Imputation and polygenic score performances of human genotyping arrays in diverse populations

Dat Thanh Nguyen [1,*], **Trang Tran** [1], **Mai Tran** [1], **Khai Tran** [1], **Duy Pham** [2], **Nguyen Thuy Duong** [1,3], **Quan Nguyen** [2,*], **and Nam S. Vo** [1,4,*]

[1] *Center for Biomedical Informatics, Vingroup Big Data Institute, Hanoi, Vietnam*

[2] *Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia*

[3] *Institute of Genome Research, Vietnam Academy of Science and Technology, Hanoi, Vietnam*

[4] *College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam*

Correspondence*:
Dat Thanh Nguyen or Quan Hoang Nguyen or Nam Sy Vo
n.dat@outlook.com or quan.nguyen@imb.uq.edu.au or v.namvs@vinbigdata.org

ABSTRACT

Regardless of the overwhelming use of next-generation sequencing technologies, microarray-based genotyping combined with the imputation of untyped variants remains a cost-effective means to interrogate genetic variations across the human genome. This technology is widely used in genome-wide association studies (GWAS) at bio-bank scales, and more recently, in polygenic score (PGS) analysis to predict and to stratify disease risk. Over the last decade, human genotyping arrays have undergone a tremendous growth in both number, and content making a comprehensive evaluation of their performances became more important. Here, we performed a comprehensive performance assessment for 23 available human genotyping arrays in 6 ancestry groups using diverse public, and in-house datasets. The analyses focus on performance estimation of derived imputation (in terms of accuracy and coverage) and PGS (in term of concordance to PGS estimated from whole genome sequencing data) in three different traits and diseases. We found that the arrays with a higher number of SNPs are not necessarily the ones with higher imputation performance, but the arrays that are well-optimized for the targeted population could provide very good imputation performance. In addition, PGS estimated by imputed SNP array data is highly correlated to PGS estimated by whole genome sequencing data in most of cases. When optimal arrays are used, the correlations of key PGS metrics between two types of data can be higher than 0.97, but interestingly, arrays with high density can result in lower PGS performance. Our results suggest the importance of properly selecting a suitable genotyping array for PGS applications. Finally, we developed a web tool that provide interactive analyses of tag SNP contents and imputation performance based on population and genomic regions of interest. This study would act as a practical guide for researchers to design their genotyping arrays-based studies. The tool is available at: `https://genome.vinbigdata.org/tools/saa/`

Keywords: keyword, keyword, keyword, keyword, keyword, keyword, keyword, keyword

1 INTRODUCTION

Over the last decade, low-cost, robust genotyping platforms and large-scale genome variation projects such as the 1000 Genomes Project (Auton et al., 2015) have facilitated genome-wide association studies (GWAS) on numerous human phenotypes, ranging from height to diseases (Bycroft et al., 2018). To date, thousands of DNA loci that are significantly associated with complex traits and diseases have been discovered (Buniello et al., 2019). Among numerous possible applications of GWAS results, disease risk prediction is rapidly gaining broad interest recently (Torkamani et al., 2018; Lewis and Vassos, 2020; Lambert et al., 2019). A polygenic score (PGS) or polygenic risk score (PRS) is an estimate of an individual's genetic liability to a trait or disease, calculated based on their genotype profile and relevant GWAS data (Choi et al., 2020). In its most common form, a PGS is computed as the sum of allele count of risk alleles (0, 1, or 2) that are weighted by its effect size (i.e. log odd ratio or beta coefficient) of hundreds-to-thousands of associated SNPs. The outcome is a single score that aggregate each individual's genetic loading proportional to the risk of a given disease or a quantitative trait (Lambert et al., 2019). Although clinical utility of PGS has yet to be established, recent works have suggested that PGS may be used for disease risk stratification that potentially facilitates early disease detection, assists in diagnosis or informs treatment choices (Torkamani et al., 2018; Lewis and Vassos, 2020). For example, PGS of coronary artery disease, type 2 diabetes, and breast cancer at the top 8, 3.5, and 1.5% are risk equivalent to a monogenic mutation risk that confers an odd ratio of 3 (Khera et al., 2018).

Similar to GWAS analysis, PGS can be derived from various types of genotyping data such as those obtained by single-nucleotide polymorphism (SNP) microarrays or whole genome sequencing (WGS). While WGS is attractive by the ability to interrogate variations across the entire human genome, SNP arrays are the dominant assays to obtain genetic data for PGS calculation. They come up with several advantages such as cost-effectiveness and light computational requirement which are preferable for population-sale screening, where PGS would be most useful (Chen et al., 2020). Because the coverage of SNP arrays is typically limited to lower than a million SNPs, a procedure involving haplotye phasing and genotype imputing of missing sites is usually employed to add more genotyping information that can increase powers of these genetic studies (Howie et al., 2012; Marchini and Howie, 2010; Choi et al., 2020). The imputation performance is affected by three main factors, including algorithms of choice (Das et al., 2016), imputation reference panels (Huang et al., 2015; McCarthy et al., 2016), and the SNP array designs (Nelson et al., 2013).

In principle, genotyping SNP arrays are designed by selecting a set of SNPs, commonly reffered to as “tag SNPs”, which maximize coverage of ungenotyped DNA variants through associations between these alleles in the population (known as linkage disequilibrium, LD) (Gibbs et al., 2003; Carlson et al., 2004).

Based on the target population, human genotyping SNP arrays can be classified into three categories that optimized for global, super population or specific to targeted populations. In the early phase of development, genotyping SNP arrays were focused on common genetic variations of the whole world population (minor allele frequency, MAF, of 0.10 or greater) based on the HapMap catalog (Consortium et al., 2007). The second generation of SNP arrays were designed to cover variants with MAF as low as 0.01 by providing SNP arrays specifically for European, East Asian, African American, and Latino race/ethnicity populations based on the 1000 Genomes Project (1KGP) catalog (Hoffmann et al., 2011a,b). However, the fact that the majority of human genetic variants are rare and population-specific demands customizing SNP arrays to improve over those designed for global or super populations (Consortium et al., 2015; Tam et al., 2019). Indeed, population-specific genotyping arrays such as the UK Biobank Axiom Array (Bycroft et al., 2018), the Axiom-NL Array (Ehli et al., 2017), the Japonica and Japonica NEO Arrays (Kawai et al., 2015; Sakurai-Yageta et al., 2020), and the Axiom KoreanChip (Moon et al., 2019) have been developed on top of the many existing commercial arrays. These arrays are not only optimized for genomic coverage based on their unique variant catalogs, but also include a large amounts of functional variants. For example, the Axiom KoreanChip contains more than 200,000 nonsynonymous loci and the new Japonica NEO Arrays is designed with abundant disease risk variants (Moon et al., 2019; Sakurai-Yageta et al., 2020).

The development of customized arrays together with commercial arrays provided by genotyping platform producers result in a large number of genotyping arrays. Each of these arrays has specific properties and contents, and thus, there is an urgent demand for a systematic guideline to determine which array best suits specific research questions and populations. Although there are SNP array comparative studies, they are either not updated with the many recent arrays (Nelson et al., 2013; Ha et al., 2014), or limited in only testing for a small set of populations, and some studies focused on LD coverage (Ha et al., 2014; Verlouw et al., 2021) that may not be relevant to current imputation practice for use in association studies and PGS analysis (Marchini and Howie, 2010; Choi et al., 2020). Moreover, although PGS is gaining increasing attention, practical evaluation of performance for PGS applications by current genotyping arrays is still lacking. Here, we provide a comprehensive evaluation of imputation-based genomic coverage (Lindquist et al., 2013; Nelson et al., 2013) and PGS performance of 23 human genotyping arrays in diverse populations. These analyses are intended to be a practical guide for researchers in selecting the most suitable genotyping array for their genetic studies.

2 MATERIALS AND METHODS

2.1 Genotyping arrays

In this study, we benchmarked 23 different human genotyping arrays including 14 arrays from Illumina and 9 arrays from Affymetrix. The examined arrays contain the numbers of tag SNPs (array size) ranging from approximately 300,000 (Infinium HumanCytoSNP-12 v2.1) up to more than 4,300,000 (Infinium Omni5 v1.2). They can be classified as old arrays such as the Genome-Wide Human SNP Array 6.0; population specific optimized arrays such as Axiom UK Biobank Array and Axiom Japonica Array NEO; multiple population optimized arrays such as Infinium Multi-Ethnic Global v1.0 and Infinium Global Diversity Array v1.0; cytogenetics and cancer applications optimized arrays such as Infinium CytoSNP-850K v1.2. Recently developed arrays include Infinium Global Screening Array v3.0, Axiom Precision Medicine Research Array, and Axiom Precision Medicine Diversity Array. Manifests of the 23 examined arrays were obtained from respective manufacturers' websites. Genomic positions were further harmonized to the UCSC hg38 reference genome coordinate with CrossMap v0.2.6 for those require lifted over (Zhao et al., 2014). Details and component statistics of these arrays are shown in Table 1.

2.2 Genomic datasets and pipelines

An overview of our evaluation pipeline is presented in Figure 1. In brief, the phased genomic data in Variant Call Format (VCF) of 2,504 and 1,008 unrelated individuals from the 1000 Genomes Project samples that were re-sequenced by New York Genome Center (1KGP) (Byrska-Bishop et al., 2021) and the 1000 Vietnamese Genomes Project (1KVG) (Tran et al., 2022) were used to estimate imputation-based coverage and PGS performance of 23 different genotyping arrays by the 10-fold cross-validation

Table 1. Details of 23 human genotyping arrays used in this study.

Array	Short name	No.Assays	No.Positions	No.Autosomal	No.X	No.y	No.MT
Infinium HumanCytoSNP-12 v2.1	CytoSNP-12	293552	293467	276248	15082	1444	0
Infinium Core-24 v1.2	Infinium_Core	304151	304111	293850	8097	2003	161
Infinium OncoArray-500K v1.0	Infinium_OncoArray	497191	496203	481495	14276	312	120
Infinium PsychArray v1.3	PsychArray	592414	584233	567619	14221	2051	342
Axiom Genome-Wide ASI	Axiom_GW_ASI	629494	629492	609774	17263	2222	233
Infinium Global Screening Array v3.0	Infinium_GSA	654027	648380	616080	26635	3822	987
Axiom Genome-Wide CHB	Axiom_GW_CHB	656638	656625	631283	24267	980	95
Axiom Japonica Array NEO	Axiom_JAPONICA	671123	666782	652237	13336	779	409
Axiom Genome-Wide EUR	Axiom_GW_EUR	674287	673449	659956	13104	290	99
Infinium Chinese Genotyping Array v1.0	Infinium_Chinese	695116	682199	647335	27668	6210	986
Infinium Japanese Screening Array v1.0	Infinium_JSA	719938	707559	675012	26223	4686	948
Axiom UK Biobank Array	Axiom_UKB	843755	820407	798493	20827	813	274
Infinium CytoSNP-850K v1.2	CytoSNP-850K	845050	842682	811217	29666	1097	0
Axiom Precision Medicine Research Array	Axiom_PMRA	919099	900406	864096	36132	8	170
Axiom Precision Medicine Diversity Array	Axiom_PMDA	921664	900770	837511	62039	448	714
Genome-Wide Human SNP Array 6.0	Affymetrix_6.0	931991	929011	889847	37894	859	411
Infinium OmniZhongHua v1.4	OmniZhongHua	1170268	1165100	1134324	28444	2220	112
Infinium Multi-Ethnic EUR/EAS/SAS v1.0	Multi-Ethnic_EUR_EAS_SAS	1471475	1471475	1429754	39479	1598	644
Infinium Multi-Ethnic Global v1.0	Multi-Ethnic_Global	1748250	1733356	1673788	50914	3569	776
Infinium Global Diversity Array v1.0	Infinium_GDA	1904599	1825277	1752897	60512	5744	1115
Axiom Genome-Wide PanAFR	Axiom_GW_PanAFR	2264666	2264432	2195556	65949	2647	280
Infinium Omni2.5 v1.5	Infinium_Omni2.5	2373357	2363610	2311073	50841	1515	181
Infinium Omni5 v1.2	Infinium_Omni5	4327108	4245106	4131134	106418	2396	207

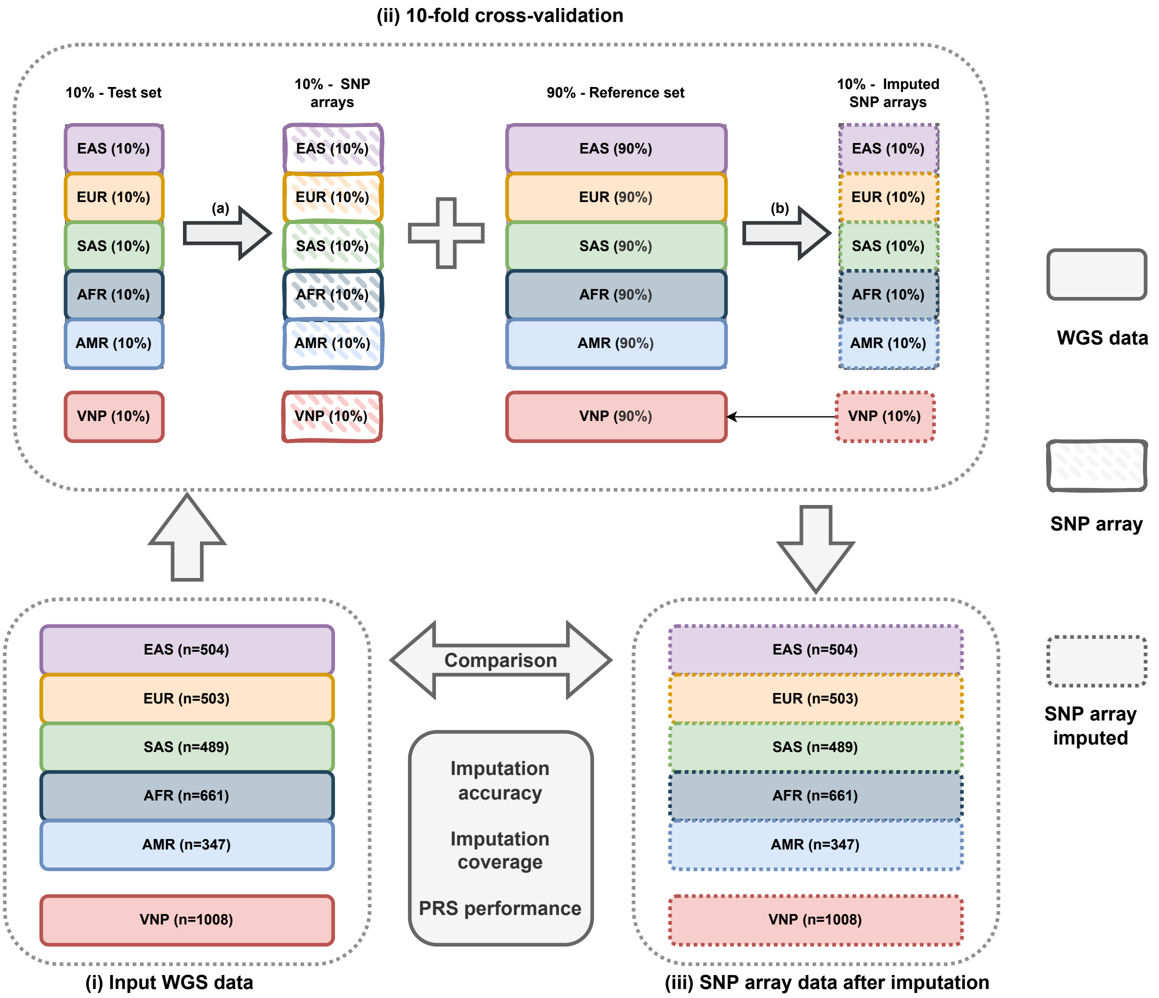

Figure 1. Overview of evaluation pipeline. (i) Two input genetic datasets, including the 1KGP and 1KVG were randomly divided into 10 batches that are equally distributed by populations. (ii) 10-fold cross-validation procedure. In each turn, variants of 10% samples were extracted based on arrays' manifest to generate simulated array genotyping data (arrow a) as input for phasing and imputation with the remaining 90% samples used as the reference set to generate the imputed SNP array data (arrow b). (iii) SNP array data after imputation. Imputed SNP array data of 10 batches were merged according to populations after 10-fold cross-validation, and were then benchmarked by treating the input WGS data as the golden standard.

approach. Specifically, genomic datasets were randomly divided into 10 batches (equally distributed across populations in the 1KGP dataset). In each turn, one batch was used as the test set and the remaining samples as the reference set. For each array, variants in the test set with the same position as variants on the array were extracted with vcftools v0.1.17 (Danecek et al., 2011) and phasing information was removed to generate the pseudo SNP array genotyped data, while variants in reference data were used as the pre-phasing and imputation reference panel. The pre-phasing and imputation were performed with SHAPEIT v4.1.3 (Delaneau et al., 2019) and Minimac4 v1.0.2 (Das et al., 2016) respectively. Finally, the imputed genotyping data of 10 batches were combined to estimate imputation and PGS performance according to their populations, including 504, 503, 489, 661, 347, and 1,008 individuals in East Asian (EAS), European (EUR), South Asian (SAS), African (AFR), American (AMR) and Vietnamese (VNP) populations, respectively. This approach is similar to the strategy used previously to estimate imputation-based genomic coverage (Lindquist et al., 2013; Nelson et al., 2013; Nguyen et al., 2021).

2.3 Imputation performance evaluation

Both GWAS and PGS often require genotype imputation that involves prediction of untyped variants in the genome. While GWAS benefits from the boosting the number of imputed SNPs that can be tested for association (Marchini and Howie, 2010), computation of PGS is conducted by summing the product of risk allele count (0, 1, or 2) and its effect size derived from the GWAS. Thus, imputation performance is expected to play a key role in PGS derivation. Here, we focus on imputation r^2 metric although there are several other criteria that can be used to assess imputation performance such as allele concordance (Nelson et al., 2013), imputation quality (Verlouw et al., 2021), LD coverage (Barrett and Cardon, 2006). We choose imputation r^2 as the evaluation metric for following reasons. First, it is more relevant to the context of GWAS and PGS analysis because the imputation r^2 at a given variant is proportional to its χ^2 statistic that results from an association test (Pritchard and Przeworski, 2001; Chapman et al., 2003; Marchini, 2019; Li et al., 2021). This leads to the interpretation that an increase in mean imputation r^2 at genome wide scale directly corresponds to the increase of statistical power (Pritchard and Przeworski, 2001; Li et al., 2021). Second, it is less sensitive to allele frequency than concordance (Nelson et al., 2013). Third, it incorporates imputation uncertainty by using expected allele dosage rather than the most likely genotype (Nelson et al., 2013). Finally, imputation r^2 can be computed on a site-by-site basis, which enables more detailed evaluation than at the allele frequency level (Li et al., 2021). In this evaluation setting, we treated genotypes derived from WGS datasets as gold-standard. Imputation performance is measured as imputation r^2 that is SNP-wise squared Pearson's correlation between the imputed dosages and the WGS genotypes, and imputation coverage that is defined as proportion of SNPs with imputation r^2 passing the cut-off of 0.8. These metrics were stratified into three minor allele frequency (MAF) bins, including (0-0.01], (0.01-0.05], (0.05-0.5]. To reduce the data noise, variants with allele count ¡ 2 are excluded in the bin of (0-0.01]. Of note, the MAF bin of (0.01-0.5], which is the most common cutoff for GWAS and PGS analysis, was also considered in the analysis (Marees et al., 2018; Choi et al., 2020).

2.4 PGS performance assessment

Instead of using pre-tuned PGS models as other studies (Li et al., 2021; Chen et al., 2020)), in this study PGS was computed with a standard P+T (Prunning and Thresholding) approach implemented in PRSice-2 (Choi and O'Reilly, 2019). The main reason for using this approach is that we tried to mimic the real-life practice of PGS analysis that involves running a PGS computational method with multiple parameters and selecting the best one (Choi et al., 2020). Another reason was that using pre-built PGS models may introduce a potential bias for some specific arrays as they were used in tuning, and we tried to avoid training using the same array twice. Using summary statistics for three phenotypes, namely height, body mass index (BMI), and type 2 diabetes (T2D), obtained from previous GWAS meta analyses (Yengo et al., 2018; Xue et al., 2018), a PGS for an individual i was calculated as:

$$PGS_i(P_T) = \sum_{j=1}^{M} \mathbf{1}_{\{P_j<P_T\}} x_{ij}\hat{\beta}_j, \tag{1}$$

where P_T is the p-value threshold values (5e-08, 1e-07, 1e-06, 1e-05, 0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.5, and 1); M is number of SNPs after clumping with *"–clump-kb 250kb"* and *"–clump-r2 0.1"*; x_{ij} and $\hat{\beta}_j$ is the allele count and the marginal effect size derived from GWAS summary statistics of SNP_j.

Similar to imputation performance evaluation, we treat PGSs derived from WGS as the "gold-standard". PGSs derived from 23 different SNP arrays were evaluated using Pearson's correlation and absolute percentile differences compared to the gold-standard under the same PRSice-2 parameter settings. In addition, absolute difference of PGS percentile ranking generated by array-imputed and the gold-standard was also evaluated.

3 RESULTS

3.1 Imputation performance

Overall, we found two main factors affecting the imputation accuracy and imputation coverage, including array sizes and population-specific optimization. The Infinium Omni2.5 v1.5 and Infinium Omni5 v1.2 with approximately 2.4 and 4.3 minion tag SNPs yielded the highest imputation performance. In contrast, low density SNP arrays with approximately 300,000 tag SNPs such as Infinium HumanCytoSNP-12 v2.1 and Infinium Core-24 v1.2 obtain the poorest imputation performance in all all six examined populations. At the MAF bin of (0.01-0.5], the Infinium Omni5 v1.2 yielded the mean imputation accuracy r^2 of 0.9032, 0.9144, 0.8644, 0.9176, 0.8873, 0.9499 and the imputation coverage of 0.8721, 0.8813, 0.8019, 0.8885, 0.8344, 0.9207 while the Infinium HumanCytoSNP-12 v2.1 obtained 0.6682, 0.7708 0.7112, 0.7608 0.7218, 0.8635 for mean imputation accuracy r^2 and 0.4031, 0.6265, 0.5879, 0.6297, 0.5731, 0.7655 for imputation coverage in six populations AFR, AMR, EAS, EUR, SAS, and VNP respectively. Details are reported in Figure 2 and Table 2, 3.

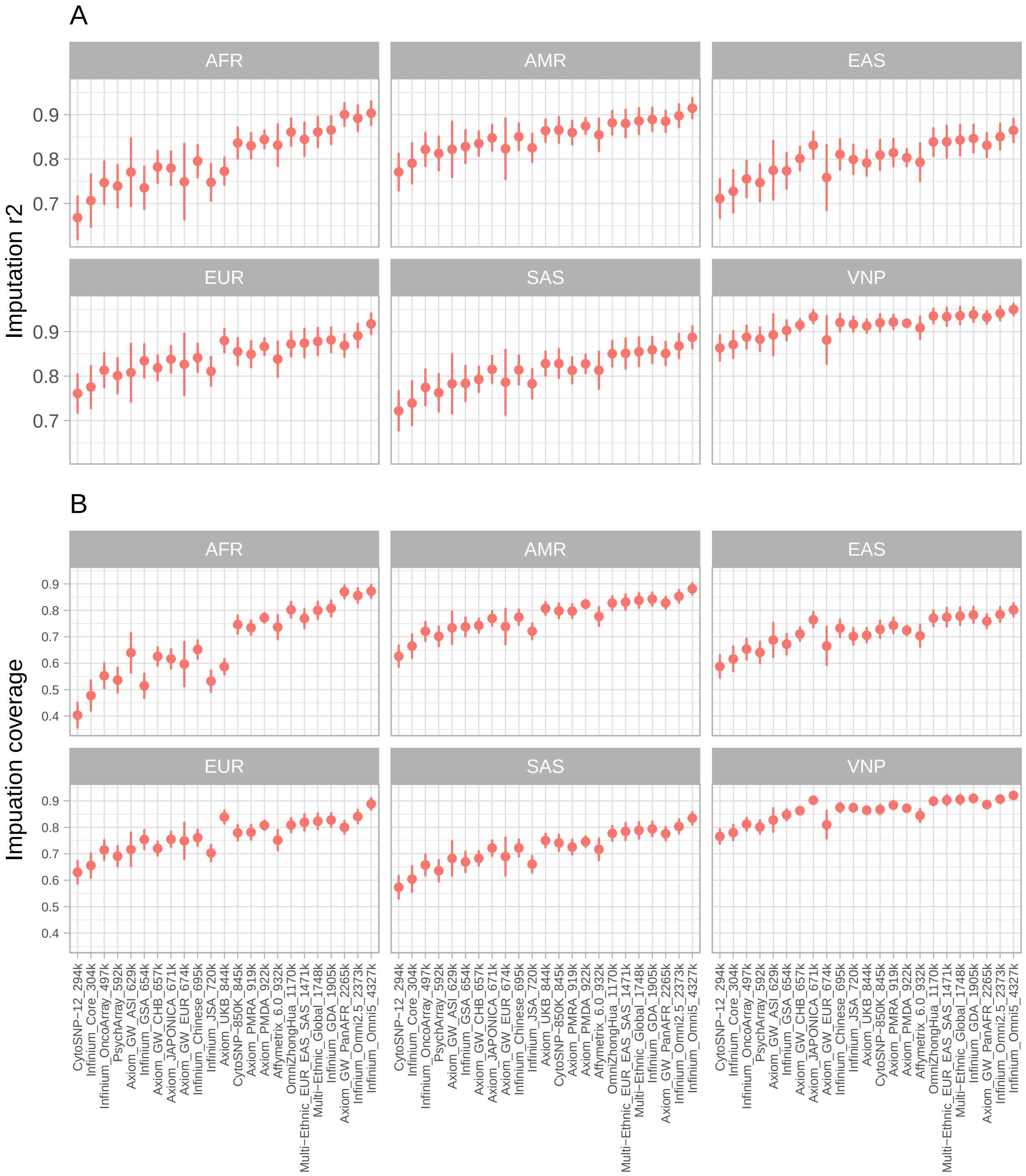

Figure 2. A. Mean imputation r^2, and B. Imputation coverage across 22 autosomes of 23 SNP arrays in the MAF bin of (0.01-0.5]. The dots and the vertical lines present the mean and the standard deviation of imputation accuracy, and imputation coverage values in 22 autosomes respectively.

Regarding population optimization, imputation performance are generally better for those arrays optimized specifically for their closely related populations. The Axiom_UKB, which is optimized for the British population, performs superior for the EUR than most other arrays, except for the ultra high density arrays Infinium Omni2.5 v1.5 and Infinium Omni5 v1.2. In detail, at the MAF bin of (0.01-0.5], The Axiom_UKB with size of 844k SNPs obtained the mean imputation coverage of 0.8389 which was higher than globally optimized, higher density arrays such as Axiom_PMRA (919k), Axiom_PMDA (922k), Affymetrix_6.0 (932k), Multi-Ethnic_Global (1784k), and Infinium_GDA (1905k), with imputation coverage of 0.7814, 0.8078, 0.7513, 0.8228, 0.8277, respectively and lower 0.8409, and 0.8885 that were obtained by Infinium Omni2.5 v1.5 and Infinium Omni5 v1.2 arrays with 2373k and 4327k SNPs. Similarly, the Axiom_JAPONICA (671k) which was designed for Japanese population also performed well against global optimized, higher-density arrays. These two arrays yielded mean imputation accuracy of 0.831, 0.9333 and imputation coverage of 0.7642, and 0.9024 in EAS and VNP populations. These performances were higher than those of multi-ethics SNP arrays, even with higher density including Axiom_PMRA (919k), Axiom_PMDA (922k), Affymetrix_6.0 (932k) as showed in Figure 2 and Table 2, 3. Notably, the Chinese GWAS array, OmniZhongHua, was also out performed in the EAS and VNP populations. Regarding the AFR population, an optimized population that is Axiom_GW_PanAFR with 2265k SNPs. The performance of this array is nearly equivalent the Infinium Omni5 v1.2 array with 4327k SNPs (0.9002 versus 0.9032 for mean imputation accuracy, and 0.8700 versus 0.8721 interns of imputation coverage). There were no SNP arrays with superior performances in the two remaining populations (AMR and SAS), although the Axiom_UKB and the Axiom_GW_AS obtained slightly better performance than other arrays with the same size, when applied for the AMR and SAS populations. In this case, we focused on the MAF bin of (0.01-0.5] as this is most common cutoff allele frequency in both GWAS and PGS analysis (Visscher et al., 2017; Choi et al., 2020). However, the results are also generalized for other bins as shown in Figure S.1 and Table S.1-6.

Table 2. Mean and the standard deviation of imputation accuracy r^2 measured in 22 autosomes at the MAF bin of (0.01-0.5].

Array name	AFR	AMR	EAS	EUR	SAS	VNP
CytoSNP-12	0.6682±0.0498	0.7708±0.043	0.7112±0.0452	0.7608±0.045	0.7218±0.0458	0.8635±0.0304
Infinium_Core	0.7066±0.0606	0.7906±0.047	0.7275±0.0494	0.7755±0.0491	0.7391±0.0512	0.871±0.0321
Infinium_OncoArray	0.7469±0.0498	0.8215±0.0388	0.7553±0.0428	0.8132±0.0402	0.7745±0.0422	0.8878±0.0275
PsychArray	0.7392±0.0496	0.8126±0.0398	0.7469±0.0437	0.801±0.0415	0.7625±0.0437	0.883±0.028
Axiom_GW_ASI	0.7707±0.0779	0.8218±0.0641	0.7749±0.0677	0.8078±0.0668	0.7826±0.0687	0.8927±0.0477
Infinium_GSA	0.7355±0.0491	0.8281±0.0386	0.7735±0.042	0.8342±0.0391	0.7835±0.0414	0.9026±0.0243
Axiom_GW_CHB	0.7824±0.0375	0.835±0.0292	0.8011±0.0281	0.8187±0.0299	0.7926±0.0303	0.9147±0.0152
Axiom_JAPONICA	0.7797±0.0392	0.8476±0.0307	0.831±0.0321	0.8379±0.0314	0.8149±0.0324	0.9333±0.0161
Axiom_GW_EUR	0.7489±0.0867	0.8233±0.0698	0.7584±0.0747	0.8264±0.0712	0.7861±0.0749	0.8813±0.0553
Infinium_Chinese	0.7953±0.0378	0.8502±0.0318	0.8106±0.0361	0.8412±0.0334	0.8139±0.0346	0.9205±0.0214
Infinium_JSA	0.7477±0.0431	0.8254±0.0332	0.7992±0.0346	0.8108±0.0342	0.7829±0.0349	0.9168±0.0184
Axiom_UKB	0.7726±0.0321	0.864±0.027	0.7913±0.0303	0.8797±0.0278	0.8281±0.0289	0.9126±0.0167
CytoSNP-850K	0.8363±0.0369	0.8654±0.0308	0.8091±0.0353	0.8552±0.032	0.8283±0.0339	0.9199±0.021
Axiom_PMRA	0.8299±0.0302	0.8597±0.0286	0.8143±0.0325	0.8493±0.0308	0.8128±0.0311	0.9217±0.0177
Axiom_PMDA	0.8443±0.0216	0.8744±0.0194	0.8032±0.021	0.8665±0.0203	0.8276±0.0226	0.9189±0.0104
Affymetrix_6.0	0.8315±0.0485	0.8545±0.0383	0.7929±0.0443	0.8382±0.0411	0.8131±0.0436	0.9087±0.0269
OmniZhongHua	0.8609±0.0326	0.8817±0.0283	0.8386±0.0327	0.8721±0.0295	0.8503±0.0307	0.9353±0.0185
Multi-Ethnic_EUR_EAS_SAS	0.8446±0.039	0.8801±0.0325	0.8388±0.0379	0.8742±0.0339	0.851±0.0357	0.9335±0.0222
Multi-Ethnic_Global	0.8611±0.0358	0.8856±0.0309	0.8428±0.0361	0.8779±0.0324	0.8548±0.0341	0.936±0.0212
Infinium_GDA	0.8654±0.0331	0.8893±0.028	0.8463±0.0329	0.8816±0.0295	0.8589±0.0308	0.9384±0.0182
Axiom_GW_PanAFR	0.9002±0.0273	0.8849±0.0257	0.831±0.0292	0.8689±0.0273	0.8509±0.0278	0.9322±0.0159
Infinium_Omni2.5	0.8919±0.0308	0.8975±0.0273	0.8503±0.031	0.891±0.0286	0.8678±0.0294	0.9412±0.0174
Infinium_Omni5	0.9032±0.0281	0.9144±0.0244	0.8644±0.0278	0.9176±0.0248	0.8873±0.0263	0.9499±0.0146

Table 3. Mean and standard deviation of imputation coverage (defined by the proportion of variants with $r^2 \geq 0.8$ over total number of variants in each chromosome) measured in 22 autosomes at the MAF bin of (0.01-0.5].

Array name	AFR	AMR	EAS	EUR	SAS	VNP
CytoSNP-12	0.4031±0.0578	0.6265±0.0621	0.5879±0.0553	0.6297±0.0603	0.5731±0.0584	0.7655±0.0508
Infinium_Core	0.4776±0.0883	0.6645±0.0688	0.6162±0.0604	0.6557±0.0645	0.6041±0.0656	0.7801±0.0494
Infinium_OncoArray	0.5525±0.0756	0.7205±0.0558	0.6533±0.052	0.7142±0.0529	0.6576±0.0546	0.8119±0.0419
PsychArray	0.5359±0.0744	0.7015±0.058	0.6405±0.0532	0.6909±0.0545	0.6358±0.0564	0.8012±0.0426
Axiom_GW_ASI	0.6394±0.1148	0.7339±0.0908	0.6876±0.0827	0.716±0.0863	0.6824±0.0872	0.8275±0.0749
Infinium_GSA	0.5145±0.0805	0.7373±0.0588	0.6719±0.0552	0.754±0.0558	0.6692±0.0579	0.8476±0.0404
Axiom_GW_CHB	0.6257±0.0572	0.7428±0.0376	0.7102±0.0312	0.7198±0.0351	0.6826±0.0356	0.8628±0.0209
Axiom_JAPONICA	0.6165±0.0648	0.7694±0.0444	0.7642±0.0396	0.7547±0.0415	0.7215±0.0425	0.9024±0.0271
Axiom_GW_EUR	0.5963±0.1214	0.7386±0.0991	0.665±0.0889	0.7489±0.0939	0.6893±0.0953	0.8096±0.0836
Infinium_Chinese	0.6518±0.0606	0.774±0.043	0.7324±0.043	0.7611±0.0416	0.7218±0.0435	0.8752±0.0317
Infinium_JSA	0.5316±0.0674	0.7207±0.0494	0.7017±0.0462	0.703±0.0468	0.6603±0.0476	0.8745±0.032
Axiom_UKB	0.5866±0.0478	0.8072±0.0352	0.7058±0.0348	0.8389±0.0343	0.75±0.035	0.8646±0.0252
CytoSNP-850K	0.7459±0.0568	0.7979±0.0401	0.7283±0.0415	0.7796±0.0393	0.7413±0.0423	0.8673±0.0312
Axiom_PMRA	0.7342±0.0423	0.7969±0.0374	0.7432±0.0388	0.7814±0.0385	0.7253±0.0377	0.8842±0.0274
Axiom_PMDA	0.7717±0.0295	0.8234±0.0232	0.724±0.0256	0.8078±0.0243	0.7458±0.0291	0.8721±0.0176
Affymetrix_6.0	0.7364±0.0816	0.7768±0.0574	0.704±0.0559	0.7513±0.0564	0.7168±0.0593	0.8445±0.0454
OmniZhongHua	0.8019±0.0461	0.8275±0.0349	0.7701±0.0373	0.8085±0.0348	0.7773±0.0362	0.899±0.0273
Multi-Ethnic_EUR_EAS_SAS	0.7692±0.0608	0.8309±0.0414	0.774±0.045	0.8189±0.042	0.7846±0.0444	0.9029±0.0323
Multi-Ethnic_Global	0.7997±0.0556	0.8376±0.0395	0.7781±0.0434	0.8228±0.0405	0.7886±0.0428	0.9057±0.0307
Infinium_GDA	0.8074±0.0505	0.8425±0.036	0.7826±0.0396	0.8277±0.0369	0.7938±0.0391	0.9094±0.0275
Axiom_GW_PanAFR	0.87±0.0334	0.8281±0.0301	0.7588±0.0314	0.8001±0.031	0.7759±0.0316	0.8862±0.0228
Infinium_Omni2.5	0.856±0.0393	0.8529±0.0327	0.7842±0.0344	0.8409±0.0338	0.8034±0.0342	0.907±0.0253
Infinium_Omni5	0.8721±0.0349	0.8813±0.0288	0.8019±0.0315	0.8885±0.0294	0.8344±0.0315	0.9207±0.0219

3.2 PGS performance

We evaluated PGS performance of these arrays based on two criteria: (i) Pearson's correlation of PGSs estimated by using imputed SNP array data compared to the PGSs estimated by using WGS data - hereafter we refer as PGS correlation for short, (ii) absolute difference of percentile ranking (ADPR) between PGSs generated by array-imputed and gold standard WGS. Both comparisons are set under various p-value cutoffs that allows us unbiased evaluate PGS performance of these arrays. In general, we found that PGS performance were highly concordant with imputation performance, i.e. SNP arrays with better imputation performance showed higher correlation with WGS PGSs and less ADPR than the arrays with poor imputation performance.

The summary results of Pearson's correlation values of PGSs from 23 genotyping SNP arrays for three different phenotypes are shown in Figure 3 and in Tables S.7-9. In general, all examined arrays yielded high PGS correlations. Notably, the vast of majority PGS correlations ranged from 0.90 to 0.99, except for the Infinium HumanCytoSNP-12 v2.1 which had the lowest values. Interestingly, when optimal arrays for populations were used, the PGS correlation to WGS was higher than 0.95. The PGS correlation patterns were also highly concordant in all three evaluated traits with comparable performances. As expected, SNP arrays with larger sizes had higher PGS correlations. The lowest performer was the Infinium HumanCytoSNP-12 v2.1 with the correlation of 0.8731 in the height phenotype in the AFR population while the highest performance was obtained by the Infinium Omni5 v1.2 with the correlation higher than 0.99 in all populations and traits. We also examined the deviation of PGS correlation in various

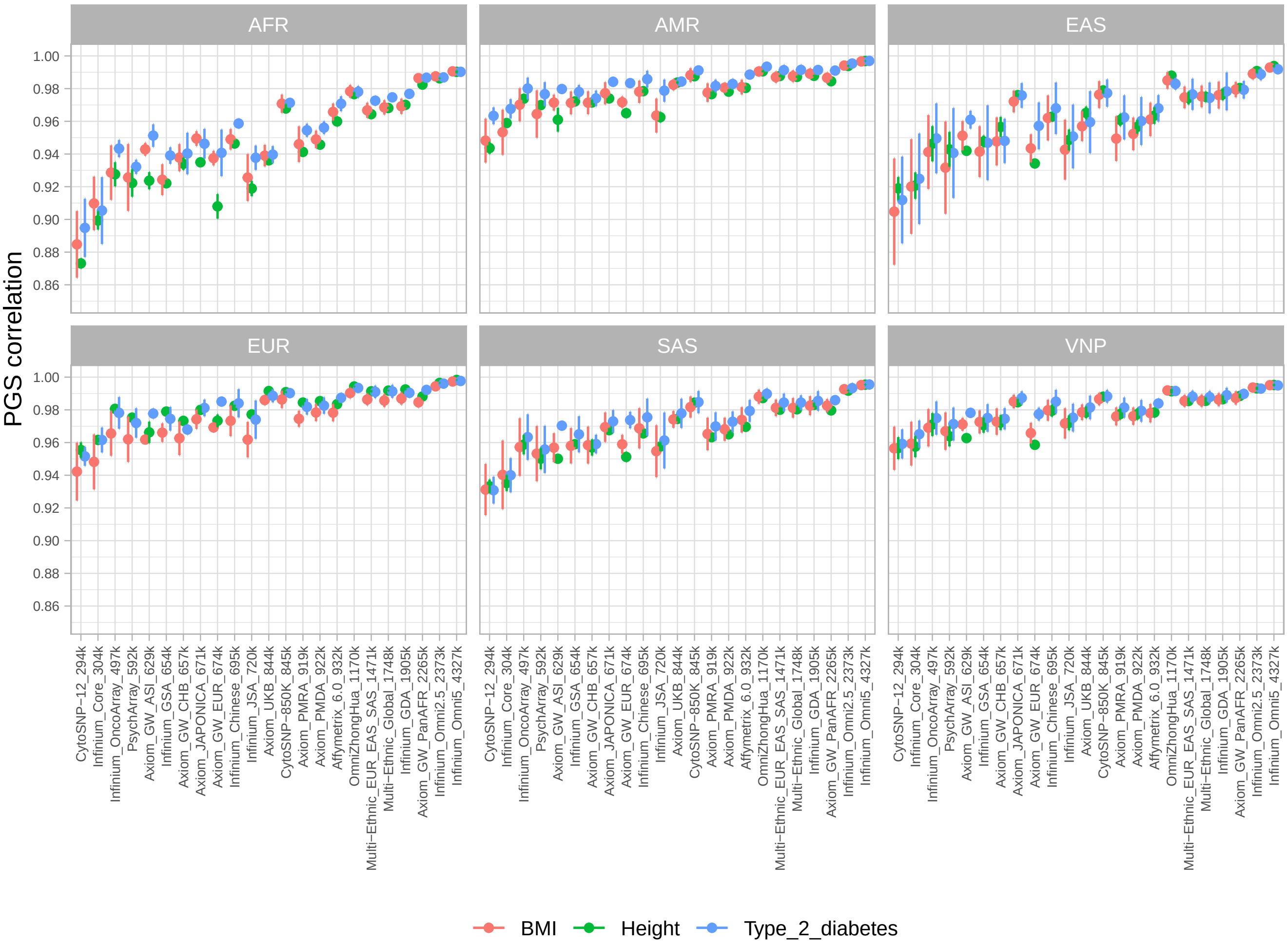

Figure 3. Correlations between PGSs estimated from imputed genotyping data of 23 SNP arrays and PGSs estimated from WGS in six different populations with three phenotypes including height, BMI, and type 2 diabetes. The dots and the vertical lines present the mean and standard deviation of PGS correlation at various p-value settings including 5e-08, 1e-07, 1e-06, 1e-05, 0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.5, and 1.

p-value settings. The results showed that SNP array with lower PGS correlation had higher PGS correlation standard deviation than the high performance arrays, A possible explanation for this observation is the PGSs estimated from low imputation performance are more vulnerable by the random pruning process than the high imputation performance arrays (Choi and O'Reilly, 2019). Notably, we also observed the higher standard deviations of PGS correlation in EAS than other populations.

In agreement with imputation performance, SNP arrays optimized specifically for targeted populations showed supperior PGS correlation in the targeted/closely related populations. For instance, Axiom Japonica Array NEO and Infinium OmniZhongHua v1.4 that was optimized for Japanese, and Chinese showed clear advantages in the populations of EAS, and VNP while Axiom UK Biobank Array yielded higher PGS correlation in the EUR population than the other size-comparable genotyping arrays. Taking height as a typical trait of interest, PGS correlations of the Japonica Array NEO were 0.9760, and 0.9847, while

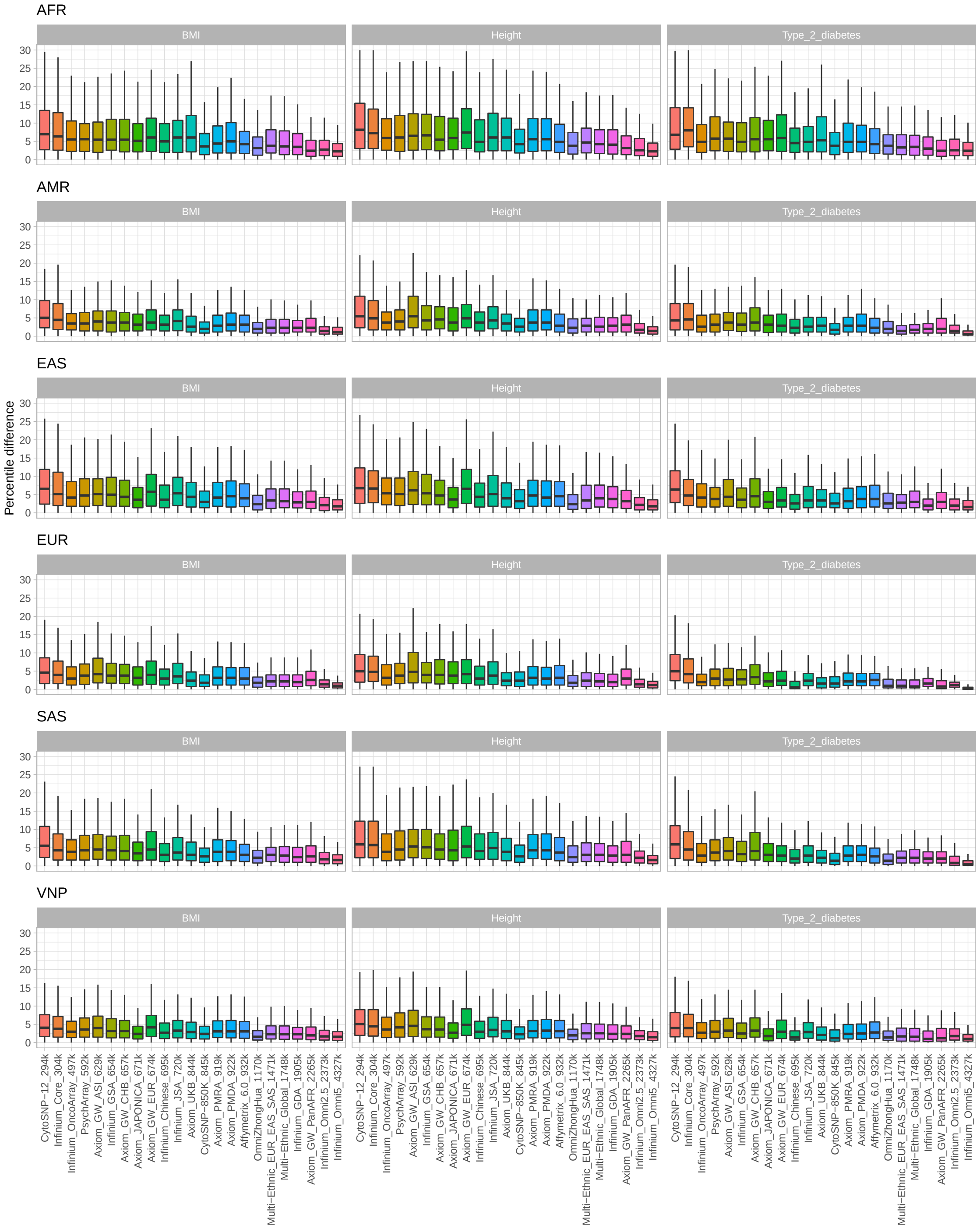

Figure 4. The absolute difference of percentile ranking between PGSs estimated from imputed genotyping data of 23 SNP arrays and PGSs estimated from WGS in six different populations. The figure shows results of three phenotypes including height, BMI, and type two diabetes with PRsice p-value setting of 5e-08.

the Infinium OmniZhongHua v1.4 had 0.9879, and 0.9914 in EAS and VNP respectively. Interestingly, we observed that the Infinium CytoSNP-850K v1.2 was the array with superior PGS correlations in all populations, for all the three evaluated traits. For example, the PGS correlation for this array for height phenotype in AFR, AMR, EAS, EUR, SAS and VNP were 0.9679, 0.9876, 0.9789, 0.9908, 0.9844, 0.988, respectively.

Regarding the ADPR metric, the performance of arrays was in an agreement with the trend from comparing PGS correlation. ADPR in different PRSice-2 p-values settings are shown in Figure **??**, Figure S.2-12 and reported in detail in Table S.10-21. Most of the arrays yielded mean ADPR less than 10 in all three traits. An exception was the AFR populations with low density arrays. The highest density array, i.e. Infinium Omni5 v1.2, had the highest performance with ADPR less than 4. Notably, ADPRs varied by populations. Under-represented populations like AFR, and EAS tended to exhibit higher ADPRs than the others. Taking the p-value cutoff at 5e-8 for the height phenotype as an example (Figure **??**), Infinium Omni5 v1.2 obtained ADPRs of 3.8600, 2.4774, 2.8884, 1.9758, 2.8391, and 2,3699 in AFR, AMR, EAS, EUR, SAS and, VNP respectively. A consistent trend was also observed in other traits, with the lowest performance in AFR and the highest performance in EUR with ADPRs of 3.5974 and 1.8489 in BMI, and of 3.7206 and 1.6592 in type 2 diabetes. Similar to the other experiments, population specific arrays and the Infinium CytoSNP-850K v1.2 also illustrated their advantages when comparing the ADPR metric. The Axiom UK Biobank Array obtained good performance for the EUR population with ADPR of 3.0584, 3.1714, and 2.2734 in height, BMI, and type 2 diabetes respectively. This trend was also observed in the cases of Axiom Japonica Array NEO, and Infinium OmniZhongHua v1.4 applied for the EAS and VNP populations. Regarding the Infinium CytoSNP-850K v1.2 array, good performances in all traits and populations were observed. Specifically, ADPRs for height were 5.7141, 3.4914, 4.3753, 3.2501, 3.7638, 3.0267; for BMI at 4.9872, 2.5463, 4.1560, 2.6272, 3.5409, 3.1523; and for type 2 diabetes at 5.2000, 2.5762, 3.7687, 2.6066, 2.4707, 2.3812 in AFR, AMR, EAS, EUR, SAS and, VNP, respectively, all at the same p-value cutoff.

3.3 Performance evaluation using real genotyping data

We further utilize the availability of real genotyping data in the 1KVG dataset (95 out of the 1008 samples were also genotyped by Affymetrix PMRA array) to investigate how our simulated array data performed relative to the real array data. In brief, we simulated genotyping data of 95 samples by extracting variants from WGS data that matched with PMRA manifest, excluding phasing information. We then applied the same evaluation pipeline to compare the performance using simulated genotyping data against the results from the real genotyping data. In details, both simulated and real PMRA genotyping data were phased with SHAPEIT v4.1.3 (Delaneau et al., 2019), and imputed with Minimac4 v1.0.2 (Das et al., 2016). Reference data for both phasing and imputation were the remaining 913 WGS samples. Finally, imputation performance of both simulated and real PMRA arrays were estimated as described in the “Imputation performance evaluation” section. As expected, the imputation accuracies of simulated and real PMRA were

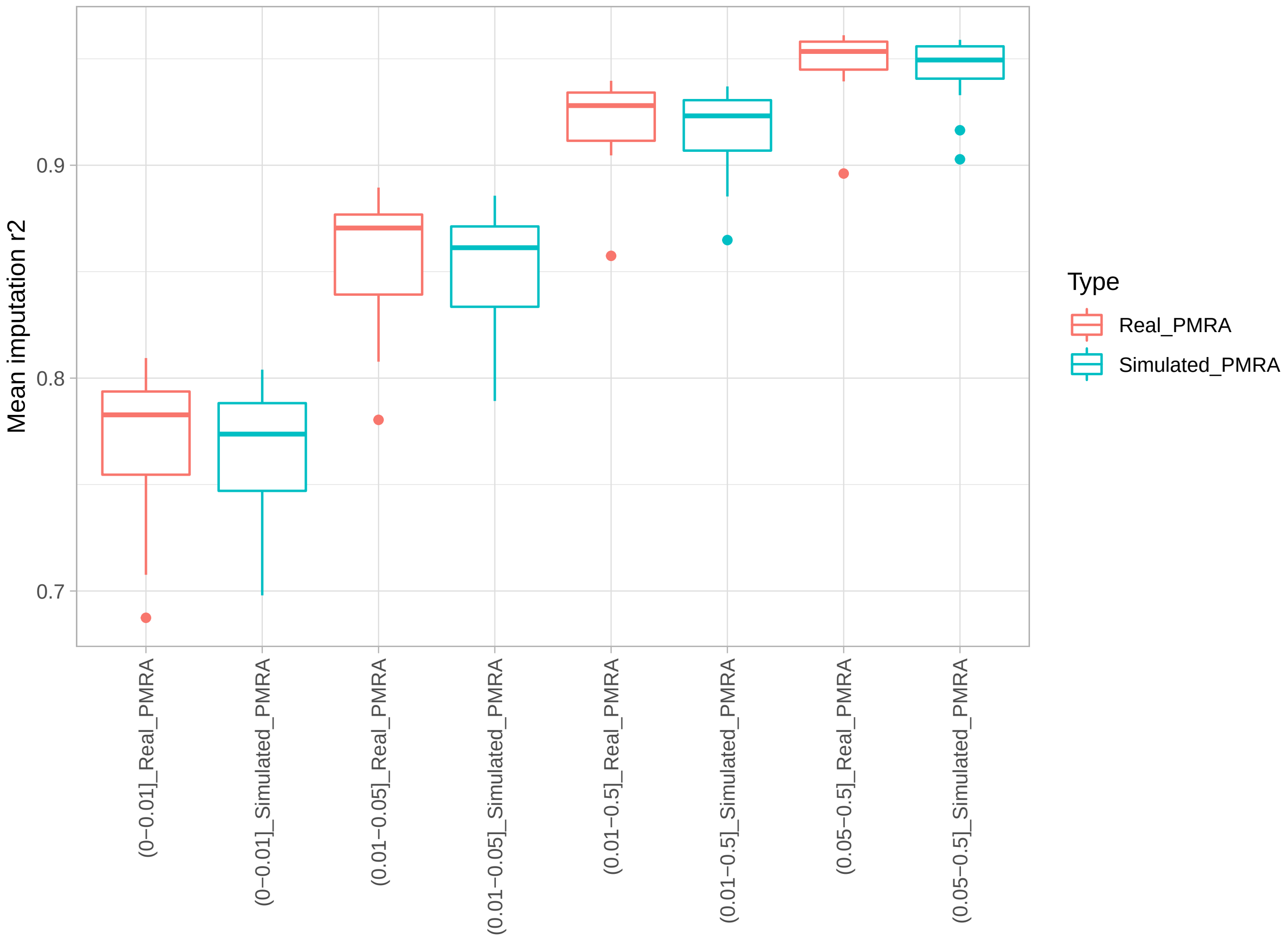

Figure 5. Mean imputation accuracy comparisons of simulated PMRA and real genotyped PMRA of 95 VNP samples at various MAF bins measured in 22 autosomes.

highly concordant as shown in Figure 5 and Table 4. In details, mean and standard deviation of imputation accuracies of simulated PMRA are 0.7664±0.0309, 0.8522±0.0269, 0.9165±0.0179, and 0.9453±0.014; real PMRA are 0.7721±0.032, 0.8592±0.0278, 0.9217±0.0181, and 0.9497±0.0138 in four MAF bins of (0-0.01], (0.01-0.05], (0.01-0.5], and (0.05-0.5], respectively. These results indicated the robustness of our simulation approach for estimating imputation performances of genotyping arrays in reality.

Table 4. Mean and the standard deviation of imputation accuracies of simulated PMRA and real genotyped PMRA of 95 VNP samples at various MAF bins measured in 22 autosomes.

MAF range	Simulated PMRA	Real PMRA
(0-0.01]	0.7664±0.0309	0.7721±0.032
(0.01-0.05]	0.8522±0.0269	0.8592±0.0278
(0.01-0.5]	0.9165±0.0179	0.9217±0.0181
(0.05-0.5]	0.9453±0.014	0.9497±0.0138

4 DISCUSSIONS AND CONCLUSIONS

Even in a booming time of next-generation sequencing technologies, current big genotyping projects are still using SNP arrays as the work-horse for generating valuable data, especially for bio-bank scale projects (Bycroft et al., 2018; Moon et al., 2019; Sakurai-Yageta et al., 2020). Moreover, genotyping by SNP arrays produce the exact information type required for PGS analysis, which is based on summarizing effect sizes from individual SNPs. A promising application of genomic research that is gaining increasing interest recently across the health-care system, and direct-to-consumer genomic services based on polygenic scoring like 23andMe (Lewis and Vassos, 2020; Folkersen et al., 2020). SNP arrays are clearly more economical in data generation and analysis, an important factor in designing project with a large sample size and/or with a limited budget. Given that there are many available human genotyping arrays optimized for various purposes, a comprehensive guideline for choosing the most suitable SNP arrays in multiple ancestry groups is still lacking. To address this gap, we have developed a systematic approach and tool to assess a large range of SNP arrays across multiple datasets. We performed imputation and PGS performance assessments for 23 human available genotyping arrays in six ancestry groups using both public and in-house datasets by various metrics. By comparing relative performance of SNP arrays relative to that of WGS with 4 metrics including imputation accuracy, imputation coverage, PGS correlation, and ADPR, we found expected trends and our results suggested suitable arrays that can maximize PGS performance for specific populations, especially for under-represented populations.

Overall, we found that all 23 assessed arrays had high performances in both imputation, and especially in PGS. These commercial arrays differ markedly in designs, i.e. the number of markers on the arrays and targeted ancestry groups that would cause performance deviations. An important finding in our analysis was that in order to obtain high imputation performances, the choice of array is not necessarily about getting higher density, but small to moderately-sized arrays (approximately 650k-850k tag SNPs), accompanied by well optimization for the targeted population could also produce high PGS performances. For example, the Japonica Array NEO, and the UK Biobank Array showed the highest performance when comparing with other arrays with the same sizes for EAS, and EUR populations respectively. This indicates that using customized, small-size SNP arrays at population-specific level can be a cost-effective genotyping solution without loosing PGS performance (Tam et al., 2019; Nguyen et al., 2022). We also observed that there are no specific arrays with moderate sizes that had superior imputation performances in AFR, and SAS, suggesting the need for genotyping arrays optimized for these populations. PGS performances were concordant to imputation performances in general. However, CytoSNP-850K v1.2 was an interesting array that showed superior PGS performances in all populations. This superior performance may be explained by the enrichment of cytogenomic regions in the design of the Infinium CytoSNP-850K v1.2 array (Illumina,

????). In agreement with previous studies (Martin et al., 2019; Sirugo et al., 2019), our analyses also show that underrepresented populations such as AFR, and SAS exhibited lower PGS performances (and ADPRs tended to be higher in AFR, and SAS) than other well-studied populations regardless the sample sizes of these populations are not significant different.

Notably, PGS performances of array constructed from imputed genotypes were very high in comparison with the original WGS PGS. The majority PGS correlations ranged from 0.90 to 0.99. In cases of optimal arrays for targeted populations in used (UK Biobank Array is used for EUR, Japonica Array NEO is used for EAS), the PGS correlation to WGS was higher than 0.97. In addition, PGS ranking differences between WGS and imputed array genotypes were not high with the majority of differences were under 5 percentile when optimal arrays were used. The possible reason for this observation was that current GWAS summary statistics were mostly generated by imputed array genotypes (Xue et al., 2018; Yengo et al., 2018) that are limited to detect rare associated markers. This indicates that using WGS for PGS analysis does not provide significant improvement interm of disease risk stratification at this time although this trend can change in the future when GWAS summary statistics at higher resolution become widely available (Wainschtein et al., 2022).

Finally, to make this analysis capability available to a broad audience, we developed a web application that provides interactive analyses SNP array contents and performances. As researchers may be interested in specific variants or regions, the application is aimed to support researchers to analyze SNP array contents and imputation performance based on population and genomic regions of interest. We hope that application will facilitate researchers in designing their genetic studies.

CONFLICT OF INTEREST STATEMENT

There is NO Competing Interest.

AUTHOR CONTRIBUTIONS

DTN initiated the study, designed experiments, analyzed data, interpreted results, developed the web tool, and drafted the manuscript. TT, MT, and NTD contributed to the 1KVG data generation and preprocessing. KT, DP, QN, and NSV contributed to the discussion, design and interpretation. QN and NSV critically revised the manuscript, coordinated the project, and supervised the study. All authors have read and approved the final manuscript.

FUNDING

This work is funded by Vingroup Big Data Institute internal funding, and partly supported by the Vingroup Innovation Foundation under grant VINIF.DA.2020.02.

ACKNOWLEDGMENTS

We especially thank Nguyen T. Nguyen for his kindly help in downloading the 1KGP-NYGC datasets, Mr Hoang H. Ho for the help with deploying the web tool. We also thank the Vingroup Big Data Institute for providing computational resources.

DATA AVAILABILITY STATEMENT

The 1KGP-NYGC datasets are freely available at IGSR data portal (`https://www.internationalgenome.org`). The 1KVG WGS and PMDA array datasets are available under agreement at MASH data portal (`https://genome.vinbigdata.org/`). Data and source codes to generate figures of this study are available at: `https://github.com/datngu/SNP_array_comparison`. SNP array analyzing tool is available online at: `https://genome.vinbigdata.org/tools/saa/`. SNP-wise imputation performance estimation based on 1KGP-NYGC are freely available at: `https://zenodo.org/record/6548098`. SNP-wise imputation performance estimation based on 1KVG are available and can be supplied under ethical policy agreement.

REFERENCES

Auton, A., Abecasis, G., Altshuler, D., Durbin, R., Abecasis, G., Bentley, D., et al. (2015). A global reference for human genetic variation. nature [internet]

Barrett, J. C. and Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nature genetics* 38, 659–662

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* 47, D1005–D1012

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., et al. (2021). High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *bioRxiv*

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *The American Journal of Human Genetics* 74, 106–120

Chapman, J. M., Cooper, J. D., Todd, J. A., and Clayton, D. G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Human heredity* 56, 18–31

Chen, S.-F., Dias, R., Evans, D., Salfati, E. L., Liu, S., Wineinger, N. E., et al. (2020). Genotype imputation and variability in polygenic risk score estimation. *Genome medicine* 12, 1–13

Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* 15, 2759–2772

Choi, S. W. and O'Reilly, P. F. (2019). Prsice-2: Polygenic risk score software for biobank-scale data. *Gigascience* 8, giz082

Consortium, G. P., Auton, A., Brooks, L., Durbin, R., Garrison, E., and Kang, H. (2015). A global reference for human genetic variation. *Nature* 526, 68–74

Consortium, I. H. et al. (2007). A second generation human haplotype map of over 3.1 million snps. *Nature* 449, 851

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and vcftools. *Bioinformatics* 27, 2156–2158

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation genotype imputation service and methods. *Nature genetics* 48, 1284–1287

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., and Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature communications* 10, 1–10

Ehli, E. A., Abdellaoui, A., Fedko, I. O., Grieser, C., Nohzadeh-Malakshah, S., Willemsen, G., et al. (2017). A method to customize population-specific arrays for genome-wide association testing. *European Journal of Human Genetics* 25, 267–270

Folkersen, L., Pain, O., Ingason, A., Werge, T., Lewis, C. M., and Austin, J. (2020). Impute. me: an open-source, non-profit tool for using data from direct-to-consumer genetic testing to calculate and interpret polygenic risk scores. *Frontiers in genetics* 11, 578

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., et al. (2003). The international hapmap project

Ha, N.-T., Freytag, S., and Bickeboeller, H. (2014). Coverage and efficiency in current snp chips. *European Journal of Human Genetics* 22, 1124–1130

Hoffmann, T. J., Kvale, M. N., Hesselson, S. E., Zhan, Y., Aquino, C., Cao, Y., et al. (2011a). Next generation genome-wide association tool: design and coverage of a high-throughput european-optimized snp array. *Genomics* 98, 79–89

Hoffmann, T. J., Zhan, Y., Kvale, M. N., Hesselson, S. E., Gollub, J., Iribarren, C., et al. (2011b). 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, 422–430

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics* 44, 955–959

Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., et al. (2015). Improved imputation of low-frequency and rare variants using the uk10k haplotype reference panel. *Nature communications* 6, 1–9

[Dataset] Illumina (????). Infinium cytosnp 850k genotyping array

Kawai, Y., Mimori, T., Kojima, K., Nariai, N., Danjoh, I., Saito, R., et al. (2015). Japonica array: improved genotype imputation by designing a population-specific snp array with 1070 japanese individuals. *Journal of human genetics* 60, 581–587

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics* 50, 1219–1224

Lambert, S. A., Abraham, G., and Inouye, M. (2019). Towards clinical utility of polygenic risk scores. *Human molecular genetics* 28, R133–R142

Lewis, C. M. and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. *Genome medicine* 12, 1–11

Li, J. H., Mazur, C. A., Berisa, T., and Pickrell, J. K. (2021). Low-pass sequencing increases the power of gwas and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Research* 31, 529–537

Lindquist, K. J., Jorgenson, E., Hoffmann, T. J., and Witte, J. S. (2013). The impact of improved microarray coverage and larger sample sizes on future genome-wide association studies. *Genetic epidemiology* 37, 383–392

Marchini, J. (2019). Haplotype estimation and genotype imputation. *Handbook of Statistical Genomics: Two Volume Set* , 87–114

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* 11, 499–511

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., et al. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research* 27, e1608

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics* 51, 584–591

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* 48, 1279–1283

Moon, S., Kim, Y. J., Han, S., Hwang, M. Y., Shin, D. M., Park, M. Y., et al. (2019). The korea biobank array: design and identification of coding variants associated with blood biochemical traits. *Scientific reports* 9, 1–11

Nelson, S. C., Doheny, K. F., Pugh, E. W., Romm, J. M., Ling, H., Laurie, C. A., et al. (2013). Imputation-based genomic coverage assessments of current human genotyping arrays. *G3: Genes, Genomes, Genetics* 3, 1795–1807

Nguyen, D. T., Dinh, H. Q., Vu, G. M., Nguyen, D. T., and Vo, N. S. (2021). A comprehensive imputation-based evaluation of tag snp selection strategies. In *2021 13th International Conference on Knowledge and Systems Engineering (KSE)* (IEEE), 1–6

Nguyen, D. T., Nguyen, Q., Nguyen, D. T., and Vo, N. S. (2022). Lmtag: functional-enrichment and imputation-aware tag snp selection for population-specific genotyping arrays. *bioRxiv*

Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics* 69, 1–14

Sakurai-Yageta, M., Kumada, K., Gocho, C., Makino, S., Uruno, A., Tadaka, S., et al. (2020). Japonica array neo with increased genome-wide coverage and abundant disease risk snps. *bioRxiv*

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell* 177, 26–31

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* 20, 467–484

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* 19, 581–590

Tran, H. et al. (2022). 1000 vietnamese genome project paper. *In-preparation*

Verlouw, J. A., Clemens, E., de Vries, J. H., Zolk, O., Verkerk, A. J., am Zehnhoff-Dinnesen, A., et al. (2021). A comparison of genotyping arrays. *European Journal of Human Genetics* , 1–14

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of gwasn discovery: biology, function, and translation. *The American Journal of Human Genetics* 101, 5–22

Wainschtein, P., Jain, D., Zheng, Z., Cupples, L. A., Shadyab, A. H., McKnight, B., et al. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics* 54, 263–273

Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K. E., Zheng, Z., et al. (2018). Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature communications* 9, 1–14

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. *Human molecular genetics* 27, 3641–3649

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., and Wang, L. (2014). Crossmap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 30, 1006–1007