Cell Genomics, Volume *3*

Supplemental information

Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations

Dhriti Sengupta, Gerrit Botha, Ayton Meintjes, Mamana Mbiyavanga, AWI-Gen Study, H3Africa Consortium, Scott Hazelhurst, Nicola Mulder, Michèle Ramsay, and Ananyo Choudhury

Supplementary Figures

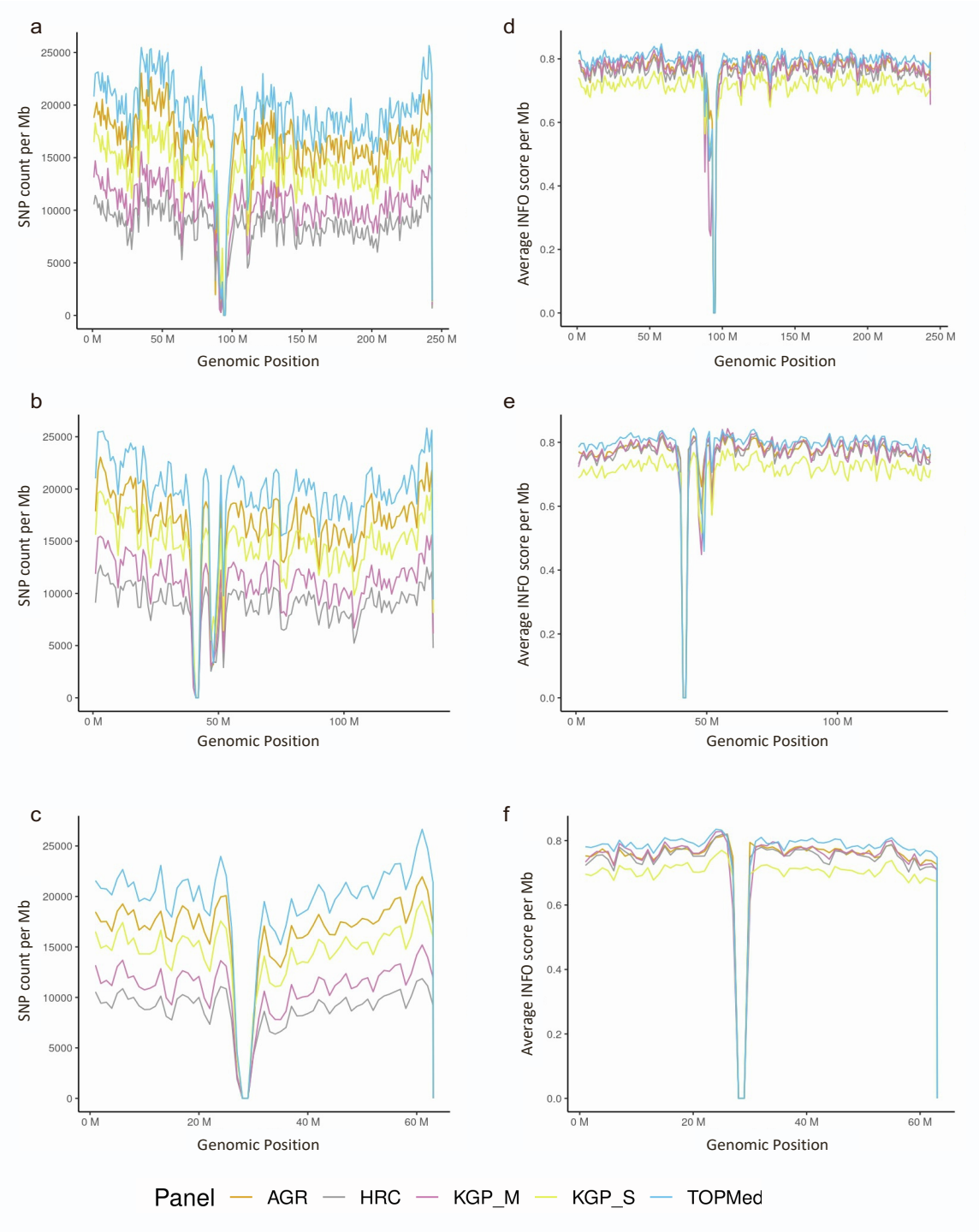

Figure S1. Chromosome wise evaluation of the AWI-Gen dataset imputed by the five reference panels, related to Figure 2. Density of SNPs imputed per Mb for (a) Chromosome 2, (b) Chromosome 10 and (c) Chromosome 20, Average imputation score per Mb for (d) Chromosome 1, (e) Chromosome 10 and (f) Chromosome 20. Panel codes: AGR (African Genome Resource hosted at Sanger Imputation Server (SIS)), KGP_S (1000 Genomes Project hosted at SIS), HRC (Haplotype Reference Consortium hosted at SIS), KGP_M (1000 Genomes Project hosted at Michigan Imputation Server) and TOPMed (hosted at TOPMed Imputation Server)

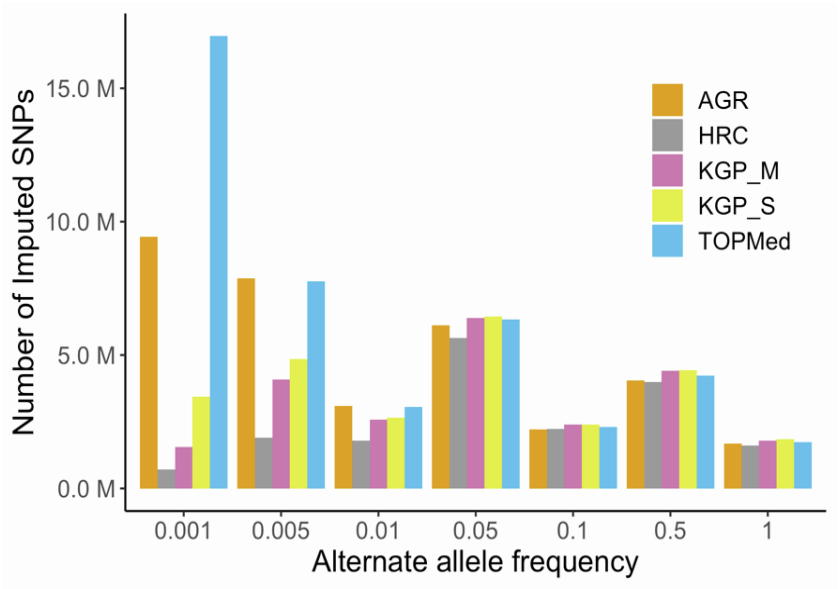

Figure S2. Number of well imputed SNPs with INFO score (or R2) over 0.6 imputed by the five panels across allele frequency bins, related to Figure 2. Panel codes are as in Figure S1.

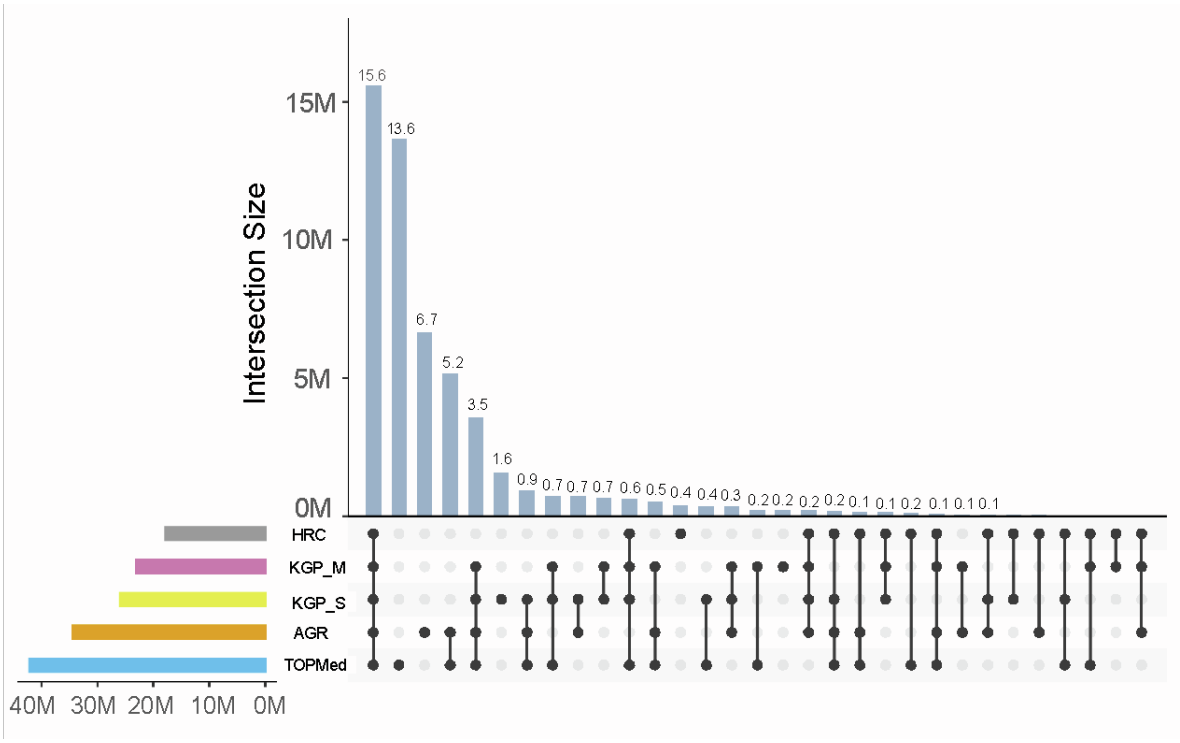

Figure S3. UPSET plots showing the intersection and union of SNPs imputed with INFO score (or R2) over 0.6 by the five imputation panels, related to Figure 3. Panel codes are as in Figure S1.

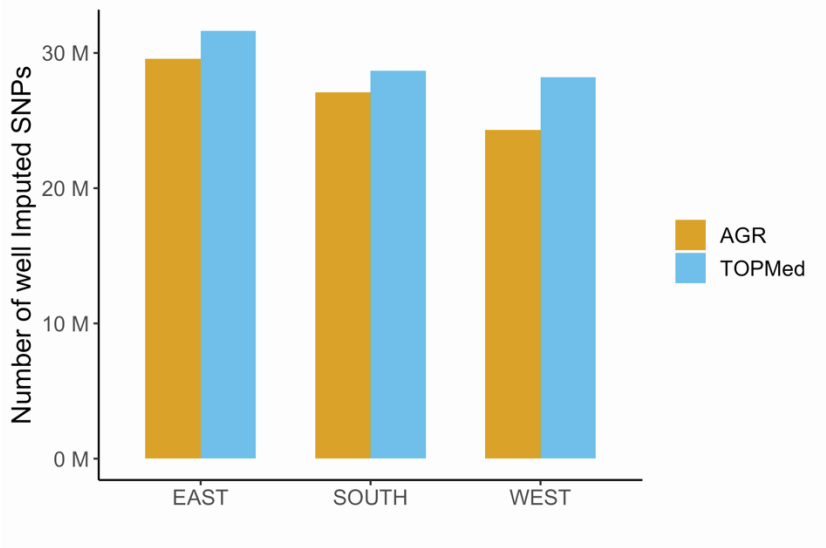

Figure S4. Imputation of SNPs with INFO score (or R2) over 0.6 for East, West and South African samples, related to Figure 4. Panel codes are as in Figure S1.

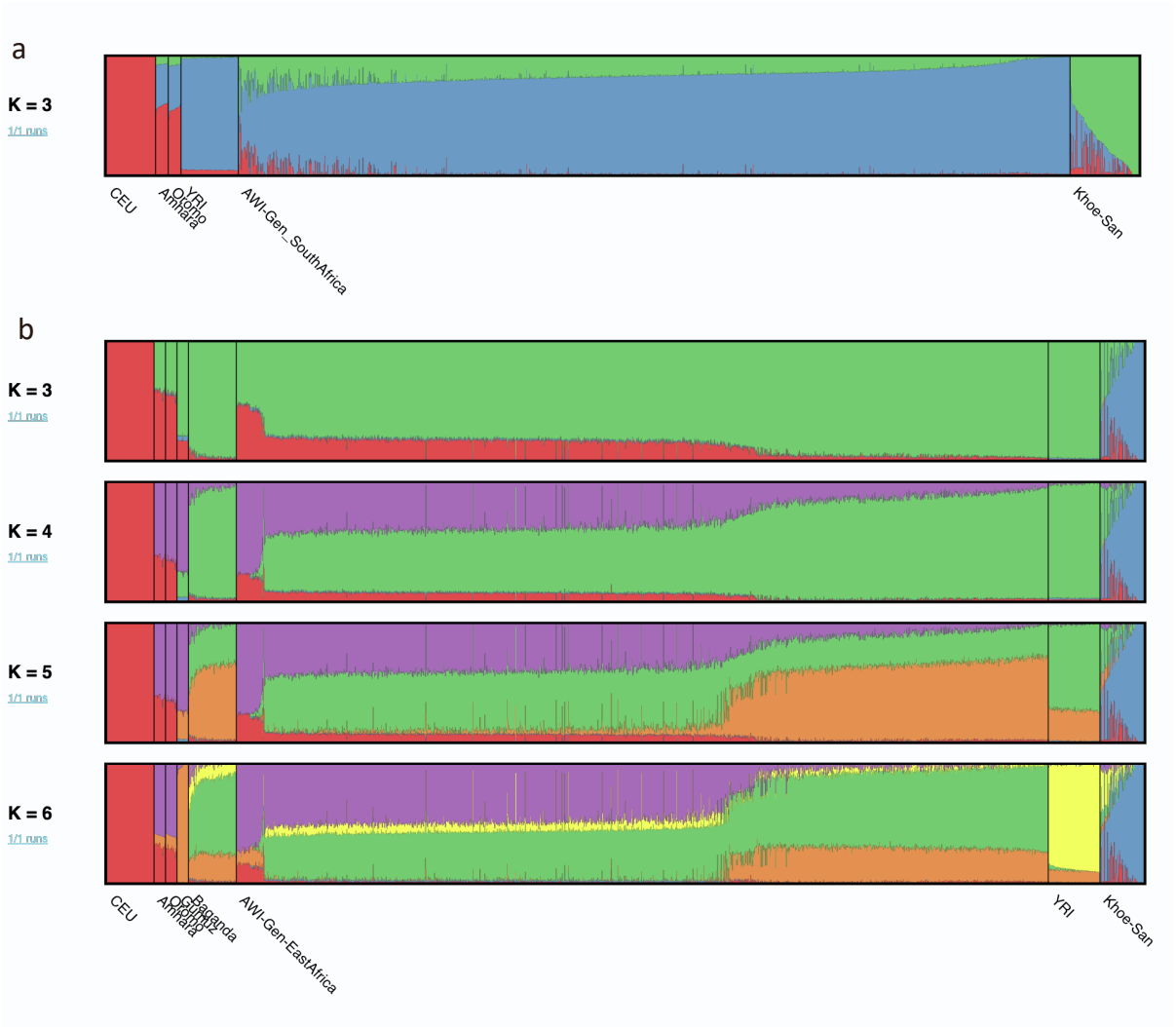

Figure S5. Non-Niger-Congo gene flow into South African and East African populations, related to Figure 4. (a) ADMIXTURE plot at *K*=3 based on the merged dataset with subset of South African individuals from AWI-Gen dataset, Khoe-San populations from Schlebusch et al. (Schlebusch et al., *Science* 2012), and the other populations including Amhara and Oromo from Gurdasani et al. (Gurdasani et al., *Nature* 2015). The plot shows differences in the level of Khoe-San gene flow (shown in green) into the South African populations. (b) ADMIXTURE plot at *K*=3-6 based on the merged dataset with East African (from Kenya) individuals from AW-Gen dataset, Khoe-San populations from Schlebusch et al. (Schlebusch et al., *Science* 2012), and the other populations including CEU, Amhara, Oromo and Gumuz from Gurdasani et al. (Gurdasani et al., *Nature* 2015). The plot shows differences in the level of non-Niger Congo (Afro-Asiatic/Nilo-Saharan/Eurasian) ancestry (shown in red) into the East African populations.

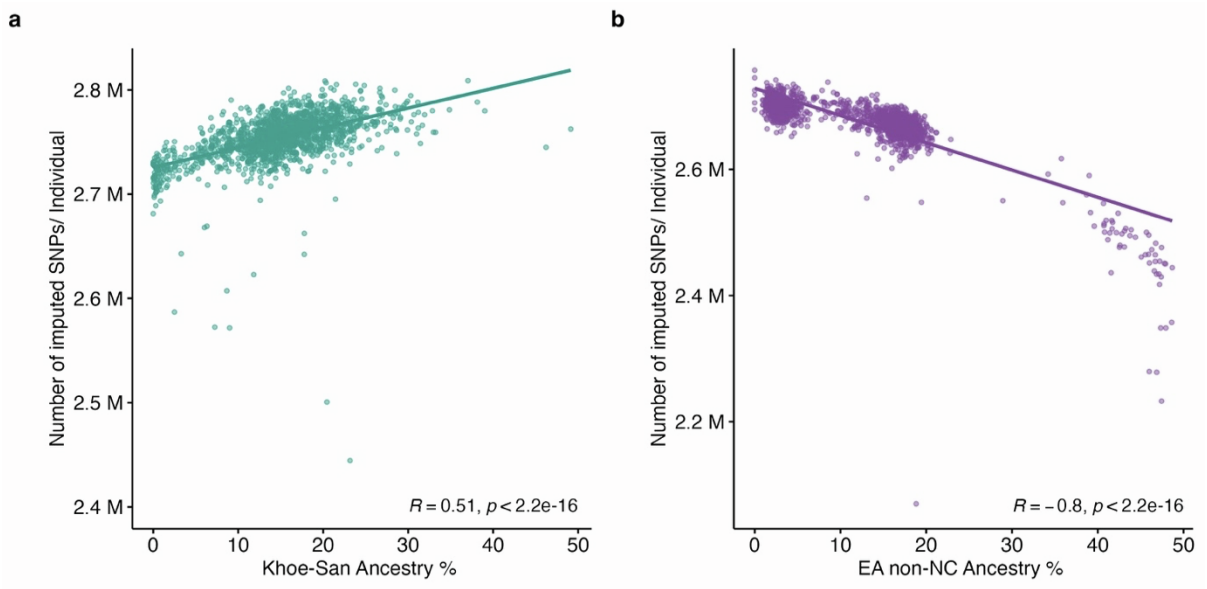

Figure S6. Impact of ancestry on number of SNP imputed using TOPMed, related to Figure 4. Correlation between the number of SNPs imputed per individual and (a) the level of Khoe-San ancestry in South African participants. (b) the level of East African non-Niger-Congo (EA non-NC) ancestry (Afro-Asiatic/Nilo-Saharan/Eurasian) in the East African participants. The regression line along with correlation coefficient (R) and p value (Pearson correlation) are shown. The ancestry proportions were inferred using ADMIXTURE (see Figure S5).

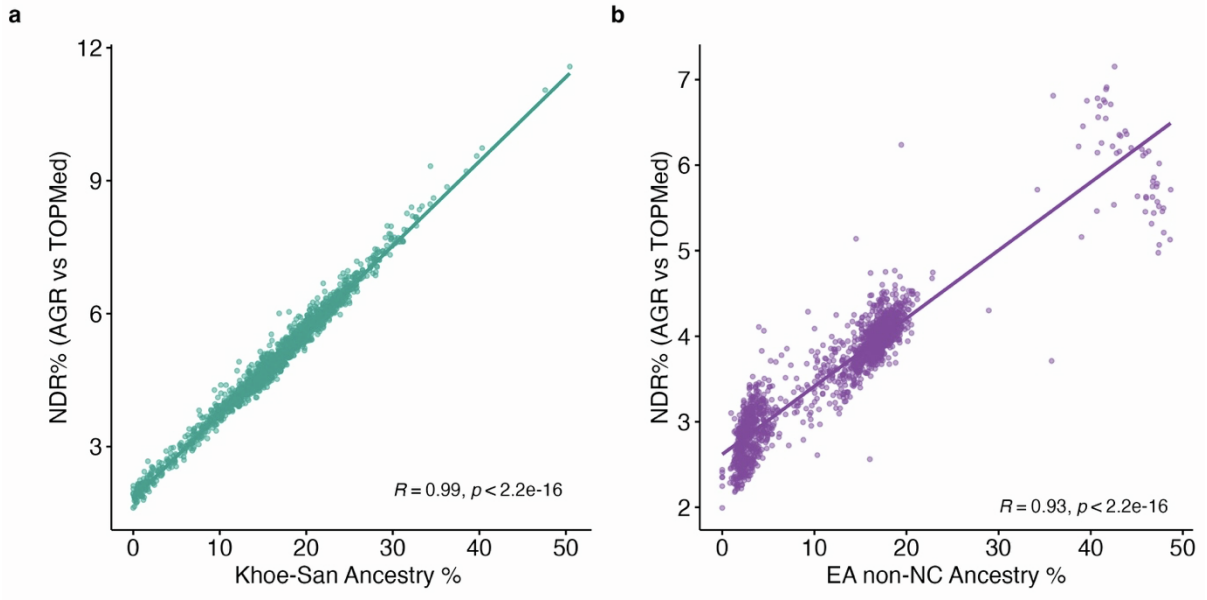

Figure S7. Impact of ancestry on non-reference discordance rate (NDR) between genotypes imputed using AGR and TOPMed, related to Figure 4. Correlation between NDR and (a) the level of Khoe-San ancestry in South African participants (b) the level of east African non-Niger-Congo (EA non-NC) ancestry (Afro-Asiatic or Nilo-Saharan or Eurasian ancestry) in the east African participants. The regression line along with correlation coefficient (R) and p value (Pearson correlation) are shown. The ancestry proportions were inferred using ADMIXTURE (see Figure S5).

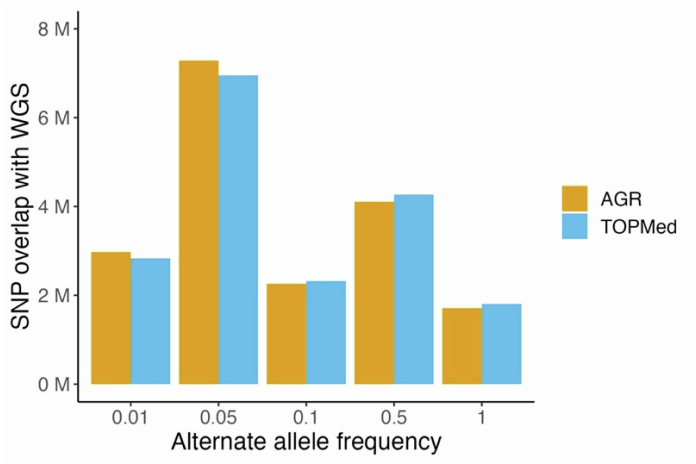

Figure S8. Number of SNPs imputed by AGR and TOPMed that overlap with WGS data across allele frequency bins, related to Figure 5. Panel codes are as in Figure S1.

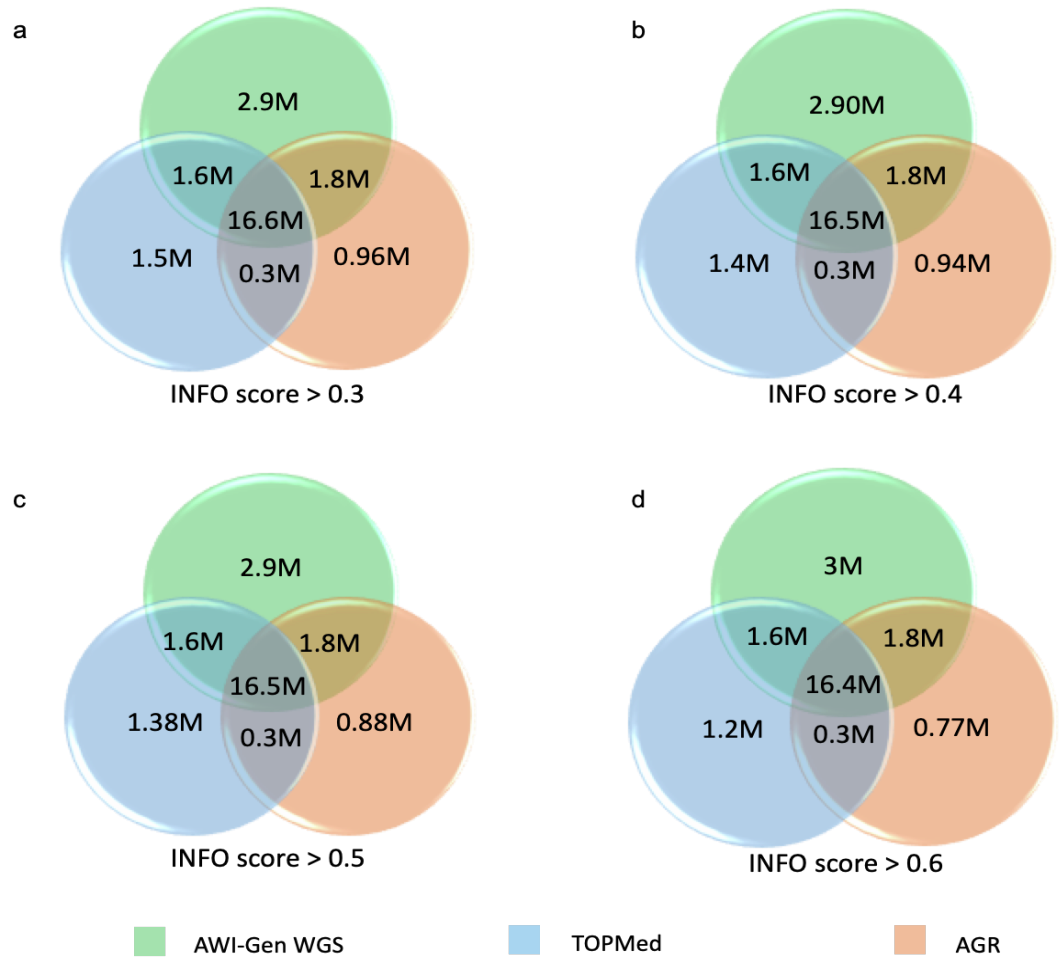

Figure S9. Overlap of SNPs between 95 high-coverage WGS data and datasets imputed using AGR and TOPMed at different INFO score (or R2) cutoffs, related to Figure 5. Panel codes are as in Figure S1.

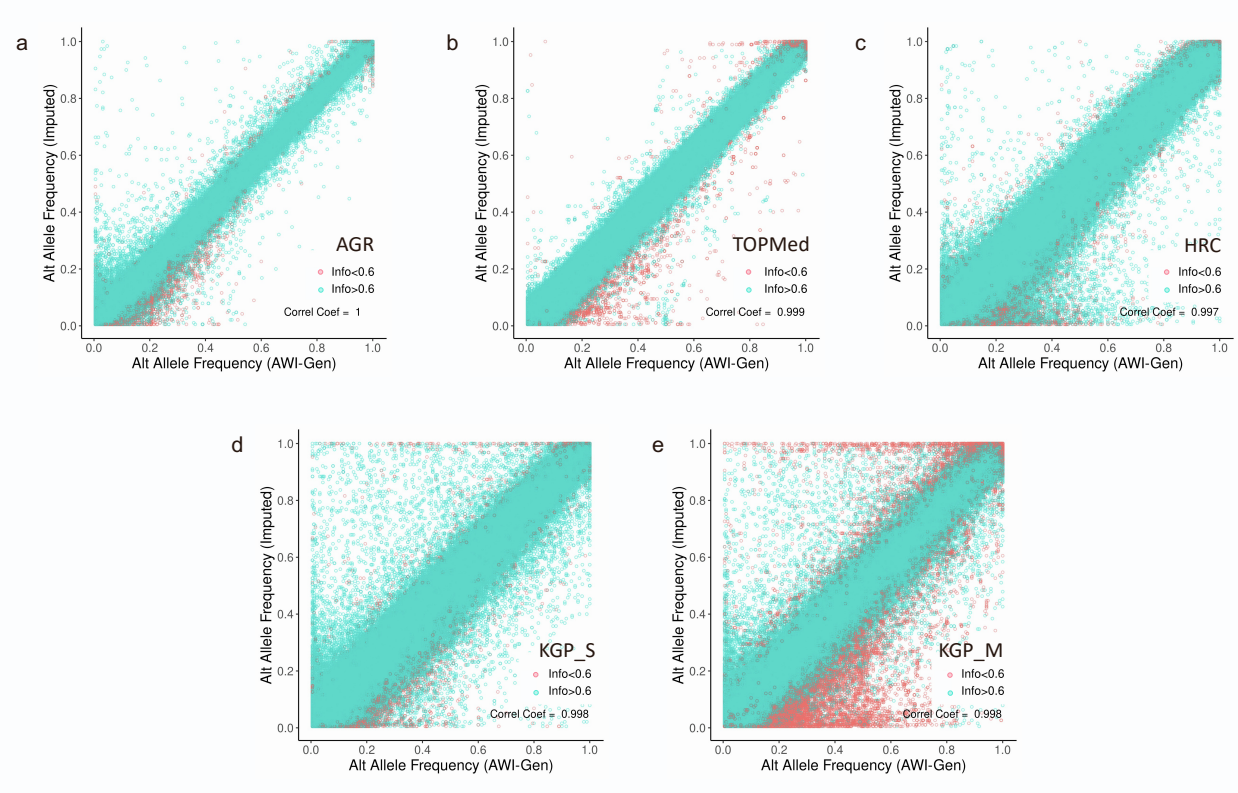

Figure S10. Comparison of WGS based allele frequencies to allele frequencies in the datasets imputed, related to Figure 5. (a) AGR (b) TOPMed, (c) HRC, (d) KGP_S and (e) KGP_M panels. Green open circles show SNPs with INFO score (or R2) greater than 0.6 and the red open circles represent SNPs with INFO score (or R2) less than 0.6. Panel codes are as in Figure S1.

Supplementary Tables

Table S1. Self-reported ethnic distribution of AWI-Gen participants across the four countries, related to STAR Methods.

Country (study centre)	Ethnolinguistic group
South Africa (Agincourt, Dikgale and Soweto)	Tsonga, BaPedi, Zulu, Sotho, Tswana, Xhosa, Swati, Venda, Ndebele, Other[a], Unknown[b]
Burkina Faso (Nanoro)	Mossi, Gourounsi, Peulh, Dagara, Dioula, Samo, Gourmatche, Other[a], Unknown[b]
Ghana (Novrongo)	Kassena, Nankana, Bulsa, Mampruga, Frafra, Kantosi, Mossi, Other[a], Unknown[b]
Kenya (Nairobi)	Kikuyu, Kamba, Luo, Luhya, Kisii, Somali, Meru, Embu, Borana, Gari, Kalenjin, Maasai, Other[a]

[a] Only one or two individuals in a specific ethnic category.
[b] Person did not provide information on ethnicity.

Table S2. Number of SNPs showing differences in allele frequencies (AF>0.01) in datasets imputed using different panels, related to Figure 3.

	AGR	KGP_S	HRC	KGP_M
TOPMed	86698	328150	572538	258225
AGR		404233	651655	359029
KGP_S			343108	92256
HRC				557591

Panel codes are as in Figure S1

Table S3. Change in the number of SNPs (in millions) with increase in INFO or R2 score cut-offs, related to Table 3.

INFO score	AGR	TOPMed	KGP_S	KGP_M	HRC
>0	18.34	18.19	16.25	16.07	14.09
> 0.3	18.34	18.15	16.24	16.01	14.08
> 0.4	18.33	18.13	16.22	15.97	14.05
> 0.5	18.30	18.11	16.15	15.91	13.97
> 0.6	18.22	18.05	15.99	15.79	13.79
> 0.7	18.02	17.93	15.65	15.51	13.41
> 0.8	17.47	17.6	14.86	14.83	12.6
> 0.9	15.59	16.43	12.68	12.79	10.54

Panel codes are as in Figure S1