0.766	0.874	0.946	0.978	0.130	0.941
Russian vs. Han South	0.935	0.940	0.948	0.965	0.075	0.937
Surui vs. Yakut	0.673	0.711	0.737	0.852	0.150	0.922
East Asia vs. Africa	0.678	0.802	0.870	0.939	0.131	0.921
Europe vs. Amazon	0.606	0.679	0.703	0.746	0.141	0.914
Papuan vs. Melanesian	0.367	0.602	0.731	0.950	0.071	0.914
Mbuti vs. Biaka	0.465	0.699	0.822	0.878	0.043	0.911
Surui vs. Karitiana	0.336	0.521	0.577	0.803	0.132	0.898
Europe vs. Africa	0.611	0.792	0.873	0.941	0.112	0.895
Surui vs. Pima	0.378	0.514	0.573	0.760	0.123	0.865
Karitiana vs. Pima	0.300	0.537	0.629	0.851	0.106	0.857
Kalash vs. Uygur	0.677	0.770	0.803	0.888	0.034	0.846
Europe vs. East Asia	0.755	0.821	0.861	0.904	0.071	0.833
Karitiana vs. Colombian	0.324	0.523	0.568	0.880	0.080	0.826
Surui vs. Colombian	0.396	0.498	0.512	0.789	0.095	0.825
Russian vs. Yakut	0.822	0.825	0.857	0.924	0.058	0.825
Surui-A vs. Surui-B	0.390	0.532	0.559	0.770	0.175	0.819
Papuan-A vs. Papuan-B	0.039	0.286	0.317	0.731	0.025	0.801
Burusho vs. Kalash	0.337	0.542	0.610	0.793	0.024	0.789
Lahu vs. Naxi	-0.747	0.009	0.648	0.777	0.020	0.768
Colombian vs. Pima	0.092	0.305	0.371	0.767	0.063	0.752
Russian vs. Uygur	0.793	0.825	0.847	0.886	0.021	0.735
Balochi vs. Kalash	0.380	0.549	0.639	0.768	0.023	0.713
Russian vs. Burusho	0.489	0.621	0.677	0.801	0.017	0.682
Uygur vs. Tuscan	0.809	0.851	0.877	0.919	0.027	0.681
Russian vs. Sardinian	0.559	0.648	0.688	0.776	0.015	0.677
Brahui vs. Kalash	0.377	0.478	0.577	0.751	0.025	0.663
Karitiana vs. Maya	0.334	0.532	0.626	0.866	0.070	0.654
Bantu S. Afr vs. San	0.751	0.820	0.873	0.930	0.057	0.648
Surui vs. Maya	0.364	0.459	0.520	0.739	0.086	0.647
Naxi vs. Yi	-1.324	-0.250	0.686	0.791	0.004	0.638
Hazara vs. Uygur	-0.505	0.175	0.602	0.745	0.002	0.586
Bantu S. Afr vs. Bantu Kenya	-0.881	-0.056	0.479	0.809	0.006	0.553
Mandenka vs. Yoruba	-0.701	0.175	0.421	0.681	0.007	0.549
Italian vs. Orcadian	-0.615	0.106	0.455	0.699	0.005	0.547
Cambodian vs. Naxi	-0.503	0.209	0.794	0.878	0.014	0.537
Pima vs. Maya	0.068	0.279	0.427	0.720	0.042	0.507
Druze vs. Bedouin	-0.531	-0.053	0.189	0.660	0.009	0.485
French vs. Sardinian	0.273	0.396	0.477	0.644	0.007	0.465
Palestinian vs. Bedouin	-0.627	-0.187	0.023	0.627	0.006	0.420
Russian vs. Adygei	0.348	0.421	0.504	0.663	0.009	0.415
Druze vs. Palestinian	-0.566	-0.075	0.102	0.565	0.007	0.375
Colombian vs. Maya	-0.283	-0.071	0.021	0.675	0.027	0.229
Japanese vs. Han South	0.190	0.238	0.365	0.597	0.006	0.155
French vs. Russian	0.092	0.201	0.319	0.524	0.004	0.152
Cambodian vs. Mongola	0.465	0.581	0.634	0.791	0.013	0.130
Italian vs. Tuscan	-0.078	0.135	0.245	0.540	0.001	0.101



Number of SNPs	10	100	1000	10000	100000	660755
Surui Karitiana	0.287402	0.255082	0.527655	0.612226	0.645473	0.642203
Russian Chinese	0.169798	0.346408	0.63601	0.880498	0.926977	0.948114
Russian Sardinian	0.099688	0.127379	0.144685	0.461724	0.64308	0.687639
Balochi Kalash	0.068614	0.141008	0.292738	0.527545	0.607674	0.639284
Africa Europe	0.086934	0.452973	0.727686	0.83431	0.872332	0.880404
Mean	0.1424872	0.26457	0.4657548	0.6632606	0.7391072	0.7595288
Mean as % of Max	19%	35%	61%	87%	97%	100%

Figure S1. E_{ST}median as a function of SNP sample size. E_{ST} in various population pairs with increasing SNP sample size ranging from 10 to 660,755. As expected, E_{ST} initially rises rather steeply, however the results tend to plateau around 100,000 SNPs. This suggests that we are approaching the maximal resolving power, and adding more markers beyond this point is unlikely to significantly affect E_{ST} and cluster separation.

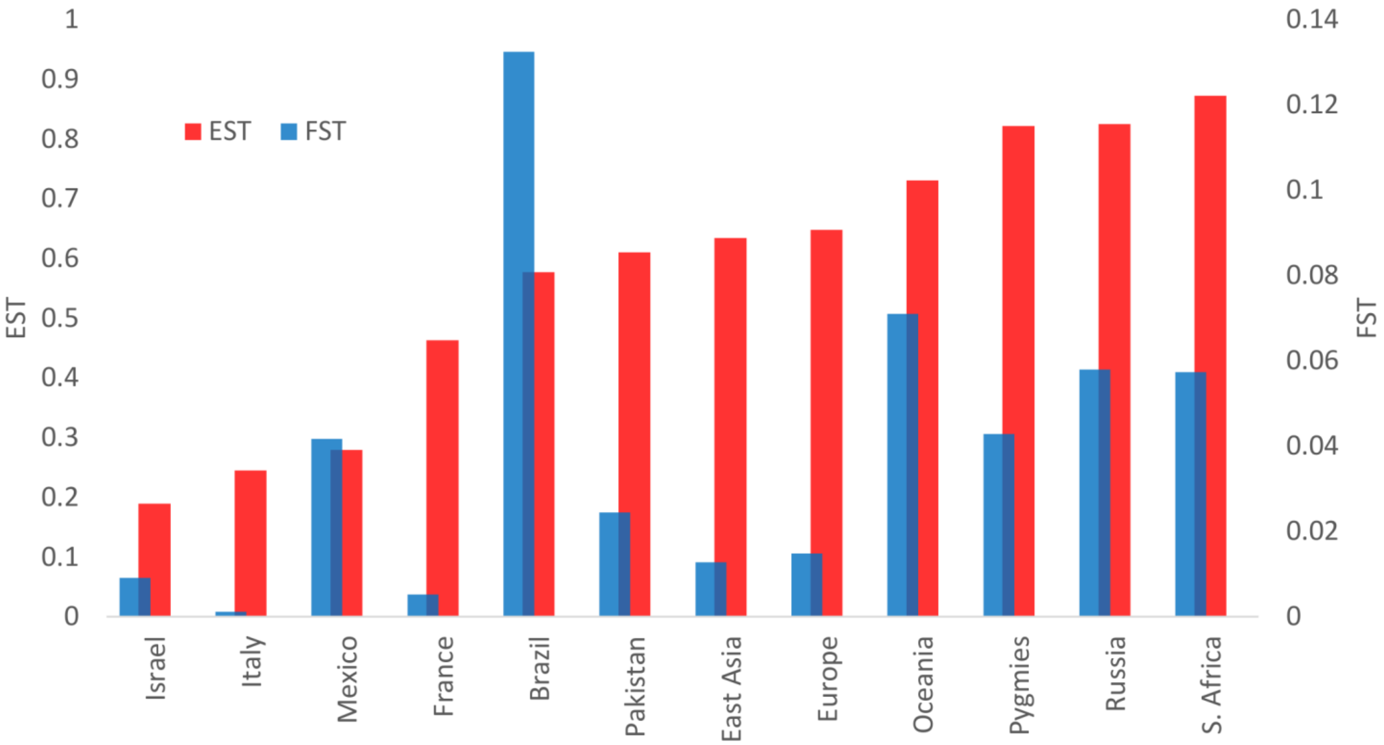


Figure S2. F_{ST} and E_{ST} within 12 local regions. Calculated from a single pair of HGDP populations per region: Israel (Bedouins vs. Druze), Italy (North Italians vs. Tuscans), Mexico (Maya vs. Pima), France (Basque vs. French), Brazil (Karitiana vs. Surui), Pakistan (Burusho vs. Kalash) East Asia (Cambodian vs. Mongola), Europe (Russians vs. Sardinians), Oceania (Melanesians vs. Papuans), Pygmies (Biaka vs. Mbuti), Russia (Russians vs. Yakut) and Southern Africans (South African Bantu vs. San). The most obvious discrepancy between F_{ST} and E_{ST} is in Brazil, with a high F_{ST} and moderate E_{ST}, and to a lesser extent between the Druze and Bedouins. Note that these Druze and Bedouins live within a few hundred kilometers of each other, speak the same language, and have the lowest E_{ST} among these 12 pairs, yet have a somewhat higher F_{ST} (several times higher than between the two Italian populations from Northern Italy and Tuscany and almost twice as high as between the French and Basques).

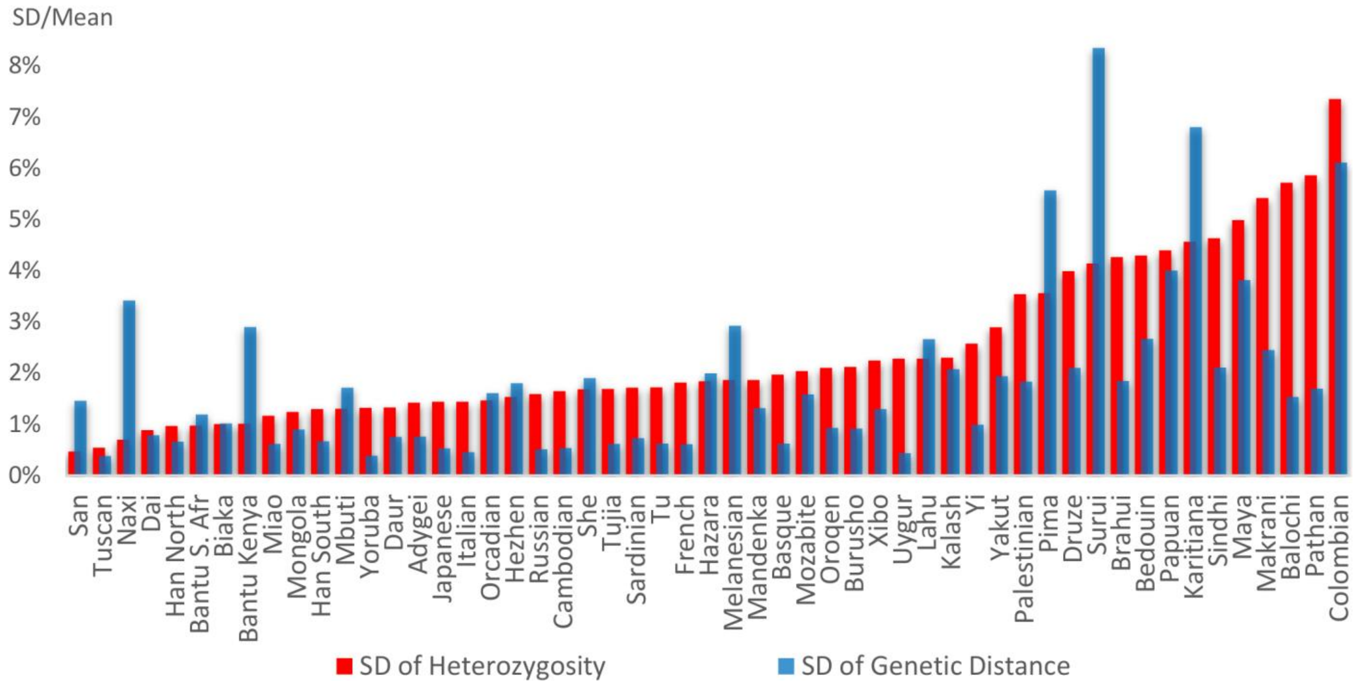


Figure S3. Standard deviations (SD) of heterozygosity and pairwise distances (from 660,755 SNPs in 53 populations). Excessive SD of genetic distance (blue) compared to SD of heterozygosity (red), as in the San and Naxi samples, implies the inclusion of relatives.

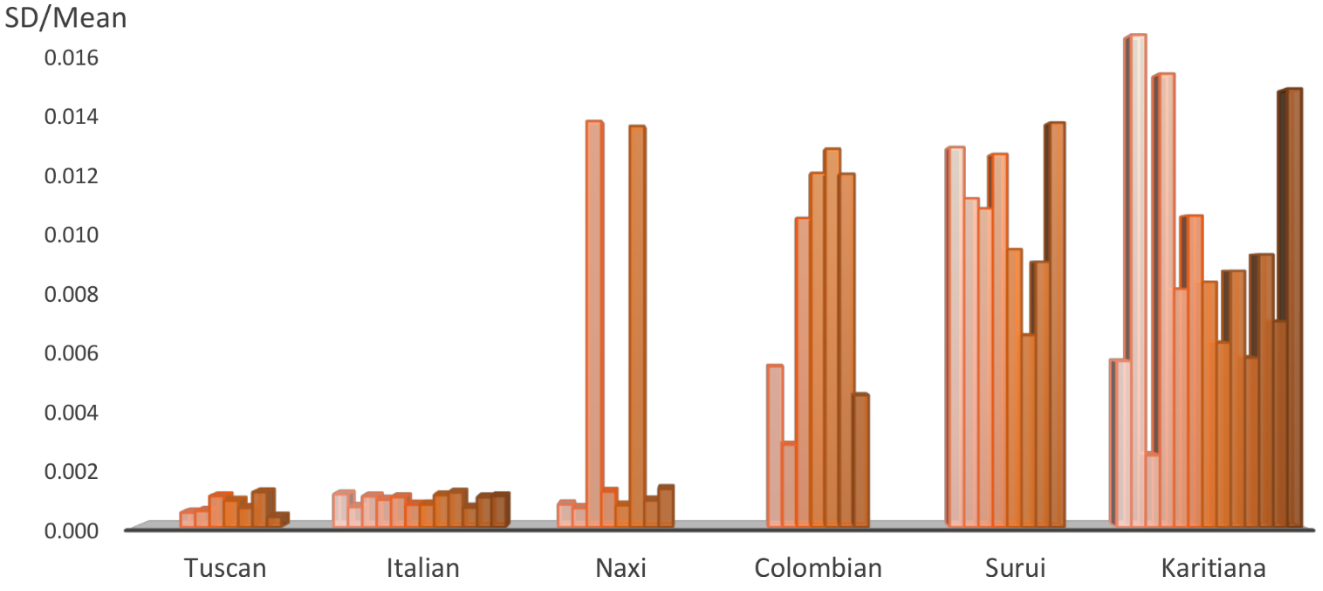


Figure S4. Individual standard deviations in six HGDP populations. Each column represents the standard deviation (SD) between a single individual and all other samples within the given population. Tuscans (n=7), Italians (n=12), Naxi (n=8), Colombian (n=7), Surui (n=8), and Karitiana (n=13). The "twin towers" in the Naxi batch are inferred to be a pair of close relatives in an otherwise panmictic population sample. These two individuals stick out like a sore thumb, while similarly related individuals are not nearly as obvious among the Native American samples due to a higher base-level of structure in these population samples. This can be used as a quick way of visually identifying ascertainment bias.

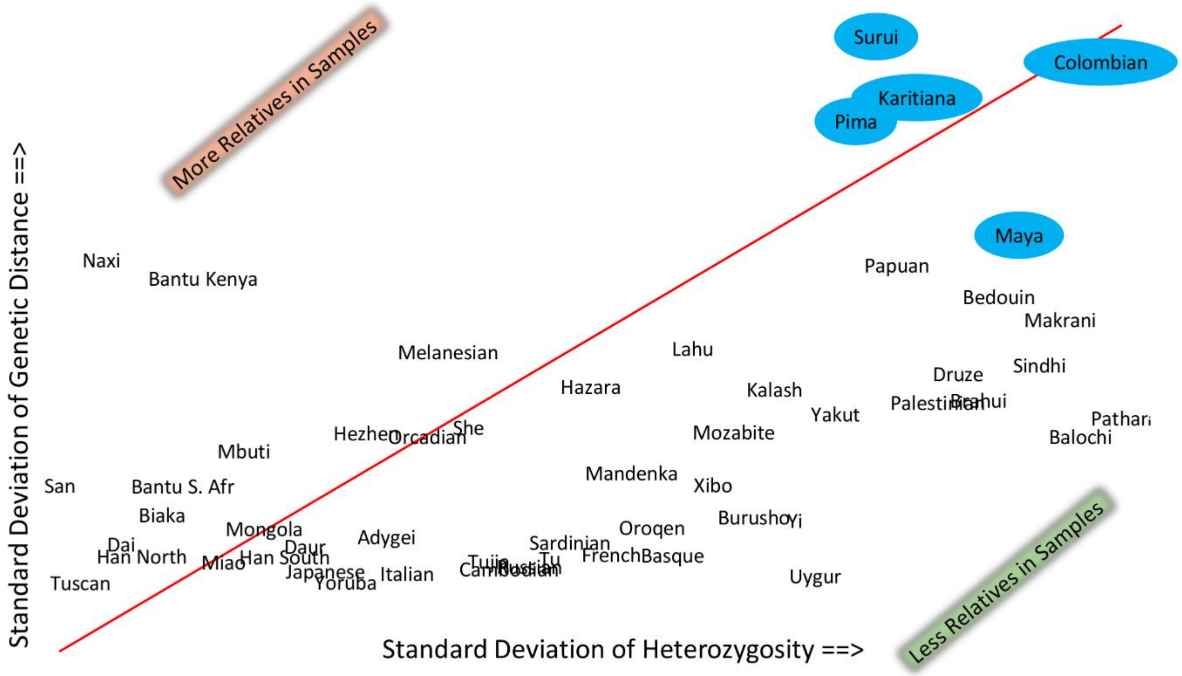


Figure S5. The SD of pairwise distance plotted against the SD of heterozygosity (938 individuals from 53 populations). The red diagonal line represents the linear trend line of the standard deviation of heterozygosity. Populations above this line are inferred to have more genetic structure than expected from heterozygosity, implying that relatives may have been included in the samples. Native American populations, highlighted in light blue, appear to have moderate or moderately high levels of relatives included among their samples.

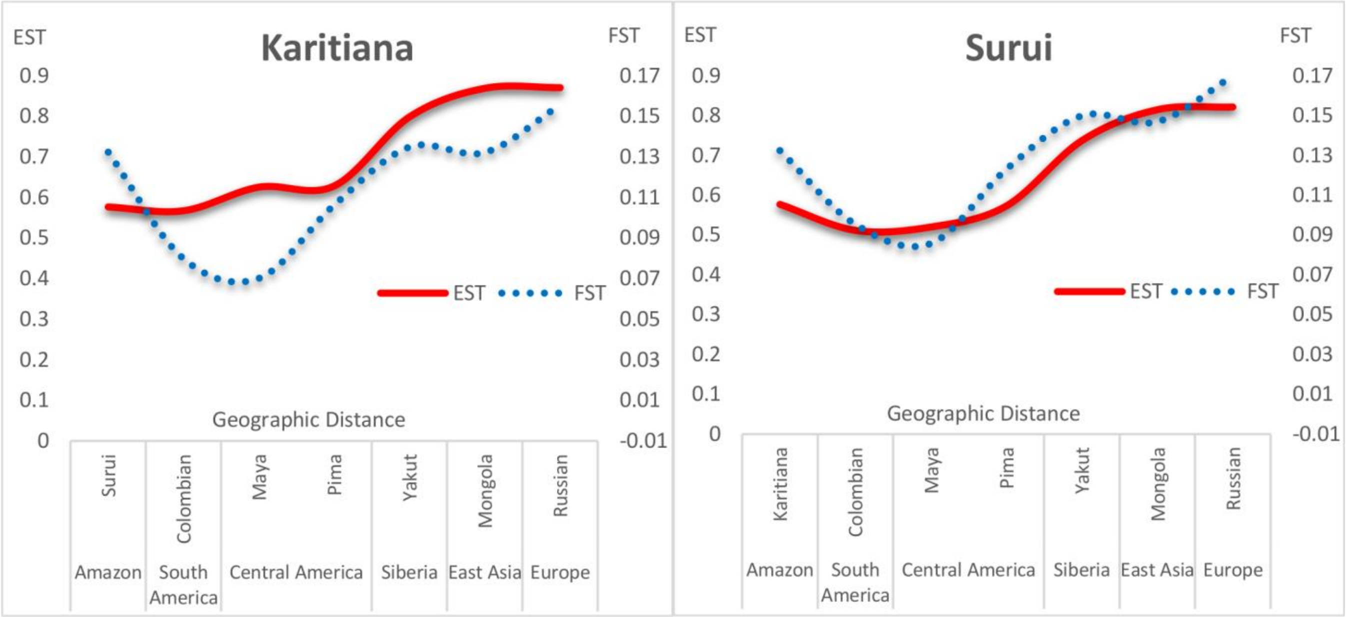


Figure S6. Pairwise F_{ST} and E_{ST} vs. geographic distance from the two Amazonian tribes to various global populations with increasing distance from the Amazon. Note the initial steep drop in F_{ST} with increased distance from the Amazon.

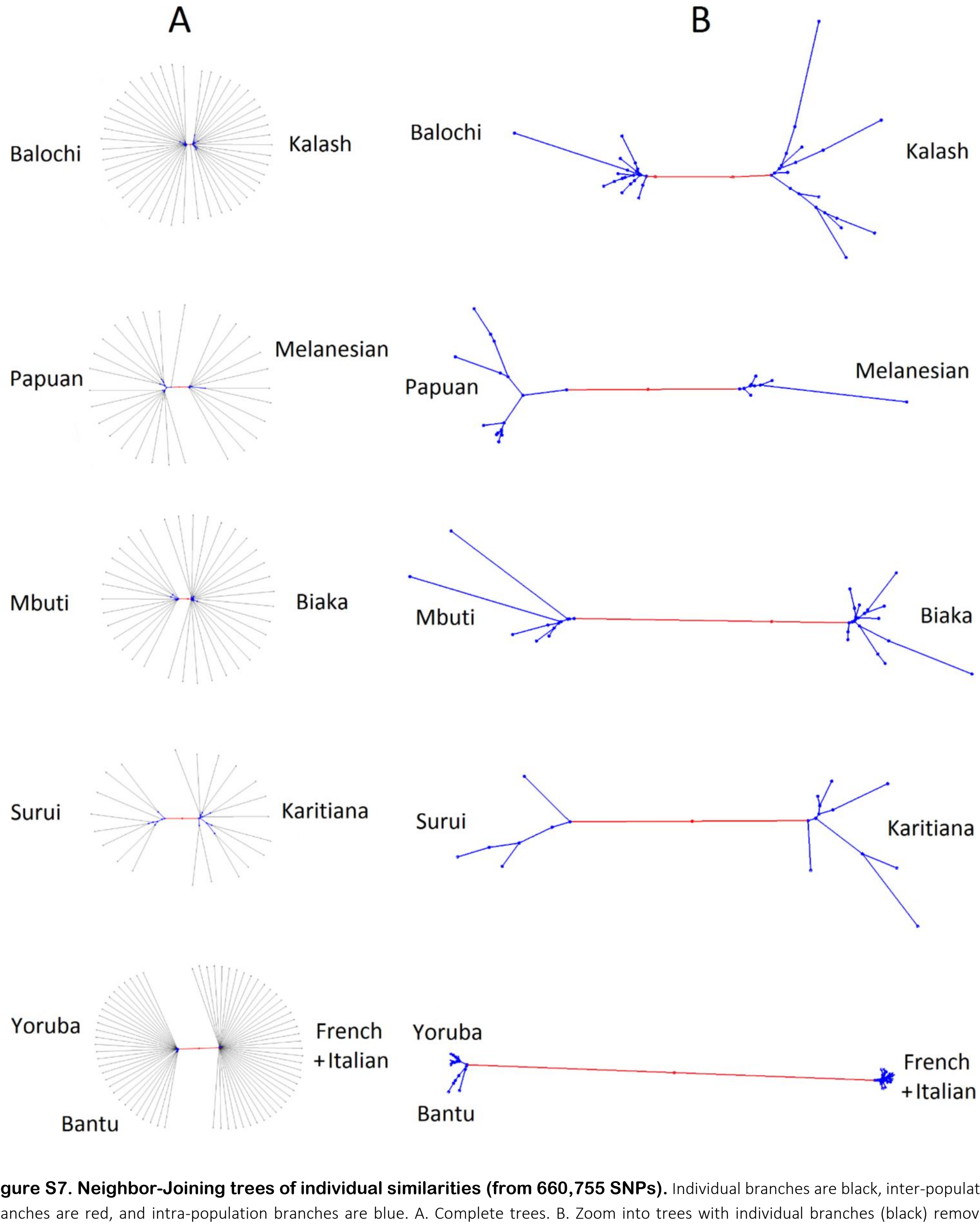


Figure S7. Neighbor-Joining trees of individual similarities (from 660,755 SNPs). Individual branches are black, inter-population branches are red, and intra-population branches are blue. A. Complete trees. B. Zoom into trees with individual branches (black) removed.

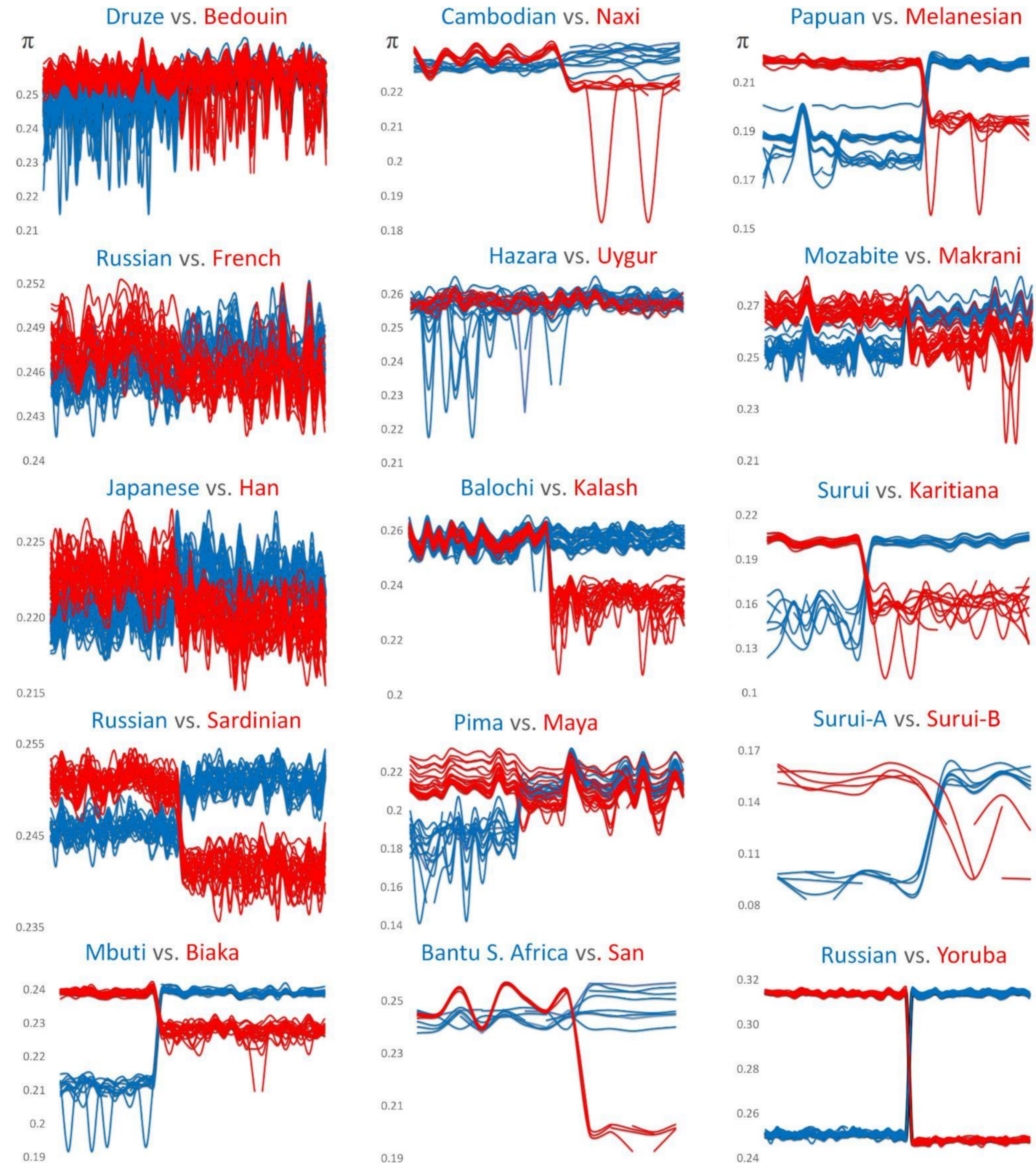


Figure S8. Pairwise population distance charts. Each sample is represented by a red or blue string and each point on each string reflects distance between a pair of samples. Points that fall far below the rest (as in the Naxi and Melanesians) are inferred to reflect close relatives.

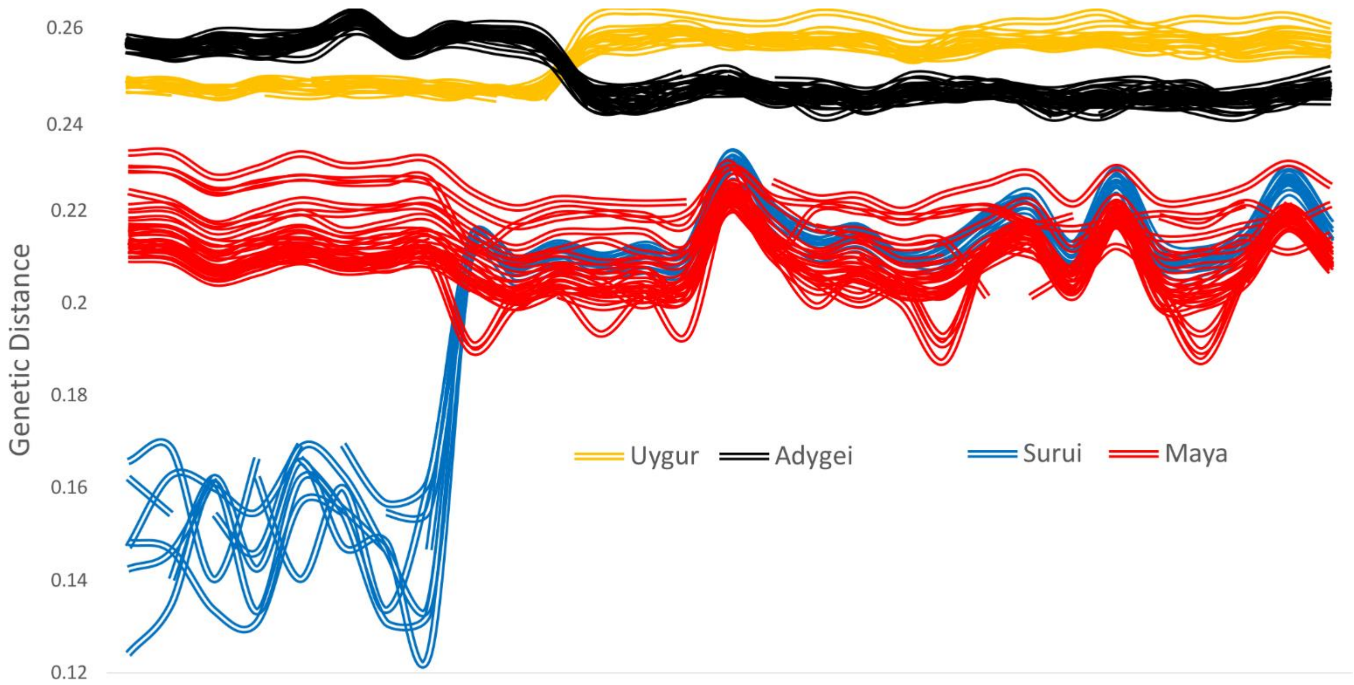


Figure S9. Superimposed distance plots of Uygur and Adygei (top) and Surui and Maya (bottom). This is the same kind of plot as in Figure S8, with each string representing a single individual and each population represented by a different color. Despite a high F_{ST} of 0.09 (E_{ST} = 0.52), some Mayan individuals (red) are genetically closer to some Surui individuals (blue) than to some fellow Mayan individuals ($\omega > 0$), presumably due to outbreeding (some Mayan individuals have significant European admixture, which increases distances among Mayans). There is no such overlap between Uygur (yellow) and Adygei (black) samples ($\omega = 0$) despite a much lower pairwise F_{ST} of 0.02 (E_{ST} = 0.79).

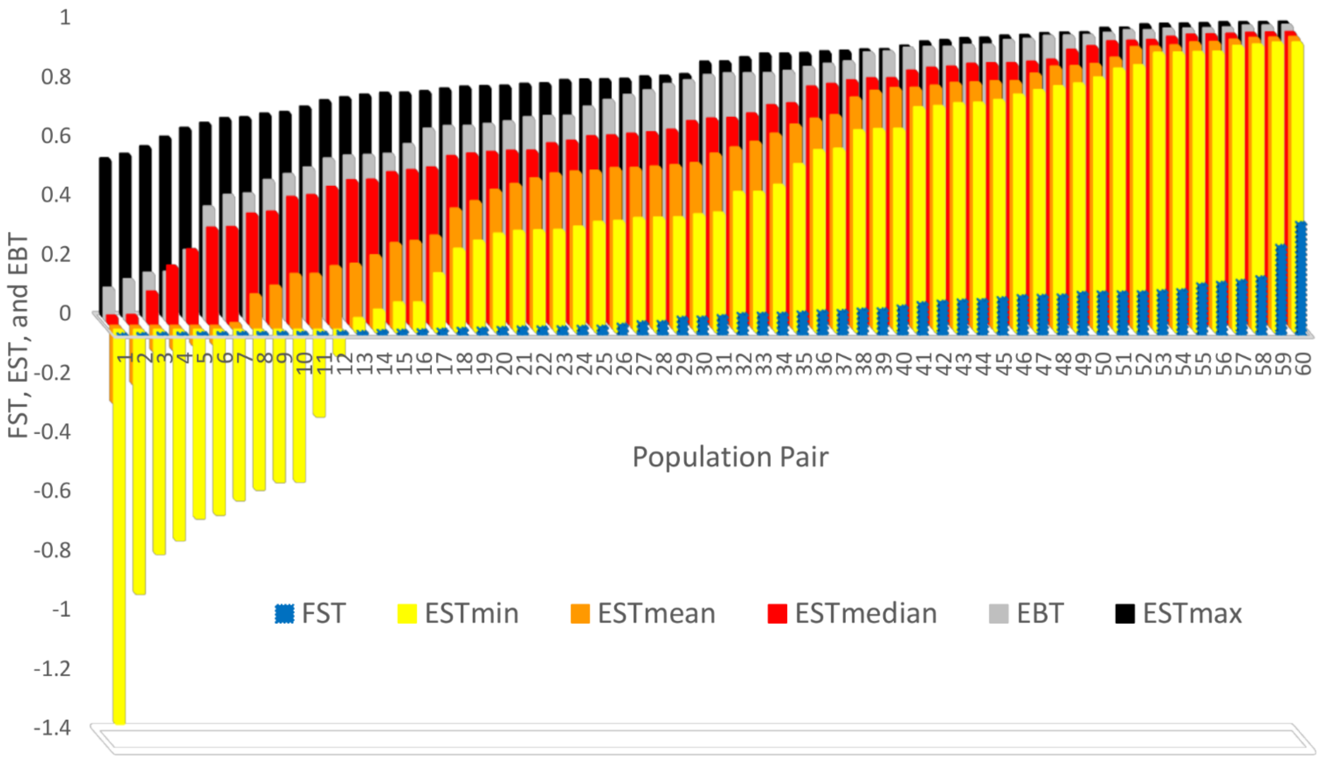


Figure S10. Overview of F_{ST}, E_{ST} and E_{BT} among 60 HGDP population pairs (660,755 SNPs). Negative E_{ST}min (yellow) and E_{ST}mean (orange) would imply that close relatives were included among these samples. Of the 60 population pairs, 12 (20%) have negative E_{ST}min and 6 have negative E_{ST}mean values. E_{ST}median, E_{ST}max, and E_{BT} cover virtually the entire 0-1 range with no negative values in these samples. The general trend is F_{ST} < E_{ST}min < E_{ST}mean < E_{ST}median < E_{ST}max. E_{BT} (gray) is usually somewhere between E_{ST}median (red) and E_{ST}max (black).

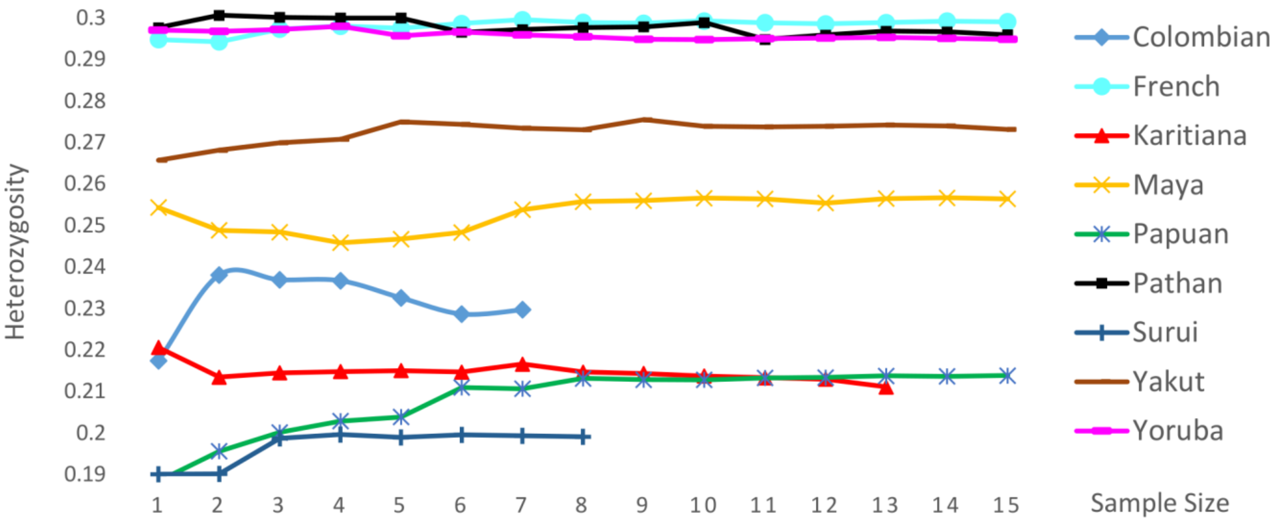


Figure S11. Mean heterozygosity as a function of sample size. Heterozygosity in various HGDP populations with sample size increasing from 1 to 15. All samples were included in populations with less than 15 samples (7 in Colombians, 8 in Surui, and 13 in Karitiana).

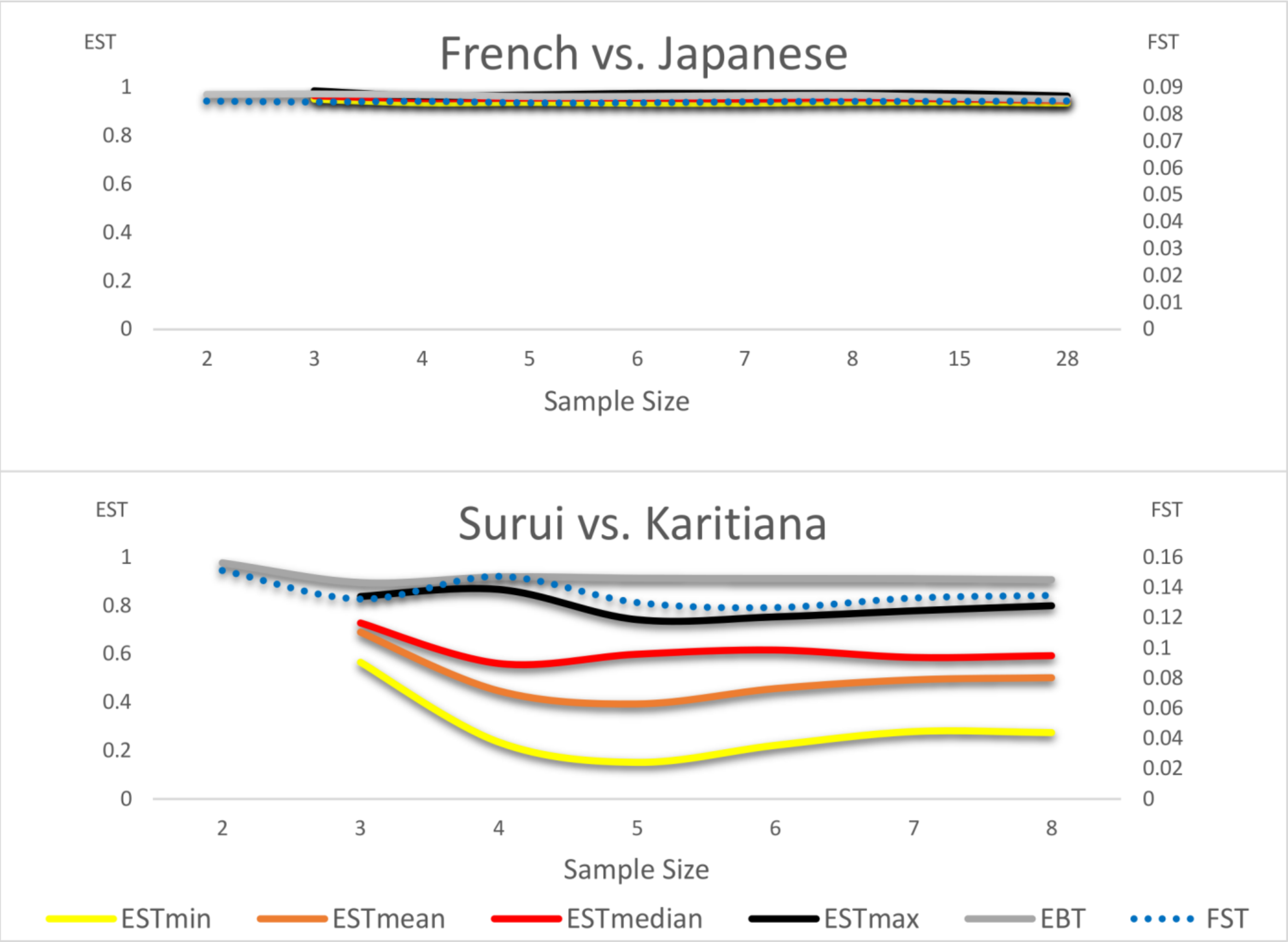


Figure S12. Pairwise F_{ST}, E_{ST} and E_{BT} as a function of sample size. Differentiation was estimated in two population pairs: French-Japanese and Surui-Karitiana, with population sample sizes ranging from n=2 to n=8. French-Japanese estimates were also taken at n=15 and n=28 due to their larger samples. F_{ST} and E_{BT} start at n=2; E_{ST} starts at n=3, the minimal sample size for estimating the SD of pairwise distances.