Supplemental Data

Genotype-Imputation Accuracy across Worldwide Human Populations

Lucy Huang, Yun Li, Andrew B. Singleton, John A. Hardy, Gonçalo Abecasis, Noah A. Rosenberg, and Paul Scheet

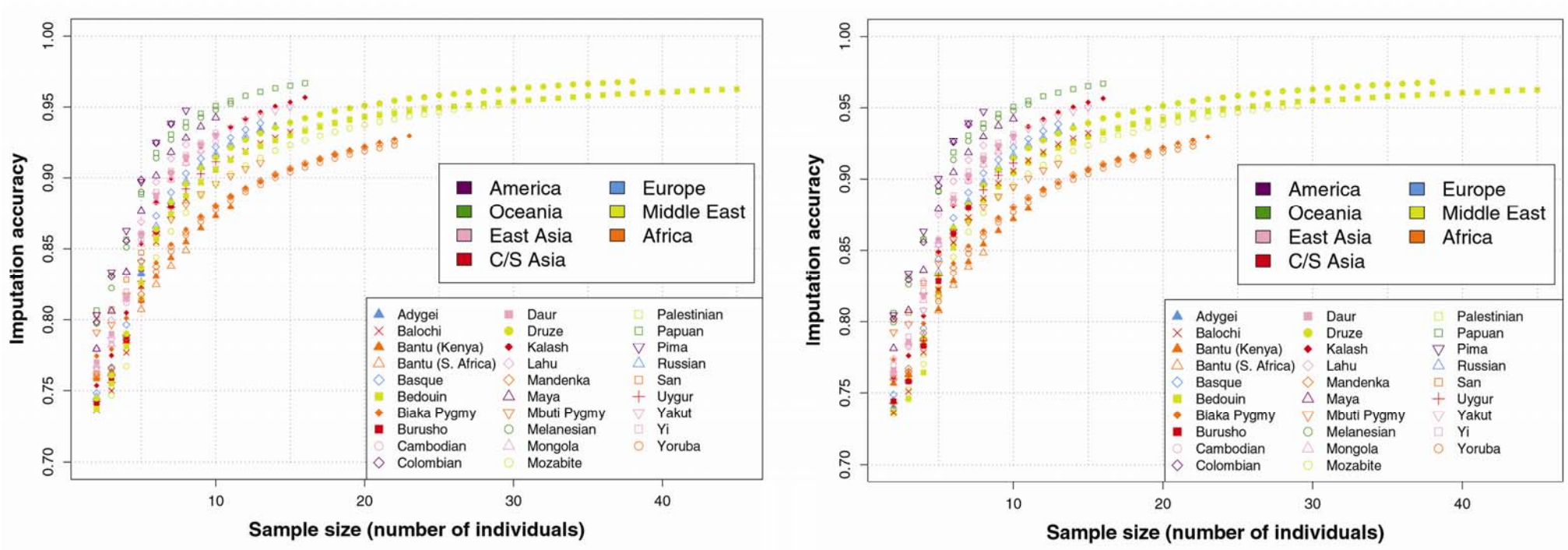

Figure S1. Imputation Accuracy versus Sample Size, in Each of 29 Populations

The two plots are based on two different subsets of individuals of the sample. The plot on the left is identical to Figure 3.

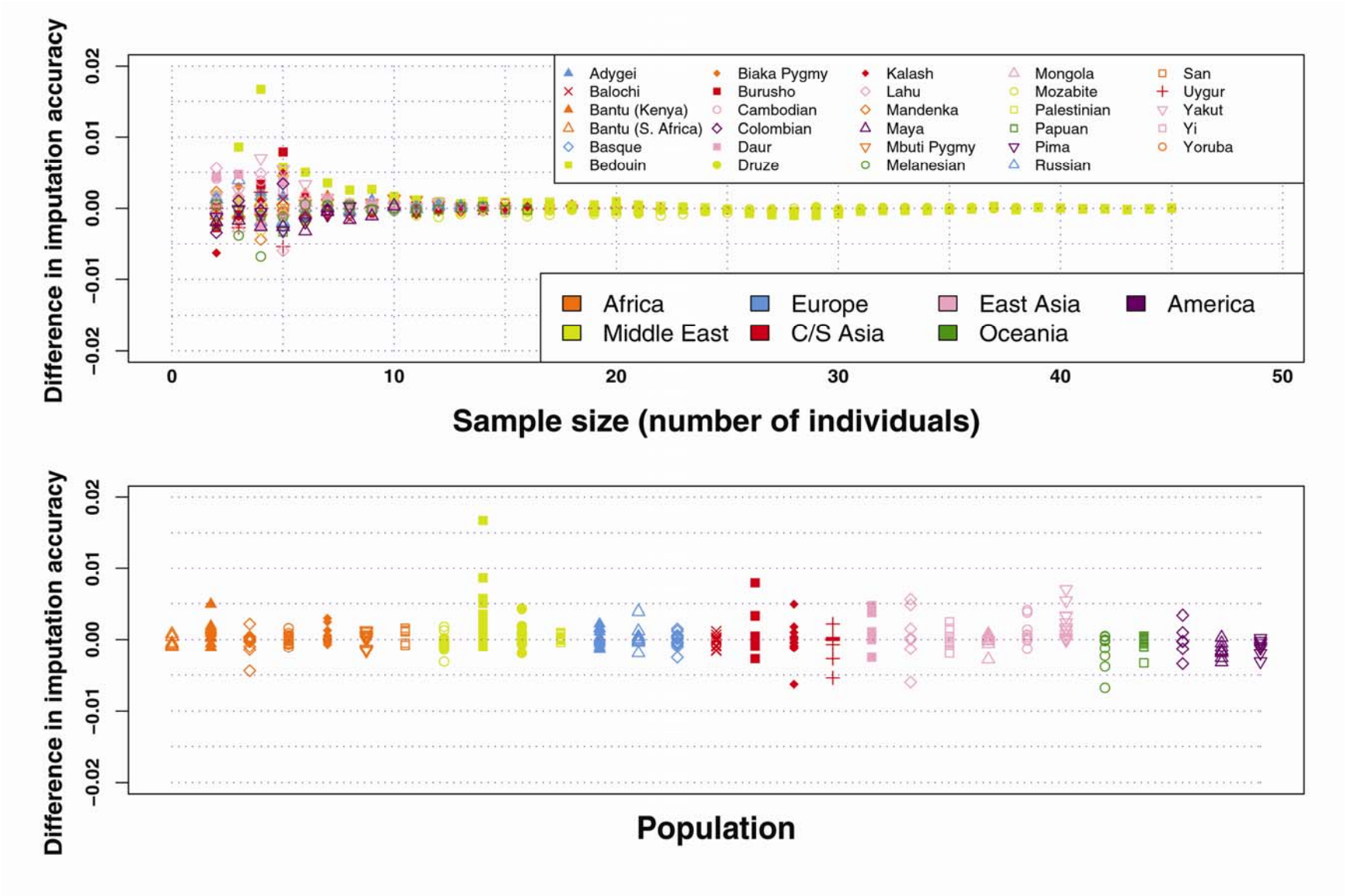

Figure S2. Difference in Imputation Accuracy Assessed with One Subset of Individuals Compared to a Second Subset Based on Another Permutation of the Individuals, in Each of 29 Populations

The points correspond to point-wise differences between the values in the two plots in Figure 7 (i.e., subtracting values in the right plot from corresponding values in the left plot).

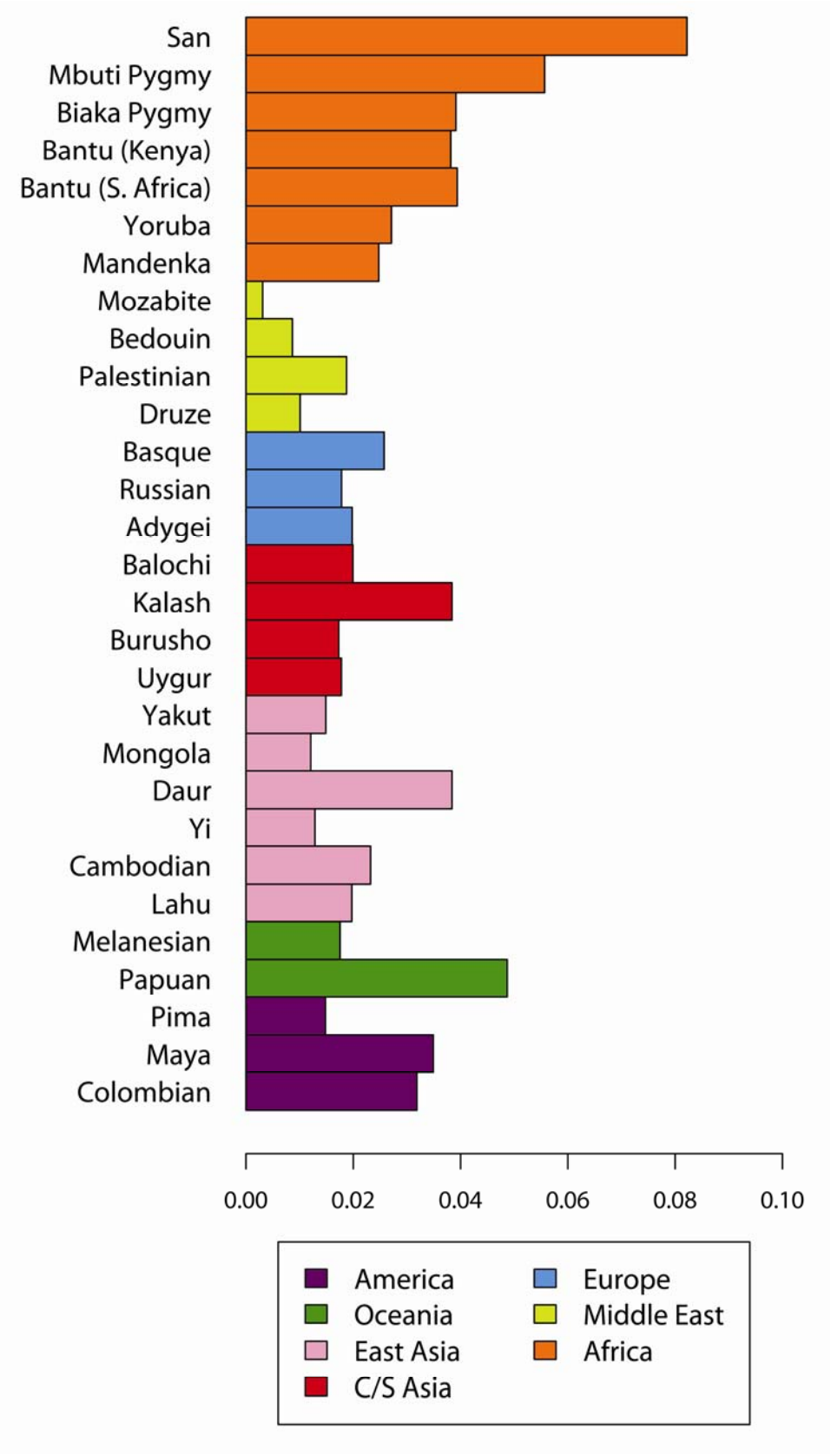

Figure S3. Difference in Maximal Imputation Accuracy for Two Sets of SNPs (MAF > 0.2 vs MAF ≤ 0.2), Based on Data in Figure 9

Bars are colored by geographic locations of the populations.

Table S1. Imputation Accuracy for Inference of Genotypes of Untyped Markers in the Data of Jakobsson et al.,[1] Based on Any One or Two or All Three HapMap Reference Panels (with Their Original Size)

	YRI	CEU	CHB+JPT	YRI/CEU	YRI/ CHB+JPT	CEU/ CHB+JPT	ALL
San	**0.8912**	0.8226	0.8205	0.8873	0.8873	0.8379	0.8879
Mbuti Pygmy	**0.9018**	0.8342	0.8224	0.8999	0.8994	0.8411	0.8988
Biaka Pygmy	0.9145	0.8559	0.8459	**0.9176**	0.9126	0.8629	0.9150
Bantu (Kenya)	0.9360	0.8752	0.8705	**0.9396**	0.9362	0.8875	0.9382
Bantu (S. Africa)	0.9322	0.8559	0.8536	**0.9329**	0.9322	0.8647	0.9325
Yoruba	**0.9457**	0.8667	0.8600	0.9448	0.9437	0.8790	0.9448
Mandenka	**0.9419**	0.8692	0.8650	0.9414	0.9412	0.8854	0.9408
Mozabite	0.9301	0.9180	0.9032	**0.9458**	0.9415	0.9226	0.9455
Bedouin	0.9279	0.9407	0.9215	0.9486	0.9421	0.9423	**0.9486**
Palestinian	0.9342	0.9502	0.9296	**0.9550**	0.9476	0.9507	0.9543
Druze	0.9300	0.9552	0.9341	0.9562	0.9500	0.9562	**0.9569**
Basque	0.9337	0.9577	0.9361	0.9570	0.9503	0.9576	**0.9579**
Russian	0.9338	0.9597	0.9389	0.9590	0.9524	**0.9600**	0.9599
Adygei	0.9315	0.9593	0.9372	0.9593	0.9512	**0.9605**	0.9600
Balochi	0.9249	0.9511	0.9337	**0.9527**	0.9473	0.9516	0.9524
Kalash	0.9165	0.9469	0.9287	0.9488	0.9425	**0.9492**	0.9477
Burusho	0.9301	0.9487	0.9342	0.9487	0.9465	0.9506	**0.9509**
Uygur	0.9270	0.9471	0.9427	0.9503	0.9487	**0.9537**	0.9534
Yakut	0.9249	0.9454	0.9466	0.9468	0.9505	**0.9516**	0.9513
Mongola	0.9302	0.9444	0.9517	0.9509	0.9545	0.9549	**0.9553**
Daur	0.9278	0.9434	0.9543	0.9493	0.9552	0.9565	**0.9571**
Yi	0.9268	0.9431	0.9522	0.9510	0.9533	0.9543	**0.9545**
Cambodian	0.9301	0.9413	0.9487	0.9455	0.9509	**0.9518**	0.9511
Lahu	0.9345	0.9480	**0.9599**	0.9531	0.9598	0.9588	0.9597
Melanesian	0.9207	0.9332	0.9454	0.9471	0.9477	0.9475	**0.9497**
Papuan	0.9212	0.9268	0.9399	0.9407	0.9436	0.9419	**0.9444**
Pima	0.9321	0.9487	0.9481	0.9540	0.9552	0.9542	**0.9554**
Maya	0.9305	0.9539	0.9495	0.9543	0.9558	**0.9588**	0.9582
Colombian	0.9305	0.9493	0.9398	0.9499	0.9507	0.9517	**0.9539**

These values were used in the scatter plot of Figure 7. For each population, the highest imputation accuracy obtained among the seven possible reference panels is highlighted in bold.

Table S2. Squared Correlation Coefficient, r^2, between the Genotypes Imputed from the Data of Jakobsson et al.[1] and Those Directly Measured in the Data of Conrad et al.[2] and Pemberton et al.[3]

	YRI	CEU	CHB+JPT	YRI/CEU	YRI/ CHB+JPT	CEU/ CHB+JPT	ALL
San	**0.7633**	0.6116	0.6200	0.7443	0.7470	0.6416	0.7341
Mbuti Pygmy	0.7340	0.6299	0.6235	**0.7397**	0.7331	0.6570	0.7346
Biaka Pygmy	0.7804	0.6708	0.6397	**0.7882**	0.7716	0.6785	0.7795
Bantu (Kenya)	0.8611	0.7178	0.7212	**0.8726**	0.8553	0.7387	0.8672
Bantu (S. Africa)	0.8452	0.6825	0.6833	**0.8510**	0.8454	0.6842	0.8492
Yoruba	**0.8999**	0.7155	0.7049	0.8951	0.8924	0.7350	0.8957
Mandenka	**0.8744**	0.7087	0.6978	0.8670	0.8731	0.7327	0.8671
Mozabite	0.8541	0.8253	0.7833	0.8959	0.8911	0.8339	**0.8973**
Bedouin	0.8574	0.8830	0.8296	**0.9067**	0.8871	0.8843	0.9062
Palestinian	0.8678	0.9015	0.8526	**0.9157**	0.8984	0.9002	0.9095
Druze	0.8525	0.9107	0.8588	0.9123	0.8943	0.9156	**0.9161**
Basque	0.8686	0.9240	0.8869	0.9214	0.9023	0.9234	**0.9262**
Russian	0.8682	**0.9310**	0.8689	0.9241	0.9093	0.9292	0.9291
Adygei	0.8620	0.9277	0.8749	0.9296	0.9006	**0.9307**	0.9269
Balochi	0.8614	0.9099	0.8724	0.9175	0.8943	**0.9185**	0.9155
Kalash	0.8585	0.9058	0.8562	0.9116	0.8931	**0.9135**	0.9069
Burusho	0.8896	0.9082	0.8699	0.9067	0.9059	0.9127	**0.9161**
Uygur	0.8675	0.9114	0.8986	0.9175	0.9188	**0.9300**	0.9282
Yakut	0.8633	0.9036	0.9102	0.9078	0.9205	**0.9239**	0.9225
Mongola	0.8803	0.9066	0.9212	0.9180	**0.9265**	0.9236	0.9257
Daur	0.8612	0.8968	0.9289	0.9147	0.9286	0.9300	**0.9305**
Yi	0.8665	0.8947	0.9127	0.9069	0.9181	0.9199	**0.9212**
Cambodian	0.8752	0.8858	0.9102	0.8962	0.9156	**0.9165**	0.9114
Lahu	0.8832	0.8978	0.9332	0.9098	**0.9335**	0.9291	0.9323
Melanesian	0.8504	0.8597	0.8986	0.9002	0.9035	0.9042	**0.9057**
Papuan	0.8581	0.8638	0.8796	0.8762	0.8855	**0.8884**	0.8839
Pima	0.9271	0.9486	0.9423	0.9604	0.9509	0.9572	**0.9618**
Maya	0.8738	0.9210	0.9146	0.9051	0.9182	**0.9243**	0.9196
Colombian	0.8858	0.9309	0.9101	0.9209	0.9302	0.9296	**0.9331**

These values were used in the scatter plot of Figure 8. For each population, the highest r^2 value obtained among the seven possible reference panels is highlighted in bold.

Table S3. Summary Statistics for Minor Allele Frequencies of 513 SNP Loci in the Data of Conrad et al.[2] and Pemberton et al.[3]

	All		MAF < 0.2			MAF ≥ 0.2		
	Mean	Standard deviation	No. of SNPs	Mean	Standard deviation	No. of SNPs	Mean	Standard deviation
San	0.1861	0.1639	294	0.0620	0.0672	219	0.3528	0.0912
Mbuti Pygmy	0.1988	0.1521	271	0.0751	0.0616	242	0.3372	0.0921
Biaka Pygmy	0.2270	0.1570	252	0.0886	0.0631	261	0.3607	0.0905
Bantu (Kenya)	0.2474	0.1487	206	0.0925	0.0625	307	0.3513	0.0859
Bantu (S. Africa)	0.2270	0.1445	253	0.1000	0.0707	260	0.3506	0.0731
Yoruba	0.2419	0.1444	213	0.0985	0.0564	300	0.3437	0.0916
Mandenka	0.2370	0.1453	227	0.0998	0.0581	286	0.3460	0.0914
Mozabite	0.2670	0.1259	168	0.1196	0.0521	345	0.3387	0.0807
Bedouin	0.2564	0.1284	179	0.1181	0.0599	334	0.3305	0.0875
Palestinian	0.2539	0.1378	204	0.1159	0.0649	309	0.3450	0.0886
Druze	0.2468	0.1431	200	0.1002	0.0603	313	0.3404	0.0935
Basque	0.2299	0.1503	255	0.1008	0.0602	258	0.3575	0.0923
Russian	0.2235	0.1400	223	0.0935	0.0549	290	0.3235	0.0966
Adygei	0.2383	0.1446	231	0.1016	0.0527	282	0.3502	0.0888
Balochi	0.2459	0.1408	173	0.0877	0.0549	340	0.3264	0.0955
Kalash	0.2490	0.1460	202	0.0959	0.0659	311	0.3484	0.0850
Burusho	0.2628	0.1521	188	0.0882	0.0573	325	0.3638	0.0822
Uygur	0.2636	0.1410	167	0.0961	0.0512	346	0.3444	0.0900
Yakut	0.2383	0.1484	182	0.0692	0.0507	331	0.3313	0.0913
Mongola	0.2507	0.1515	198	0.0903	0.0631	315	0.3515	0.0923
Daur	0.2303	0.1473	206	0.0823	0.0604	307	0.3297	0.0959
Yi	0.2481	0.1541	177	0.0715	0.0585	336	0.3412	0.0967
Cambodian	0.2510	0.1528	216	0.0943	0.0738	297	0.3649	0.0741
Lahu	0.2223	0.1603	261	0.0843	0.0745	252	0.3652	0.0798
Melanesian	0.2155	0.1761	261	0.0589	0.0646	252	0.3777	0.0838
Papuan	0.2113	0.1626	240	0.0590	0.0622	273	0.3453	0.0889
Pima	0.2047	0.1828	237	0.0267	0.0445	276	0.3576	0.0987
Maya	0.2142	0.1633	256	0.0703	0.0656	257	0.3575	0.0878
Colombian	0.2157	0.1617	227	0.0607	0.0599	286	0.3387	0.0991

The statistics reported here correspond to those of the marker sets that yielded the imputation accuracy plotted in Figure 9.

Supplemental References

1. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype, and copy-number variation in worldwide human populations. Nature 451, 998–1003.

2. Conrad, D.F., Jakobsson, M., Coop, G., Wen, X., Wall, J.D., Rosenberg, N.A., and Pritchard, J.K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet. 38, 1251–1260.

3. Pemberton, T.J., Jakobsson, M., Conrad, D.F., Coop, G., Wall, J.D., Pritchard, J.K., Patel, P.I., and Rosenberg, N.A. (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.