

**Imputation-Based Genomic Coverage Assessments of Current Genotyping Arrays:
Illumina HumanCore, OmniExpress, Multi-Ethnic global array and sub-arrays, Global
Screening Array, Omni2.5M, Omni5M, and Affymetrix UK Biobank**

Sarah C. Nelson¹, Jane M. Romm², Kimberly F. Doheny²,
Elizabeth W. Pugh², Cathy C. Laurie¹

1 - Genetic Analysis Center, Dept. of Biostatistics, University of Washington

2 - Center for Inherited Disease Research, Johns Hopkins University School of Medicine

I. Introduction

Genotyping arrays have been widely adopted as an efficient means to interrogate variation across the human genome. Genetic variants may be observed either directly, via genotyping, or indirectly, through linkage disequilibrium (LD) with a genotyped variant. The total proportion of genomic variation captured by an array, either directly or indirectly, is referred to as “genomic coverage.” Assessments of genomic coverage generally take one of two forms. (1) LD-based genomic coverage estimates describe the fraction of variants that are in LD with array variants at a given pairwise correlation value ($r^2 \geq 0.8$, e.g.)^{1,2}. (2) Imputation-based genomic coverage leverages array variants to impute into a more densely genotyped or sequenced reference panel, such as the HapMap Project³ or 1000 Genomes Project^{4,5}. In this approach, genomic coverage is quantified as the proportion of variants with an imputation r^2 above a given threshold (again, usually $r^2 \geq 0.8$)⁶.

In response to both the evolving needs of the genetic research community and technological advances in genotyping, new arrays from both Illumina (www.illumina.com) and Affymetrix (www.affymetrix.com) continue to be developed. While vendors often provide genomic coverage estimates in product documentation, the methods may not be comparable and/or well-described, making it difficult to objectively compare across arrays. Previously we published imputation-based genomic coverage for a series of arrays using Phase 1 of the 1000 Genomes Project as a reference⁷. Here we report an extension of those previous coverage analyses, using 1000 Genomes Phase 3⁵ and an updated set of arrays, including the Illumina Multi-Ethnic arrays and Global Screening Array. The original and present analyses are intended to function as practical guides to researchers planning genetic studies. Both coverage and accuracy information are presented.

Product literature from the manufacturers details the design and intended uses of the arrays we assess here. In brief, Illumina Multi-Ethnic arrays were designed using content from multiple sources, including Phase 3 of the 1000 Genomes Project, the Consortium on Asthma among African ancestry Populations in the Americas (CAAPA), and Population Architecture using Genomes and Epidemiology (PAGE). The Multi-Ethnic global array is intended for use in studies with diverse ancestry, while the sub-arrays have genome-wide tagging content tailored to certain populations. Specifically, the AFR/AMR sub-array is optimized for African American and Hispanic study samples, while the EUR/EAS/SAS sub-array is optimized for European, East Asian, and South Asian samples. For more details, see the array-specific data sheets available at <http://www.illumina.com/products/infinium-multi-ethnic-global-array.html>. The Illumina Global Screening Array (GSA) was developed by a consortium of human genetics researchers to function as

a cost-efficient GWAS array with additional clinically-relevant content (see <https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/infinium-commercial-gsa-data-sheet-370-2016-016.pdf>). The Affymetrix UK Biobank array is optimized for “populations of European and British ancestry” and indeed was initially designed for use in the UK Biobank (see http://www.affymetrix.com/support/technical/datasheets/uk_axiom_biobank_genotyping_arrays_datasheet.pdf). The Illumina HumanCore, OmniExpress, Omni2.5M and Omni5M are included as earlier-generation arrays to serve as reference points for coverage.

II. Materials and Methods

For these genomic coverage assessments, we used the publicly available 1000 Genomes Project Phase 3 integrated variant set, in variant call format (VCF, available at <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>). Samples are grouped into continental panels or “super-populations”⁵: African (AFR), Americas (AMR), East Asian (EAS), European (EUR), and South Asian (SAS) — see Table 1. Following our published methods⁷, we first randomly assigned each of the 2,504 Phase 3 samples into one of ten batches, balancing samples across populations. We then created subsets of the data in the 1000 Genomes VCF files, which are separated by chromosome. For each sample batch, we extracted genotypes from the VCF files for just the samples in the batch and the variants on the array. Array variants were identified within the VCF using chromosome and base pair positions from the most current array manifests available from vendor websites (see Table 2). A graphical overview of our methods is shown in Figure 1.

We removed the VCF phasing information when converting to PLINK binary format (via VCFtools⁸) to better approximate array genotypes, which are usually at first unphased. We then used SHAPEIT2⁹ (v2.r790) software to pre-phase each dataset. The resulting phased haplotypes were then imputed with IMPUTE2¹⁰ (v2.3.2) where, for each batch, the remaining 1000 Genomes samples served as a worldwide imputation reference panel. We restricted imputation to variants with at least two copies of the minor allele in any one of the five super-populations. Following imputation, the output (genotype probabilities file) for each of the ten sample batches was then combined. We restricted these experiments to chromosomes 1 and 22.

To assess imputation accuracy and by extension genomic coverage, we compared imputed results at all the non-array variants to observed genotypes from the original VCF files. This comparison was performed separately by super-population and restricted to variants with at least two copies of the minor allele in the given panel, yielding 2,701,990 variants in the AFR panel; 1,709,647 in AMR; 1,304,742 in EAS; 1,401,215 in EUR; and 1,556,550 in SAS (note counts are across chromosomes 1 and 22). This variant restriction was done to avoid an excess of missing imputation metrics, which occur when a variant is either imputed or observed to be monomorphic. Imputation included all three variant types in the 1000 Genomes Project release: single nucleotide, insertion/deletion, and structural variants, and thus all three variant types were included in these metrics calculations.

For each variant we calculated three metrics: (1) the squared correlation between observed and imputed allelic dosage, which we call “dosage r^2 ”; (2) the concordance between observed and most likely imputed genotype, the “genotype concordance”; and (3) the concordance between observed

and most likely imputed genotype, when at least one of those two genotypes contains one or two copies of the minor allele, which we call “minor allele (MA) concordance.” All metrics were calculated using R statistical and graphing software¹¹. Array (observed) variants are included in these metrics summaries and are given dosage r^2 , genotype concordance, and MA concordance values of 1.

III. Results and Discussion

Figure 2 shows the fraction of variants passing a dosage r^2 threshold of 0.8. Each quadrant represents a different minor allele frequency (MAF) bin, where MAF was calculated in each super-population. Genomic coverage assessments commonly use the MAF groupings of >1% and >5%, which we have shown in panels C and D, respectively. The position of arrays along the x-axis indicates the total number of unique positions assayed by each array. Thus the least dense array is to the far left (Illumina HumanCore) and the densest array to the far right (Illumina Omni5M). Figure 3 shows mean MA concordance, by MAF bin and super-population. Similarly Figure 4 shows mean dosage r^2 , and Figure 5 shows genotype concordance. All the data plotted in Figures 2 – 5 are presented numerically in Tables 3A-E, which also include tabular summaries of mean dosage r^2 and genotype concordance. Tables 3A-E also indicate the counts of variants in each MAF bin, as this count differs by super-population. The same metrics are also presented in barplots in Appendix A, for an alternative visualization.

As one might expect, quality metrics generally improve with increasing array density, regardless of either super-population or MAF bin. Rare variants (MAF < 1%) are generally not well imputed by either low or high density arrays. Common (MAF > 5%) variants are generally well imputed by all arrays, though there is still a distinction between the sparsest and densest arrays. Most of the line plots are monotonic. Some notable exceptions are the transitions between (1) OmniExpress and Affy UK Biobank in the AFR super-population and between (2) Affy UK Biobank and the Multi-Ethnic arrays in the EUR super-population.

The Multi-Ethnic global array and sub-arrays each assay a similar number of variants, though the global array has a slightly higher density. According to product literature, the Multi-Ethnic AMR/AFR array (“ME_amr” in plots) is designed for use in Hispanic and African American populations, while the Multi-Ethnic EUR/EAS/SAS array (“ME_eur” in plots) is designed for use in European, East Asian, and South Asian populations. Notably, according to these experiments, the Multi-Ethnic AMR/AFR array does not consistently outperform Multi-Ethnic EUR/EAS/SAS array for the AFR and AMR 1000 Genomes panels. However, this may be because populations in AMR and AFR are imperfect proxies for contemporary Hispanic and African American populations in the US. In contrast, for EUR, EAS, and SAS panels, the Multi-Ethnic EUR/EAS/SAS does perform slightly better than Multi-Ethnic AMR/AFR, which is consistent with the stated purpose of the arrays. While denser, the Multi-Ethnic global array does not consistently outperform the subarrays, though the relative performances vary by super-population and MAF bin.

The GSA results are generally consistent with what one would expect based on array densities: coverage is better than the less dense HumanCore but not as good as for the OmniExpress, the next densest assessed array after the GSA. One situation in which GSA outperforms the OmniExpress is for MAF 1-5% variants in the EUR and, to a lesser extent, the EAS super-populations. Notably, in

contrast to the mostly tagging-focused OmniExpress, the GSA devotes ~10% of its content to clinically-relevant markers. These clinically-relevant markers may be underperforming as an imputation basis but derive value from their clinical associations. There are several potential limitations to these analyses. First, by assessing genomic coverage in the context of genome-wide imputation, we can only assess array variants that are also present in the chosen reference panel (1000 Genomes Phase 3 integrated variant set). Thus not all array variants can inform the imputation, for one of two reasons: (1) the array variant is not present in this 1000 Genomes Project dataset or (2) the array variant is present but without two or more copies of the minor allele in any one of the five super-populations (the MAF filtering threshold for imputation). In Table 2 we show for each array the percent overlap with 1000 Genomes at any allele frequency and the percent overlap with the 1000 Genomes with the requisite minor allele count to be included in imputation. Notably, the Multi-Ethnic arrays have less overlap than the other arrays.

A second caveat is that we used chromosomes 1 and 22 to estimate genome-wide coverage. Above we present coverage for the two chromosomes combined. To assess whether coverage patterns were consistent across these two chromosomes, we also evaluated them separately. We observed a statistically significant difference between the two chromosomes for all four coverage metrics. However, the trends in coverage for the two chromosomes are quite consistent across arrays, for each of the super-populations and MAF bins. These patterns are illustrated in Figure 6. While the absolute coverage metrics may differ, the relative coverage is similar — evidence that supports the reliability of our approach here. Furthermore, it is not surprising that coverage may differ by chromosome, given that chromosomes are heterogeneous with respect to structure, gene density, and CG content¹². This heterogeneity of features may lead to differential destiny of genotyping assays on a SNP array and thus differences in coverage across chromosomes. Other factors, such as LD patterns, may also vary among chromosomes. Our primary goal here is to get a sense of how arrays compare to each other and to observe the trade-offs in assay density and genomic coverage. (Note GSA is absent from these chromosome-specific analyses, due to timing of coverage experiments. However, given the consistency of the chromosome-specific trends across the remaining arrays, it is reasonable to expect the same trends would hold for GSA.)

In summary, our current and previous⁷ imputation-based genomic coverage analyses are intended to help researchers weigh the costs and benefits of different array choices across different genetic ancestry groups. Ultimately, the genomic coverage afforded by a given array will likely be affected by many study-specific factors, such as genotyping quality, genetic ancestry profile of study samples, and imputation procedures. These findings, however, can serve as a robust starting point for researchers evaluating their genotyping array options.

References

1. Barrett, J.C. & Cardon, L.R. Evaluating coverage of genome-wide association studies. *Nat Genet* **38**, 659-62 (2006).
2. Li, C., Li, M., Long, J.R., Cai, Q. & Zheng, W. Evaluating cost efficiency of SNP chips in genome-wide association studies. *Genet Epidemiol* **32**, 387-95 (2008).
3. Frazer, K. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61 (2007).
4. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
5. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
6. Hoffmann, T.J. *et al.* Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* **98**, 422-30 (2011).
7. Nelson, S.C. *et al.* Imputation-Based Genomic Coverage Assessments of Current Human Genotyping Arrays. *G3-Genes Genomes Genetics* **3**, 1795-1807 (2013).
8. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8 (2011).
9. Delaneau, O., Zagury, J.F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6 (2013).
10. B. Howie, J.M., and M. Stephens. Genotype Imputation with Thousands of Genomes. *G3: Genes, Genomics, Genetics* **1**, 457-470 (2011).
11. R Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria, 2015).
12. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

Table 1. An overview of the 2,504 samples in the 1000 Genomes Project Phase 3 reference used for these coverage assessments. The Project assigned each population to one of five super-populations: African (AFR), Americas (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). Each sample was assigned to one of ten test batches at random, yet keeping an even representation of populations across batches.

Full Population Name	Abbreviation	Number of Samples	Number of samples per batch*
African Caribbean in Barbados	ACB	96	10
African Ancestry in Southwest US	ASW	61	6
Esan in Nigeria	ESN	99	10
Gambian in Western Division, The Gambia	GWD	113	11
Luhya in Webuye, Kenya	LWK	99	10
Mende in Sierra Leone	MSL	85	8
Yoruba in Ibadan, Nigeria	YRI	108	11
<i>Total African ancestry</i>	<i>AFR</i>	<i>661</i>	<i>66</i>
Colombian in Medellin, Colombia	CLM	94	9
Mexican ancestry in Los Angeles, California	MXL	64	6
Peruvian in Lima, Peru	PEL	85	8
Puerto Rican in Puerto Rico	PUR	104	10
<i>Total Americas ancestry</i>	<i>AMR</i>	<i>347</i>	<i>33</i>
Chinese Dai in Xishuangbanna, China	CDX	93	9
Han Chinese in Beijing, China	CHB	103	10
Southern Han Chinese, China	CHS	105	10
Japanese in Tokyo, Japan	JPT	104	10
Kinh in Ho Chi Minh City, Vietnam	KHV	99	10
<i>Total East Asian ancestry</i>	<i>EAS</i>	<i>504</i>	<i>49</i>
Bengali in Bangladesh	BEB	86	9
Gujarati Indian in Houston, TX	GIH	103	10
Indian Telugu in the UK	ITU	102	10
Punjabi in Lahore, Pakistan	PJL	96	10
Sri Lankan Tamil in the UK	STU	102	10
<i>Total South Asian ancestry</i>	<i>SAS</i>	<i>489</i>	<i>49</i>
Utah residents with Northern and Western European ancestry	CEU	99	10
Finnish in Finland	FIN	99	10
British in England and Scotland	GBR	91	9
Iberian populations in Spain	IBS	107	11
Toscani in Italia	TSI	107	11
<i>Total European ancestry</i>	<i>EUR</i>	<i>503</i>	<i>51</i>

*Sample counts are for batches 1-9; batch 10 took the remainder, and thus each population contributed slightly more or fewer samples.

Table 2. Summaries of each array included in these genomic coverage analyses. “Product information” refers to the manifest version; “Date manufactured” and “Loci Count” were extracted from the array manifest headers. The last two columns indicate the percentage of unique positions assayed by the array (mapped to chromosomes 1 through 22 and the non-pseudoautosomal portion of the X chromosome) that are also found in the 1000 Genomes Phase 3 (“1000G Ph3”) integrated variant set. The second to last column is over all 1000 Genomes variants while the last column is restricted to those with a minimum of two minor allele copies seen in at least one of the five super-populations. Because the last column is over a restricted set of 1000 Genomes variants, the percentage overlap decreases. Note coverage analyses were restricted to chromosomes 1 and 22. Arrays are manufactured by Illumina, with the exception of the Affymetrix UK Biobank.

Array Name	Product information	Date Manufactured	Loci Count	% overlap w/ 1000G Ph3	% overlap w/1000G Ph3, MAF filtered
HumanCore	24v1-0_A	21-Aug-14	306,670	99%	99%
Global Screening Array	24v1-0_A	22-Oct-16	642,824	98%	96%
OmniExpress	24v1-1_A	14-Oct-14	713,014	99%	99%
Affymetrix UK Biobank	na 35	9-Mar-15	820,967	94%	92%
Multi-ethnic AMR/AFR	8v1-0_A1	7-Dec-15	1,430,141	85%	81%
Multi-ethnic EUR/EAS/SAS	8v1-0_A1	7-Dec-15	1,475,140	88%	85%
Multi-ethnic global	_A1	20-Oct-15	1,779,819	86%	83%
Omni2.5M	8v1-2_A1	6-Mar-15	2,338,671	95%	94%
Omni 5M	4-v1-1_A	11-Feb-15	4,284,426	96%	95%

Tables 3A-3E. Genome-wide coverage estimates for all evaluated arrays, separately by MAF bin and super-population. “Number of variants” gives the count of 1000 Genomes Phase 3 variants per MAF bin for the given super-population. This count includes “observed” (i.e., included on the assessed array) variants, which have dosage r^2 and concordance values of 1. The “geno conc” is the genotype concordance; “MA conc” is the minor allele concordance (see narrative for details).

Table 3A. AFR super-population

Array	Super-population	AFR			
	MAF bin	0.2%<MAF<1%	1%<MAF<5%	MAF>1%	MAF>5%
	Number of variants	1,159,896	650,607	1,458,219	807,612
<u>HumanCore</u>	dosage r2 (mean)	0.319	0.558	0.660	0.742
	dosage r2 ≥ 0.8 (%)	9.2%	22.7%	39.9%	53.8%
	geno conc (mean)	0.994	0.974	0.945	0.922
	MA conc (mean)	0.313	0.560	0.672	0.762
<u>GSA</u>	dosage r2 (mean)	0.360	0.607	0.697	0.770
	dosage r2 ≥ 0.8 (%)	11.8%	25.1%	42.1%	55.8%
	geno conc (mean)	0.994	0.977	0.950	0.929
	MA conc (mean)	0.351	0.603	0.703	0.784
<u>OmniExpress</u>	dosage r2 (mean)	0.475	0.758	0.822	0.875
	dosage r2 ≥ 0.8 (%)	22.3%	54.9%	68.8%	80.0%
	geno conc (mean)	0.995	0.986	0.972	0.961
	MA conc (mean)	0.460	0.749	0.823	0.882
<u>Affy_UKBio</u>	dosage r2 (mean)	0.390	0.641	0.728	0.797
	dosage r2 ≥ 0.8 (%)	14.4%	29.5%	46.2%	59.6%
	geno conc (mean)	0.994	0.979	0.956	0.938
	MA conc (mean)	0.380	0.636	0.731	0.808
<u>ME_amr</u>	dosage r2 (mean)	0.479	0.789	0.844	0.888
	dosage r2 ≥ 0.8 (%)	22.4%	60.6%	73.1%	83.2%
	geno conc (mean)	0.996	0.988	0.975	0.964
	MA conc (mean)	0.464	0.780	0.842	0.893
<u>ME_eur</u>	dosage r2 (mean)	0.482	0.774	0.834	0.882
	dosage r2 ≥ 0.8 (%)	23.0%	57.3%	71.0%	81.9%
	geno conc (mean)	0.996	0.987	0.974	0.963
	MA conc (mean)	0.467	0.764	0.833	0.888
<u>ME_global</u>	dosage r2 (mean)	0.493	0.799	0.851	0.893
	dosage r2 ≥ 0.8 (%)	24.0%	62.8%	74.8%	84.4%
	geno conc (mean)	0.996	0.989	0.976	0.966
	MA conc (mean)	0.478	0.790	0.850	0.898
<u>Omni2.5M</u>	dosage r2 (mean)	0.563	0.861	0.900	0.932
	dosage r2 ≥ 0.8 (%)	33.2%	77.6%	84.9%	90.8%
	geno conc (mean)	0.996	0.993	0.984	0.978
	MA conc (mean)	0.547	0.854	0.900	0.937
<u>Omni5M</u>	dosage r2 (mean)	0.589	0.879	0.913	0.940
	dosage r2 ≥ 0.8 (%)	37.2%	81.3%	87.2%	91.9%
	geno conc (mean)	0.997	0.994	0.986	0.980
	MA conc (mean)	0.573	0.873	0.913	0.945

Table 3B. AMR super-population

Array	<i>Super-population</i>	<i>AMR</i>			
	<i>MAF bin</i>	<i>0.3%<MAF<1%</i>	<i>1%<MAF<5%</i>	<i>MAF>1%</i>	<i>MAF>5%</i>
	<i>Number of variants</i>	681,000	372,245	985,301	613,056
<u>HumanCore</u>	dosage r2 (mean)	0.505	0.651	0.778	0.855
	dosage r2 ≥ 0.8 (%)	26.6%	41.3%	63.7%	77.3%
	geno conc (mean)	0.994	0.982	0.962	0.951
	MA conc (mean)	0.480	0.645	0.783	0.867
<u>GSA</u>	dosage r2 (mean)	0.563	0.745	0.829	0.880
	dosage r2 ≥ 0.8 (%)	31.8%	54.2%	71.7%	82.3%
	geno conc (mean)	0.995	0.987	0.968	0.957
	MA conc (mean)	0.534	0.734	0.830	0.888
<u>OmniExpress</u>	dosage r2 (mean)	0.616	0.773	0.859	0.911
	dosage r2 ≥ 0.8 (%)	39.4%	61.4%	77.3%	86.9%
	geno conc (mean)	0.996	0.989	0.976	0.969
	MA conc (mean)	0.593	0.766	0.861	0.919
<u>Affy_UKBio</u>	dosage r2 (mean)	0.615	0.791	0.867	0.914
	dosage r2 ≥ 0.8 (%)	37.8%	62.6%	78.9%	88.9%
	geno conc (mean)	0.996	0.990	0.977	0.969
	MA conc (mean)	0.587	0.781	0.867	0.920
<u>ME_amr</u>	dosage r2 (mean)	0.599	0.786	0.865	0.913
	dosage r2 ≥ 0.8 (%)	35.9%	62.7%	78.5%	88.1%
	geno conc (mean)	0.995	0.989	0.977	0.969
	MA conc (mean)	0.573	0.779	0.867	0.921
<u>ME_eur</u>	dosage r2 (mean)	0.633	0.811	0.879	0.920
	dosage r2 ≥ 0.8 (%)	41.0%	67.9%	81.1%	89.1%
	geno conc (mean)	0.996	0.991	0.979	0.971
	MA conc (mean)	0.608	0.804	0.881	0.928
<u>ME_global</u>	dosage r2 (mean)	0.621	0.807	0.877	0.919
	dosage r2 ≥ 0.8 (%)	38.9%	66.8%	80.6%	89.0%
	geno conc (mean)	0.996	0.990	0.978	0.971
	MA conc (mean)	0.597	0.800	0.878	0.926
<u>Omni2.5M</u>	dosage r2 (mean)	0.656	0.830	0.895	0.934
	dosage r2 ≥ 0.8 (%)	43.9%	71.5%	83.4%	90.6%
	geno conc (mean)	0.996	0.992	0.982	0.976
	MA conc (mean)	0.635	0.825	0.897	0.941
<u>Omni5M</u>	dosage r2 (mean)	0.705	0.873	0.916	0.942
	dosage r2 ≥ 0.8 (%)	51.2%	79.5%	87.1%	91.7%
	geno conc (mean)	0.997	0.994	0.985	0.979
	MA conc (mean)	0.684	0.870	0.919	0.949

Table 3C. EAS super-population

Array	<i>Super-population</i>	<i>EAS</i>			
	<i>MAF bin</i>	<i>0.2%<MAF<1%</i>	<i>1%<MAF<5%</i>	<i>MAF>1%</i>	<i>MAF>5%</i>
	<i>Number of variants</i>	495,616	214,790	738,128	523,338
<u>HumanCore</u>	dosage r2 (mean)	0.202	0.517	0.737	0.828
	dosage r2 ≥ 0.8 (%)	8.3%	34.9%	63.2%	74.8%
	geno conc (mean)	0.992	0.971	0.950	0.941
	MA conc (mean)	0.198	0.526	0.754	0.848
<u>GSA</u>	dosage r2 (mean)	0.248	0.652	0.797	0.856
	dosage r2 ≥ 0.8 (%)	10.7%	49.7%	70.2%	78.6%
	geno conc (mean)	0.992	0.980	0.957	0.948
	MA conc (mean)	0.241	0.655	0.808	0.871
<u>OmniExpress</u>	dosage r2 (mean)	0.278	0.639	0.818	0.891
	dosage r2 ≥ 0.8 (%)	13.7%	48.8%	74.0%	84.3%
	geno conc (mean)	0.993	0.979	0.967	0.962
	MA conc (mean)	0.272	0.644	0.829	0.904
<u>Affy_UKBio</u>	dosage r2 (mean)	0.263	0.624	0.808	0.883
	dosage r2 ≥ 0.8 (%)	12.3%	46.1%	72.5%	83.3%
	geno conc (mean)	0.992	0.978	0.965	0.959
	MA conc (mean)	0.257	0.628	0.817	0.895
<u>ME_amr</u>	dosage r2 (mean)	0.279	0.653	0.823	0.893
	dosage r2 ≥ 0.8 (%)	14.3%	51.3%	75.1%	84.9%
	geno conc (mean)	0.993	0.980	0.967	0.962
	MA conc (mean)	0.274	0.658	0.834	0.906
<u>ME_eur</u>	dosage r2 (mean)	0.304	0.726	0.852	0.904
	dosage r2 ≥ 0.8 (%)	15.6%	59.0%	78.5%	86.5%
	geno conc (mean)	0.993	0.984	0.971	0.966
	MA conc (mean)	0.299	0.727	0.861	0.916
<u>ME_global</u>	dosage r2 (mean)	0.304	0.725	0.852	0.904
	dosage r2 ≥ 0.8 (%)	15.6%	59.0%	78.5%	86.6%
	geno conc (mean)	0.993	0.984	0.971	0.966
	MA conc (mean)	0.299	0.727	0.861	0.916
<u>Omni2.5M</u>	dosage r2 (mean)	0.323	0.710	0.861	0.923
	dosage r2 ≥ 0.8 (%)	17.2%	57.2%	79.7%	89.0%
	geno conc (mean)	0.993	0.984	0.976	0.973
	MA conc (mean)	0.317	0.713	0.870	0.934
<u>Omni5M</u>	dosage r2 (mean)	0.343	0.736	0.874	0.931
	dosage r2 ≥ 0.8 (%)	19.1%	60.1%	81.3%	90.0%
	geno conc (mean)	0.994	0.985	0.978	0.975
	MA conc (mean)	0.338	0.739	0.882	0.941

Table 3D. EUR super-population

Array	Super-population	EUR			
	MAF bin	0.2%<MAF<1%	1%<MAF<5%	MAF>1%	MAF>5%
	Number of variants	500,548	263,719	846,170	582,451
<u>HumanCore</u>	dosage r2 (mean)	0.310	0.593	0.773	0.854
	dosage r2 ≥ 0.8 (%)	13.4%	37.9%	65.1%	77.5%
	geno conc (mean)	0.993	0.976	0.959	0.951
	MA conc (mean)	0.299	0.596	0.784	0.869
<u>GSA</u>	dosage r2 (mean)	0.383	0.767	0.846	0.882
	dosage r2 ≥ 0.8 (%)	17.7%	62.3%	75.9%	82.1%
	geno conc (mean)	0.994	0.987	0.967	0.958
	MA conc (mean)	0.367	0.764	0.852	0.892
<u>OmniExpress</u>	dosage r2 (mean)	0.411	0.724	0.854	0.912
	dosage r2 ≥ 0.8 (%)	21.7%	54.6%	76.9%	87.1%
	geno conc (mean)	0.994	0.985	0.974	0.969
	MA conc (mean)	0.397	0.722	0.860	0.922
<u>Affy_UKBio</u>	dosage r2 (mean)	0.435	0.817	0.891	0.925
	dosage r2 ≥ 0.8 (%)	21.7%	69.9%	84.0%	90.5%
	geno conc (mean)	0.994	0.990	0.978	0.973
	MA conc (mean)	0.418	0.811	0.895	0.933
<u>ME_amr</u>	dosage r2 (mean)	0.421	0.758	0.869	0.919
	dosage r2 ≥ 0.8 (%)	22.2%	59.6%	79.5%	88.5%
	geno conc (mean)	0.994	0.987	0.976	0.971
	MA conc (mean)	0.406	0.755	0.874	0.928
<u>ME_eur</u>	dosage r2 (mean)	0.446	0.787	0.881	0.924
	dosage r2 ≥ 0.8 (%)	25.0%	64.6%	81.6%	89.3%
	geno conc (mean)	0.995	0.988	0.977	0.972
	MA conc (mean)	0.432	0.783	0.886	0.933
<u>ME_global</u>	dosage r2 (mean)	0.441	0.786	0.881	0.924
	dosage r2 ≥ 0.8 (%)	24.2%	64.5%	81.6%	89.3%
	geno conc (mean)	0.995	0.988	0.977	0.972
	MA conc (mean)	0.426	0.783	0.886	0.933
<u>Omni2.5M</u>	dosage r2 (mean)	0.471	0.808	0.898	0.938
	dosage r2 ≥ 0.8 (%)	27.4%	67.8%	83.8%	91.0%
	geno conc (mean)	0.995	0.990	0.981	0.978
	MA conc (mean)	0.456	0.805	0.903	0.947
<u>Omni5M</u>	dosage r2 (mean)	0.560	0.890	0.929	0.947
	dosage r2 ≥ 0.8 (%)	38.7%	82.9%	89.4%	92.3%
	geno conc (mean)	0.996	0.994	0.984	0.980
	MA conc (mean)	0.545	0.889	0.934	0.955

Table 3E. SAS super-population

Array	<i>Super-population</i>	<i>SAS</i>			
	<i>MAF bin</i>	<i>0.2%<MAF<1%</i>	<i>1%<MAF<5%</i>	<i>MAF>1%</i>	<i>MAF>5%</i>
	<i>Number of variants</i>	609,704	283,990	903,657	619,667
<u>HumanCore</u>	dosage r2 (mean)	0.241	0.536	0.739	0.832
	dosage r2 ≥ 0.8 (%)	8.2%	31.4%	60.7%	74.1%
	geno conc (mean)	0.992	0.974	0.952	0.942
	MA conc (mean)	0.230	0.538	0.751	0.849
<u>GSA</u>	dosage r2 (mean)	0.310	0.639	0.793	0.864
	dosage r2 ≥ 0.8 (%)	12.9%	42.7%	67.7%	79.1%
	geno conc (mean)	0.993	0.981	0.960	0.950
	MA conc (mean)	0.296	0.636	0.801	0.876
<u>OmniExpress</u>	dosage r2 (mean)	0.328	0.675	0.830	0.901
	dosage r2 ≥ 0.8 (%)	14.2%	48.5%	73.8%	85.4%
	geno conc (mean)	0.993	0.983	0.971	0.965
	MA conc (mean)	0.314	0.672	0.837	0.912
<u>Affy UKBio</u>	dosage r2 (mean)	0.350	0.697	0.837	0.901
	dosage r2 ≥ 0.8 (%)	16.0%	50.0%	75.0%	86.4%
	geno conc (mean)	0.994	0.984	0.971	0.964
	MA conc (mean)	0.335	0.691	0.841	0.910
<u>ME_amr</u>	dosage r2 (mean)	0.339	0.709	0.845	0.907
	dosage r2 ≥ 0.8 (%)	15.4%	54.0%	76.4%	86.7%
	geno conc (mean)	0.994	0.985	0.972	0.966
	MA conc (mean)	0.326	0.706	0.850	0.917
<u>ME_eur</u>	dosage r2 (mean)	0.365	0.744	0.860	0.913
	dosage r2 ≥ 0.8 (%)	17.6%	58.9%	78.7%	87.8%
	geno conc (mean)	0.994	0.987	0.974	0.968
	MA conc (mean)	0.351	0.741	0.866	0.923
<u>ME_global</u>	dosage r2 (mean)	0.364	0.744	0.860	0.913
	dosage r2 ≥ 0.8 (%)	17.5%	58.9%	78.7%	87.8%
	geno conc (mean)	0.994	0.987	0.974	0.968
	MA conc (mean)	0.351	0.741	0.866	0.923
<u>Omni2.5M</u>	dosage r2 (mean)	0.387	0.753	0.876	0.932
	dosage r2 ≥ 0.8 (%)	19.7%	59.7%	80.6%	90.2%
	geno conc (mean)	0.994	0.988	0.979	0.975
	MA conc (mean)	0.373	0.750	0.881	0.941
<u>Omni5M</u>	dosage r2 (mean)	0.442	0.801	0.897	0.941
	dosage r2 ≥ 0.8 (%)	26.5%	67.6%	83.9%	91.4%
	geno conc (mean)	0.995	0.990	0.982	0.978
	MA conc (mean)	0.428	0.799	0.902	0.949

Figure 1. Schematic of the study design used to assess genomic coverage of select genotyping arrays, with 1000 Genomes Project Phase 3 as a source of both test and reference data. Dashed lines indicate imputed genotypes; solid lines indicate observed genotypes. Adapted from Figure 1 in Nelson et al.⁷.

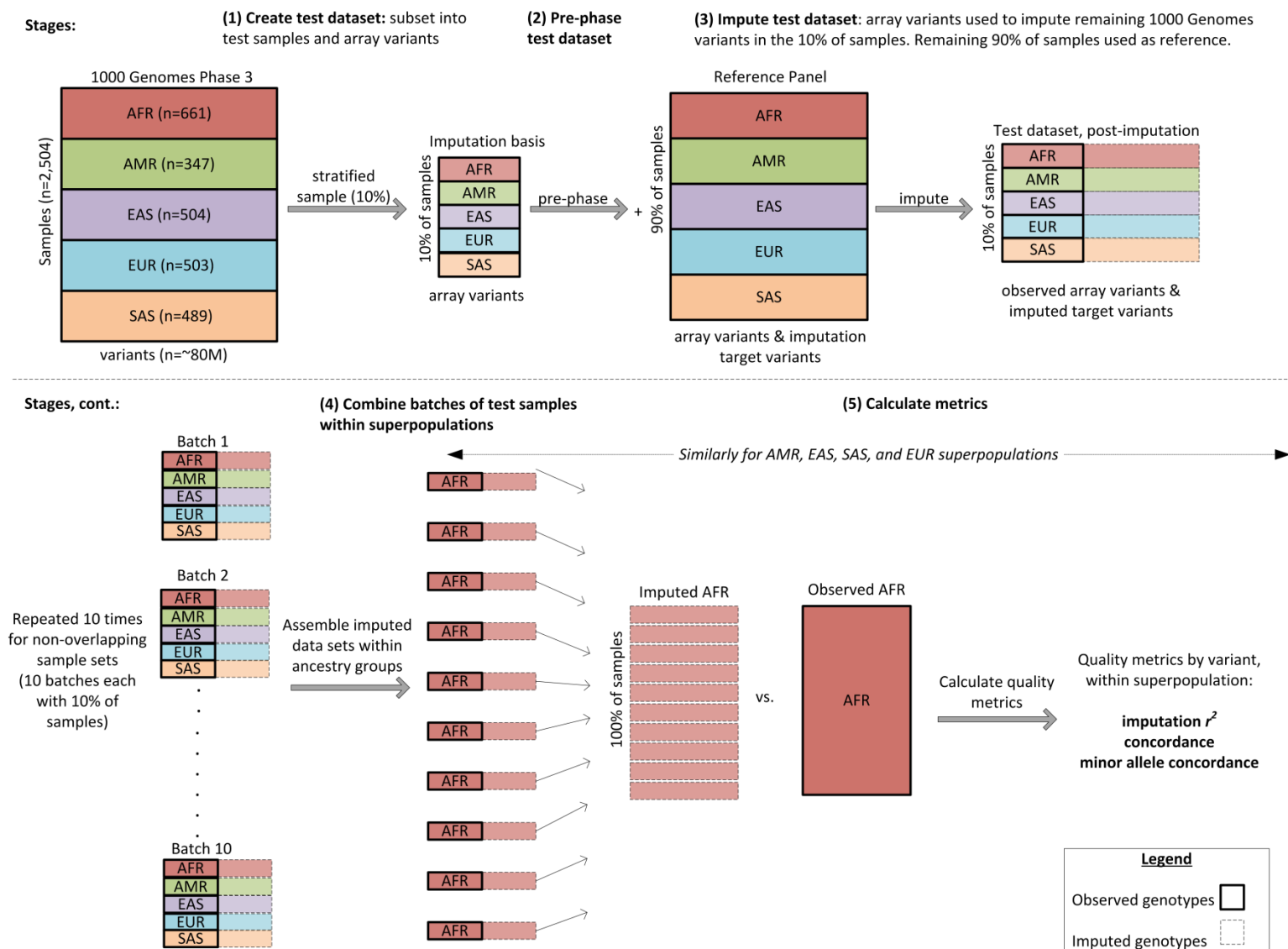


Figure 2. Fraction of variants passing a dosage r^2 threshold of 0.8, by MAF bin and super-population. Each of the five super-populations, indicated by a different colored line, was divided into ten batches and imputed using all remaining samples as a worldwide reference panel. The dosage r^2 metric plotted here is the squared correlation between imputed and observed allelic dosage in the samples comprising the super-population. The y-axis is the proportion of variants (imputed and observed) with dosage $r^2 \geq 0.8$, restricted to variants with at least two copies of the minor allele in the given super-population. The x-axis position of each array corresponds to the number of unique positions assayed by that array. Thus the order of the arrays on each axis is as follows: HumanCore, Global Screening Array, OmniExpress, Affymetrix UK Biobank, Multi-Ethnic AMR/AFR, Multi-Ethnic EUR/EAS/SAS, Multi-Ethnic global, Omni2.5M, and Omni5M.

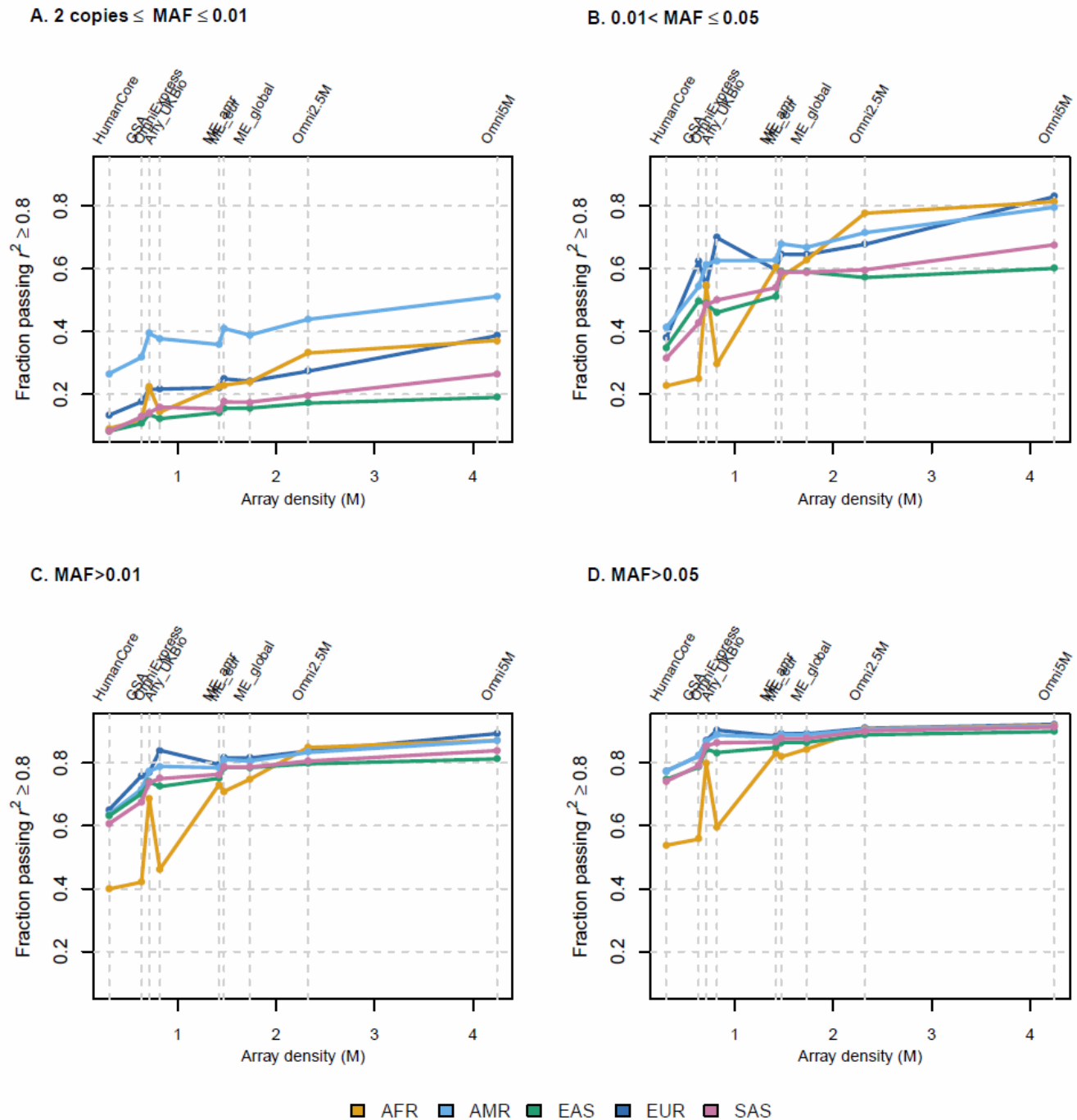


Figure 3. Mean minor allele (MA) concordance, by MAF bin and super-population. Each of the five super-populations, indicated by a different colored line, was divided into ten batches and imputed using all remaining samples as a worldwide reference panel. The y-axis values are mean MA concordance in samples comprising the given super-population. MA concordance is defined as the concordance (percent agreement) between observed and most likely imputed genotype, when at least one of those two genotypes contains one or two copies of the minor allele. Variants were restricted to those with at least two copies of the minor allele in the given super-population. Both imputed and observed variants are included in this average; the latter with MA concordance values of 1. The x-axis position of each array corresponds to the number of unique positions assayed by that array. Thus the order of the arrays on each axis is as follows: HumanCore, Global Screening Array, OmniExpress, Affymetrix UK Biobank, Multi-Ethnic AMR/AFR, Multi-Ethnic EUR/EAS/SAS, Multi-Ethnic global, Omni2.5M, and Omni5M.

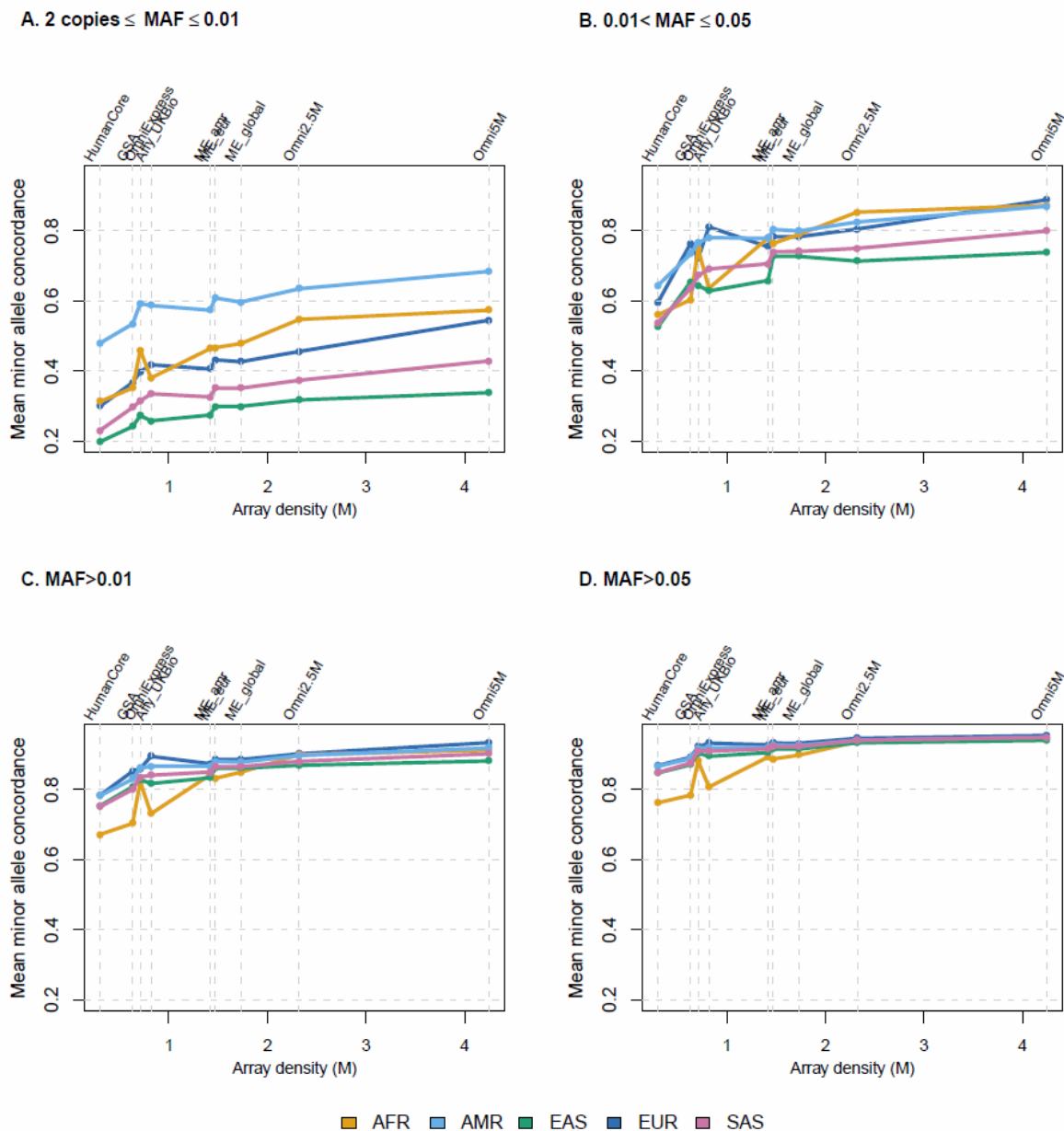


Figure 4. Mean dosage r^2 , by MAF bin and super-population. Each of the five super-populations, indicated by a different colored line, was divided into ten batches and imputed using all remaining samples as a worldwide reference panel. The y-axis is the mean dosage r^2 metric: the squared correlation between imputed and observed allelic dosage in the samples comprising the super-population. The x-axis position of each array corresponds to the number of unique positions assayed by that array. Thus the order of the arrays on each axis is as follows: HumanCore, Global Screening Array, OmniExpress, Affymetrix UK Biobank, Multi-Ethnic AMR/AFR, Multi-Ethnic EUR/EAS/SAS, Multi-Ethnic global, Omni2.5M, and Omni5M.

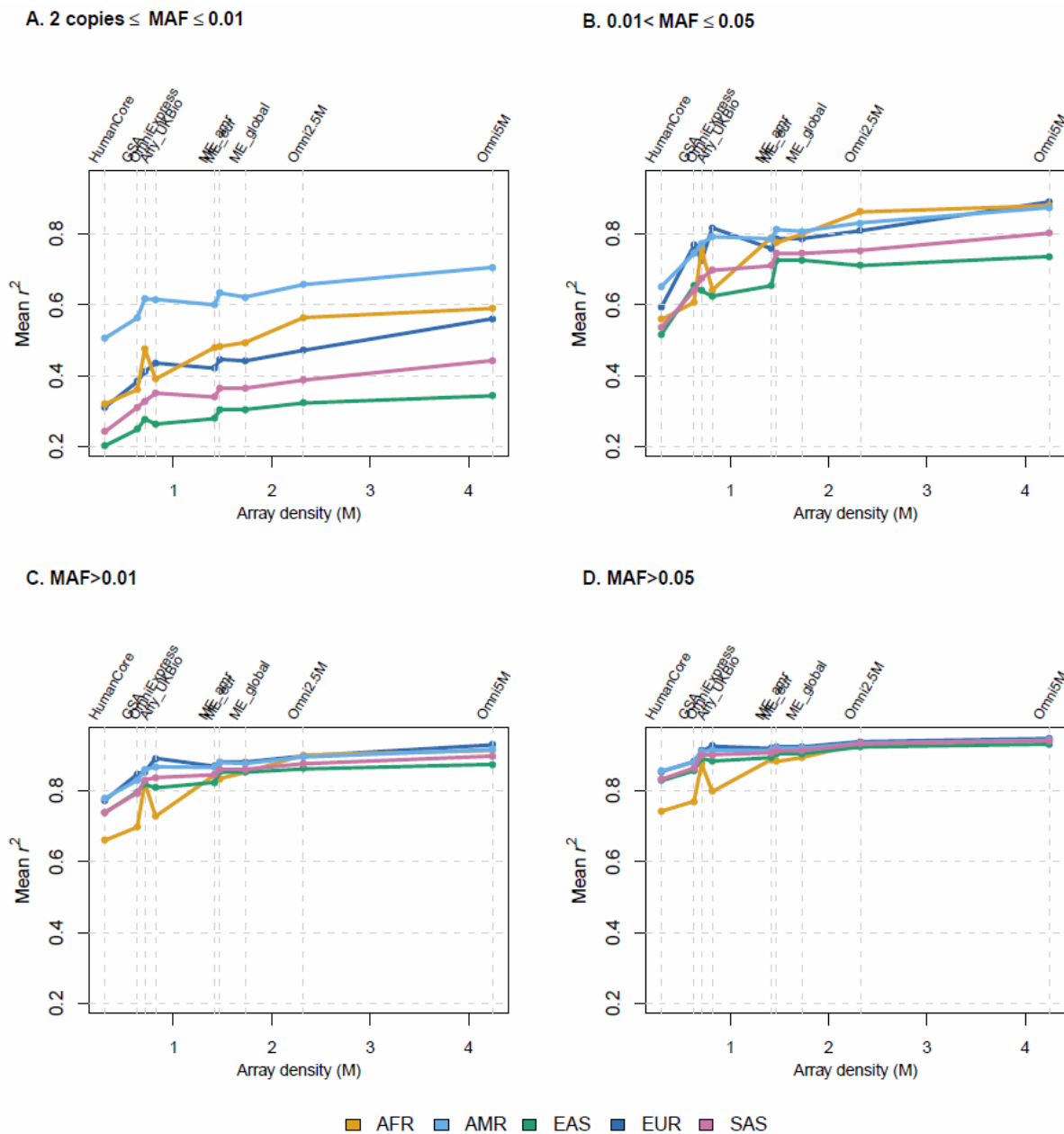


Figure 5. Mean genotype concordance, by MAF bin and super-population. Each of the five super-populations, indicated by a different colored line, was divided into ten batches and imputed using all remaining samples as a worldwide reference panel. The y-axis values are mean genotype concordance in samples comprising the given super-population. Variants were restricted to those with at least two copies of the minor allele in the given super-population. Both imputed and observed variants are included in this average; the latter with concordance values of 1. The x-axis position of each array corresponds to the number of unique positions assayed by that array. Thus the order of the arrays on each axis is as follows: HumanCore, Global Screening Array, OmniExpress, Affymetrix UK Biobank, Multi-Ethnic AMR/AFR, Multi-Ethnic EUR/EAS/SAS, Multi-Ethnic global, Omni2.5M, and Omni5M.

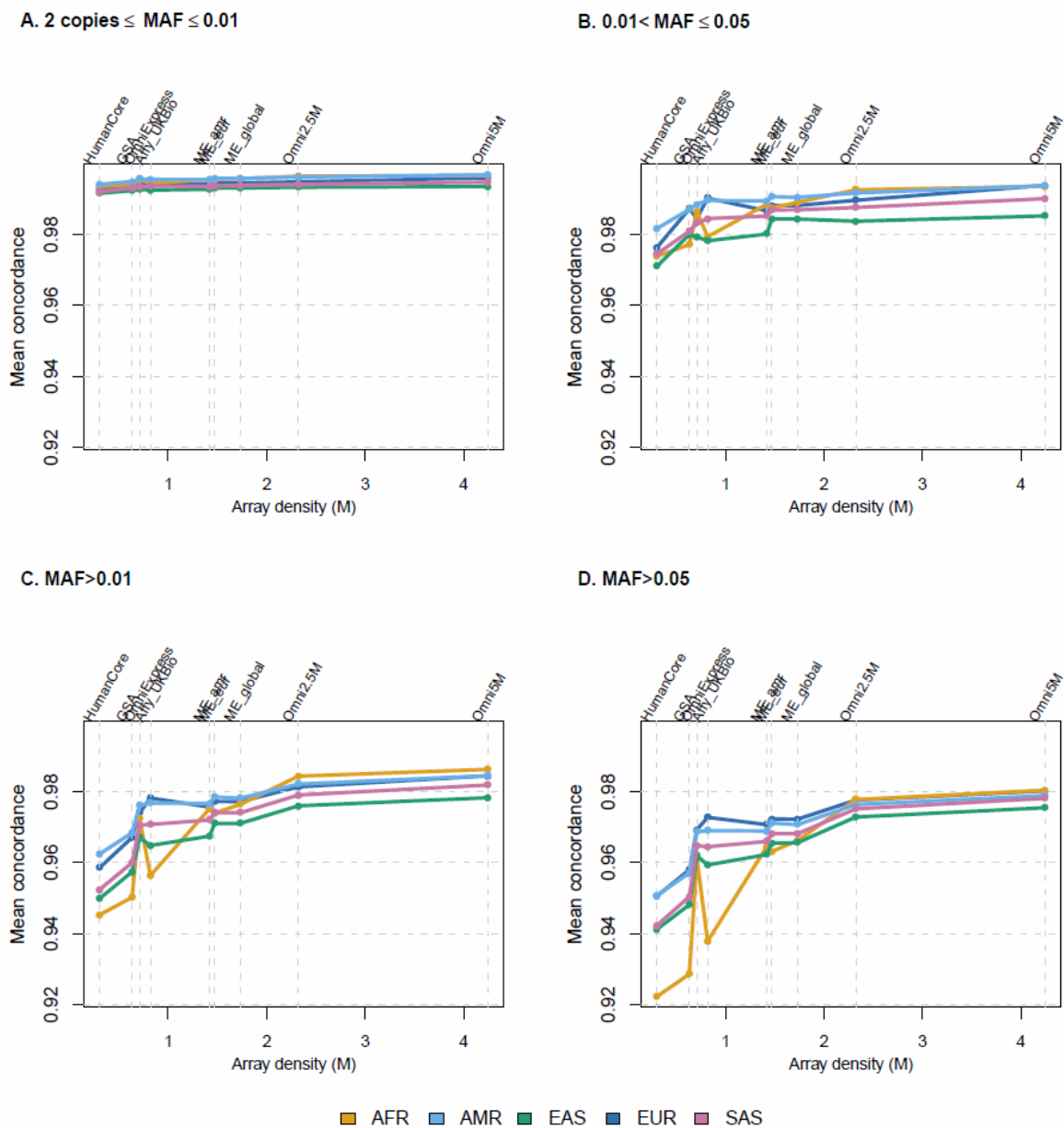
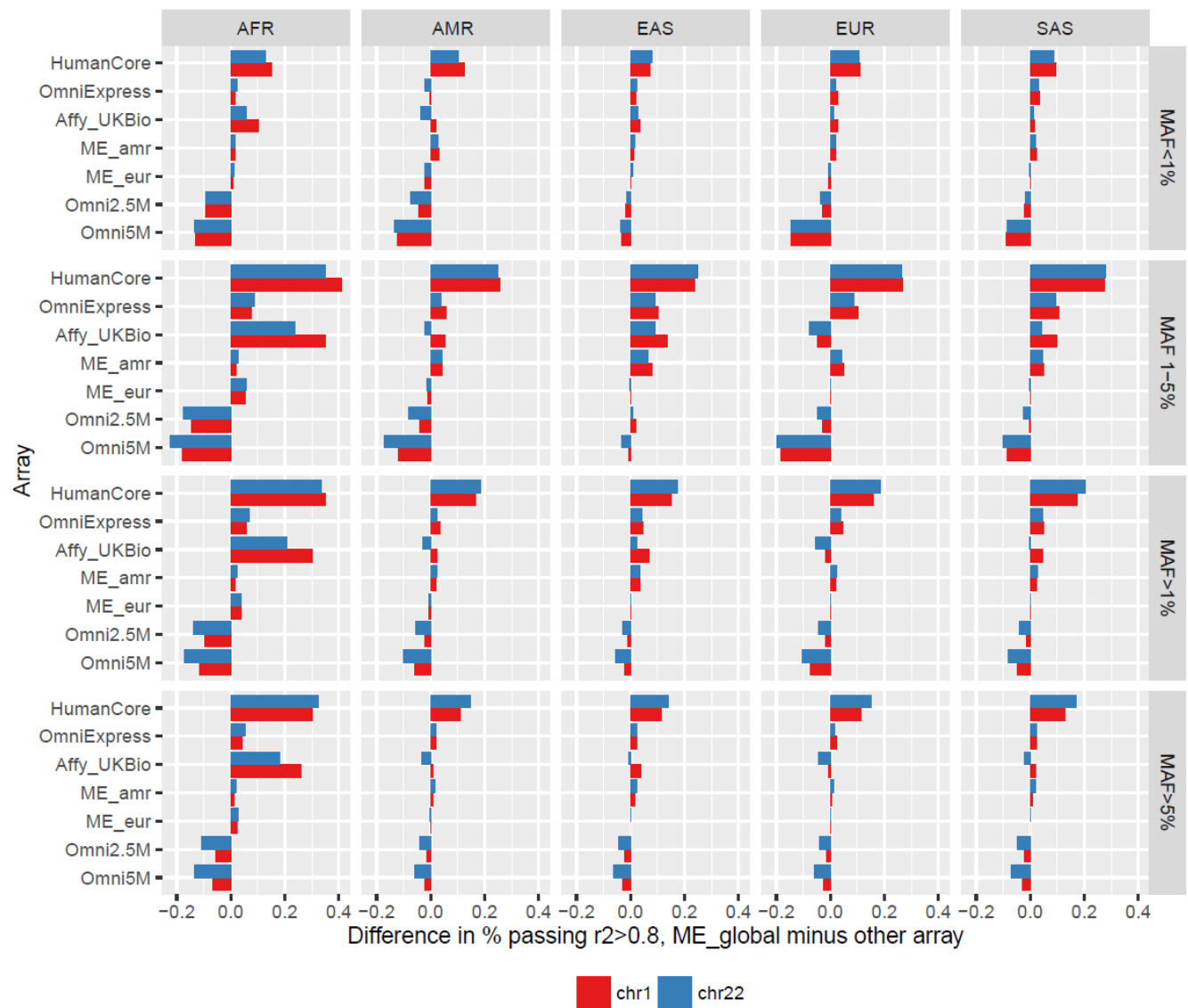
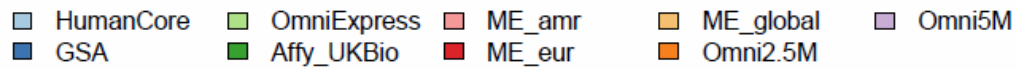


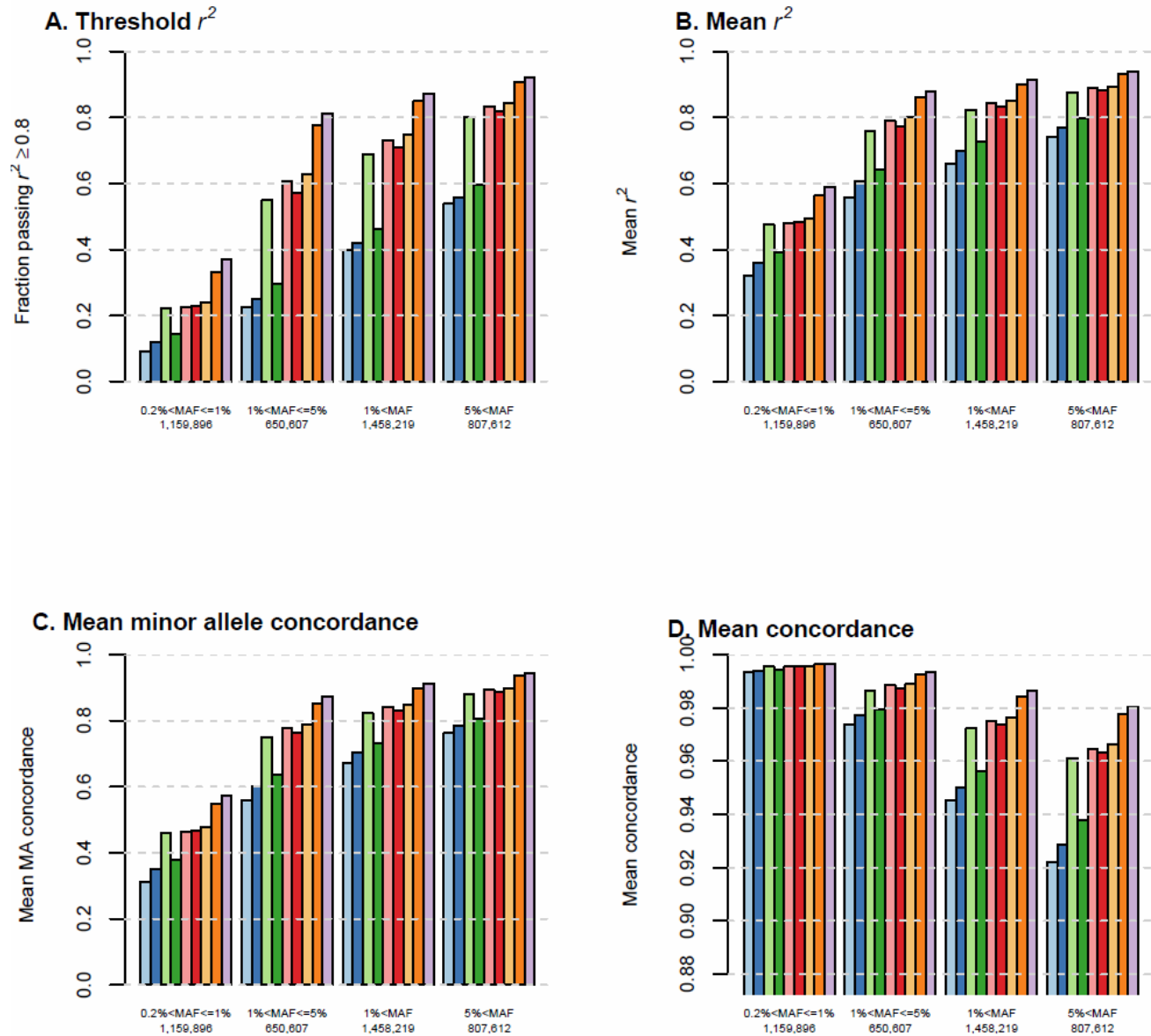
Figure 6. Comparison of genomic coverage on chromosomes 1 and 22. Each plot in the grid is for a given MAF group (as rows) and super-population (as columns). The bars show the difference in fraction of variants with $r^2 > 0.8$, subtracting the value for the array indicated from the Multi-Ethnic global array. Note the Multi-Ethnic global array is omitted from the plot, as the comparison is always 0% difference. Positive values on the x-axis mean Multi-Ethnic global has higher coverage; negative values means the labelled array has higher coverage. The Global Screening Array was not assessed for chromosome-specific metrics and is thus not represented on this plot.



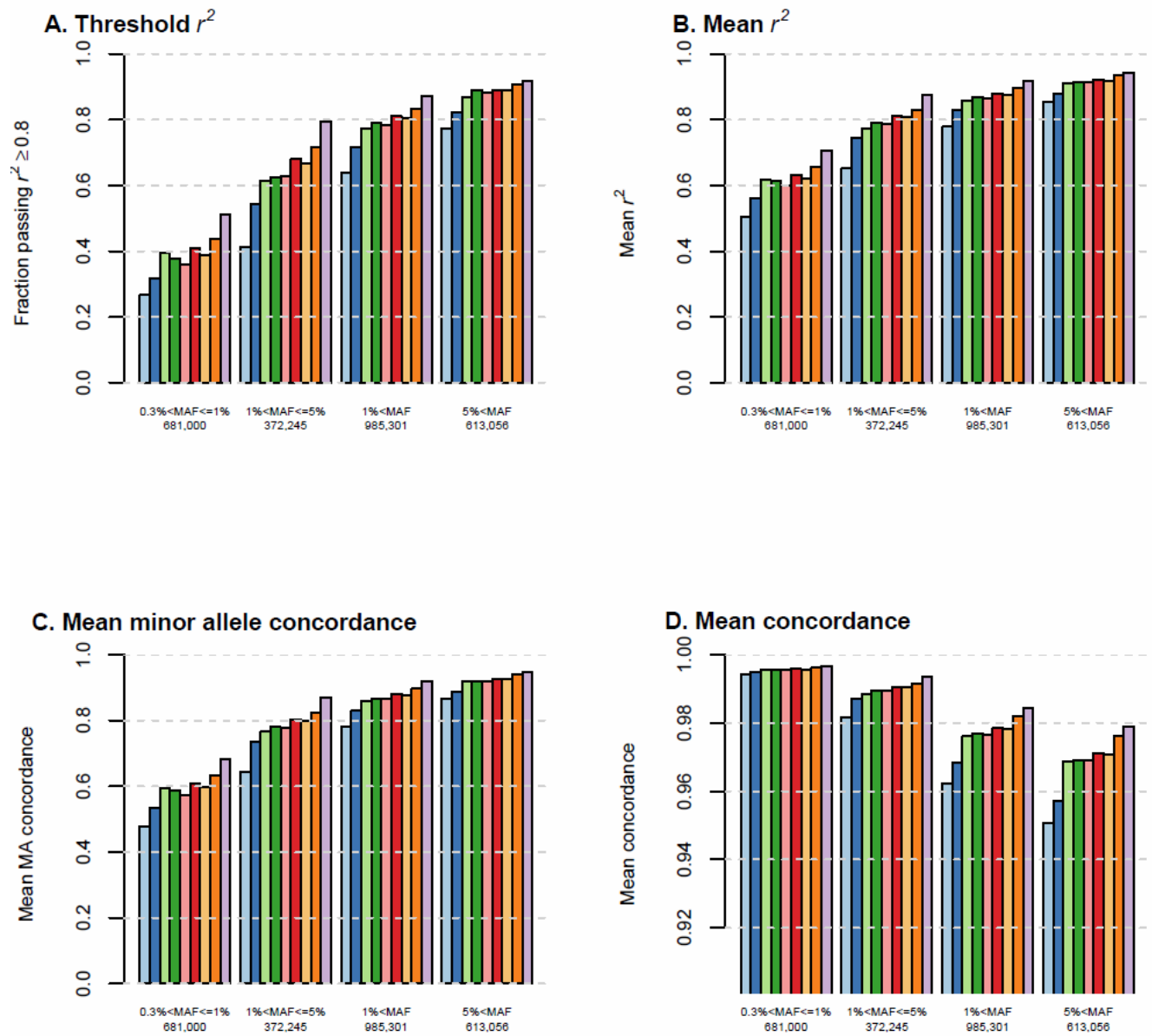
Appendix A. Barplots of coverage metrics, by super-population. Arrays are ordered by density within each MAF group, with least dense to the left (HumanCore) and most dense to the right (Omni5M). The array color coding is as follows:



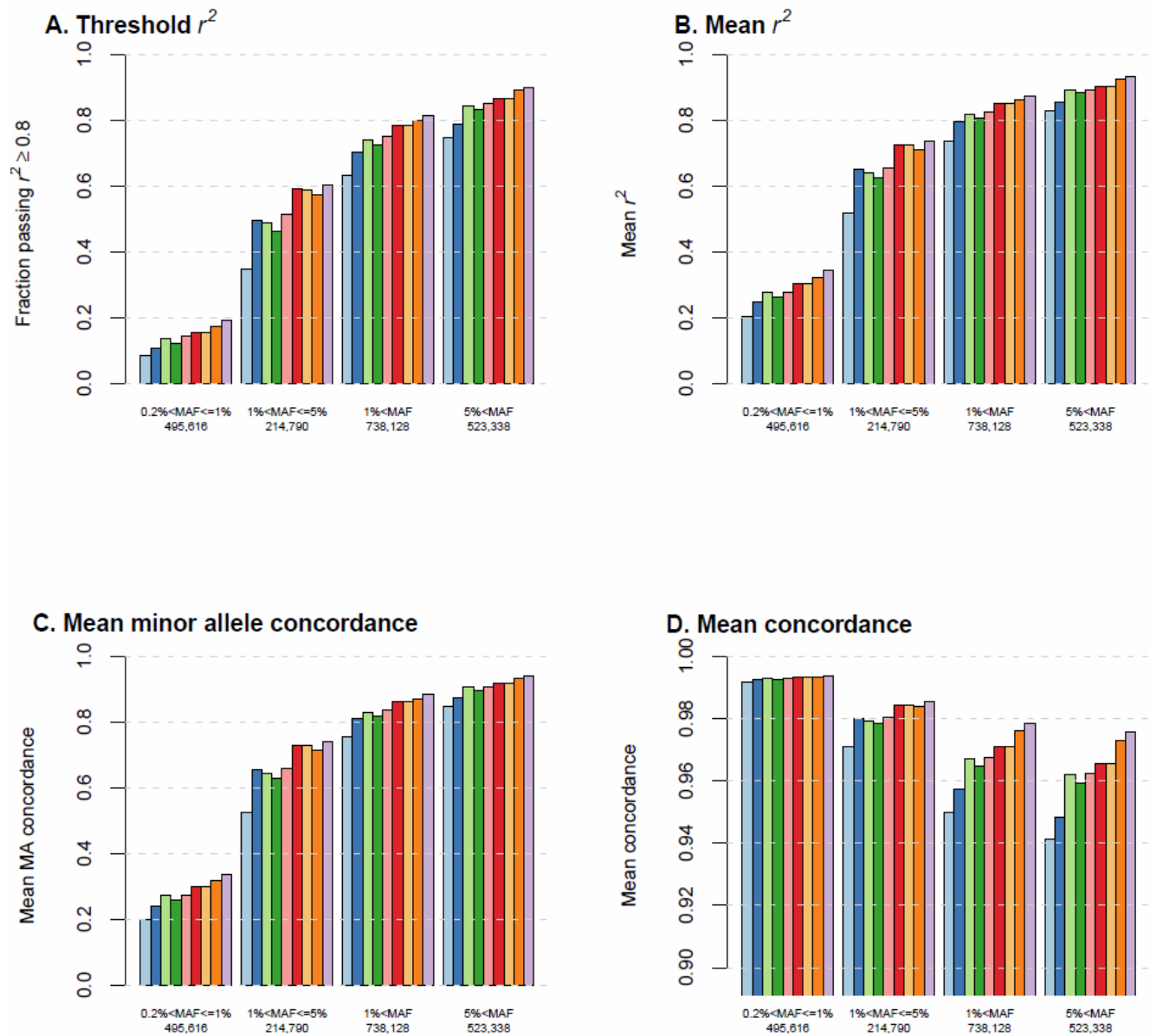
AFR



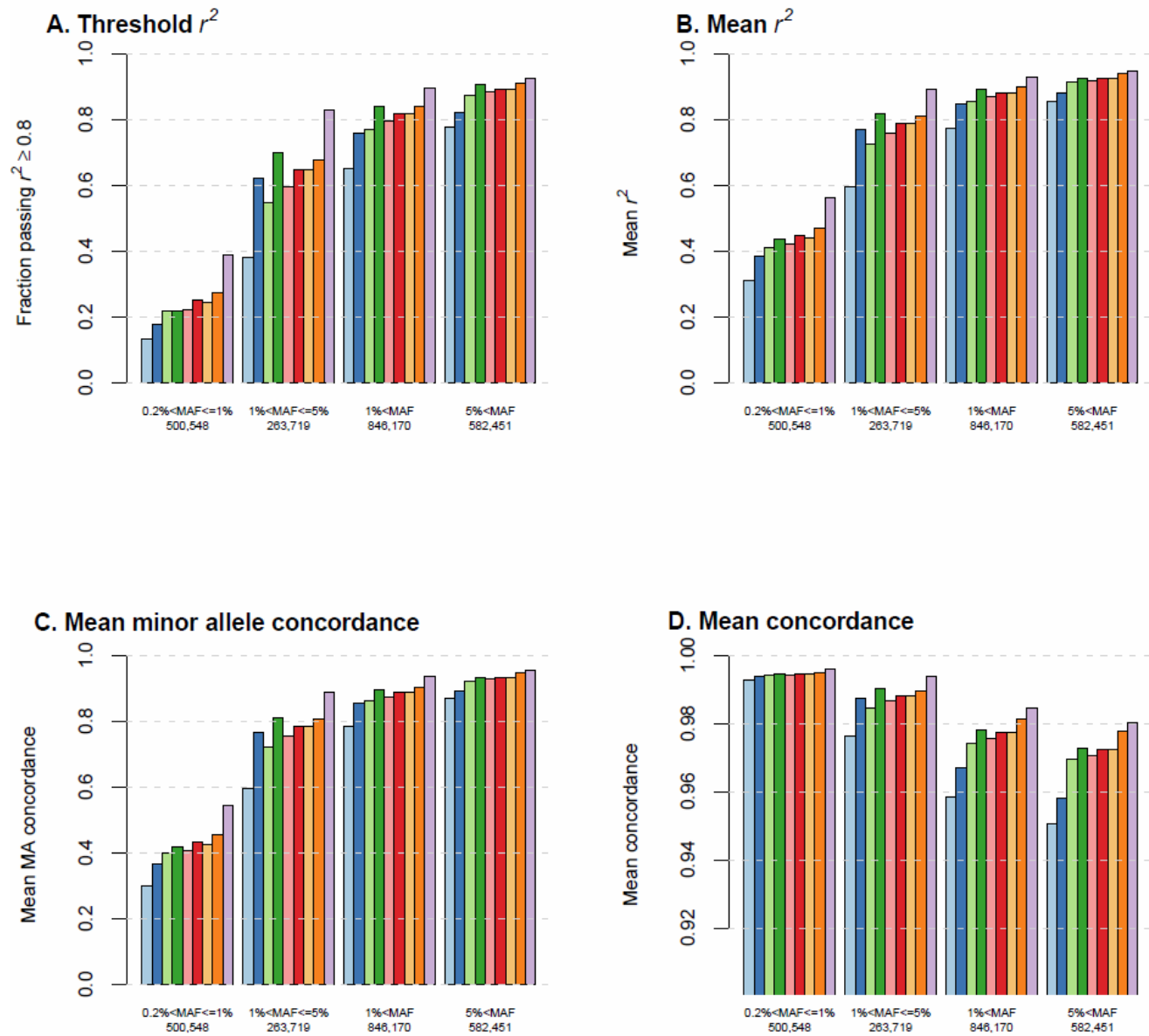
AMR



EAS



EUR



SAS

