The American Journal of Human Genetics, Volume *109*

Supplemental information

Extent to which array genotyping and imputation with large reference panels approximate deep whole-genome sequencing

Sarah C. Hanks, Lukas Forer, Sebastian Schönherr, Jonathon LeFaive, Taylor Martins, Ryan Welch, Sarah A. Gagliano Taliun, David Braff, Jill M. Johnsen, Eimear E. Kenny, Barbara A. Konkle, Markku Laakso, Ruth F.J. Loos, Steven McCarroll, Carlos Pato, Michele T. Pato, Albert V. Smith, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Michael Boehnke, Laura J. Scott, and Christian Fuchsberger

Supplemental Figures

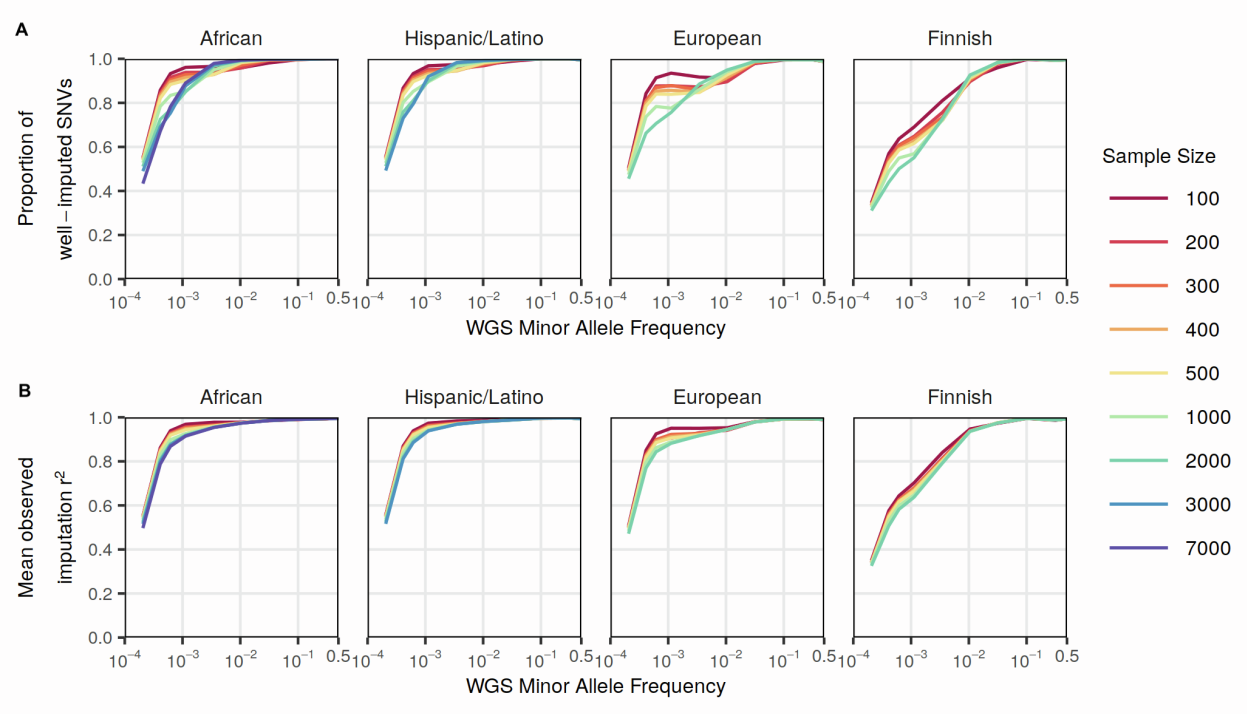


Figure S1. Effect of sample size on imputation quality metrics. Random subsets of individuals were taken from each of the WGS studies as the total sample size of unrelated individuals allowed (up to 7,000 for African, 3,000 for Hispanic/Latino, and 2,000 for European and Finnish). Imputation was performed with the Omni 2.5M array and the TOPMed imputation reference panel. A. The proportion of sequenced biallelic SNVs that are well-imputed (r^2>0.8) by sample size. B. The mean r2 by sample size. In both plots, the x-axes show minor allele frequency (MAF) calculated separately by study based on the 2,429 samples used in the main analyses. Sequenced biallelic SNVs not present in reference panels were assigned r^2=0. Biallelic SNVs were then aggregated by MAF bins of width 0.00025 MAF for MAF between 0.0002 and 0.002 and of size 0.001 MAF for MAF > 0.002; those plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.001, 0.0032, 0.01, 0.032, 0.1, 0.32, and 0.5.

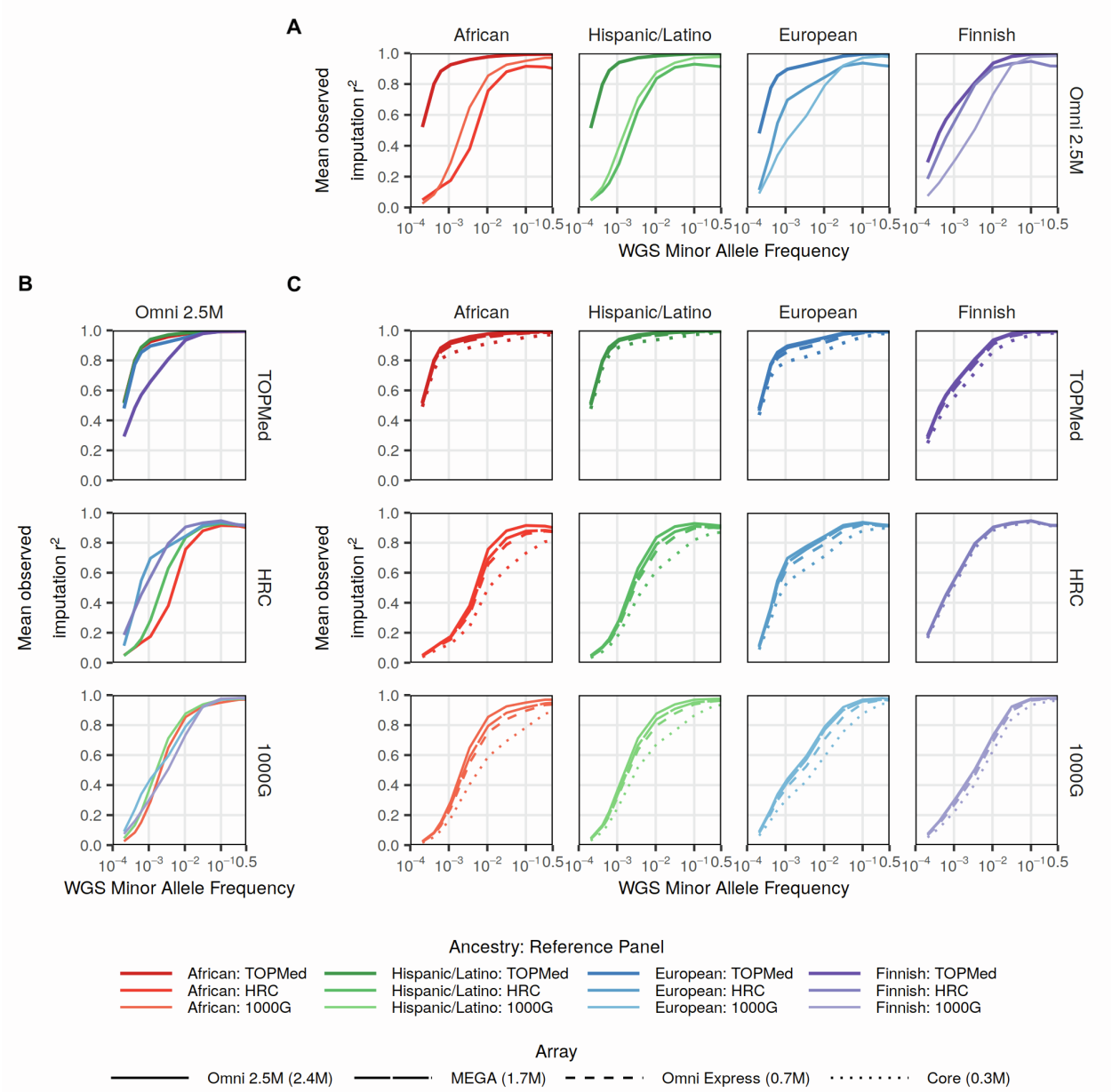


Figure S2. Mean observed imputation r^2 of biallelic SNVs by reference panel, study ancestry, and genotyping array. The mean observed imputation r^2 with the TOPMed, HRC, and 1000G imputation reference panels. A. Comparison across the reference panels using the Illumina Omni 2.5M array. B. Comparison across the four studies using the Illumina Omni 2.5M array. C. Comparison across four Illumina genotyping arrays: Omni 2.5M, MEGA, Omni Express, and Core by ancestry (columns) and imputation reference panels (rows). In all plots, the x-axes show minor allele frequency (MAF) calculated separately by study. Sequenced biallelic SNVs not present in reference panels were assigned r^2=0. Biallelic SNVs were then aggregated by MAF bins of width 0.00025 MAF for MAF between 0.0002 and 0.002 and of size 0.001 MAF for MAF > 0.002; those plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.001, 0.0032, 0.01, 0.032, 0.1, 0.32, and 0.5.

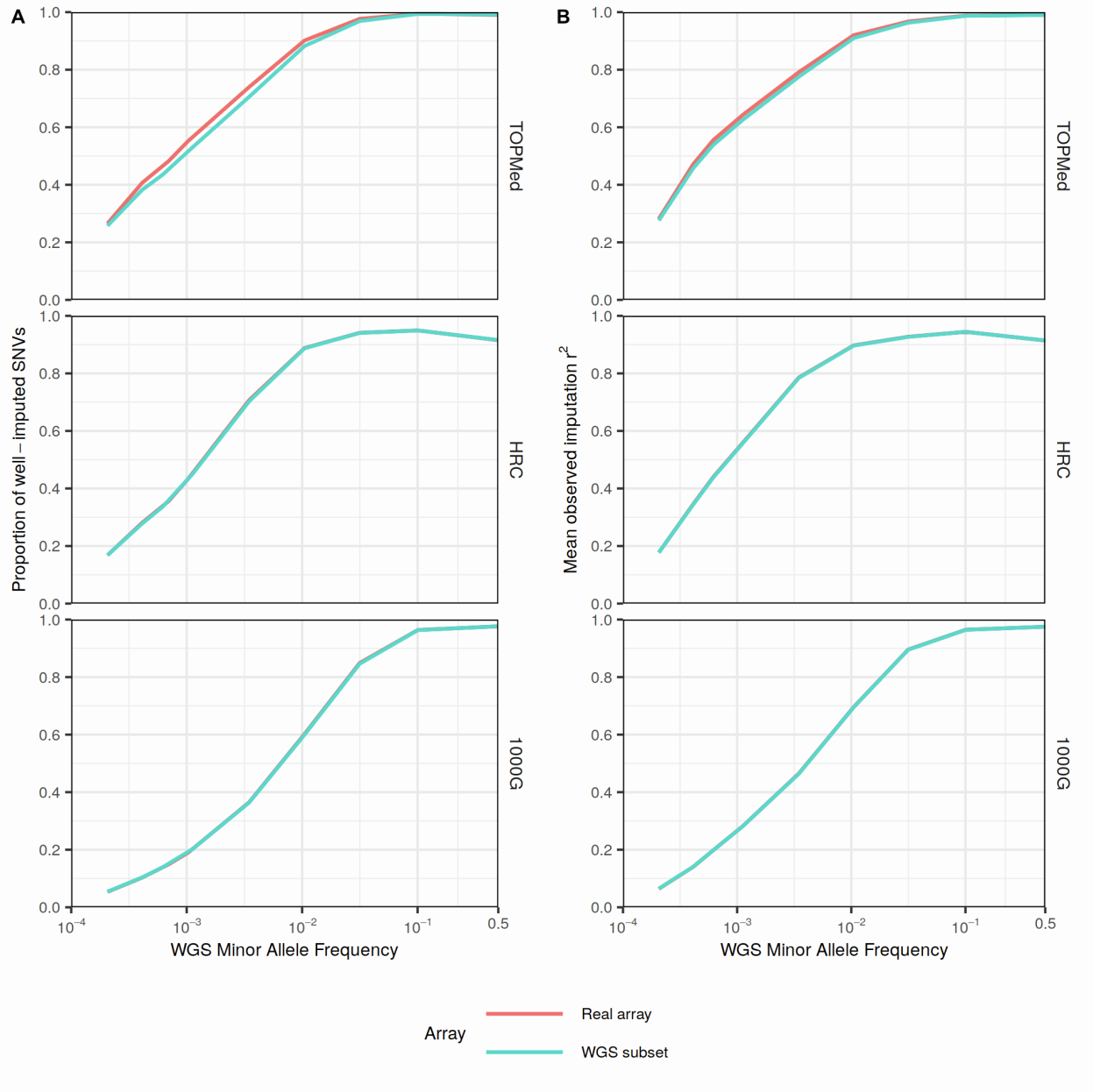


Figure S3. Imputation quality of biallelic SNVs by reference panel using WGS-based and real Illumina OmniExpress arrays. A. The proportion of sequenced biallelic SNVs imputed from real array data (red line) or from WGS-based array (blue line) in the Finnish study that are well-imputed (r^2>0.8) by imputation reference panel. B. The mean observed imputation r^2 for the same variants. In all plots, the x-axes show minor allele frequency (MAF) calculated separately by study. Variants were aggregated by MAF bins of size 0.00025 MAF for MAF between 0.0002 and 0.002 and of size 0.001 MAF for MAF > 0.002; those plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.001, 0.0032, 0.01, 0.316, 0.1, and 0.5. The lines appear entirely overlapping for the HRC and 1000G reference panels.

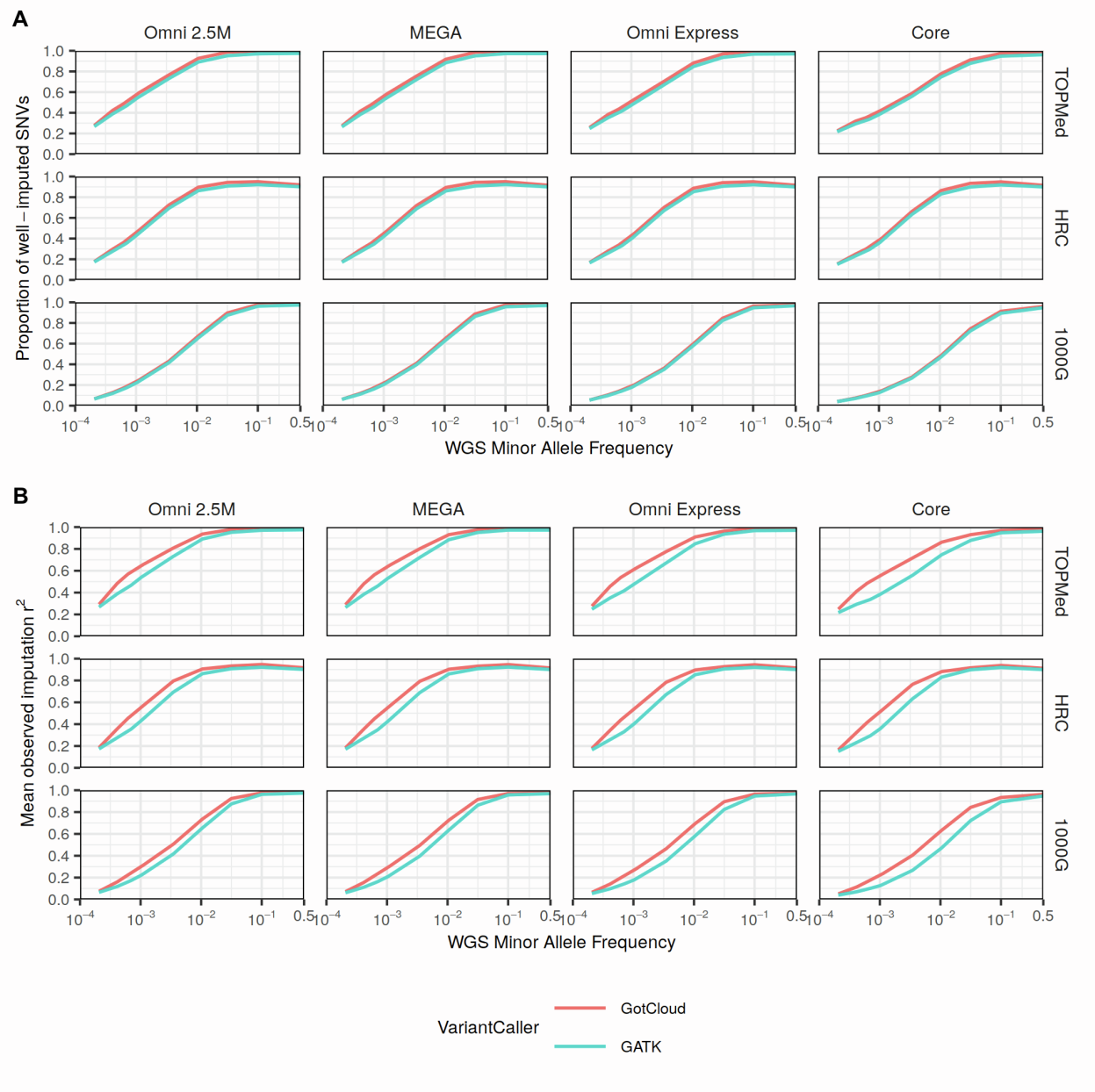


Figure S4. Proportion of well-imputed (r^2>0.8) biallelic SNVs by reference panel, genotyping array, and variant caller in Finnish study. The proportion of sequenced biallelic SNVs called with the GotCloud pipeline (red line) or GATK pipeline (blue line) in the Finnish study that are well-imputed (r^2>0.8) by reference panel (rows) and genotyping array (columns). In all plots, the x-axes show minor allele frequency (MAF) calculated separately by study. Variants were aggregated by MAF bins of size 0.00025 MAF for MAF between 0.0002 and 0.002 and of size 0.001 MAF for MAF > 0.002; those plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.001, 0.0032, 0.01, 0.316, 0.1, and 0.5.

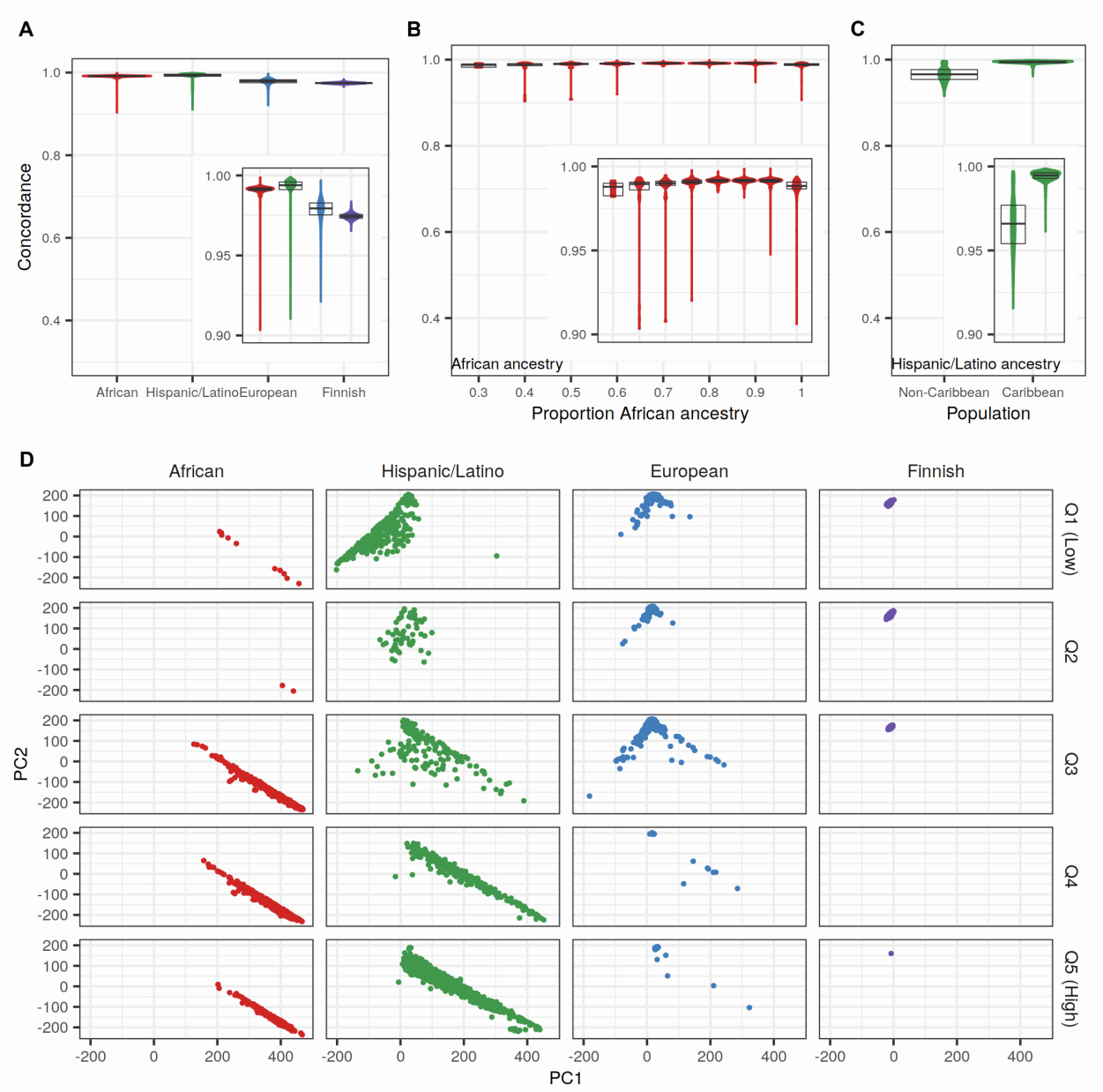


Figure S5. Heterozygous genotype concordance rates for low-frequency variants by ancestry with TOPMed panel imputation. Heterozygous concordance rates were calculated between sequenced and TOPMed imputed genotypes for low-frequency (0.5%<MAF<5%, calculated separately in each study) biallelic SNVs with the Omni2.5M array. A. Distribution of concordance rates in each of the four studies. Boxplots correspond to 25th, 50th, and 75th percentiles. B. Distribution of concordance rates by bins of estimated proportion of African ancestry in the admixed African study. C. Distribution of concordance rates in Caribbean and non-Caribbean populations in the Hispanic/Latino study. The inset figures in panels A-C show the same distributions with a restricted y-axis. D. Principal component analysis (PCA) by genotype concordance quintile and ancestry. PCA was performed by projecting onto the Human Genome Diversity Project reference samples. Genotype concordance quintiles were calculated across all four studies and correspond to concordance rates of 0.903-0.964 (Q1), 0.964-0.971 (Q2), 0.971-0.973 (Q3), 0.973-0.974 (Q4), and 0.974-0.974 (Q5). Points are colored by ancestry.

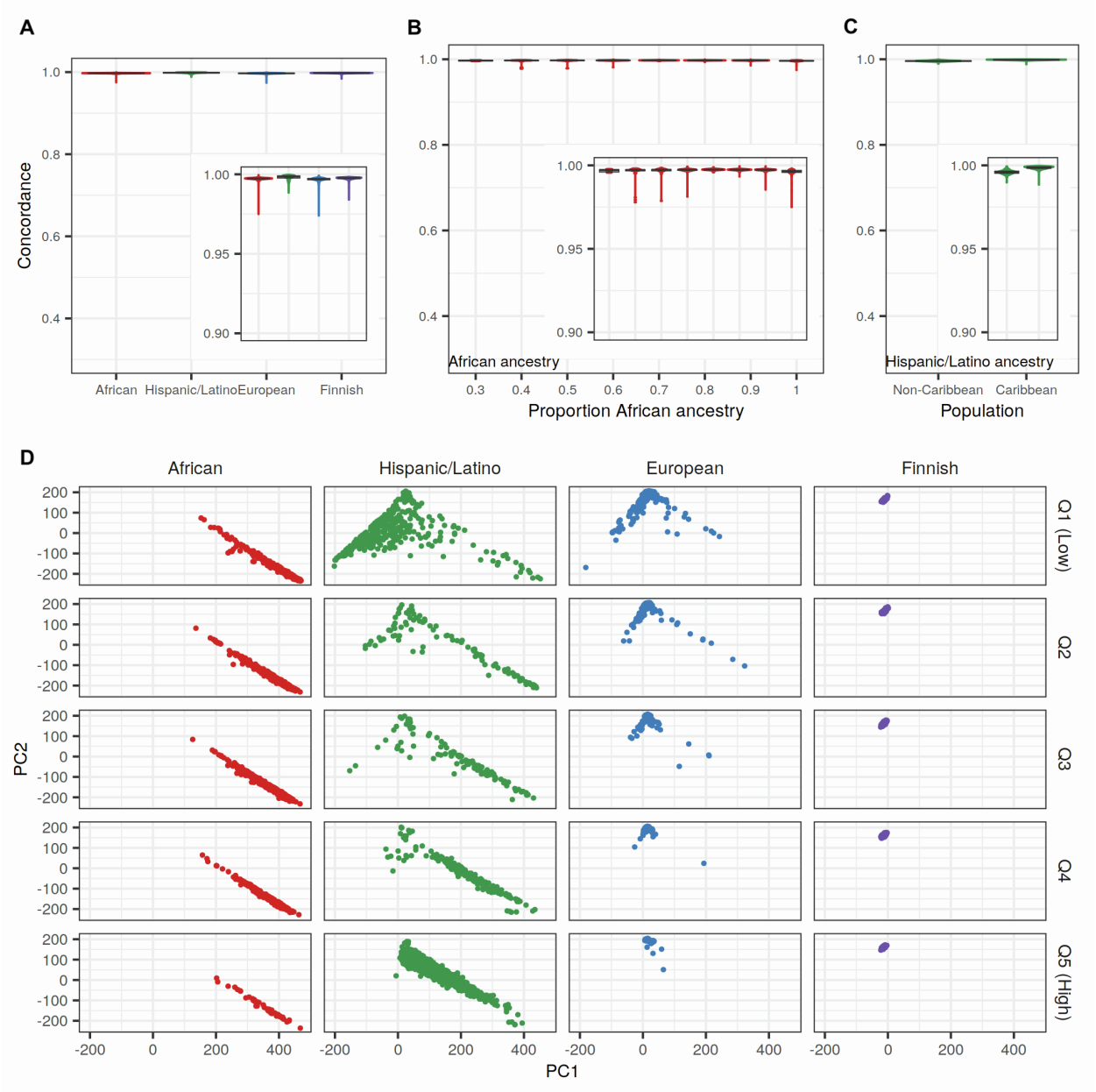


Figure S6. Heterozygous genotype concordance rates for common variants by ancestry with TOPMed panel imputation. Heterozygous concordance rates were calculated between sequenced and TOPMed imputed genotypes for common (MAF>5%, calculated separately in each study) biallelic SNVs with the Omni2.5M array. A. Distribution of concordance rates in each of the four studies. Boxplots correspond to 25th, 50th, and 75th percentiles. B. Distribution of concordance rates by bins of estimated proportion of African ancestry in the admixed African study. C. Distribution of concordance rates in Caribbean and non-Caribbean populations in the Hispanic/Latino study. The inset figures in panels A-C show the same distributions with a restricted y-axis. D. Principal component analysis (PCA) by genotype concordance quintile and ancestry. PCA was performed by projecting onto the Human Genome Diversity Project reference samples. Genotype concordance quintiles were calculated across all four studies and correspond to concordance rates of 0.974-0.995 (Q1), 0.995-0.996 (Q2), 0.996-0.996 (Q3), 0.996-0.997 (Q4), and 0.997-0.997 (Q5). Points are colored by ancestry.

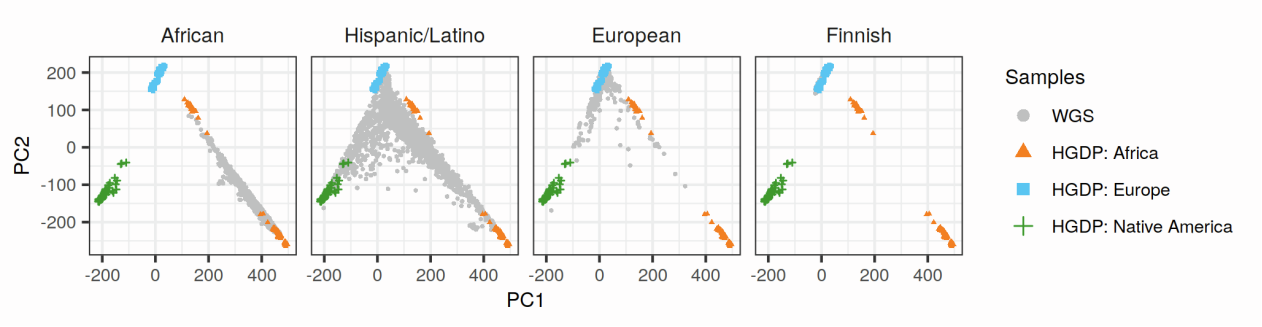
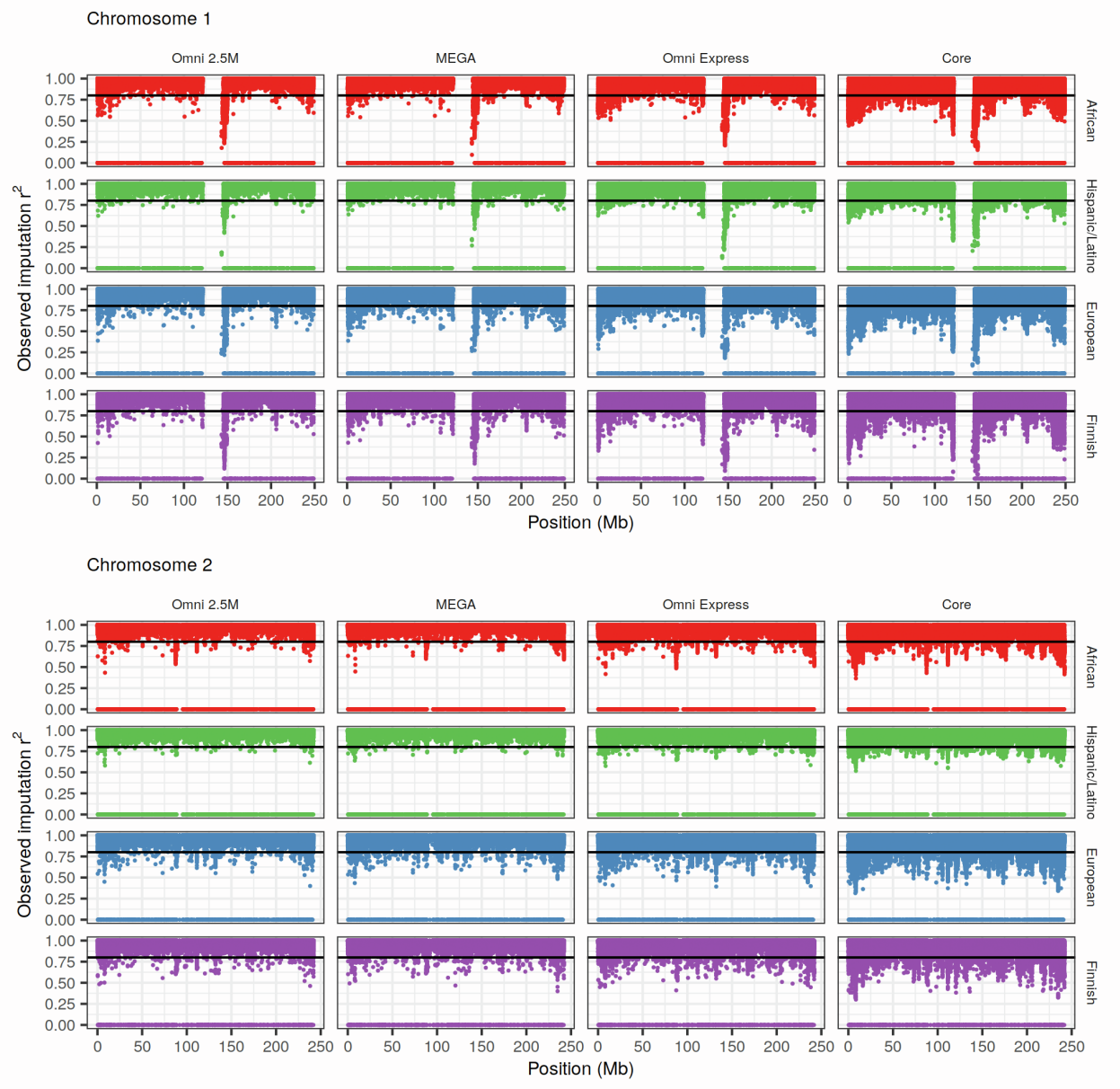
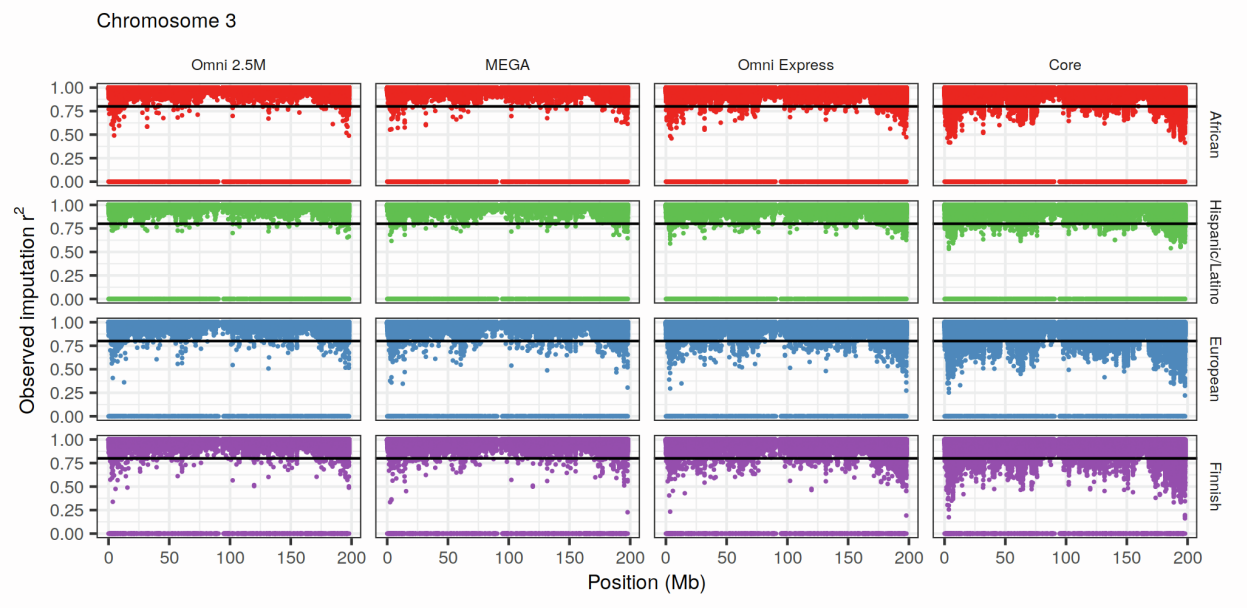
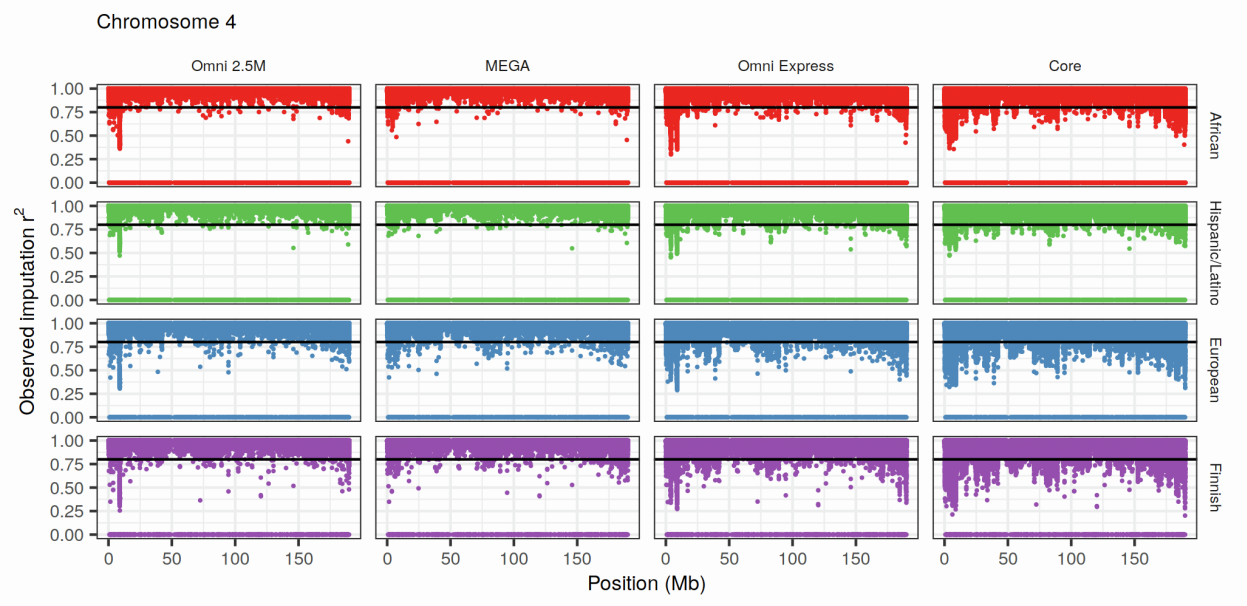
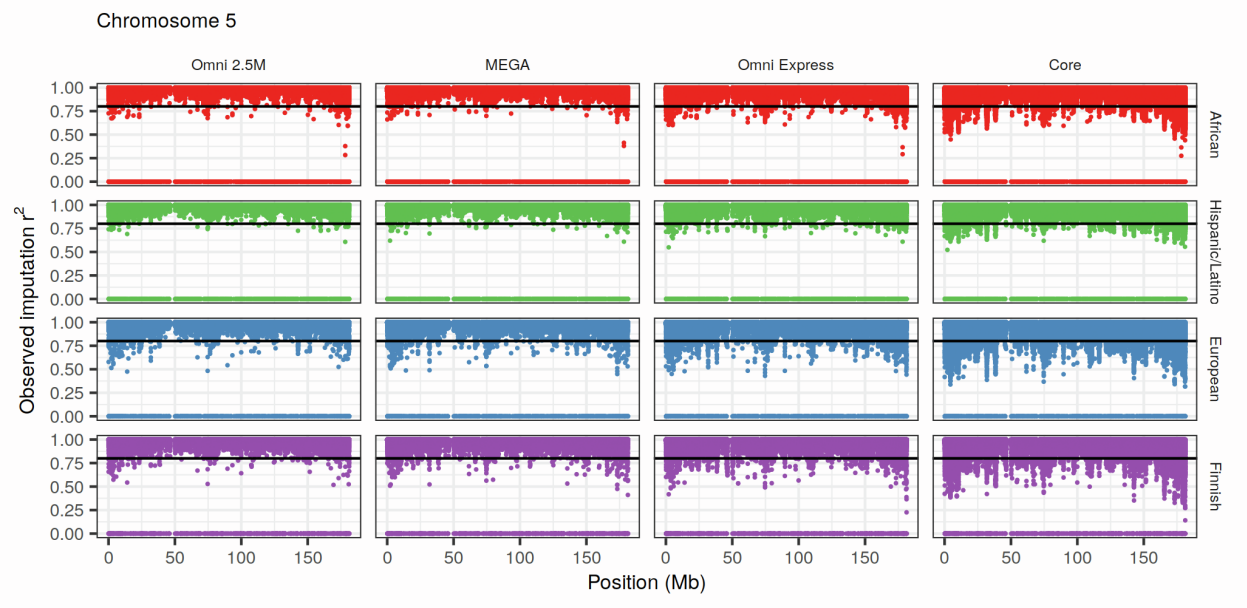


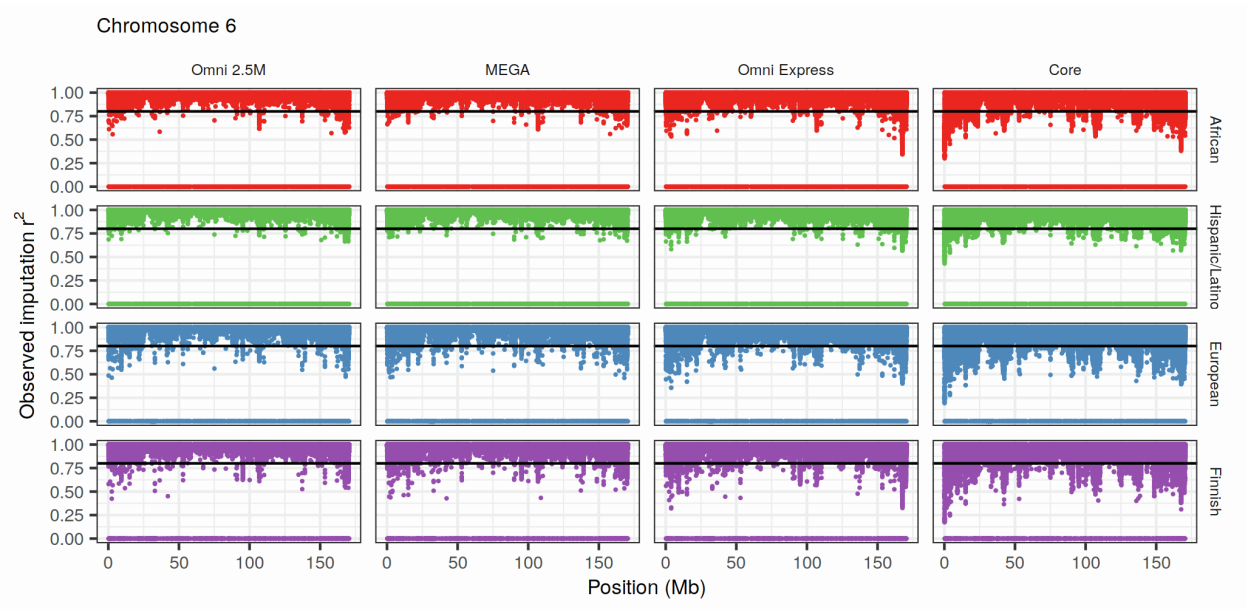
Figure S7. Principal component analysis of WGS samples. PC1 and PC2 for the four WGS studies and Human Genome Diversity Project (HGDP) reference samples from Africa (n=129), Europe (n=156), and Native America (n=63). PCA was performed by projecting onto all HGDP reference samples (n=938).

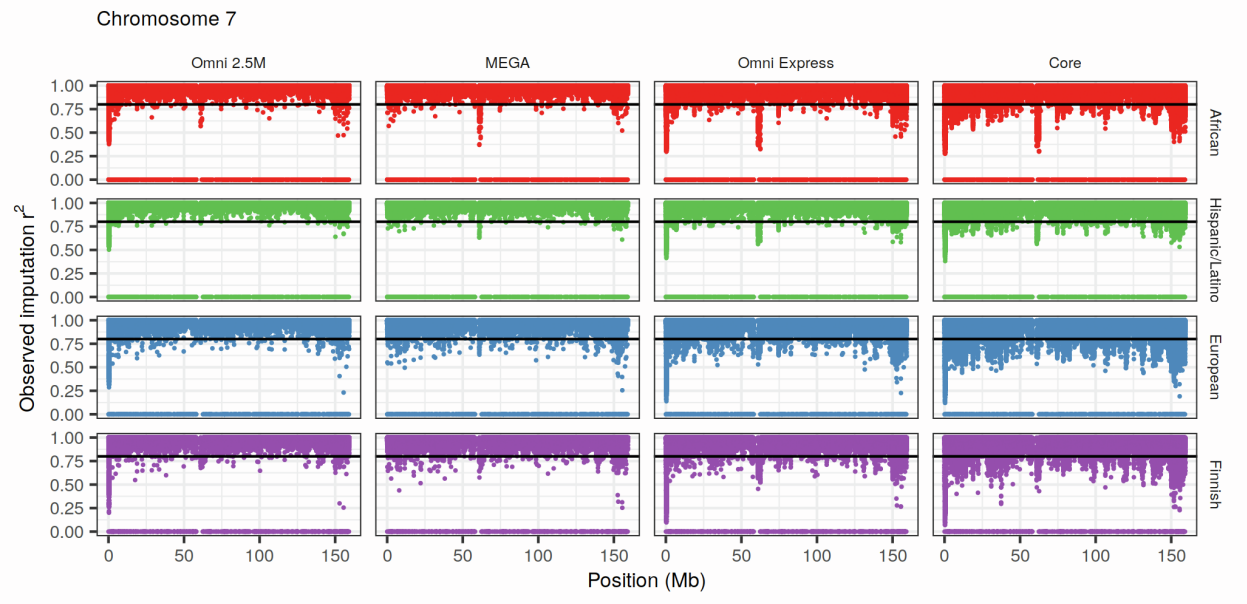

Chromosome 1
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r^2
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
250
Position (Mb)
Chromosome 2
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r^2
Position (Mb)

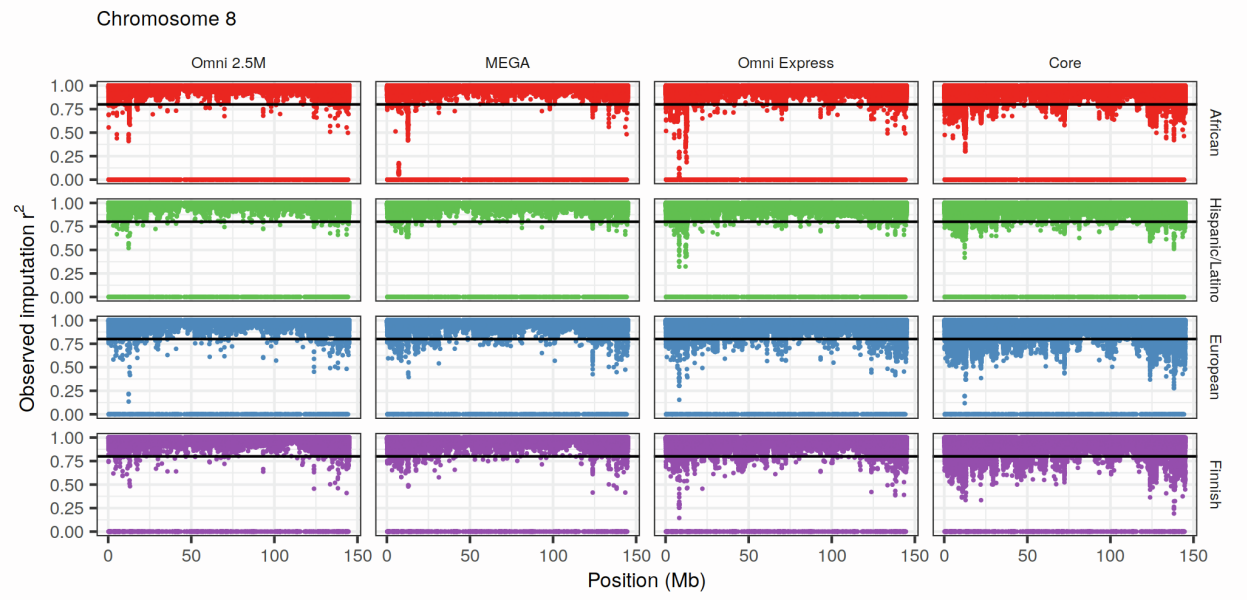

Chromosome 3
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
Position (Mb)

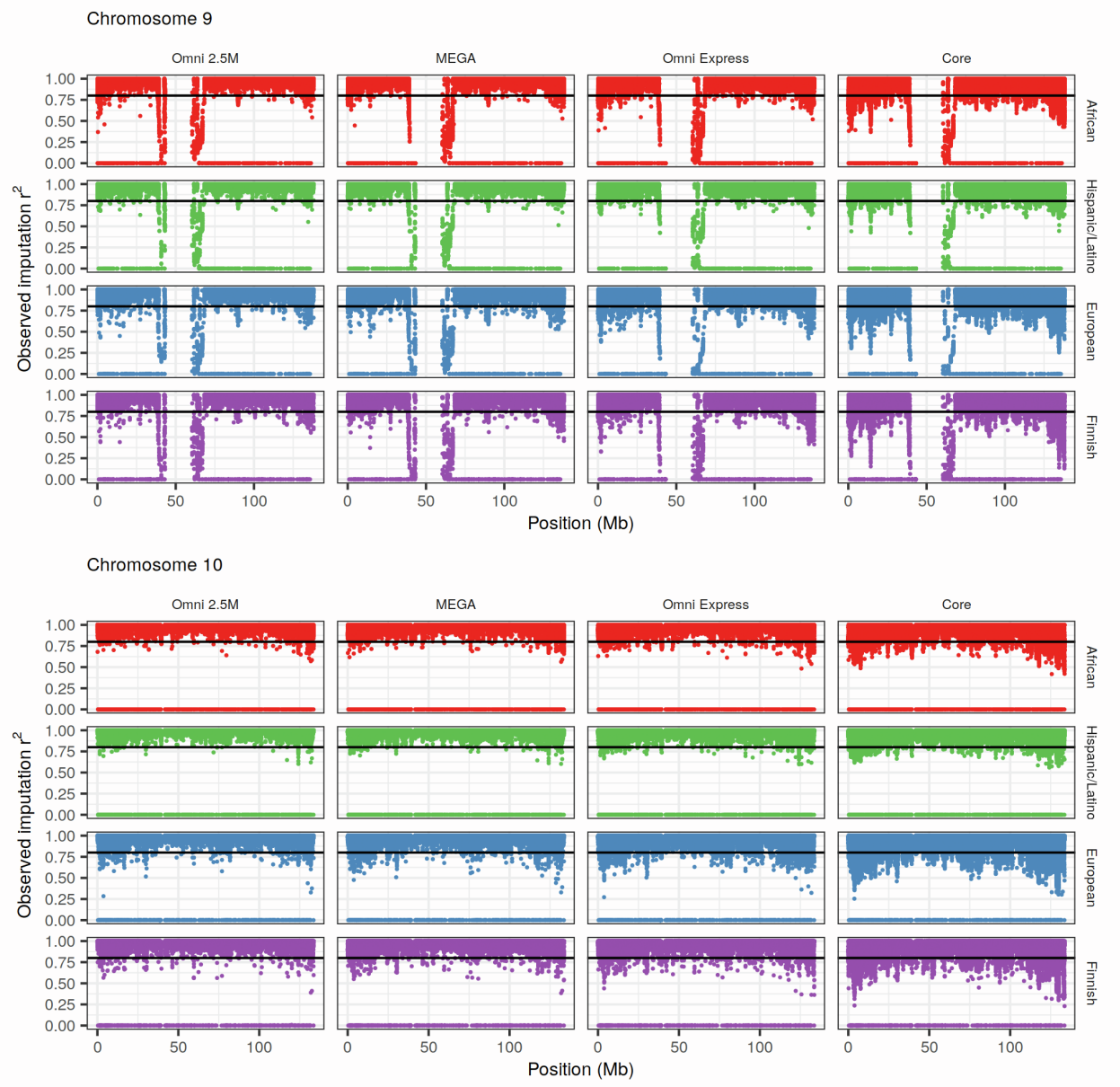

Chromosome 4
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
1.00
0.75
0.50
0.25
0.00
0
50
100
150
Position (Mb)

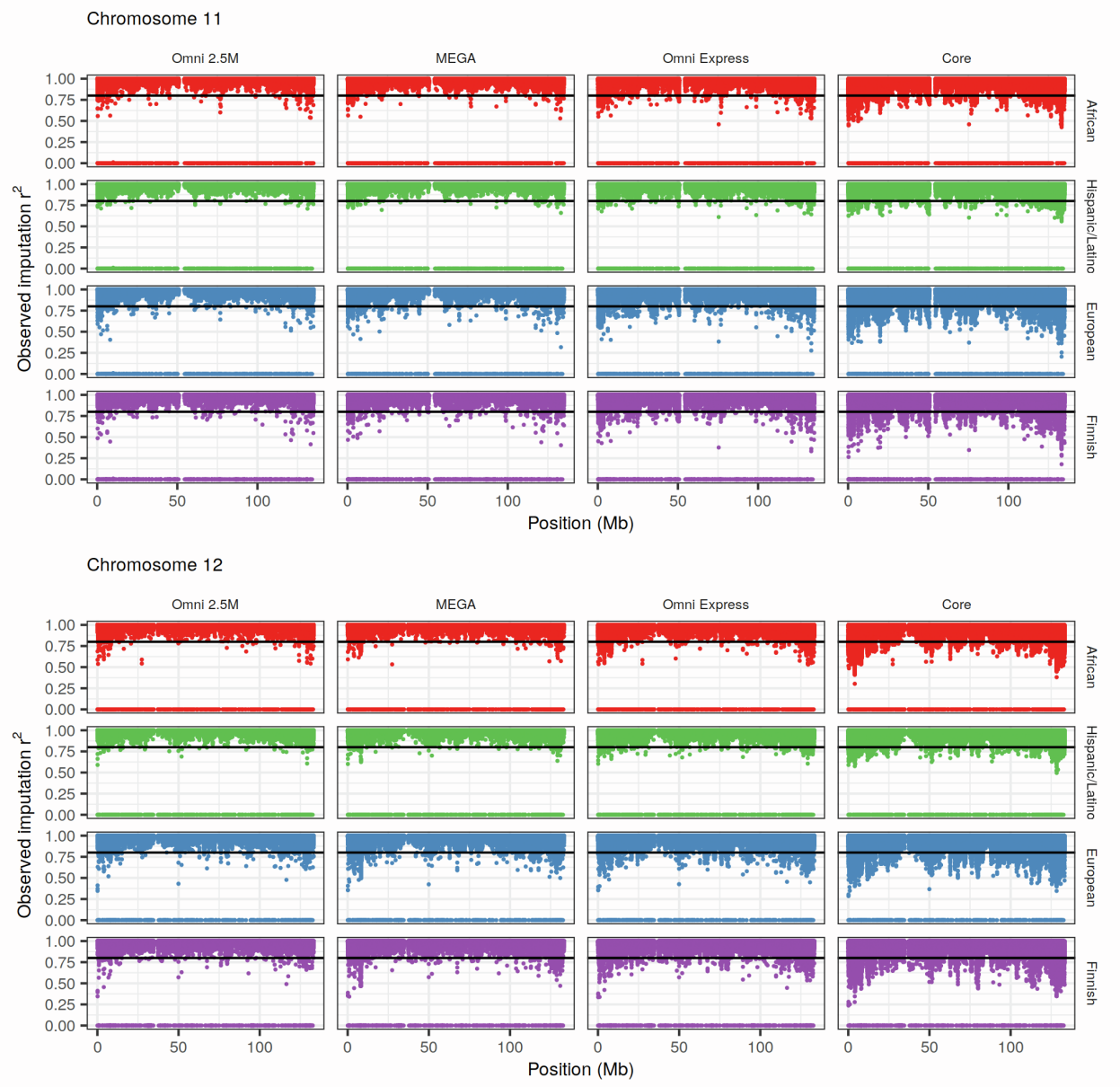

Chromosome 5
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
1.00
0.75
0.50
0.25
0.00
0
50
100
150
Position (Mb)

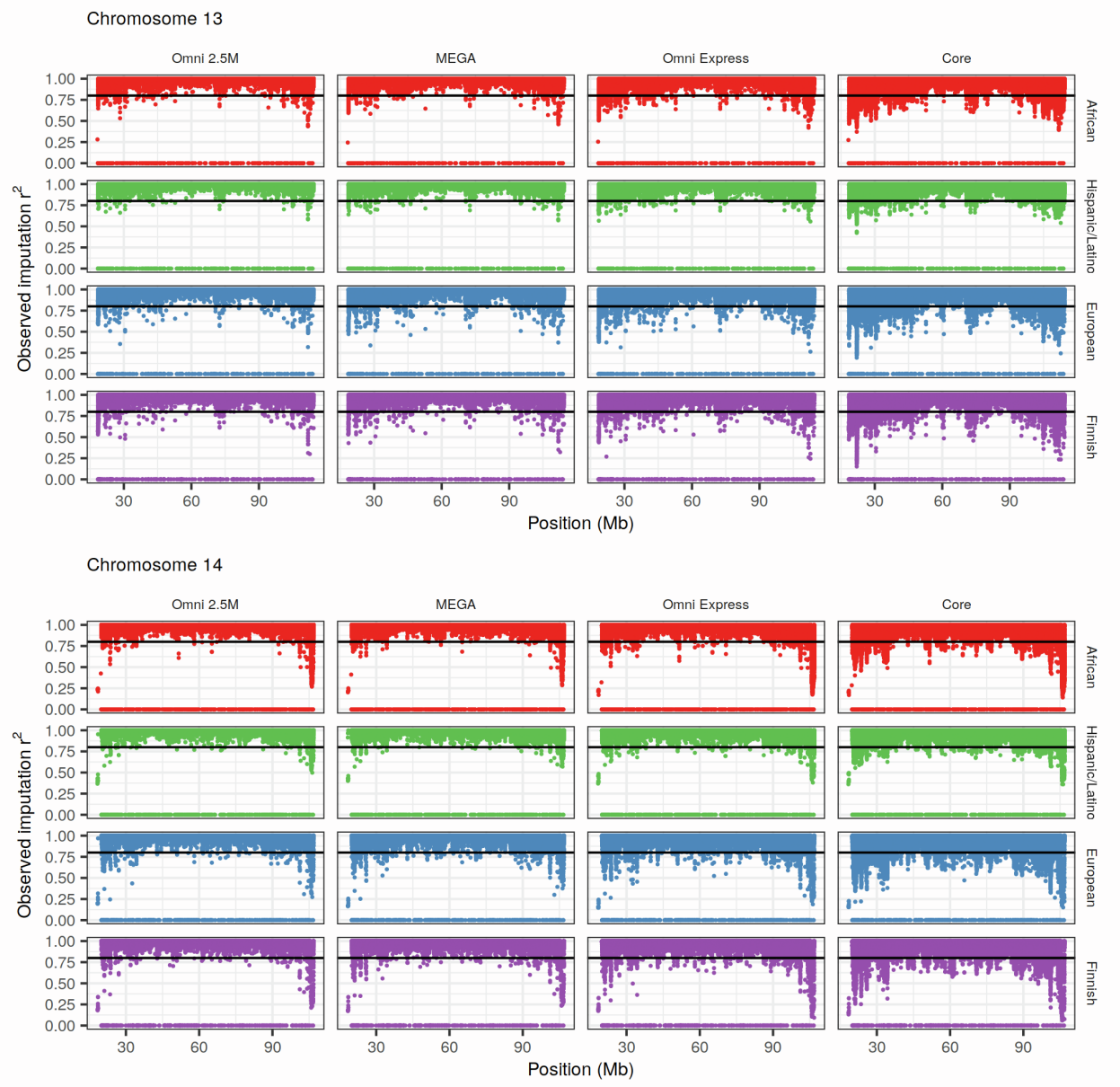

Chromosome 6
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
1.00
0.75
0.50
0.25
0.00
0
50
100
150
Position (Mb)

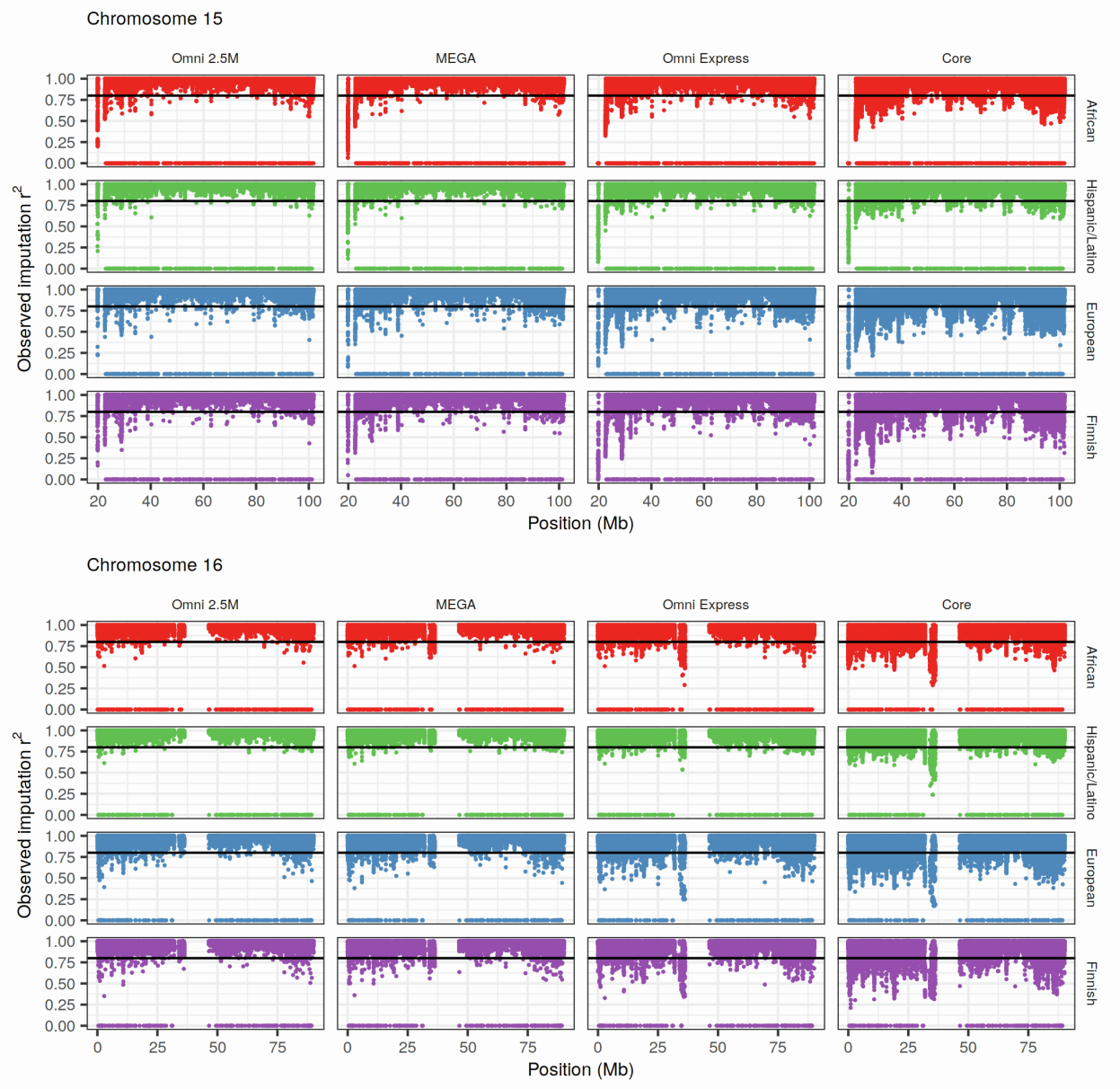

Chromosome 7
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
1.00
0.75
0.50
0.25
0.00
0
50
100
150
Position (Mb)

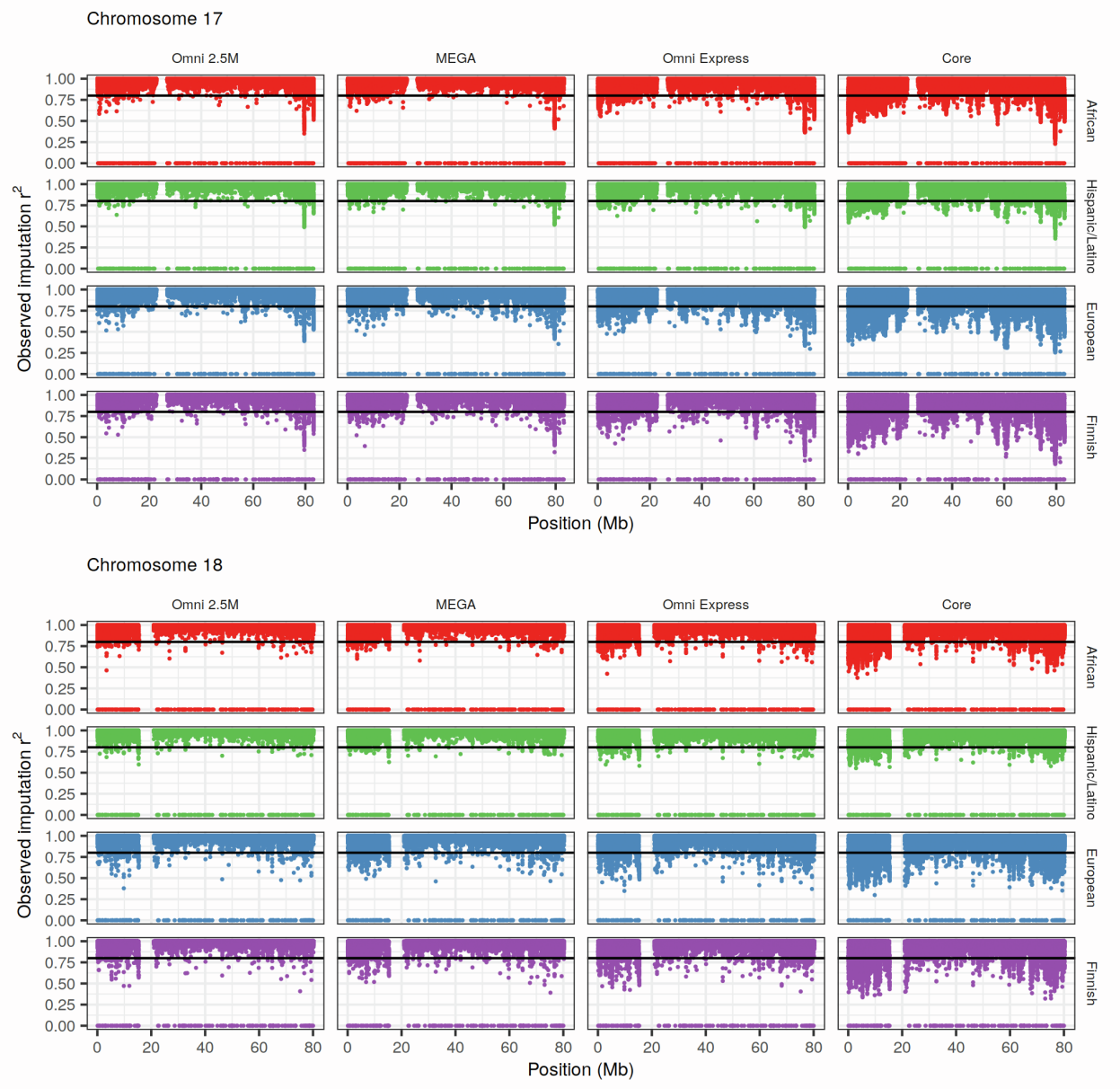

Chromosome 8
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
1.00
0.75
0.50
0.25
0.00
0
50
100
150
Position (Mb)

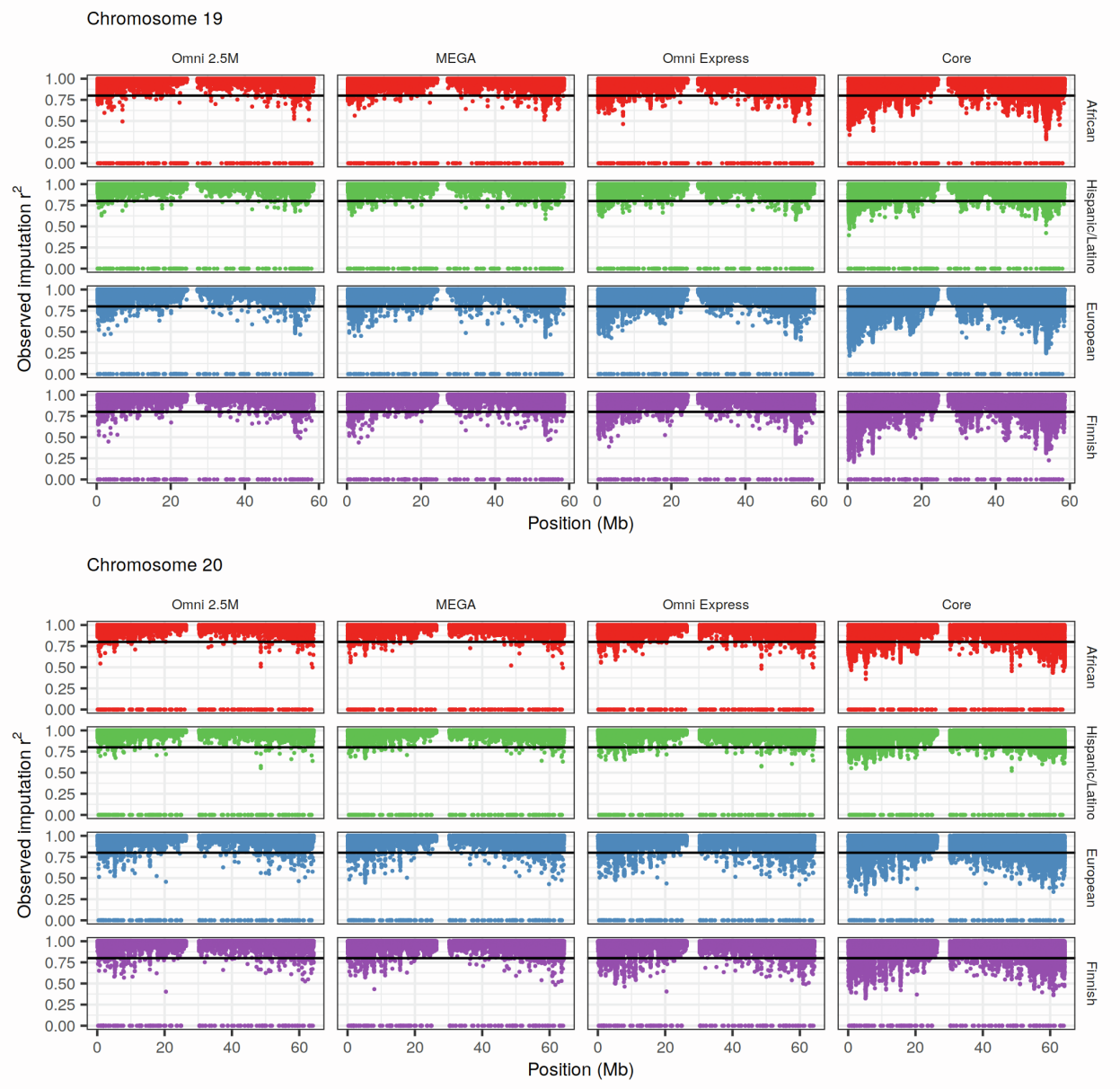

Chromosome 9
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
1.00
0.75
0.50
0.25
0.00
0
50
100
Position (Mb)
Chromosome 10
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
Position (Mb)


Chromosome 11
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
1.00
0.75
0.50
0.25
0.00
0
50
100
Position (Mb)
Chromosome 12
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
Position (Mb)


Chromosome 13
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
1.00
0.75
0.50
0.25
0.00
30
60
90
Position (Mb)
Chromosome 14
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
Position (Mb)


Chromosome 15
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
1.00
0.75
0.50
0.25
0.00
20
40
60
80
100
Position (Mb)
Chromosome 16
0
25
50
75


Chromosome 17
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Position (Mb)
Chromosome 18
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
Position (Mb)


Chromosome 19
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
1.00
0.75
0.50
0.25
0.00
0
20
40
60
Position (Mb)
Chromosome 20
Omni 2.5M
MEGA
Omni Express
Core
African
Hispanic/Latino
European
Finnish
Observed imputation r2
Position (Mb)

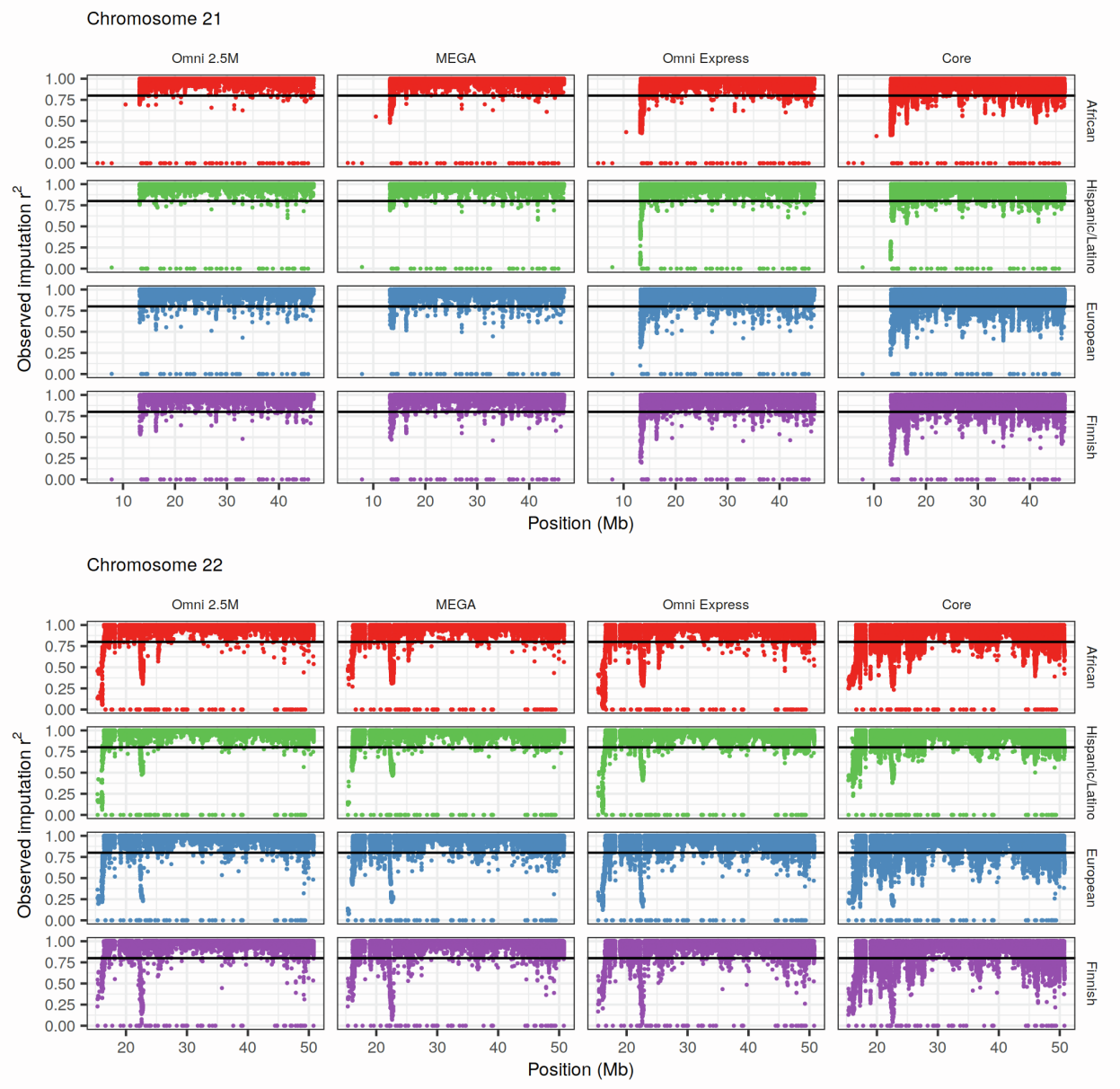


Figure S8. Regional variability in imputation quality of common variants with the TOPMed reference panel by genotyping array and ancestry across all chromosomes. Observed imputation r^2 by genomic position (Mb) for common (MAF>0.05) biallelic SNVs across all chromosomes by genotyping array (columns) and ancestry (rows). Variants above the horizontal black lines are well-imputed (observed imputation r^2>0.08).

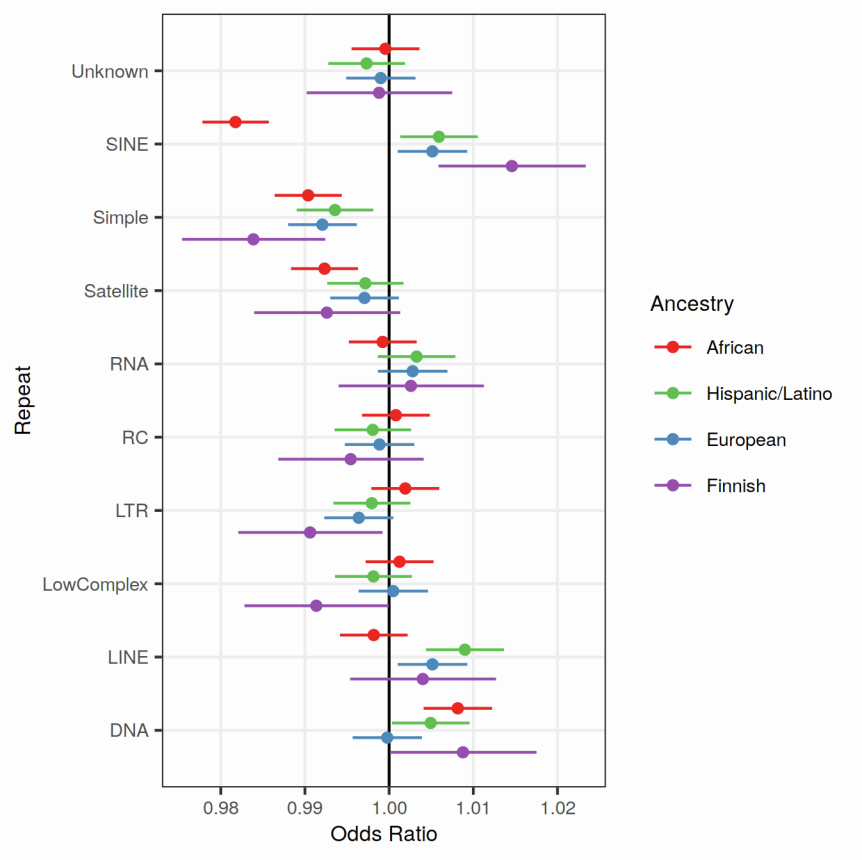


Figure S9. Repeat classes associated with TOPMed imputation quality of biallelic SNVs by ancestry. The odds ratios and corresponding 95% confidence intervals from logistic regression models. Estimates are from separate models testing the associations between each repeat class and whether or not a variant is well-imputed (observed imputation r2>0.8) adjusting for variant MAF. Repeat classes as defined by RepeatMasker include DNA repeat elements (DNA), long interspersed repeated elements (LINE), low complexity repeats (LowComplex), long terminal repeat elements including retrotransposons (LTR), rolling circle repeats (RC), RNA repeats (RNA), satellite repeats, microsatellites (Simple), short interspersed repeat elements including ALUs (SINE), and repeats of unknown class.

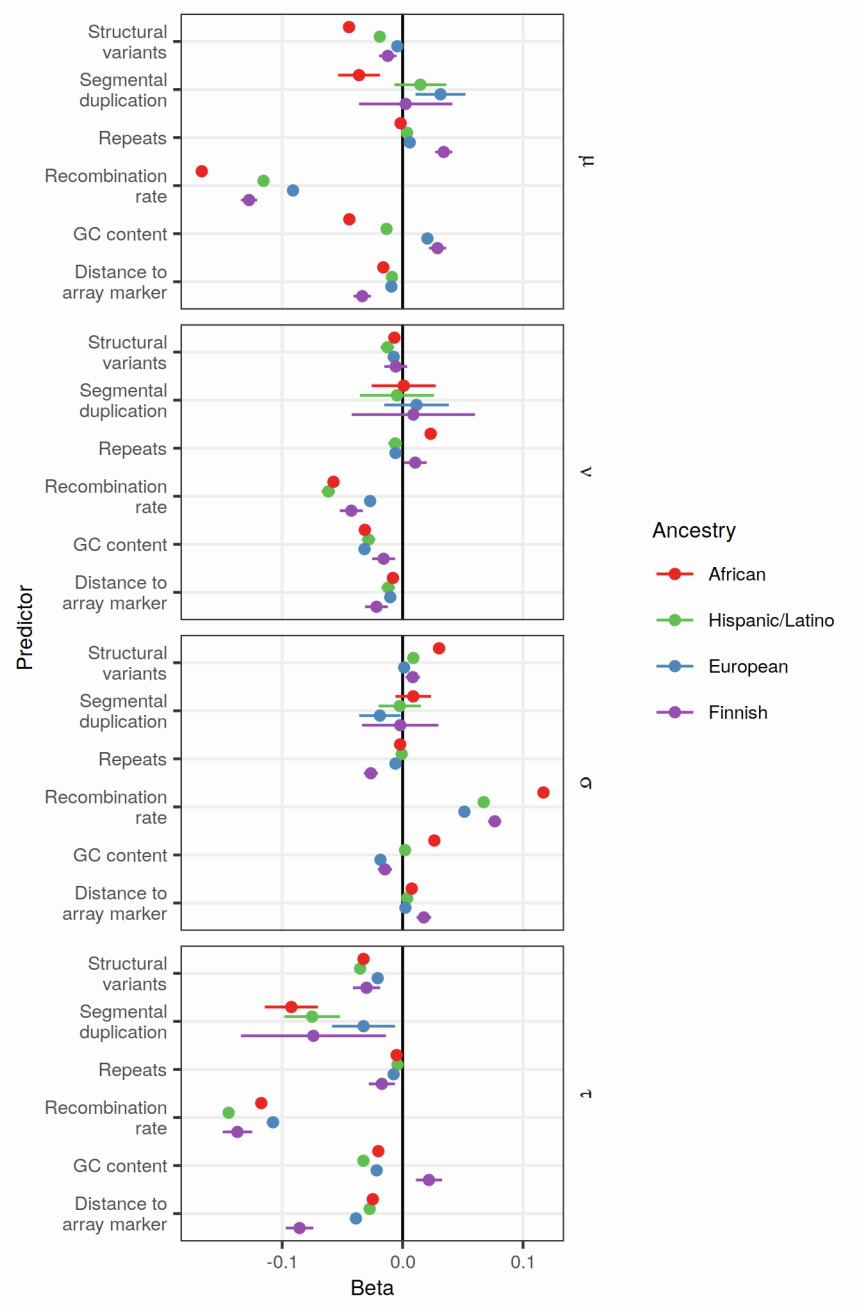


Figure S10. Genomic features associated with TOPMed imputation quality of biallelic SNVs by ancestry. The odds ratios and corresponding 95% confidence intervals from zero-one inflated beta regression models testing the association of genomic features with the observed imputation r^2 in the open interval $0<r^2<1$ (mean μ and variance-related parameter σ) and the probabilities of observed imputation $r^2=0$ (ν) or $r^2=1$ (τ). Estimates are from separate models testing the associations between characteristics of regional genomic features and imputation quality (observed imputation r^2) adjusting for variant MAF.

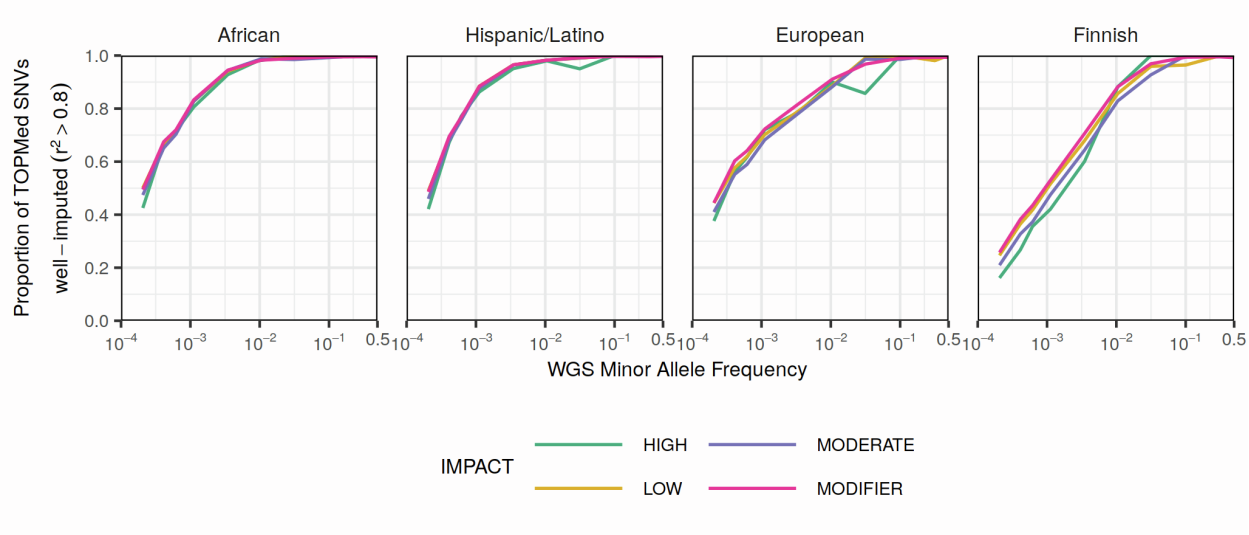


Figure S11. Proportion of well-imputed (r^2>0.8) biallelic SNVs by predicted functional impact and ancestry. The predicted functional impact of all sequenced biallelic SNVs was determined with VEP. The x-axes show minor allele frequency (MAF) calculated separately by study. Biallelic SNVs were then aggregated by MAF bins of width 0.00025 MAF for MAF between 0.0002 and 0.002 and of size 0.001 MAF for MAF > 0.002; those plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.001, 0.0032, 0.01, 0.032, 0.1, 0.32, and 0.5.

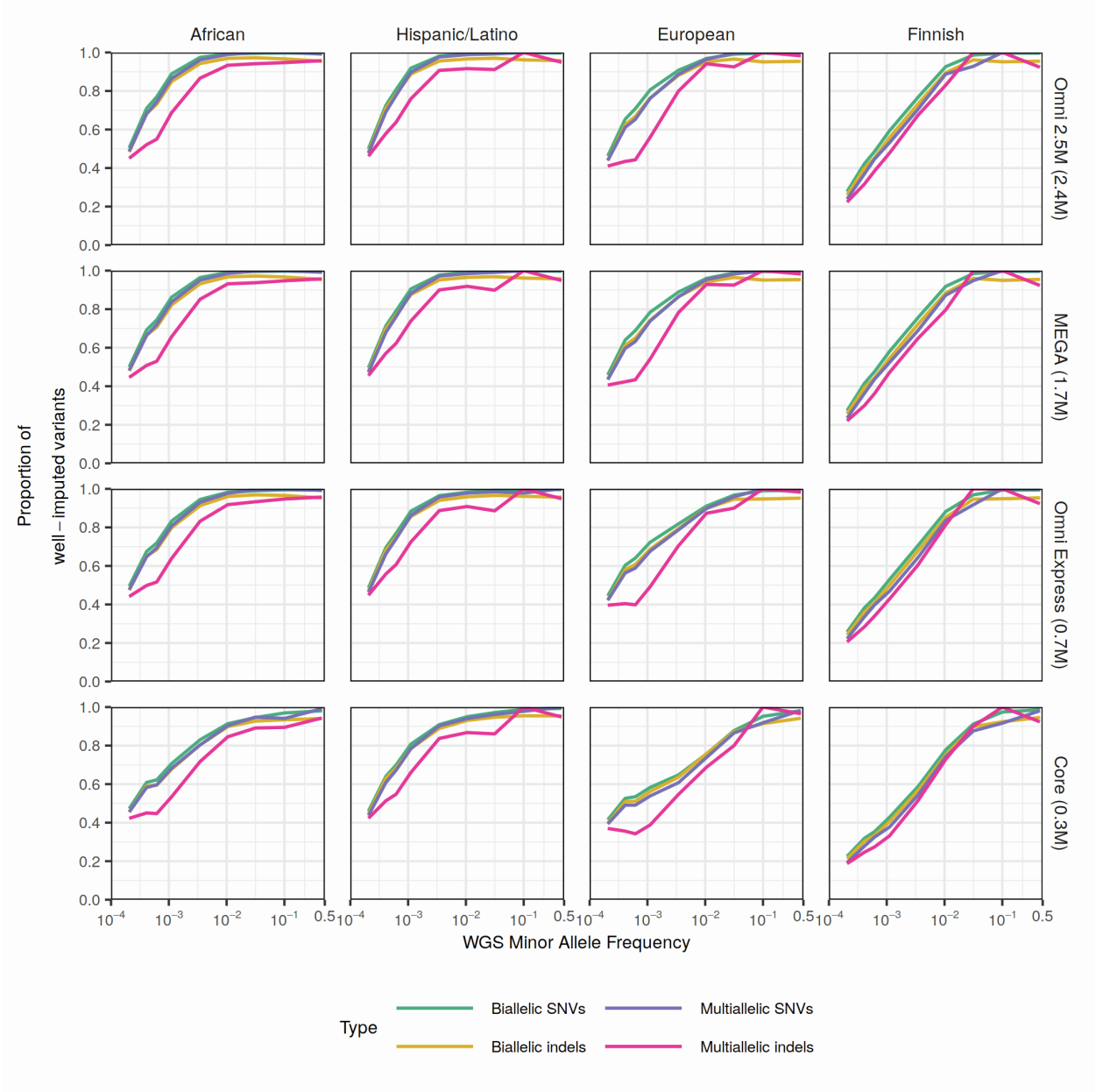


Figure S12. Proportion of well-imputed (r^2>0.8) variants by variant type, genotyping array, and ancestry with the TOPMed panel. The proportion of sequenced variants that are well-imputed by genotyping array (rows) and ancestry (columns). X-axes show minor allele frequency (MAF) calculated separately in each study. Sequenced variants not present in reference panels were assigned r^2=0. Variants were then aggregated by MAF bins of width 0.00025 MAF for MAF between 0.0002 and 0.002, bins of width 0.001 MAF for MAF between 0.002 and 0.4, and one bin of width 0.1 MAF for MAF between 0.4 and 0.5. MAF bins plotted here correspond to singletons, doubletons, and tripletons in each study, as well as those with mean MAF closest to the values 0.01, 0.0032, 0.01, 0.316, 0.1, and 0.5.

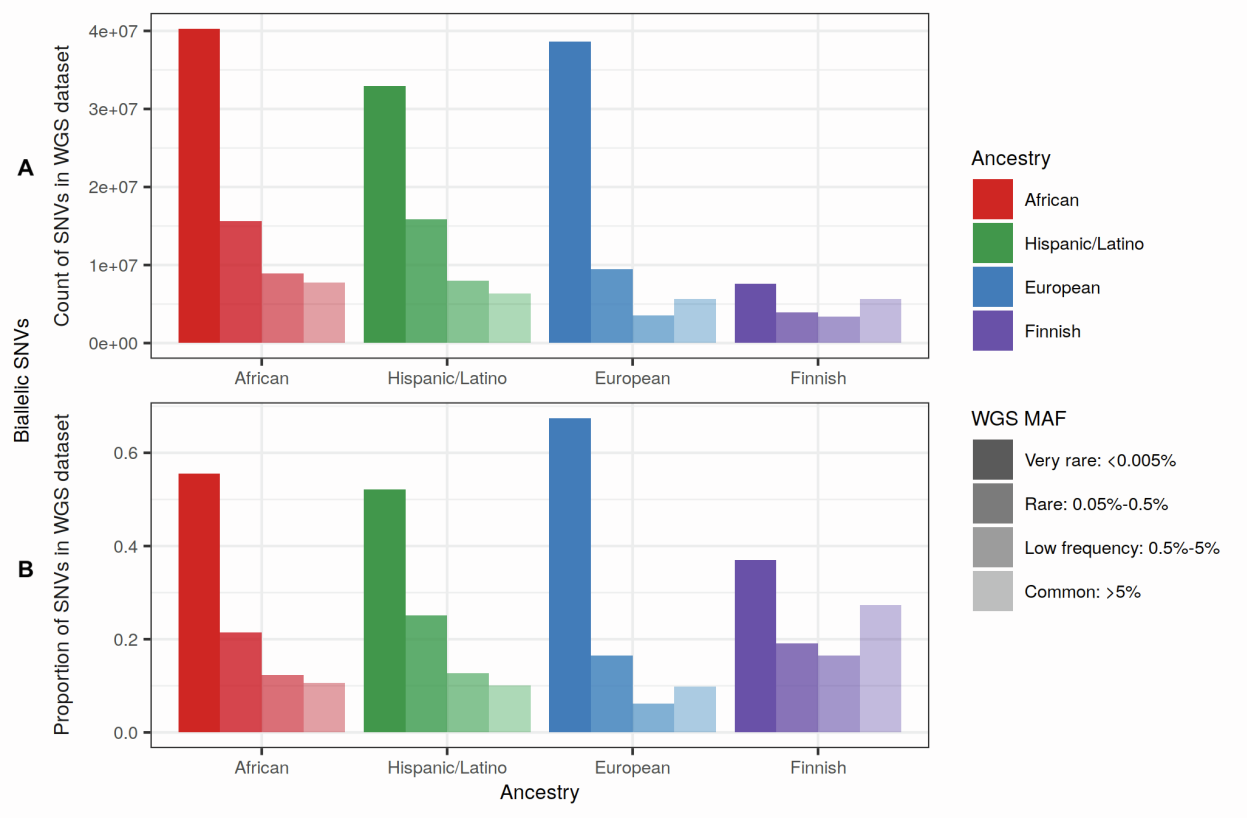


Figure S13. Distribution of MAF for biallelic SNVs by ancestry. A. Barplots of the number of biallelic SNVs in each MAF category for each WGS dataset. B. Barplots of the proportion of biallelic SNVs in each MAF category for each WGS dataset.

Supplemental Tables

Array	Number of variants	African	Hispanic/ Latino	European	Finnish
Omni 2.5M	2,381,000	2,132,501	2,330,998	2,330,998	2,264,709
MEGA	1,780,000	1,415,237	1,759,171	1,759,171	1,676,050
OmniExpress	710,000	680,234	706,652	706,652	698,865
Core	307,000	266,727	288,599	288,599	302,423

Table S1. Whole genome sequencing (WGS)-based genotype arrays. The numbers of variants included on the Illumina arrays and the actual number of WGS variants in each study used to create the WGS-based arrays.

Reference panel	Array	MAF	African	Hispanic/ Latino	European	Finnish
TOPMed	Omni 2.5M	Common	7.7M	6.3M	5.6M	5.6M
		Low frequency	8.9M	8.0M	3.4M	3.2M
		Rare	35.6M	32.4M	26.5M	4.8M
	MEGA	Common	7.7M	6.3M	5.6M	5.6M
		Low frequency	8.9M	8.0M	3.4M	3.2M
		Rare	35.0M	32.0M	26.1M	4.7M
	OmniExpress	Common	7.7M	6.3M	5.6M	5.6M
		Low frequency	8.8M	7.9M	3.3M	3.1M
		Rare	34.2M	31.4M	24.9M	4.4M
	Core	Common	7.5M	6.3M	5.5M	5.5M
		Low frequency	8.2M	7.7M	2.8M	2.8M
		Rare	31.2M	29.2M	22.2M	3.7M
HRC	Omni 2.5M	Common	7.1M	5.9M	5.2M	5.2M
		Low frequency	6.0M	6.2M	2.9M	3.1M
		Rare	4.0M	5.1M	9.4M	3.6M
	MEGA	Common	6.7M	5.8M	5.2M	5.2M
		Low frequency	4.9M	5.4M	2.8M	3.1M
		Rare	3.6M	4.3M	8.6M	3.6M
	OmniExpress	Common	6.5M	5.7M	5.2M	5.2M
		Low frequency	4.1M	4.7M	2.5M	3.1M
		Rare	3.1M	3.7M	7.8M	3.4M
	Core	Common	4.7M	5.0M	4.9M	5.2M
		Low frequency	1.9M	2.7M	1.9M	3.0M
		Rare	2.0M	2.3M	5.7M	3.1M
1000G	Omni 2.5M	Common	7.5M	6.2M	5.5M	5.5M
		Low frequency	7.2M	6.6M	2.4M	2.6M
		Rare	4.4M	6.5M	7.0M	1.7M
	MEGA	Common	7.2M	6.1M	5.4M	5.5M
		Low frequency	6.1M	6.0M	2.3M	2.5M
		Rare	3.5M	5.4M	6.3M	1.6M
	OmniExpress	Common	6.9M	6.0M	5.3M	5.4M
		Low frequency	5.3M	5.4M	2.0M	2.4M
		Rare	2.9M	4.8M	5.6M	1.4M
	Core	Common	5.4M	5.4M	5.0M	5.2M
		Low frequency	2.6M	3.3M	1.4M	2.0M
		Rare	1.4M	2.9M	3.7M	1.1M

Table S2. Number of well-imputed biallelic single nucleotide variants (SNVs) in each whole genome sequencing (WGS) study by reference panel, genotype array, ancestry, and minor allele frequency (MAF) category.

Reference panel	Array	MAF	African	Hispanic/ Latino	European	Finnish
TOPMed	Omni 2.5M	Common	0.997	0.997	0.996	0.996
		Low frequency	0.993	0.992	0.974	0.945
		Rare	0.637	0.664	0.552	0.415
	MEGA	Common	0.997	0.997	0.996	0.996
		Low frequency	0.992	0.992	0.967	0.939
		Rare	0.626	0.656	0.543	0.408
	OmniExpress	Common	0.994	0.996	0.992	0.993
		Low frequency	0.984	0.985	0.927	0.913
		Rare	0.613	0.642	0.517	0.379
	Core	Common	0.973	0.990	0.969	0.978
		Low frequency	0.922	0.954	0.800	0.830
		Rare	0.559	0.598	0.461	0.318
HRC	Omni 2.5M	Common	0.921	0.926	0.929	0.933
		Low frequency	0.668	0.772	0.812	0.908
		Rare	0.071	0.104	0.195	0.314
	MEGA	Common	0.871	0.914	0.926	0.933
		Low frequency	0.546	0.679	0.784	0.907
		Rare	0.065	0.088	0.180	0.310
	OmniExpress	Common	0.834	0.894	0.917	0.932
		Low frequency	0.463	0.591	0.701	0.901
		Rare	0.055	0.076	0.162	0.298
	Core	Common	0.609	0.792	0.875	0.931
		Low frequency	0.208	0.338	0.539	0.886
		Rare	0.036	0.047	0.119	0.273
1000G	Omni 2.5M	Common	0.970	0.976	0.974	0.977
		Low frequency	0.801	0.828	0.692	0.760
		Rare	0.079	0.134	0.145	0.150
	MEGA	Common	0.936	0.965	0.965	0.974
		Low frequency	0.679	0.752	0.658	0.745
		Rare	0.063	0.110	0.131	0.142
	OmniExpress	Common	0.895	0.946	0.948	0.966
		Low frequency	0.590	0.667	0.559	0.697
		Rare	0.052	0.098	0.116	0.125
	Core	Common	0.691	0.851	0.880	0.930
		Low frequency	0.286	0.409	0.401	0.590
		Rare	0.025	0.058	0.077	0.091

Table S3. Proportion of biallelic single nucleotide variants (SNVs) in each whole genome sequencing (WGS) study that are well-imputed (r^2>0.8) by reference panel, genotype array, ancestry, and minor allele frequency (MAF) category.

Reference panel	Array	African	Hispanic/ Latino	European	Finnish
TOPMed	Omni 2.5M	0.0014	0.0011	0.0035	0.0084
	MEGA	0.0016	0.0011	0.0045	0.0095
	OmniExpress	0.0024	0.0014	0.0095	0.0126
	Core	0.0084	0.0035	0.0395	0.0275
HRC	Omni 2.5M	0.0485	0.0364	0.0276	0.0115
	MEGA	0.3065	0.0565	0.0346	0.0115
	OmniExpress	NA	0.1055	0.0585	0.0135
	Core	NA	NA	0.2015	0.0154
1000G	Omni 2.5M	0.0245	0.0235	0.0385	0.0325
	MEGA	0.0665	0.0364	0.0455	0.0365
	OmniExpress	0.1395	0.0675	0.0705	0.0515
	Core	NA	0.2225	0.1704	0.0945

Table S4. Minor allele frequency (MAF) threshold above which array genotyping and imputation can approximate whole genome sequencing (WGS) for biallelic single nucleotide variants (SNVs) by reference panel, genotype array, and ancestry. Threshold is the smallest MAF for which >90% of biallelic SNVs are well-imputed (observed imputation r^2>0.8).

	Array	MAF	African				Hispanic/ Latino			European	Finnish
			All	0.25-0.5	0.5-0.75	0.75-1.0	All	NC	C	All	All
TOPMed	Omni 2.5M	Common	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		Low frequency	0.99	0.98	0.99	0.99	0.99	0.99	0.97	0.98	0.97
		Rare	0.93	0.89	0.93	0.93	0.93	0.79	0.96	0.86	0.82
	MEGA	Common	0.92	0.92	0.92	0.92	0.96	0.97	0.96	0.99	1.00
		Low frequency	0.98	0.97	0.99	0.99	0.99	0.99	0.96	0.97	0.97
		Rare	0.91	0.87	0.91	0.91	0.92	0.78	0.95	0.84	0.81
	Omni Express	Common	0.92	0.92	0.92	0.92	0.96	0.97	0.96	0.99	1.00
		Low frequency	0.98	0.97	0.99	0.99	0.99	0.99	0.96	0.97	0.97
		Rare	0.91	0.86	0.91	0.91	0.91	0.75	0.95	0.82	0.78
	Core	Common	0.98	0.97	0.98	0.98	0.97	0.97	0.96	0.98	0.99
		Low frequency	0.98	0.97	0.99	0.99	0.99	0.99	0.96	0.97	0.97
		Rare	0.86	0.80	0.86	0.86	0.88	0.68	0.92	0.76	0.71
HRC	Omni 2.5M	Common	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	1.00
		Low frequency	0.91	0.91	0.91	0.91	0.92	0.92	0.90	0.94	0.97
		Rare	0.70	0.73	0.72	0.69	0.70	0.67	0.71	0.70	0.82
	MEGA	Common	0.96	0.97	0.96	0.96	0.98	0.98	0.98	0.99	1.00
		Low frequency	0.81	0.82	0.82	0.81	0.89	0.89	0.88	0.93	0.97
		Rare	0.65	0.71	0.69	0.64	0.65	0.63	0.66	0.68	0.81
	Omni Express	Common	0.95	0.96	0.96	0.95	0.98	0.98	0.98	0.98	1.00
		Low frequency	0.83	0.85	0.84	0.83	0.84	0.85	0.82	0.90	0.97
		Rare	0.62	0.67	0.66	0.61	0.60	0.57	0.61	0.65	0.80
	Core	Common	0.89	0.92	0.90	0.89	0.95	0.95	0.95	0.97	0.99
		Low frequency	0.68	0.73	0.70	0.67	0.73	0.73	0.70	0.84	0.95
		Rare	0.50	0.58	0.55	0.49	0.51	0.47	0.51	0.55	0.77
1000G	Omni 2.5M	Common	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.97	0.99
		Low frequency	0.93	0.92	0.93	0.93	0.91	0.92	0.89	0.88	0.90
		Rare	0.73	0.69	0.71	0.73	0.73	0.66	0.75	0.62	0.71
	MEGA	Common	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.97	0.99
		Low frequency	0.88	0.87	0.88	0.88	0.89	0.90	0.87	0.86	0.91
		Rare	0.65	0.64	0.65	0.65	0.68	0.61	0.69	0.58	0.68
	Omni Express	Common	0.96	0.96	0.96	0.96	0.98	0.98	0.98	0.96	0.99
		Low frequency	0.85	0.85	0.85	0.85	0.86	0.87	0.82	0.81	0.88
		Rare	0.61	0.58	0.61	0.62	0.64	0.56	0.66	0.54	0.63
	Core	Common	0.90	0.92	0.91	0.90	0.95	0.95	0.95	0.94	0.97
		Low frequency	0.70	0.72	0.71	0.70	0.74	0.75	0.69	0.69	0.78
		Rare	0.44	0.44	0.45	0.44	0.51	0.42	0.53	0.40	0.52

Table S5. Mean heterozygous concordance rates by reference panel, genotype array, ancestry, and MAF category. Summary statistics are further broken down for the African ancestry study by estimated proportion of African ancestry (0.25-0.5, 0.5-0.75, 0.75-1.00) and for the Hispanic/Latino ancestry study by Caribbean (C) and non-Caribbean (NC) origin.

	Array	MAF	Number of consecutively well-imputed (r2>0.8) biallelic SNVs											
			African			Hispanic/ Latino			European			Finnish		
			25th	50th	75th	25th	50th	75th	25th	50th	75th	25th	50th	75th
TOPMed	Omni 2.5M	Common	41	277	750	52	295	777	33	197	576	35	210	592
		Low frequency	9	85	287	18	66	186	4	12	41	4	11	25
		Rare	1	2	4	1	2	4	1	2	3	1	1	2
	MEGA	Common	21	243	715	41	276	753	16	139	473	17	157	505
		Low frequency	5	45	205	14	57	166	3	9	30	4	10	23
		Rare	1	2	4	1	2	4	1	2	3	1	1	2
	Omni Expres	Common	4	106	512	17	193	616	7	56	267	8	72	328
		Low frequency	3	15	98	7	30	92	2	6	16	3	7	16
		Rare	1	2	3	1	2	4	1	2	3	1	1	2
	Core	Common	1	3	20	2	17	194	2	10	46	2	11	61
		Low frequency	1	4	13	2	8	28	1	3	7	2	4	9
		Rare	1	2	3	1	2	3	1	1	2	1	1	2
HRC	Omni 2.5M	Common	2	9	28	3	13	33	4	14	35	4	15	38
		Low frequency	1	2	5	1	3	7	2	4	8	3	9	19
		Rare	1	1	1	1	1	1	1	1	1	1	1	2
	MEGA	Common	2	4	15	3	10	28	4	14	33	4	15	38
		Low frequency	1	2	3	1	2	5	1	3	7	3	8	18
		Rare	1	1	1	1	1	1	1	1	1	1	1	2
	Omni Expres	Common	1	4	12	2	8	22	3	12	29	4	15	38
		Low frequency	1	2	3	1	2	4	1	2	5	3	8	17
		Rare	1	1	1	1	1	1	1	1	1	1	1	2
	Core	Common	1	2	5	1	3	10	2	7	19	4	15	37
		Low frequency	1	1	2	1	1	3	1	2	4	3	6	15
		Rare	1	1	1	1	1	1	1	1	1	1	1	2
1000G	Omni 2.5M	Common	3	15	68	5	27	85	6	25	72	7	30	88
		Low frequency	2	4	8	2	4	9	1	2	5	1	3	6
		Rare	1	1	1	1	1	1	1	1	1	1	1	1
	MEGA	Common	2	4	20	3	12	54	4	16	53	5	24	74
		Low frequency	1	2	5	1	3	6	1	2	4	1	3	5
		Rare	1	1	1	1	1	1	1	1	1	1	1	1
	Omni Expres	Common	1	4	14	2	9	37	3	13	36	4	18	55
		Low frequency	1	2	4	1	2	5	1	2	3	1	2	5
		Rare	1	1	1	1	1	1	1	1	1	1	1	1
	Core	Common	1	2	5	1	3	11	2	5	17	2	8	27
		Low frequency	1	1	2	1	2	3	1	1	3	1	2	4
		Rare	1	1	1	1	1	1	1	1	1	1	1	1

Table S6. 25th, 50th, and 75th percentiles of the number of consecutive well-imputed (observed imputation r^2>0.8) biallelic single nucleotide variants (SNVs) by reference panel, genotype array, ancestry, and minor allele frequency (MAF) category.

	Array	MAF	Length in kb of consecutively well-imputed (r2>0.8) biallelic SNVs											
			African			Hispanic/ Latino			European			Finnish		
			25th	50th	75th	25th	50th	75th	25th	50th	75th	25th	50th	75th
TOPMed	Omni 2.5M	Common	10.4	84.8	253.2	15.7	109.6	315.3	11.6	80.1	256.5	12.5	87.2	267.4
		Low frequency	2.3	23.9	84.3	5.2	20.7	60.1	1.6	7.8	29.8	2.3	8.0	19.2
		Rare	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.2
	MEGA	Common	5.1	73.3	241.6	11.4	101.3	302.2	5.3	55.7	210.5	6.3	64.9	224.4
		Low frequency	1.0	12.4	59.8	4.0	17.3	53.5	1.1	5.7	21.4	1.9	7.1	17.4
		Rare	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.2
	Omni Express	Common	0.7	31.3	162.8	4.6	70.4	245.3	2.5	23.9	113.6	3.2	31.1	144.6
		Low frequency	0.3	3.8	27.9	1.7	9.2	29.7	0.5	3.4	11.2	1.0	4.6	12.0
		Rare	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.2
	Core	Common	0.0	0.6	5.9	0.3	6.2	73.1	0.3	4.1	21.2	0.4	5.0	28.1
		Low frequency	0.0	0.7	3.4	0.3	2.1	8.9	0.0	1.2	4.5	0.1	1.9	6.0
		Rare	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1
HRC	Omni 2.5M	Common	0.2	2.3	8.5	0.6	4.0	12.3	0.8	5.0	14.9	0.8	5.3	16.2
		Low frequency	0.0	0.3	1.0	0.0	0.6	1.9	0.1	1.7	5.2	1.4	5.6	13.8
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	MEGA	Common	0.0	0.9	4.3	0.3	2.9	10.1	0.7	4.7	14.2	0.8	5.3	16.2
		Low frequency	0.0	0.1	0.7	0.0	0.3	1.2	0.0	1.4	4.4	1.3	5.5	13.5
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	Omni Express	Common	0.0	0.7	3.4	0.2	2.2	8.2	0.7	4.2	12.4	0.8	5.3	16.1
		Low frequency	0.0	0.0	0.5	0.0	0.2	1.0	0.0	0.8	2.9	1.1	5.1	12.7
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	Core	Common	0.0	0.2	1.1	0.0	0.8	3.6	0.3	2.4	8.1	0.8	5.2	15.9
		Low frequency	0.0	0.0	0.2	0.0	0.0	0.5	0.0	0.3	1.7	0.8	4.0	10.6
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1000G	Omni 2.5M	Common	0.3	4.0	21.2	1.2	9.7	33.4	1.9	10.6	31.6	2.1	12.9	40.0
		Low frequency	0.0	0.6	2.0	0.1	0.9	2.6	0.0	0.7	2.5	0.0	1.2	3.7
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MEGA	Common	0.0	1.0	6.0	0.4	4.2	20.4	0.9	6.7	23.2	1.5	9.9	33.1
		Low frequency	0.0	0.3	1.1	0.0	0.5	1.7	0.0	0.5	2.2	0.0	1.1	3.4
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Omni Express	Common	0.0	0.7	4.0	0.3	3.0	14.3	0.9	5.4	16.5	1.3	8.1	25.0
		Low frequency	0.0	0.2	0.9	0.0	0.3	1.3	0.0	0.2	1.5	0.0	0.8	2.7
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Core	Common	0.0	0.2	1.3	0.0	0.8	4.2	0.1	1.9	7.6	0.4	3.4	12.5
		Low frequency	0.0	0.0	0.4	0.0	0.0	0.6	0.0	0.0	1.0	0.0	0.3	1.9
		Rare	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table S7. 25th, 50th, and 75th percentiles of the length in kilobases (kb) of consecutively well-imputed (observed imputation r^2>0.8) variants by reference panel, genotype array, ancestry, and minor allele frequency (MAF) category.

Table S8. Associations of genomic features with dichotomous TOPMed imputation quality status from logistic regressions by genotype array and ancestry. MAF refers to the model with 9 MAF categories as predictors. ALL refers to the model with 9 MAF categories and all 6 genomic features as predictors. All other models refer to a model with 9 MAF categories and the named genomic feature as predictors. The features include mean GC content (GC), number of repeats (REP), number of structural variants (SV), presence of segmental duplications (SEG), mean recombination rate (RECOMB), and distance to nearest genotyped marker (DIST).

Table S9. Associations of repeats with dichotomous TOPMed imputation quality status from logistic regressions by genotype array and ancestry. All model names refer to the model with 9 MAF categories and membership in the named repeat class as predictors. Repeat classes as defined by RepeatMasker include DNA repeat elements (DNA), long interspersed repeated elements (LINE), low complexity repeats (LowComplex), long terminal repeat elements including retrotransposons (LTR), rolling circle repeats (RC), RNA repeats (RNA), satellite repeats, microsatellites (Simple), short interspersed repeat elements including ALUs (SINE), and repeats of unknown class.

Table S10. Associations of genomic features with continuous TOPMed imputation quality status from zero-one inflated beta regressions by genotype array and ancestry. All models used the same set of predictors for each of the four parameters (see methods). MAF refers to the model with 9 MAF categories as predictors. ALL refers to the model with 9 MAF categories and all 6 genomic features as predictors. All other models refer to a model with 9 MAF categories and the named genomic feature as predictors. The features include mean GC content (GC), number of repeats (REP), number of structural variants (SV), presence of segmental duplications (SEG), mean recombination rate (RECOMB), and distance to nearest genotyped marker (DIST).

Reference panel	Array	Impact	African	Hispanic/ Latino	European	Finnish
TOPMed	Omni 2.5M	High	0.605	0.641	0.511	0.456
		Moderate	0.645	0.668	0.540	0.510
		Low	0.706	0.726	0.603	0.620
		Modifier	0.719	0.739	0.621	0.662
	MEGA	High	0.551	0.593	0.443	0.387
		Moderate	0.571	0.608	0.448	0.416
		Low	0.619	0.658	0.499	0.521
		Modifier	0.648	0.683	0.532	0.584
	OmniExpress	High	0.589	0.622	0.483	0.430
		Moderate	0.626	0.651	0.508	0.480
		Low	0.684	0.707	0.568	0.589
		Modifier	0.699	0.721	0.589	0.635
	Core	High	0.551	0.593	0.443	0.387
		Moderate	0.571	0.608	0.448	0.416
		Low	0.619	0.658	0.499	0.521
		Modifier	0.648	0.683	0.532	0.584
HRC	Omni 2.5M	High	0.151	0.181	0.191	0.392
		Moderate	0.163	0.191	0.204	0.442
		Low	0.222	0.259	0.280	0.555
		Modifier	0.235	0.271	0.305	0.582
	MEGA	High	0.173	0.210	0.229	0.420
		Moderate	0.209	0.251	0.268	0.485
		Low	0.202	0.238	0.271	0.555
		Modifier	0.210	0.246	0.290	0.579
	OmniExpress	High	0.123	0.147	0.166	0.382
		Moderate	0.129	0.154	0.175	0.430
		Low	0.174	0.207	0.243	0.542
		Modifier	0.188	0.224	0.270	0.571
	Core	High	0.081	0.107	0.139	0.368
		Moderate	0.074	0.099	0.133	0.406
		Low	0.099	0.135	0.189	0.518
		Modifier	0.119	0.159	0.220	0.554
1000G	Omni 2.5M	High	0.162	0.200	0.148	0.282
		Moderate	0.170	0.208	0.152	0.314
		Low	0.240	0.284	0.231	0.437
		Modifier	0.263	0.307	0.261	0.478
	MEGA	High	0.187	0.233	0.208	0.351
		Moderate	0.225	0.276	0.250	0.432
		Low	0.216	0.262	0.224	0.438
		Modifier	0.232	0.277	0.246	0.469
	OmniExpress	High	0.128	0.165	0.125	0.257
		Moderate	0.131	0.170	0.126	0.286
		Low	0.186	0.232	0.194	0.405
		Modifier	0.208	0.256	0.225	0.450
	Core	High	0.081	0.120	0.104	0.235
		Moderate	0.071	0.110	0.089	0.238
		Low	0.103	0.152	0.142	0.346
		Modifier	0.129	0.183	0.176	0.404

Table S11. Proportion of biallelic single nucleotide variants (SNVs) in each whole genome sequencing (WGS) study that are well-imputed (r^2>0.8) by reference panel, genotype array, ancestry, and predicted impact on protein coding. Predicted impact was estimated with VEP.

Array	Variant type	African	Hispanic/ Latino	European	Finnish
Omni 2.5M	Biallelic SNV	0.0014	0.0011	0.0035	0.0084
	Biallelic indel	0.0024	0.0014	0.0045	0.0115
	Multiallelic SNV	0.0016	0.0014	0.0045	0.0115
	Multiallelic indel	0.0055	0.0035	0.0075	0.0144
MEGA	Biallelic SNV	0.0017	0.0011	0.0045	0.0095
	Biallelic indel	0.0024	0.0014	0.0055	0.0126
	Multiallelic SNV	0.0024	0.0014	0.0055	0.0126
	Multiallelic indel	0.0065	0.0045	0.0105	0.0165
OmniExpress	Biallelic SNV	0.0024	0.0014	0.0095	0.0126
	Biallelic indel	0.0035	0.0019	0.0115	0.0165
	Multiallelic SNV	0.0035	0.0016	0.0115	0.0164
	Multiallelic indel	0.0074	0.0045	0.0145	0.0165
Core	Biallelic SNV	0.0084	0.0035	0.0395	0.0275
	Biallelic indel	0.0115	0.0045	0.0425	0.0336
	Multiallelic SNV	0.0105	0.0035	0.0405	0.0224
	Multiallelic indel	0.0185	0.0075	0.0284	0.0384

Table S12. Minor allele frequency (MAF) threshold above which array genotyping and imputation can approximate whole genome sequencing (WGS) with the TOPMed panel by genotype array, ancestry, and variant type. Threshold is the smallest MAF for which >90% of variants are well-imputed (observed imputation r^2>0.8).

ACKNOWLEDGEMENTS

Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Genome Sequencing for "NHLBI TOPMed: Whole Genome Sequencing in the Bio*Me* Study" (phs phs001644) was performed at the Baylor College of Medicine Human Genome Sequencing Center and the McDonnell Genome Institute. Genome Sequencing for "NHLBI TOPMed: Whole Genome Sequencing in MLOF Study" (phs001515) was performed at the Baylor College of Medicine Human Genome Sequencing Center and New York Genome Center Genomics. Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Administrative Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

The Mount Sinai Bio*Me* Biobank has been supported by The Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI (U01HG00638001; U01HG007417; X01HL134588). We thank all participants in the Mount Sinai Biobank. We also thank all our recruiters who have assisted and continue to assist in data collection and management and are grateful for the computational

resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

The My Life, Our Future samples and data are made possible through the partnership of Bloodworks Northwest, the American Thrombosis and Hemostasis Network, the National Hemophilia Foundation, and Bioverativ. We gratefully acknowledge the hemophilia treatment centers and their patients who provided biological samples and phenotypic data.

Bio-samples and data for the InPSYght study were obtained from NIMH Repository & Genomics Resource, a centralized national biorepository for genetic studies of psychiatric disorders. Contributing studies include the Genomic Psychiatry Cohort, the Consortium of the Genetics of Schizophrenia, the Biopolar Genome Study,the Lithium treatment moderate dose use study, and the Systematic Treatment Enhancement Program for Bipolar Disorder. We gratefully acknowledge the participants who provided biological samples and data for these studies.

We gratefully acknowledge the participants who provided biological samples and data for the METSIM study.

We gratefully acknowledge Hyun Min Kang for his guidance in using bed-diff to calculate genotype concordance metrics and other useful discussions.

We gratefully acknowledge Corbin Quick for his script and guidance in calculating the observed imputation r^2.

This work was supported by NHGRI grant R01 HG009976 (Boehnke). S.C.H was also supported by NHGRI grant F31 HG011186.