ARTICLE

Securing the use of existing sample collections for future human genetic research

George Kanouni, Peter Nürnberg and Michael Nothnagel*

Hundreds of thousands of individuals have been genotyped in the past decades using genotyping arrays, representing both a valuable data resource for future biomedical research and a substantial investment in human genetic research. However, novel chip designs and their altered sets of single-nucleotide polymorphisms (SNPs) pose the question of how well established data resources, such as large samples of healthy controls genotyped on legacy arrays, can be combined with newer samples genotyped on those novel arrays using genotype imputation. We exemplarily investigated this question based on genotype data of 30 European and 30 African unrelated samples from the 1000 Genomes project and on markers present on two legacy SNP arrays, namely Affymetrix's Human SNP 6.0 and Illumina's 550k array, and three newer arrays, namely two Axiom arrays from Affymetrix and an OmniExpress array from Illumina. We cross-compared the imputation accuracy as well as efficacy and assessed genotype concordance among these arrays. Although the accuracy of genotype prediction was uniformly high across all arrays, the imputation efficacy, that is, the proportion of successfully imputed markers, differed considerably between array combinations in both sample sets, with legacy arrays showing a trend towards lower efficacy values compared with newer arrays when serving as imputation basis. We conclude that, given the substantial losses of markers covered by the legacy arrays, the re-genotyping of existing samples sets, in particular those of healthy population controls, would be a worthwhile endeavor to secure their continued use in the future.

European Journal of Human Genetics (2017) **25**, 522–529; doi:10.1038/ejhg.2017.1; published online 1 February 2017

INTRODUCTION

More than a decade of genome-wide association studies (GWAS) have helped identifying several thousands of genetic variants showing robust association with a single or multiple diseases or traits, thereby often improving the understanding of their etiology. The basis for the endeavor was the genotyping of very large sample collections for hundreds of thousands of single-nucleotide polymorphisms (SNPs), producing a wealth of genetic information. Not the least do these DNA sample collections document a tremendous investment in collecting and phenotyping of samples in the past. The joint availability of phenotypic and genotypic information offers a high potential to further genetic research into human diseases even in the future.

However, most of these collections have been genotyped on SNP arrays that are being used with decreasing frequency. For example, the GeneChip Human Mapping 500 K array set and the Genome-wide Human SNP arrays 5.0 and 6.0 by Affymetrix (Santa Clara, CA, USA) as well as the HumanHap550 Bead array (550k) by Illumina (San Diego, CA, USA) have been routinely used a few years ago. In fact, this also applies to a number of DNA sample collections that have provided large sets of convenience controls for past GWAS, including for example the PopGen¹ and KORA² biobanks in Germany. These SNP arrays have been replaced by numerous newer array versions that were designed for particular tasks and that feature different SNP selections and partly different genotyping chemistries. This includes, for example, HumanOmniExpress, Infinium PsychChip (Psychiatric Genomics Consortium, pgc.unc.edu), MetaboChip³ and ImmunoChip⁴ by Illumina and the Axiom Genome-Wide Human Origins 1, Axiom

CEU and Axiom UK BioBank arrays by Affymetrix. A future study may aim at combining SNP data of patients with a particular disease, recently genotyped on the Axiom UK BioBank array, with legacy convenience control data genotyped on the Affymetrix SNP array 6.0. It is yet unclear how well genetic information obtained from these different types of chips can be combined.^{5–7} This raises the issue that those existing large DNA collections may become increasingly inaccessible to future genetic research.

A standard approach to combining SNP data obtained from different arrays is genotype imputation. A number of different software solutions (e.g., refs 8–12) and reference panels (e.g., refs 13–17) are available for this task. Past benchmarking studies and review papers documented very accurate imputation for a majority of markers in cross-comparisons between arrays for the same population (e.g., refs 18–22), although Uh *et al.*⁵ reported that some discrepancies are common to appear in such situations.

In this study, we aimed at quantifying the potential information loss due to aging chip designs used in legacy sample collections. To this end, we considered the exemplary case of three Affymetrix and two Illumina arrays, namely the Genome-wide Human SNP array 6.0 and the 550k as a legacy arrays and the Axiom CEU array, the Axiom UK BioBank array and the HumanOmniExpress array as currently used ('novel') arrays, and evaluated the imputation performance separately in 30 unrelated individuals of European ancestry and in 30 individuals of African ancestry. Data for each of the three arrays were obtained from the 1000 Genomes resource.²³ We observed substantial proportions of markers on the more recent arrays that cannot be reliably imputed from the legacy arrays. Furthermore, we found the overlap

between the quality-controlled sets of imputed markers based on a legacy and on a novel array to be only partial. Our results indicate that it may be a worthwhile endeavor to re-genotype selected existing DNA sample collections, with their wealth of phenotypic information, with newer genotyping arrays to guarantee their future accessibility and use.

MATERIALS AND METHODS

DNA sample data

All SNP genotype data were obtained from the 1000 Genomes²³ website (www.1000genomes.org; NCBI build 37 (hg19); <ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/>). We obtained data for 30 unrelated samples of CEPH Utah residents of northern and western European ancestry (CEU) as well as 30 unrelated African Yoruba samples (YRI); see Supplementary Table S1 for the corresponding identifiers. We extracted five data sets for each of the two ethnic sample sets, namely those markers that are present on the Affymetrix GeneChip Genome-Wide Human SNP 6.0 (A6), on the Illumina 550k, on the Axiom Genome-Wide CEU (ACEU), on the Axiom UK-BioBank (UKBB) and on the HumanOmniExpress v24.1 (HOE), respectively, using in-house scripts. SNP sets only partially overlapped between the arrays (Supplementary Table S2).

Quality control

Each data set was separately quality-controlled using PLINK v1.07,²⁴ although a very high quality could be expected. A call rate of >95% was observed for all individual samples and of >96% for all markers. Potential duplicates and relatedness between individuals were ruled out analyzing average identity-by-descent allele sharing between pairs of samples (PI-HAT statistic in PLINK). Markers not located on an autosome or with *P*-values <10⁻⁴ in a test for deviations from Hardy-Weinberg equilibrium were removed from further analysis. SNP markers with ambiguous strand orientation were also excluded from the study. All samples were included in the final data sets. Supplementary Table S2 lists the total numbers of markers before and after quality control for each of the different array data sets, while Supplementary Tables S3 and S4 contain the numbers of polymorphic markers and their corresponding allele frequency class for CEU and YRI, respectively.

Genotype imputation

We used the IMPUTE v2.3.2 software¹² for genotype imputation. As recommended on the IMPUTE website (https://www.mathgen.stats.ox.ac.uk/impute/impute_v2.html), we pre-phased our study data sets using the SHAPEIT v2.5 software.²⁵ We used the 1000 Genomes²³ phase 3 data as reference panel for the pre-phasing and later for the imputation (NCBI build 37, hg19, downloaded on October 2014 from the IMPUTE website). As all of the individuals in our study are part of this reference panel, imputation performance measures may be upwardly biased. We therefore manually removed the genotype information on all samples in the reference that relate to our study samples and their relatives in order to avoid such bias. More specific, we excluded 50 CEU individuals (30 study samples and 20 relatives) and 49 YRI individuals (30 study samples and 19 relatives), leaving 72 CEU and 111 YRI samples in the modified imputation reference, respectively. Thus, we eventually used information on 2454 and 2455 (out of 2504) samples as imputation reference for our study CEU and YRI samples, respectively. Using a default option in IMPUTE2, chromosomal data were split into chunks of 5 Mb size for conducting the imputation.

Assessment of imputation performance

When combining information from existing sample collections with those from new ones, one is usually interested either in one or more specific markers which may have originated as likely candidates from a previous study or in a genome-wide approach for novel hypothesis generation. To answer both requests, we evaluated the imputation performance by two different approaches. We addressed the first request using a marker-on-array imputation approach, where we assessed how well markers present on one array ('target') could be imputed using the information from another array ('base'). The marker-on-array imputation approach quantifies the information loss for a situation where

Table 1 Imputation efficacy for novel arrays from legacy arrays in 30 CEU samples

Base legacy	Target novel			ACEU			UKBB			HOE		
	R ²	MAF	IMPP	0.3	0.5	0.8	0.3	0.5	0.8	0.3	0.5	0.8
A6	0-0.05	93	89	71	90	87	70	75	74	66	90	88
	0.05-0.1	97	92	80	96	82	80	90	88	79	96	88
	0.1-0.5	94	88	75	94	88	75	92	87	75	91	82
550k	0-0.05	85	82	67	83	80	66	69	69	62	88	85
	0.05-0.1	86	83	74	86	83	74	80	79	73	90	84
	0.1-0.5	69	65	57	69	65	57	67	64	57	51	47

Abbreviations: 550k, HumanHap550 Bead; A6, GeneChip Genome-Wide Human SNP 6.0; ACEU, Axiom Genome-Wide CEU; base legacy, array whose observed genotypes were used as imputation base; CEU, CEPH Utah residents of northern and western European ancestry; HOE, HumanOmniExpress; MAF, minor allele frequency of the marker; IMPP, average maximum posterior probability; R², info score; target novel, array whose observed genotypes were used as gold standard for imputed genotypes; UKBB, Axiom UK-BioBank. Efficacy values are given in percentage. Bold values: smaller value compared with the corresponding value in Table 2.

a particular SNP has been array-genotyped in one cohort or sample collection (e.g., cases) and is to be compared against another sample collection where another array had been used for genotyping (e.g., controls). For the second request of genome-wide application, we assessed how concordant the whole-genome sets of predicted marker genotypes ('imputation spheres') were when using two different arrays as imputation basis (imputation-sphere concordance approach). This approach aims at approximating the information loss in a GWAS setting, where sample collections that have been typed on different arrays are to be combined to test all imputable SNPs for phenotypic association.

We applied commonly used measures of imputation quality and stratified our comparison by the stringency of this post-imputation quality control.^{5,9,12,18} More specific, we considered three different minimum thresholds for the info score R^2 (0.3, 0.5 and 0.8), as well as for the average maximum posterior probability (MPP) of the genotype prediction (0.9, 0.95 and 0.98) and evaluated each combination of these parameter values. These performance values were obtained from the 'Info' (R^2) and the 'Certainty' (MPP) columns of the *.impute2_info file produced by IMPUTE. Furthermore, we stratified our analysis by binning markers with respect to their minor allele frequency (MAF) into three classes, namely a MAF within (0,0.05], (0.05,0.1] and (0.1,0.5], respectively, corresponding to rare, moderately frequent or common variation. MAF values refer to all 2504 samples from the 1000 Genomes data set, not to our data sets at hand. As some linkage disequilibrium (LD) is important for successful imputation, we also estimated average local LD, as described by r^2 , between each marker on a base array and all markers on a target array that were located within 25 kb upstream or downstream of this base marker, using the UNPHASED software²⁶ v3.1.7 (<https://sites.google.com/site/fdudbridge/software/unphased-3-1>) separately for the 30 CEU and the 30 YRI samples.

In assessing the imputation performance for the marker-on-array imputation approach, we followed previous work¹⁸ and defined imputation efficacy as the proportion of markers that were successfully imputed (by exceeding the minimum thresholds for both R^2 and MPP) out of all imputable markers (ie, present on the target array as well as in the imputation reference). From those successfully imputed markers, we derived the imputation accuracy as the proportion of markers whose predicted genotype, that is, the one with the highest posterior probability, matched the observed one from the target array.

To assess the extent of discrepancies in genotype prediction for genome-wide studies if different arrays were used on the same samples, we assessed the overlap and concordance of predicted genotypes that passed post-imputation quality control between different arrays serving as imputation basis for each of the investigated markers in three different ways. First, we quantified the proportion of markers that fell into the intersection of two imputation spheres (overlap). Second, we calculated the proportion of markers that yielded the MPP for the same genotype among all those SNPs that could be successfully imputed (concordance). We then also compared the imputed genotypes against the known ones from the 1000 Genomes database as an additional means of imputation accuracy assessment. Third, we quantified the distance between the

vectors of the posterior genotype probabilities obtained from imputation based on each of the two arrays of the respective pair. We used the Euclidean vector distance, re-scaled by factor $1/\sqrt{2} \approx 1.414$ in order to ensure values between 0 and 1.

RESULTS

Features of the data for imputation

A large majority of markers (>80%) present on the five investigated arrays was found to be polymorphic in the two population samples, with the exception of the UKBB array where only about half of the markers (52%) were polymorphic in the African YRI samples (Supplementary Tables S3 and S4). The African sample set showed generally lower proportions of polymorphic markers than the European set for 550k, ACEU and UKBB, but slightly higher proportions for the A6 and HOE arrays. The allele frequency distribution for polymorphic markers (Supplementary Figure S1) showed a slight excess of modestly frequent and rare alleles for A6 and ACEU in both Europeans and Africans. UKBB also showed this modest excess in Africans, while Europeans were characterized by a strong enrichment of rare alleles. Both Illumina arrays showed a nearly uniform allele frequency distribution in both population sample sets, with a tendency to a reduced number of rare alleles in the CEU samples. In Europeans, average local LD between markers on legacy and on novel arrays was generally modest, with interquartile ranges for r^2 of approximately (0.1 and 0.3), except for UKBB that showed considerably lower average LD values with markers on A6 and on 550k (Supplementary Figures S2 and S3). As expected, the African sample set showed consistently lower local LD levels than the European one, with the third quartile not exceeding 0.2 for all legacy base arrays, again with a notable drop in average LD for UKBB (Supplementary Figures S4 and S5).

Marker-on-array imputation

In a first analysis, we quantified how accessible array-specific markers are by imputation if two sample collections were genotyped on two different arrays. Those markers that passed post-imputation quality control by exceeding pre-specified thresholds for the info score (R^2) and the average MPP were imputed correctly most of the time, always showing accuracy levels of 97% or higher between legacy and novel arrays (Supplementary Tables S5–S8).

However, substantial proportions of markers failed post-imputation quality control and could therefore not reliably be imputed. This loss

Table 2 Imputation efficacy for legacy arrays from novel arrays in 30 CEU samples

Base novel	Target legacy		A6						550k						0.8			
	R^2		0.3						0.5						0.8			
	MAF	MPP	0.90	0.95	0.98	0.90	0.95	0.98	0.90	0.95	0.98	0.90	0.95	0.98	0.90	0.95	0.98	0.98
ACEU	0–0.05	97	94	87	95	93	86	89	89	85	84	80	70	83	80	70	76	69
	0.05–0.1	98	96	91	98	96	91	95	94	90	83	80	74	82	80	74	79	74
	0.1–0.5	97	94	86	97	94	86	96	94	89	64	61	56	64	61	56	63	55
UKBB	0–0.05	97	96	91	97	96	88	91	91	89	85	84	77	85	84	77	82	77
	0.05–0.1	99	98	93	99	98	93	97	97	92	84	83	78	84	83	78	83	78
	0.1–0.5	99	98	87	99	98	87	99	97	87	67	66	57	67	66	57	67	57
HOE	0–0.05	83	80	71	82	79	70	74	74	68	83	77	60	82	77	60	71	59
	0.05–0.1	83	81	77	82	81	77	80	79	76	83	80	73	83	80	73	80	72
	0.1–0.5	67	66	62	67	66	62	66	65	62	66	64	59	66	64	59	66	59

Abbreviations: ACEU, Axiom Genome-Wide CEU; base novel, array whose observed genotypes were used as imputation basis; CEU, CEPH Utah residents of northern and western European ancestry; HOE, HumanOmniExpress; MAF, minor allele frequency of the marker; MPP, average maximum posterior probability; R^2 , info score; Target legacy, array whose observed genotypes were used as gold standard for imputed genotypes; UKBB, Axiom UK-BioBank. Efficacy values are given in percentage. Bold values: smaller value compared with the corresponding value in Table 1. For further details, see legend to Table 1.

in efficacy differed considerably between array pairs. In particular, efficacy showed a strong deterioration when the legacy arrays A6 and 550k were used as imputation basis for novel arrays or if an Affymetrix array was used to impute an Illumina array and vice versa, observable in both population sample sets (Tables 1 and 2 and Tables 3 and 4 for CEU and YRI samples, respectively). In the most extreme cases, efficacy values dropped to 54%/43% in CEU/YRI samples, respectively, for A6-based imputation and to 41%/46% for 550k-based one, while the novel arrays showed minimum efficacy values of 55–59%/28–54%, depending on the chosen thresholds for R^2 and MPP and on the chosen MAF class. Interestingly, Illumina arrays showed a tendency towards inferior performance with common variants ($MAF > 5\%$). In general, imputation of markers on legacy arrays (A6, 550k) based on novel arrays (ACEU, UKBB and HOE) yielded substantially higher efficacy in both the CEU and the YRI sample sets than vice versa. In any case, efficacy values for the African sample set (Tables 3 and 4) were somewhat lower for imputation based on the UKBB array. Somewhat surprisingly, however, was the observation that A6-based imputation of markers present on the ACEU and UKBB arrays resulted in generally higher efficacy in the YRI samples than the CEU samples for low-frequency alleles (Tables 1 and 3). Imputation across novel arrays also showed considerable drops in efficacy, albeit to a lesser degree than legacy-based imputation in general (Supplementary Tables S9 and S10).

Although stricter post-imputation quality control did not result in strong improvements in accuracy, it caused dramatic reductions in efficacy in some cases. Increasing the thresholds for MPP showed a stronger effect on efficacy in the European compared to the African sample set, especially in the class of common markers. Higher thresholds for R^2 had a stronger impact on efficacy when lower thresholds were applied to MPP values, both in the European sample set (Supplementary Figures S6–S10) and, even more pronounced, in the African sample set (Supplementary Figures S11–S16).

Efficacy values differed notably between autosomes, depending on the array used as imputation basis. In the European CEU samples, A6-based imputation showed a notably inferior performance for chromosomes 6 and 19 (Supplementary Figure S6), whereas ACEU-based imputation was problematic for autosomes 19–22 (Supplementary Figure S7). UKBB showed a uniform performance across all autosomes in Europeans (Supplementary Figure S8). Array 550k-based imputation as well as HOE as a target showed a largely uniform performance across chromosomes (Supplementary Figures S9 and S10). Results for the African YRI samples were similar for the A6 array (Supplementary Figures S11–15), although efficacy levels were notably lower. UKBB-based imputation resulted in deteriorated efficacy compared to the European samples (Supplementary Figure S13).

Imputation-sphere concordance

In a second analysis, we quantified the overlap between the imputation spheres of a sample collection that was genotyped using two different arrays, that is, the proportion of SNPs that were available for genome-wide phenotypic association analysis. In general, large numbers of markers could be reliably imputed, as defined by passing pre-specified thresholds for the post-imputation quality criteria R^2 and MPP (see Supplementary Table S11 for an example of the absolute number of markers in common for a moderate thresholds in Europeans and Africans respectively), with highly concordant most-likely genotype predictions most of the time in the overlap (>97%; data not shown). Nevertheless, substantial proportions of markers would be lost to subsequent association analysis, because they were imputed successfully only on one array, but not the other (Figures 1–4). Although the

Table 3 Imputation efficacy for novel arrays from legacy arrays in 30 YRI samples

Base legacy	Target novel			ACEU						UKBB						HOE					
	R^2			0.3			0.5			0.8			0.3			0.8			0.3		
	0.90	0.95	0.98	0.90	0.95	0.98	0.90	0.95	0.98	0.90	0.95	0.98	0.90	0.95	0.98	0.90	0.95	0.98	0.90	0.95	0.98
MAFI-MPP	0.94	89	74	92	88	73	80	80	70	89	82	64	85	79	63	67	66	58	85	82	71
0-0.05	94	89	74	92	88	73	80	80	70	89	82	64	85	79	63	67	66	58	85	82	71
0.05-0.1	94	85	67	93	85	67	84	79	65	90	78	56	89	77	56	74	69	54	82	76	64
0.1-0.5	90	78	59	90	78	59	86	76	59	85	72	53	85	71	53	78	69	52	62	55	43
0-0.05	87	84	71	86	83	71	78	77	68	87	84	69	85	82	68	74	73	65	87	84	71
0.05-0.1	85	79	65	84	79	65	79	76	64	87	78	60	86	77	60	77	73	59	84	79	65
0.1-0.5	67	61	49	67	61	49	65	60	49	68	60	47	67	59	46	64	58	46	66	61	49

Abbreviations: 550k, HumanHap550 Bead; A6, GeneChip Genome-Wide Human SNP 6.0; ACEU, Axiom Genome-Wide CEU; base legacy, array whose observed genotypes were used as imputation base; MAF, minor allele frequency of the marker; MPP, average maximum posterior probability; R^2 , info score; target novel, array whose observed genotypes were used as gold standard for imputed genotypes; UKBB, Axiom UK-BioBank; YRI, African Yoruba samples. Efficacy values are given in percentage. Bold values: the smaller value compared with the corresponding value in Table 4. For further details, see legend to Table 1.

Table 4 Imputation efficacy for legacy arrays from novel arrays in 30 YRI samples

Base novel	Target legacy		A6									550k								
	R^2		0.3			0.5			0.8			0.3			0.5			0.8		
	MAF MPP		0.90	0.95	0.98	0.90	0.95	0.98	0.90	0.95	0.98	0.90	0.95	0.98	0.90	0.95	0.98	0.90	0.95	0.98
ACEU	0–0.05		96	84	56	94	83	56	79	75	55	82	75	56	81	75	56	73	70	54
	0.05–0.1		92	80	57	92	79	57	85	76	56	80	71	54	79	70	54	72	67	53
	0.1–0.5		92	80	59	92	80	59	88	79	58	60	52	39	60	52	39	57	51	38
UKBB	0–0.05		95	78	44	93	77	44	72	66	42	82	70	45	81	69	45	68	63	43
	0.05–0.1		92	72	44	91	72	44	80	67	43	80	68	45	79	67	45	69	62	43
	0.1–0.5		92	72	44	92	72	44	87	69	43	62	48	29	61	48	29	57	46	28
HOE	0–0.05		86	83	73	85	83	73	82	81	72	84	83	77	83	83	77	81	81	76
	0.05–0.1		83	79	69	82	79	69	81	78	68	83	81	73	82	81	73	80	79	72
	0.1–0.5		66	64	56	66	64	56	66	63	56	65	63	54	65	62	54	64	62	54

Abbreviations: 550k, HumanHap550 Bead; A6, GeneChip Genome-Wide Human SNP 6.0; ACEU, Axiom Genome-Wide CEU; base novel, array whose observed genotypes were used as imputation base; HOE, HumanOmniExpress; MAF, minor allele frequency of the marker; MPP, average maximum posterior probability; R^2 , info score; target legacy, array whose observed genotypes were used as gold standard for imputed genotypes; UKBB, Axiom UK-BioBank. Efficacy values are given in percentage. Bold values: smaller value compared with the corresponding value in Table 3. For further details, see legend to Table 1.

Figure 1 Intersection of imputation spheres from legacy to novel arrays in 30 CEU samples. Proportions of those successfully imputed markers using a legacy array (row) that were also successfully imputed using a novel array (column), depicted as a function of the post-imputation criteria tuple (R^2 , MPP). First row: GeneChip Genome-Wide Human SNP 6.0 (A6); second row: 550k. solid line: variants with MAF in (0–0.05); dotted-dashed line: MAF in (0.05–0.1); dashed line: MAF in (0.1–0.5). Note that ‘(’ and ‘)’ denote open and closed interval borders, respectively.

loss of markers could be $<5\%$ for common markers ($MAF > 5\%$) for some array pairs, it could reach close to 30% for rare variants ($MAF \leq 5\%$) in the European sample set (Figures 1 and 2) and even 50% in the African sample set (Figures 3 and 4). If a marker was successfully imputed with each of the two arrays of a pair (ie, it was part of the intersection of both array-specific imputation spheres), both predictions yielded the same most likely genotype most of the time ($>97\%$; data not shown), which also matched the 'true' genotype from the 1000 Genomes database (Supplementary Tables S12 and S13), indicating that the choice of the base array did not have a strong influence on the predicted genotype in an individual. Although we based our allele frequency stratification on those in the complete 1000 Genomes data set, this can be expected to have a very modest effect on the presented numbers because marker classification results were highly similar in the CEU and YRI subsets (96–99% concordance; Supplementary Table S14). For a finer-grained resolution, we also compared the distance in posterior genotype probabilities for the three possible genotypes of a marker instead of concordance in the most likely genotype. As expected, the vast majority of markers in

the intersection of imputation spheres showed only minor distances between the posterior probability vectors regardless of the pairing of arrays (see Supplementary Figures S16 and S17 for the European samples, with very similar results for the African samples (data not shown)), showing normalized Euclidean distances smaller than 0.01. Strict MPP filtering ($MPP > 98\%$) uniformly led to very low discrepancies across array pairs and allele frequency ranges, whereas $R^2 > 0.8$ was an effective means for reducing discrepancies when less strict MPP filtering was applied and variants were of moderate to low frequency ($MAF \leq 10\%$), demonstrating the effect of strict post-imputation quality control on the accuracy of prediction.

DISCUSSION

Current human genetic research is confronted with a situation where legacy and recent sample collections have been genotyped with different generations of arrays. Besides technical details, such as different chemistries, these arrays differ in their sets of included markers as a result of different design principles and increasingly more available information in the public sphere. To investigate the impact of these changes on genome-wide studies, we considered the exemplary case of two legacy arrays, namely Affymetrix's A6 array and Illumina's 550k, and three newer and currently used arrays, namely Affymetrix's ACEU and UKBB arrays and Illumina's HumanOmniExpress array. The design of the UKBB array is special with respect to its focus on a Northwest European population and the inclusion of a large number of markers that are known to be polymorphic in this population but to carry rare alleles. Although the accuracy of successfully imputed markers was very high throughout, which is in line with previously reported results on imputation performance,¹⁸ we found that efficacy values were markedly lower in both the European and the African population sample set. Depending on the combination of legacy and novel array, on marker allele frequency class and on applied post-imputation quality control, efficacy frequently dropped below 75% and even below 50% in a number of scenarios. The UKBB array performed best as imputation basis but was hardest to impute in Europeans, whereas its performance was often inferior to other arrays in Africans. These observations are consistent with the design of this array, comprising a large number of low-frequency variants, which are notoriously hard to impute. Our analysis of the overlap of the sets of successfully imputed SNPs when using two different arrays as imputation basis with the same sample set showed a picture similar to that of the marker-on-array analysis. In particular, genotypes from both predictions were highly concordant if both arrays allowed successful imputation. Markers successfully imputed by both arrays almost always had a predicted genotype that matched the 'true' genotype from the 1000 Genomes project. However, failure of successful imputation with at least one array incurred a loss of roughly 10–30% of all markers for subsequent analyses in many pairings of arrays.

Several factors can have an impact on the imputation performance with SNP arrays. Although we have removed 50 or 49 individuals from the imputation reference for CEU or YRI respectively, this reduction can be expected to have a minor effect at most on the imputation performance, given the large remaining size of >2400 individuals and the presence of individuals from the same populations and from a number of closely related populations for each the European and the African sample set to be imputed. The two population sample sets differed in their proportions of polymorphic markers, albeit a clear trend was missing. A notable exception to this was the UKBB array with a much lower proportion in Africans, which is not surprising given the special design of this array. However, in line with our

Figure 2 Intersection of imputation spheres from novel to legacy arrays in 30 CEU samples. Proportions of those successfully imputed markers using a novel array (row) that were also successfully imputed using a legacy array (column), depicted as a function of the post-imputation criteria tuple (R^2 , MPP). First row: ACEU; second row: UKBB; and third row: HOE. For further details, see legend to Figure 1.

Figure 3 Intersection of imputation spheres from legacy to novel arrays in 30 YRI samples. Proportions of those successfully imputed markers using a legacy array (row) that were also successfully imputed using a novel array (column), depicted as a function of the post-imputation criteria tuple (R^2 , MPP). First row: GeneChip Genome-Wide Human SNP 6.0 (A6); second row: 550k. For further details, see legend to Figure 1.

observations, the impact of the proportion of polymorphic markers and also of the local average LD is difficult to predict.

Imputation-based genetic studies frequently apply only info-score filtering as post-imputation quality control, but none related to the MPP. Although this can be a sufficient approach if hypothesis generation by standard allele-based testing via imputed allele-dosage with subsequent follow-up and validation by genotyping is the ultimate aim of the imputation, filtering imputation results by requiring high MPPs, that is, an overwhelming concentration of the posterior probability mass onto a single genotype, can be beneficial in these and other situations. First, assessing the posterior probability is complementary to the use of the info score and may indicate imputation errors not picked up by the latter. Given the millions of markers subject to imputation, requiring large MPPs can then serve as an additional means of quality control in order to reduce the number of false-positives in a possibly costly follow-up. Second, some heuristic or approximate data analysis approaches may actually require best-guess genotypes. Applying strict MPP thresholds then serves as a means of reducing the bias in such analyses.

As we have shown here, powerful tools of sophisticated imputation and large reference panels, such as the 1000 Genomes resource, may still not be enough to prevent a substantial information loss in terms of markers that can be jointly analyzed when aged sample collections that have been genotyped on now outdated arrays are to be combined with newer collections subject to genotyping by new generations of arrays. A likely consequence of the observed information loss is the decreasing use of those legacy collections. Given the substantial investments not only in the genotyping, but foremost in the recruitment and phenotyping of these collections, in the past, their future neglect would represent a huge waste of public resources. Securing the continued use of these collections, in particular those of healthy, population representative individuals would require an only modest investment for their re-genotyping using up-to-date arrays. Given the still substantial costs of whole-exome and whole-genome sequencing, whole-genome genotyping will remain a standard and successful approach for years to come. We therefore strongly opine that such re-genotyping would be a worthwhile and responsible endeavor, not the least because newer array generations

Figure 4 Intersection of imputation spheres from novel to legacy arrays in 30 YRI samples. Proportions of those successfully imputed markers using a novel array (row) that were also successfully imputed using a legacy array (column), depicted as a function of the post-imputation criteria tuple (R^2 , MPP). First row: ACEU; second row: UKBB; third row: HOE. For further details, see legend to Figure 1.

have in parts been designed for improved imputation performance. Future work is necessary to investigate the optimal choice(s) of arrays for this endeavor.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

MN conceived the study. GK and MN analyzed the data. GK, PN and MN wrote the manuscript. All authors read and approved the final manuscript.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)

- Krawczak M, Nikolaus S, von Eberstein H, Croucher PJ, El Mokhtari NE, Schreiber S: PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet* 2006; **9**: 55–61.
- Lowel H, Doring A, Schneider A *et al*: The MONICA Augsburg surveys—basis for prospective cohort studies. *Gesundheitswesen* 2005; **67**(Suppl 1): S13–S18.
- Voight BF, Kang HM, Ding J *et al*: The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet* 2012; **8**: e1002793.
- International Genetics of Ankylosing Spondylitis C, Cortes A, Hadler J *et al*: Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat Genet* 2013; **45**: 730–738.
- Uh HW, Deelen J, Beekman M *et al*: How to deal with the early GWAS data when imputing and combining different arrays is necessary. *Eur J Hum Genet* 2012; **20**: 572–576.
- Hofman A, Breteler MM, van Duijn CM *et al*: The Rotterdam Study: 2010 objectives and design update. *Eur J Epidemiol* 2009; **24**: 553–572.
- Uh HW, Houwing-Duistermaat JJ, Putter H, van Houwelingen HC: Assessment of global phase uncertainty in case-control studies. *BMC Genet* 2009; **10**: 54.
- Browning BL, Browning SR: Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet Epidemiol* 2007; **31**: 365–375.
- Li Y, Abecasis GR: Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet* 2006; **579**: 2290.
- Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007; **39**: 906–913.
- Fuchsberger C, Abecasis GR, Hinds DA: minimac2: faster genotype imputation. *Bioinformatics* 2015; **31**: 782–784.
- Howie BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**: e1000529.
- Genomes Project C, Abecasis GR, Altshuler D *et al*: A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.
- International HapMap C, Altshuler DM, Gibbs RA *et al*: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–58.
- International HapMap C, Frazer KA, Ballinger DG *et al*: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
- The International HapMap Consortium: The International HapMap Project. *Nature* 2003; **426**: 789–796.
- The International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
- Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A: A comprehensive evaluation of SNP genotype imputation. *Hum Genet* 2009; **125**: 163–171.
- Anderson CA, Pettersson FH, Barrett JC *et al*: Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am J Hum Genet* 2008; **83**: 112–119.
- Li Y, Willer C, Sanna S, Abecasis G: Genotype imputation. *Annu Rev Genomics Hum Genet* 2009; **10**: 387–406.
- Howie B, Marchini J, Stephens M: Genotype imputation with thousands of genomes. *G3 (Bethesda)* 2011; **1**: 457–470.
- Marchini J, Howie B: Comparing algorithms for genotype imputation. *Am J Hum Genet* 2008; **83**: 535–539, author reply 539–540.
- Genomes Project C, Auton A, Brooks LD *et al*: A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.
- Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
- Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J: Haplotype estimation using sequencing reads. *Am J Hum Genet* 2013; **93**: 687–696.
- Dudbridge F: Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered* 2008; **66**: 87–98.