Article

Opening the Black-box of Imputation Software to Study the Impact of Reference Panel Composition on Performance

Thibault Dekeyser[1,2], Emmanuelle Génin[1,2] and Anthony F. Herzig[1,*]

[1] Inserm, Univ Brest, EFS, UMR 1078, GGB, Brest, France
Faculté de Médecine - IBRBS 22 avenue Camille Desmoulins F-29238 BREST Cedex 3 - France
[2] CHRU Brest, Brest, France
*Corresponding author: anthony.herzig@inserm.fr

Abstract: Genotype imputation is widely used to enrich genetic datasets. The operation relies of panels of known reference haplotypes with typically whole-genome sequencing data. How to choose a reference panel has been widely studied and it is essential to have a panel that is well matched to the individuals who require imputation of missing genotypes. However, it is broadly accepted that such an imputation panel will have an enhanced performance with the inclusion of diversity; haplotypes from many different populations. We investigate this observation in this work by examining in fine detail exactly which reference haplotypes are contributing at different regions of the genome. This is achieved using a novel method of inserting synthetic genetic variation into the reference panel in order to track the performance of leading imputation algorithms.
We show that while diversity may globally improve imputation accuracy, there can be occasions where incorrect genotypes are imputed following the inclusion of more diverse haplotypes in the reference panel. We however demonstrate a technique for retaining and benefitting from the diversity in the reference panel whilst avoiding the occasional adverse effects on imputation accuracy. What is more, our results elucidate more clearly the role of the diversity in a reference panel than has been shown in previous studies.

Keywords: genotype imputation; population genetics; rare-variants; reference panel; admixture

1. Introduction

Imputation of missing genotypes is a widely used technique for enriching genetic datasets where strong patterns of linkage disequilibrium (LD) between physically close genetic variants facilitate highly accurate inference of missing data points [1]. Leading algorithms for genotype imputation are based on the Li-Stephens haplotype mosaic model [2] and are implemented with ever increasingly stream-lined Hidden-Markov models (HMMs) [3–6]. Typically, when performing imputation, two datasets will be in play: a group of 'target' individuals for whom some genetic data is available and a panel of 'reference' individuals who have both genetic data in common with the target group as well as additional data for other genetic variants for which the genotypes of the target group are unknown. By comparing the target and reference individuals across the data for genetic variants that are present in both datasets, imputation algorithms attempt to infer the genotypes of the target individuals for the genetic variants that have only been measured in the reference panel.

Globally, the literature is well stocked regarding studies of genotype imputation accuracy; with comparisons of different software [7,8], the impact of the choice of reference panels [9–11], or the type of genetic data involved [12]. Studies that have examined the choice of reference panel have broadly come to a consensus in that imputation will be more accurate when using a reference panel that is closely matched in terms of ancestry to the target group [13–16]. Yet, it has also been noted that having a reference panel that contains individuals from diverse populations is also beneficial [17–19]. The idea being that 'unexpected-sharing' [17] will improve imputation and that the HMMs used in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0483.v1

imputation algorithms will only use the diversity in the reference panel of haplotypes when necessary. So the general advised strategy for imputing target individuals from a given population would be to employ a reference panel that contains individuals of the same population along with individuals from neighbouring populations and even from more distant populations (in terms of the number of generations to trace back before common ancestors could be found for pairs of individuals) [20–26]. Simply put, the imputation panel cannot be too large, adding additional samples should only improve imputation and the results presented for the largest and most cosmopolitan panels are certainly convincing [11,27].

In this study, we focus on the choice of reference panel but rather than simply measuring which reference panel provides the most accuracy, we explore in greater detail the role of the composition of a reference panel in the context of large cosmopolitan reference panels. This is achieved by tracking, for each target individual and each genomic region, the ancestry groups of the reference panel individuals used in the imputation. This sheds light on previous published observations of certain imputation panels out-performing others as well as indicating potential avenues for improving the performance of existing imputation algorithms. We find that more diversity may indeed improve the imputation of rare-variants but that there may also be many imputed 'false positives' (incorrectly imputed genotypes containing rare allele where in fact none are truly present); and we illustrate how this might be avoided.

We used leading imputation software IMPUTE5 [3] and MINIMAC4 [5] and considered only the most frequent imputation scenario where are target individuals have single-nucleotide polymorphism genotyping array (SNP array) data and the reference panel has whole-genome sequencing (WGS) data.

2. Materials and Methods

First, some background on genotype imputation methods is required. The Li-Stephens model describes a process where, given a large enough sample of N haplotypes from a population, an $N+1^{th}$ haplotype can be closely approximated as a mosaic of small haplotype segments or 'chunks' from the pool of N haplotypes. Imputation algorithms apply this through Hidden Markov modelling; each haplotype in the target group (the target data will be phased statistically or directly using family data) will be imputed in turn with the idea being to infer a mosaic of reference panel haplotypes that will approximate each target haplotype. As the reference panel will contain data on more sites than are known in the target group, the mosaics that are inferred will allow for the imputation of missing genotypes. The HMMs involved will have as hidden states the index of the reference panel haplotypes who are donating a mosaic segment at a set of points across the genome. This will be the set of genetic positions that are present in both the target dataset and the reference panel. The observed states are the phased genotype data of the target haplotype with emission probabilities allowing for differences between the donating reference haplotype and the observed genotypes that could arise from mutations or genotyping errors. Crucially, software such as IMPUTE5 and MINIMAC4 will not give details about exactly which reference panel haplotypes are donating at different points in the genome; they are rather black-box-like in nature.

For each genomic position in common between the target and the reference panel, imputation software will assign a posterior probability for each haplotype in the reference panel being the donating haplotype for the mosaic being build. When imputing target haplotype j, h_i^j will be the index of the donating haplotype at genomic site i (where i goes from 1 to S) among the N reference haplotypes, so taking a value k between 1 and N. The HMM will provide, via the forward-backward algorithm [28,29], the posterior probabilities for the different possible values of k for h_i^j, based on the observed alleles of haplotype j; which we donate simply as o^j to represent the sequence of observed alleles, $o_1^j, o_2^j, \ldots, o_S^j$. We can denote these posterior probabilities as $P(h_i^j = k \mid o^j)$. In practice, o^j will be a sequence of zeros and ones where zeros corresponds to reference alleles

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0483.v1

and ones to the alternative alleles. Similarly the reference panel haplotype data can be written as H^k (for the k^{th} reference haplotype) which is similarly a sequence of zeros and ones but across a larger set of genomic sites than for the target data.

Through linear interpolation between adjacent sites, these posterior probabilities are approximated for sites that are not present in the target data; we denote such a site as i'. If a perfect mosaic were found, only one such posterior probability would be non-zero and only one reference panel haplotype would be donating; and hence the imputed value for site i' will be $H_{i'}^{k^*}$, a zero or a one depending on the known allele in the sole donating reference haplotype k^* (k^* denoting the sole value of k for which $P\left(h_{i'}^j = k \mid o^j\right) = 1$). However, there may be multiple possible mosaics and so multiple values of k for which $P\left(h_{i'}^j = k \mid o^j\right)$ is greater than zero, and so rather than imputing missing genotypes from just one reference haplotype, a weighted sum, or 'dosage', $d_{i'}^j$ is given:

$$d_{i'}^j = \sum_{k=1}^{N} P\left(h_{i'}^j = k \mid o^j\right) H_{i'}^k$$

IMPUTE5 and MINIMAC4 give the values of $d_{i'}^j$ from which the values of $P\left(h_{i'}^j = k \mid o^j\right)$ cannot be recovered by the user. We observed that if $d_{i'}^j$ is non-zero at a position i' which is observed to be a singleton in the reference panel, one could infer that the corresponding posterior probability of the reference haplotype carrying the singleton must also be non-zero. This led to the idea of simply injecting a lot of artificial singletons in the reference panel to track the imputation. Given that the algorithms of current software have become increasing complex in a bid to achieve superior efficiency for analysing huge bio-bank scale data, this was for us a more feasible idea than attempting to unpick the existing code or to re-implement the HMMs ourselves. Tracking the role of each reference haplotype in such a way would not be impossible but would however incur a certain computation burden; we therefore set about tracking groups of reference haplotypes.

We took data from the 1000 Genomes Project [30] (1000G) as our sandbox. This dataset contains 2504 individuals grouped into 26 populations who are themselves grouped into five super-populations (Table1). We used the version made available here https://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes_phase3_v5a/b37.vcf/ by the authors of Beagle [4]; a version with no variants with a minor allele count below 5 which greatly reduces the size of the data and enables us to carry out our analyses in a respectable time-frame.

Table 1. The populations of the 1000 Genomes projects, split into five super-populations. Three populations (in bold) were chosen to be our target individuals: ACB (African Caribbean in Barbados), ASW (African Ancestry in South-West USA), and MXL (Mexican Ancestry in Los Angeles California USA). In this study, the 221 individuals from these three groups were imputed, in a variety of ways, using the other 23 populations as a reference panel.

Super Population	Populations
AFR (Africa)	LWK, GWD, MSL, **ACB, ASW**, YRI, ESN
EAS (East Asia)	CHB, KHV, CHS, JPT, CDX
EUR (Europe)	TSI, CEU, IBS, GBR, FIN
SAS (South Asia)	BEB, STU, ITU, PJL, GIH
AMR (Americas)	PEL, **MXL**, CLM, PUR

We decided to separate three populations (ACB, ASW, and MXL) and impute the 221 individuals from these populations (our target group) with a reference panel formed by the remaining 23 populations. These three populations were chosen as they are known to be the more admixed individuals of 1000G [30]. For the three target populations, we separated the sites present on the UK-biobank genotyping array

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0483.v1

(https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=149601). These sites would be supplied to the imputation software; the other sites would be retained in order to assess imputation accuracy. In order to track which reference haplotypes were being called on to impute the individuals of our target group across the genome, we injected completed synthetic variants to the imputation reference panel. These variants would serve as indicators for each of the 5 reference groups; a synthetic variant tagging the EUR population would be 0 (the 'reference' allele) for all non-EUR individuals and 1 (the 'alternative') for all EUR individuals. Hence, in this example, when interrogating the imputed dosage data for this synthetic variant, a dosage of 1 would indicate that only EUR haplotypes were used to form the mosaic to achieve the imputation; a dosage of 0 would indicate that conversely that only non-EUR haplotypes were used; and a value between 0 and 1 would show that both EUR and non-EUR haplotypes were used.

To add synthetic variants systematically, we selected locations to tag in the following manner: among the genomic sites in common between the target group and the reference panel, i.e. the genomic positions on the UK-biobank array, we kept those with a minor allele frequency (measured across the whole of the 1000G) above 0.2 and thinned with a strict pruning ($r^2 < 0.02$). We then retained from this list those sites which were not 'shoulder-to-shoulder' to another variant in 1000G; i.e. there was no other variant in 1000G at a distance of 1 base-pair. This gave us a list of 32,279 variants. We then added 5 synthetic variants 1 base-pair downstream of each of these tagged variants, one to track each of the 5 super-populations in the reference panel. We verified that adding these synthetic variants in batches on 5 (and with each batch all sharing the same physical position) had no effect on the imputation of IMPUTE5; the same output was given with and without the synthetic variants. We refer to each batch of 5 synthetic variants as an 'imputation barcode'.

Having added our 32,279 imputation barcodes, imputation was completed using IMPUTE5. We could then calculate the cumulative contributions of each of the 5 reference groups to the imputation; using the imputation barcodes. This was compared to similar estimations using the chromo-painting [31] functionality of pbwt [32], supervised ADMIXTURE [33] (which required us to filter by MAF (>5%) and to perform LD-pruning (--indep-pairwise 50 10 0.1) in plink [34]), and SOURCEFIND [35] (which used the pbwt chromo-painting output between all 2504 individuals of 1000G as input). We did a second imputation where each individual was imputed with a reference panel consisting of the super-populations that had a SOURCEFIND proportion above 0.01 for the individual in question. We only analysed the 22 autosomal chromosomes. Data manipulation was performed using R-package 'gaston' [36], shapeit2 [37], and bcftools [38].

Imputation accuracy was assessed by calculating the aggregate R^2 [27] across difference sets of variants depending on the minor allele count (MAC) in either the 1000G as a whole or measured in each of the three target groups. Finally, we explored the impact of reducing the size of the reference panel based on an informed choice after examining the imputation barcodes. In order to track the imputation at the haplotype level (and not at the individual level), we imputed 442 (221×2) pseudo-individuals which were simply each haplotype of the target group paired with itself. IMPUTE5 splits the autosomes into 400 imputation chunks and processes them separately. Chromosome 1 was split into 32 chunks, chromosome 22 into only 4 chunks. For each chunk, we chose to impute each target haplotype with only a subset of the 5 super-populations. The choice was a follows: if none of the synthetic variants across all the imputation barcodes in the chunk that tag super-population 'Z' for haplotype j had a dosage above 0.9, then population 'Z' would be removed from the reference panel for haplotype j. Simply put, if on the first imputation run using all 5 super-populations, there was no evidence that population 'Z' was making a telling contribution somewhere in the chunk, we would remove it from the reference panel. Hence, each individual haplotype would be assigned its own imputation panel for each of the 400 imputation chunks. The motivation for this procedure was to see if imputation accuracy would be affected by removing reference panel haplotypes that are

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0483.v1

only making a very small contribution; which we imagined might contribute more noise than it would improve imputation accuracy.

3. Results

Having imputed the 221 target individuals using IMPUTE5, we summed and scaled the dosages of all imputation barcodes for each individual in order to provide an estimation of the proportions of the contribution of each super-population to the imputation. In Figure 1, these proportions are compared to three other alternative estimations using software typically applied for the analysis of population genetics.

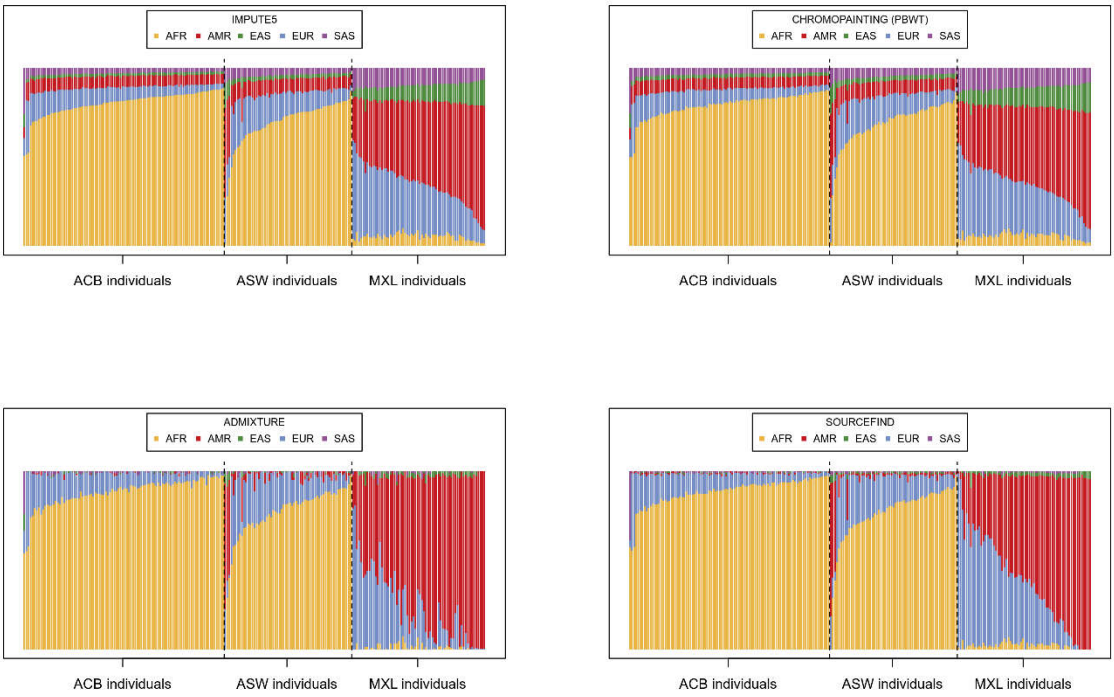
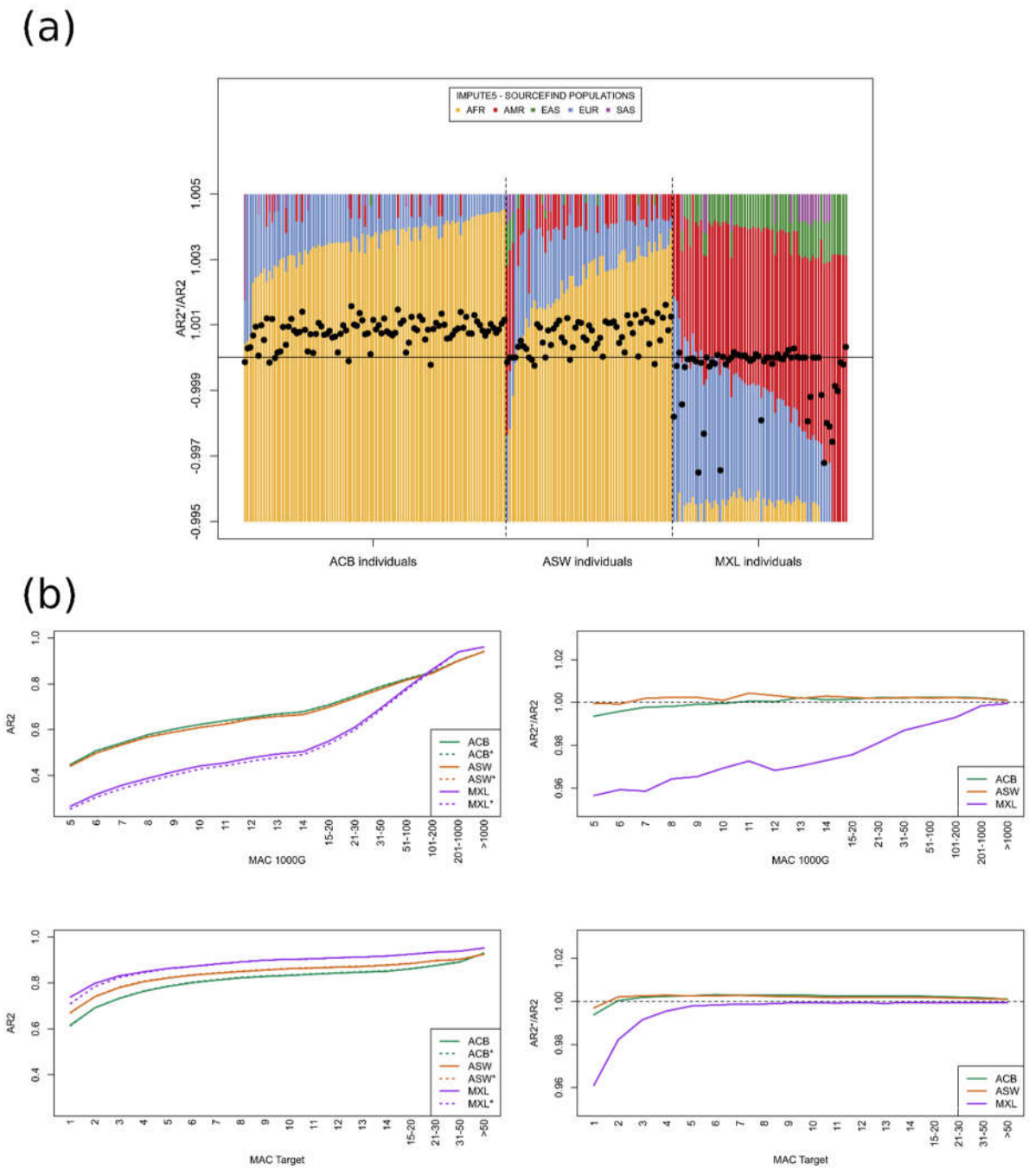


Figure 1. In these plots, each of the 221 target individuals has a vertical bar which is coloured depending on the proportion of their genome assigned to the different super-populations from the reference panel. The individuals are in the same order across the plots, arranged according to the proportions in the top right plot. Top left: proportions of the genome (autosomal chromosomes) imputed with haplotypes from the 5 super-populations AFR, EUR, EAS, SAS and AMR; as ascertained by cumulating the dosages of the imputation barcodes. Top right: Proportions are estimated from the total 'chunk length' matrix derived from the chromo-painting algorithm of software pbwt. Bottom left: ancestry proportions assigned by applying ADMIXTURE (supervised mode). Bottom right: Proportions assigned by SOURCEFIND.

The proportions derived from the imputation barcodes (Figure 1 top left) are unsurprisingly very similar to the chromo-painting based estimates (Figure 1 top right), given that chromo-painting also invokes the Li-Stephens model. ADMIXTURE (Figure 1 bottom left) and SOURCEFIND (Figure 1 bottom right), however, give slightly different estimations; individuals from ACB and ASW are described as largely having AFR and EUR as source populations and MXL as having EUR and AMR. Indeed, the reader can compare these results to the un-supervised admixture plots in [30] which show very similar patterns for MXL, ACB, and ASW. What is different is that both our IMPUTE5-derived method and chromo-painting ascribe small but distinctly larger proportions to other populations. Hence, when imputing these 221 target individuals, all the populations in the reference panel are being called upon - the diversity in the reference panel seems to be

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0483.v1

important. To test this, we simply imputed each individual with the super-populations that SOURCEFIND indicates. Specifically, we imputed each individual with only the super-populations with a SOURCEFIND contribution above 0.01 (Figure 2). Imputation accuracy was summarized by aggregate R^2 statistics for different groups of variants depending on the minor allele count in the whole of 1000G (Figure 2b top) or in the three target sub-groups (Figure 2b bottom). Here follow the details of the distribution of different reference panels used for the imputation in Figure 2: 88 individuals were imputed with a reference panel consisting of the EUR and AFR individuals of our reference panel; denoted as {EUR, AFR}. 45 were imputed with {EUR, AMR, AFR}, 38 with {EUR, EAS, AMR, AFR}, 14 with {EUR, EAS, AMR, SAS, AFR}, 10 with {EUR, SAS, AFR}, 8 with {EUR, EAS, AMR, SAS}, 6 with {EAS, AMR}, 4 with {EUR, EAS, AMR}, 3 with {EUR, AMR, SAS, AFR}, 3 with {EUR, EAS, AFR}, and 2 with {EUR, AMR}. In total, EUR was used for the imputation of 215 individuals, 201 for AFR, 120 for AMR, 73 for EAS and just 35 for SAS.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0483.v1

Figure 2. (a) The plot of proportions is calculated from the cumulated imputation barcodes after selected a reference panel for each individual based on the SOURCEFIND proportions. Here the individual level AR2*/AR2 are overlaid for each individual. AR2*/AR2 refers to the ratio of the aggregate R^2 using the reduced reference panels based on SOURCEFIND (AR2*) over the base line aggregate R^2 when all super-populations are used (AR2). Points above the horizontal line at AR2*/AR2 = 1 show that the individual's imputation accuracy improved when only using the super-populations with a SOURCEFIND proportion above 0.01. (b) Here the aggregate R^2 (AR2) statistics of imputation using all 23 non-target populations of 1000G and IMPUTE5 is compared to the aggregate R^2 when only the super-populations with a SOURCEFIND proportion above 0.01. The results are split by population (ACB, ASW, and MXL) with lines marked with and without a '*' corresponding to AR2* and AR2, respectively. On the two left panels, the aggregate R^2 are split by minor allele count (MAC) bins where MAC is either calculated in 1000G (top) or separately in each of the three target sub-groups (bottom). As the aggregate R^2 statistics are so close, we also show their ratio (AR2*/AR2) (right two plots).

Most individuals had an improved imputation accuracy when using the SOURCEFIND populations (Figure 2a), but the accuracy of imputation slightly suffered for rare-variants (Figure 2b); particularly for individuals of MXL and particular for variants with a low MAC in the 1000G as a whole. Indeed, using the populations indicated by SOURCEFIND worked well for ACB and ASW but notably less well for MXL where a few individuals in particular had less accurate imputation (Figure 2a). The 14 individuals of MXL whose imputation accuracy fell noticeably among the 20 individuals for whom the AFR super-population was not included in the reference panel. Rare-variants were, unsurprisingly, less well imputed than common variants (Figure 2b, left panels). Another interesting observation was that depending on how the minor allele count was defined, it could easily either be concluded that the MXL group was imputed better or worse than the other two groups (ACB and ASW). In Figure S1 the individual AR2 and AR2* statistics from Figure2 are given and in fact at the individual level the MXL individuals have higher imputation accuracy. Note that when regrouping variants by the minor allele count in each target group (bottom panels of Figure 2b), it is not possible to calculate the aggregate R^2 for variants that are truly monomorphic in the target group. However, the imputation of such variants may be incorrect; grouping variants by the MAC in the whole of 1000G allows such variants to be included.

To go further, we reasoned that whilst an overall improvement for rare-variants was observed when imputing using a more diverse panel, this improvement might not be uniform across the genome. For example, individuals from ACB might benefit from reference panel haplotypes aside from those coming from EUR and AFR but only sporadically. But in many parts of the genome, imputing with a panel of only EUR and AFR would be at least equivalent and perhaps even more accurate given the results of Figure 2 where the majority of the individuals of ACB had a more accurate imputation with smaller less-diverse imputation panels. We here attempt to demonstrate this in the following manner: for each imputation chunk (IMPUTE5 splits the autosomes into 400 chunks for imputation as described in the Methods) we would impute each target haplotype with a specific reference panel, the choice of which was informed by the imputation barcodes in an initial imputation run using the whole reference panel (all of 1000G aside from the three target groups). This 'informed choice' strategy is outlined in the Methods and essentially represents a more fine-grained approach than simply using the SOURCEFIND proportions. The imputation accuracy from this strategy is compared to that of the initial imputation using all populations in Figure 3. Whilst the information from the imputation barcodes comes from IMPUTE5, we mirrored the initial imputation and the informed choice imputation with MINIMAC4 (Figure S2); the results were very much alike those of IMPUTE5. Individual AR2 and AR2* statistics are given in Figure S3 where results were very similar to those of Figure S1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0483.v1

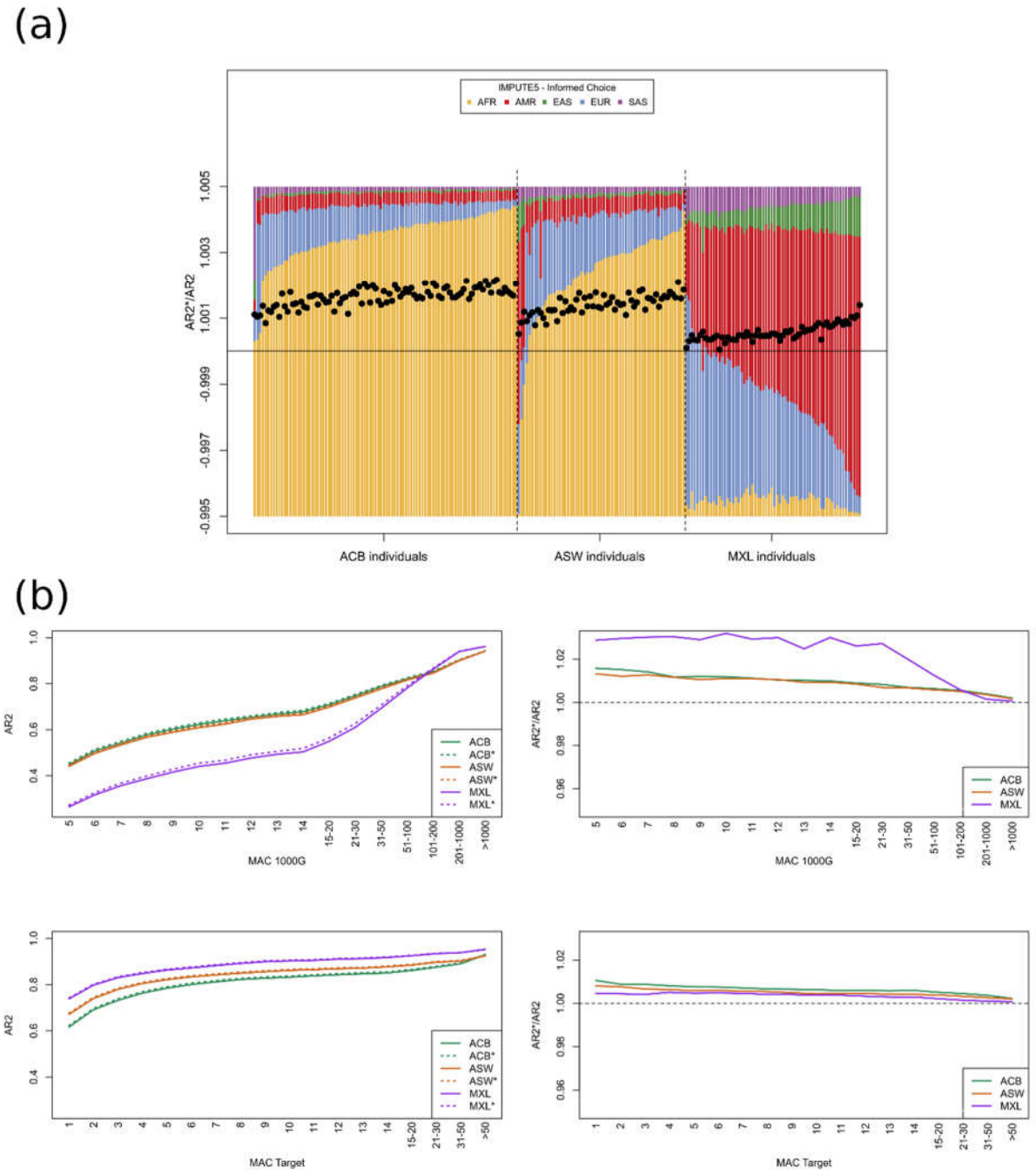


Figure 3. Similar to Figure 2, but here we compare the 'informed choice' strategy with the base line imputation. (a) The proportions plot corresponds to the cumulative proportions across imputation barcodes under the 'informed choice' strategy. The points overlaid correspond to individual AR2*/AR2 statistics as in Figure 2, but the AR2* now come from the 'informed choice' strategy. (b) As Figure 2b but again the AR2* statistics come from the 'informed choice strategy.

The results from the informed choice imputation show that the same imputation accuracy can be achieved with a reduced reference panel, and the relative contributions of the different super-populations (Figure 3a, full comparison given in Figure S4) resembled something intermediate between our first computation and those attributed by SOURCEFIND. What is more, the imputation accuracy actually increases when using the reduced imputation panels of the informed choice strategy; and unlike in Figure 2a, all individuals now have a better imputation accuracy (Figure 3a). Furthermore, even for rare-variants the imputation accuracy under the informed choice strategy is never worse

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2022 doi:10.20944/preprints202212.0483.v1

than the baseline. To observe in closer detail this small gain, Table 2 provides full details of the imputation (using IMPUTE5) of singletons, doubletons, tripletons and variants which are monomorphic in the target group. The counts of true genotypes are compared to the counts of hard-called dosages (dosages that are rounded to the nearest value out of 0,1, or 2). An equivalent table for MINIMAC4 is given in Table S1.

Table 2. Details of the imputation of genotypes for variants which were truly monomorphic in the target group, or had a minor allele count (MAC) of one, two, or three in the target group. The table compares true genotypes with hard-called imputed genotypes when either imputing with IMPUTE5 and the whole of the 1000G (minus the three target populations) as a reference panel, or with the same software and reference panel but each target haplotype imputed with a chosen subset of reference haplotypes informed by the imputation barcodes ('informed choice'). For the 'informed choice' columns, each cell also includes the percentage increase or decrease compared to the corresponding cells when using all populations. AA refers to a homozygous-for-the-reference genotypes, Aa for heterozygous genotypes, and aa for homozygous-for-the-alternative genotypes. Here MAC is measured across the whole target group, ASW, ACB and MXL together.

Hard-called dosage →		IMPUTE5			IMPUTE5, informed choice		
Truth ↓		AA	Aa	aa	AA	Aa	aa
MAC 0	AA	1437998689	920796	907	1438068970 + <0.01%	850635 -7.62%	787 -13.2%
	Aa	-	-	-	-	-	-
	aa	-	-	-	-	-	-
MAC 1	AA	626631509	738813	818	626633230 + <0.01%	737167 -0.22%	743 -9.17%
	Aa	929988	1919673	2026	915589 -1.55%	1934109 +0.75%	1989 -1.83%
	aa	-	-	-	-	-	-
MAC 2	AA	466448959	681496	618	466443614 - <0.01%	686783 +0.78%	676 +9.39%
	Aa	1208483	3032660	3071	1183935 -2.03%	3057009 +0.80%	3270 +6.48%
	aa	1862	3473	5527	1815 -2.52%	3461 -0.35%	5586 +1.07%
MAC 3	AA	343276195	601630	743	343270757 - <0.01%	607094 +0.91%	717 -3.50%
	Aa	1203342	3482444	4509	1176800 -2.21%	3508873 0.76%	4622 2.51%
	aa	2831	6730	11285	2755 -2.68%	6595 -2.01%	11496 +1.87%

In both Table 1 and Table S1, it can be noted that for the monomorphic variants (MAC=0), the informed choice method is improving the imputation by returning less 'false positives' (true genotype is AA and imputed dosage indicates Aa). For the variants with just 1,2, or 3 observed alternate alleles in the target group the improvement is less

 doi:10.20944/preprints202212.0483.v1

striking and comes more from 'false negatives' (true genotype is Aa and imputed dosage indicates AA). This illustrates the different appearances of the results in Figure 2 and 3 depending on whether variants are grouped by the MAC in each target group or in the 1000G as it is only in the latter case that the monomorphic variants are included in the calculation.

4. Discussion

Here we have shown the value of tracking the mosaics for better understanding the performance of an imputation reference panel. Whilst our technique of adding imputation barcodes is rather unwieldy, it is relatively simple to perform and allows users interested in imputation methods to explore in more detail what it is going on under the hood of leading imputation algorithms. We entered into this work with the question as to why reference panels that are more diverse generally improve on smaller reference panels. We observed this by showing that rare-variants were imputed with greater accuracy when using all super-populations of the 1000G compared to when each individual was imputed with just the super-populations that seemed relevant determined by SOURCEFIND. Also certain individuals were globally less well imputed (Figure 2a), particularly a group of 14 from the MXL group for whom the AFR super-population was not selected by SOURCEFIND. This suggested that the more distant populations (those with small SOURCEFIND proportions), whilst only having a small genome-wide contribution, were often improving imputation. We were able to show that we could essentially achieve an equivalent imputation accuracy by leaving out reference panel individuals if there had not been evidence that they were making an important contribution in a given genomic region. This 'informed-choice' approach in fact even marginally improved imputation accuracy. We would not suggest that this method put forward here would be necessarily be of use in practice, but it does help elucidate the performance of our reference panel and could provide avenues for improvements to existing imputation algorithms. It also challenges the concept that a reference panel can never be too large, at least when using current imputation software. Indeed, somewhat similar observations can be found in the literature of HLA imputation [39]; a region that requires different imputation methods.

The improvement attained by the informed choice strategy was most noticeable for rare variants; in terms of the minor allele count in the 1000G. In Table 2, the marginal increase in accuracy could be observed notably for variants that were truly monomorphic in the target group; here the informed choice method imputed over 70,000 less incorrect heterozygote genotypes compared to the baseline imputation. For variants with one, two or three alternate alleles in the target group, the increase in accuracy comes instead from slightly less truly heterozygous genotypes being imputed as homozygous for the reference allele. We also saw that the impact of changing the reference panel was greatest for the MXL population. This population may present different challenges for the imputation algorithm compared to ACB and ASW as the imputation of MXL seems to be relying on AMR haplotypes (of which there are fewer in the 1000G compared to AFR and EUR). The AMR group is also less well defined, many of the individuals have a European component (again see unsupervised admixture plot in [30]).

We have observed that the added diversity of the reference panel may indeed help in the imputation of genotypes containing rare alleles (Figure 2) but at the same time the added diversity could lead to both false positives and false negatives (Figure3, Figure S2, Table 2, Table S1) that could otherwise be avoided. As matching between target and reference alleles is made at common variants, it is reasonable to imagine that when individuals from the target and reference group come from more distant populations, their most recent common ancestor might not be particularly recent and so while they might share similar haplotypes for common (older) variants, there may be less similarity for rarer (more recent) mutations. Hence, we would propose that when the HMM encounters cases of ambiguity and several reference haplotypes will contribute to the final dosage, imputation might be slightly improved by giving priority to the haplotypes from the reference

 doi:10.20944/preprints202212.0483.v1

groups that have proved more widely relevant to the target haplotype across the genome. Such an approach would likely necessitate two passes of the HMM but given the high performance (in terms of speed and memory usage) that imputation algorithms have attained; this would not come at too high a cost. Indeed, a somewhat similar idea has already been put in place for combining imputation outputs when different reference panels have been used [40]. This meta-imputation technique also relies of an initial imputation with a second pass of an HMM to combine the inference of the different imputation realisations.

Setting aside whether or not the ideas put forward here could be practically be used to improve imputation accuracy, the results presented here certainly shed light onto the performance of imputation panels. The diversity in the sample was shown to contributing to the imputation with both positive and negative impacts. Simply removing diversity from the reference panel based on SOURCEFIND resulted in a loss of accuracy for the imputation of rare-variants. However, a more diverse panel was often shown to lead to reduced imputation accuracy; in particular leading to many false positives. This suggests that there could be situations where less could be more when choosing an imputation reference panel.

Supplementary Materials:

Figure S1: Individual AR2 (IMPUTE5 with all populations) and AR2* (IMPUTE5 with SOURCEFIND populations) statistics that correspond to those presented in Figure 2 in the main test. Individuals are coloured by their 1000 Genomes group, ACB, ASw, or MXL.

Figure S2. This figure echoes Figure 3 in the main text. The admixture plot in (a) shows the cumulative imputation barcodes used by IMPUTE5 under the informed choice approach; exactly as in Figure 3. The AR2 (aggregate R^2) overlaid in (a) and given in more detail in (b) however correspond to imputation using MINIMAC4.

Figure S3: As Figure S1 but pertaining to the 'informed choice' strategy and showing the individual aggregate R^2 statistics from Figure 3 in the main text for IMPUTE5 and Figure S2 for MINIMAC4.

Figure S4: Proportions of the genome imputed with different populations when using IMPUTE5 with the complete 23 population reference panel (top), with the informed choice strategy (middle), or the proportions as were estimated by SOURCEFIND (bottom). The top and bottom panels appear in Figure 1 in the main text, and the middle panel appears in Figure 3 in the main text, they are put together here to facilitate their comparison.

Table S1. Again, this table echoes Table 2 in the main text but the results here correspond to MINIMAC4 and not IMPUTE5.

Author Contributions: Conceptualization, AFH; methodology, TD and AFH; software, TD and AFH; validation, TD, EG, and AFH; formal analysis, TD and AFH; investigation, TD, EG, and AFH; resources, TD and AFH; data curation, TD and AFH; writing—original draft preparation, AFH; writing—review and editing, TD, EG, and AFH; visualization, TD and AFH; supervision, EG and AFH; project administration, EG and AFH; funding acquisition, EG.

Funding: This work is part of the POPGEN project supported by the French Ministry of Research in the framework of the French initiative for genomic medicine (Plan France Médecine Génomique 2025; PFMG 2025; https://www.aviesan.fr/mediatheque/fichiers/version-anglaise/actualites-en/genomic-medicine-france-2025-web). AFH is funded by POPGEN. Funding for TD came from the INSERM cross-cutting program Genomic variability 2018 GOLD (https://aviesan.fr/fr/aviesan/accueil/menu-header/instituts-thematiques-multi-organismes/genetique-genomique-et-bioinformatique/programme-transversal-gold).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data and software used here are publically available. The 1000 Genomes data was downloaded from here:

https://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes_phase3_v5a/b37.vcf/. Scripts for reproducing all results will be gladly shared on request and we plan to make them publically available.

Acknowledgments: We would like to thank Hervé Perdry and Anne-Louise Leutenegger for their insights and valuable discussions regarding this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Marchini, J.; Howie, B. Genotype Imputation for Genome-Wide Association Studies. *Nature Reviews Genetics* **2010**, *11*, 499–511, doi:10.1038/nrg2796.
2. Li, N.; Stephens, M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **2003**, *165*, 2213–2233.
3. Rubinacci, S.; Delaneau, O.; Marchini, J. Genotype Imputation Using the Positional Burrows Wheeler Transform. *PLOS Genetics* **2020**, *16*, e1009049, doi:10.1371/journal.pgen.1009049.
4. Browning, B.L.; Zhou, Y.; Browning, S.R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *The American Journal of Human Genetics* **2018**, *103*, 338–348, doi:10.1016/j.ajhg.2018.07.015.
5. Das, S.; Forer, L.; Schönherr, S.; Sidore, C.; Locke, A.E.; Kwong, A.; Vrieze, S.I.; Chew, E.Y.; Levy, S.; McGue, M.; et al. Next-Generation Genotype Imputation Service and Methods. *Nat Genet* **2016**, *48*, 1284–1287, doi:10.1038/ng.3656.
6. Rubinacci, S.; Hofmeister, R.; Mota, B.S. da; Delaneau, O. Imputation of Low-Coverage Sequencing Data from 150,119 UK Biobank Genomes 2022, 2022.11.28.518213.
7. Roshyara, N.R.; Horn, K.; Kirsten, H.; Ahnert, P.; Scholz, M. Comparing Performance of Modern Genotype Imputation Methods in Different Ethnicities. *Scientific Reports* **2016**, *6*, 34386, doi:10.1038/srep34386.
8. Marino, A.D.; Mahmoud, A.A.; Bose, M.; Bircan, K.O.; Terpolovsky, A.; Bamunusinghe, V.; Bohn, S.; Khan, U.; Novković, B.; Yazdi, P.G. A Comparative Analysis of Current Phasing and Imputation Software. *PLOS ONE* **2022**, *17*, e0260177, doi:10.1371/journal.pone.0260177.
9. Herzig, A.F.; Nutile, T.; Babron, M.-C.; Ciullo, M.; Bellenguez, C.; Leutenegger, A.-L. Strategies for Phasing and Imputation in a Population Isolate. *Genetic Epidemiology* **2018**, *42*, doi:10.1002/gepi.22109.
10. Vergara, C.; Parker, M.M.; Franco, L.; Cho, M.H.; Valencia-Duarte, A.V.; Beaty, T.H.; Duggal, P. Genotype Imputation Performance of Three Reference Panels Using African Ancestry Individuals. *Hum Genet* **2018**, *137*, 281–292, doi:10.1007/s00439-018-1881-4.
11. Kowalski, M.H.; Qian, H.; Hou, Z.; Rosen, J.D.; Tapia, A.L.; Shan, Y.; Jain, D.; Argos, M.; Arnett, D.K.; Avery, C.; et al. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium Whole Genome Sequences Improves Imputation Quality and Detection of Rare Variant Associations in Admixed African and Hispanic/Latino Populations. *PLOS Genetics* **2019**, *15*, e1008500, doi:10.1371/journal.pgen.1008500.
12. Rubinacci, S.; Ribeiro, D.M.; Hofmeister, R.J.; Delaneau, O. Efficient Phasing and Imputation of Low-Coverage Sequencing Data Using Large Reference Panels. *Nat Genet* **2021**, *53*, 120–126, doi:10.1038/s41588-020-00756-0.
13. Surakka, I.; Kristiansson, K.; Anttila, V.; Inouye, M.; Barnes, C.; Moutsianas, L.; Salomaa, V.; Daly, M.; Palotie, A.; Peltonen, L.; et al. Founder Population-Specific HapMap Panel Increases Power in GWA Studies through Improved Imputation Accuracy and CNV Tagging. *Genome Res* **2010**, *20*, 1344–1351, doi:10.1101/gr.106534.110.
14. Joshi, P.K.; Prendergast, J.; Fraser, R.M.; Huffman, J.E.; Vitart, V.; Hayward, C.; McQuillan, R.; Glodzik, D.; Polašek, O.; Hastie, N.D.; et al. Local Exome Sequences Facilitate Imputation of Less Common Variants and Increase Power of Genome Wide Association Studies. *PLOS ONE* **2013**, *8*, e68604, doi:10.1371/journal.pone.0068604.
15. Zeggini, E. Next-Generation Association Studies for Complex Traits. *Nat Genet* **2011**, *43*, 287–288, doi:10.1038/ng0411-287.
16. Pistis, G.; Porcu, E.; Vrieze, S.I.; Sidore, C.; Steri, M.; Danjou, F.; Busonero, F.; Mulas, A.; Zoledziewska, M.; Maschio, A.; et al. Rare Variant Genotype Imputation with Thousands of Study-Specific Whole-Genome Sequences: Implications for Cost-Effective Study Designs. *European Journal of Human Genetics* **2015**, *23*, 975–983, doi:10.1038/ejhg.2014.216.
17. Howie, B.; Marchini, J.; Stephens, M. Genotype Imputation with Thousands of Genomes. *G3 (Bethesda)* **2011**, *1*, 457–470, doi:10.1534/g3.111.001198.
18. Huang, J.; Howie, B.; McCarthy, S.; Memari, Y.; Walter, K.; Min, J.L.; Danecek, P.; Malerba, G.; Trabetti, E.; Zheng, H.-F.; et al. Improved Imputation of Low-Frequency and Rare Variants Using the UK10K Haplotype Reference Panel. *Nature Communications* **2015**, *6*, 8111, doi:10.1038/ncomms9111.
19. Chou, W.-C.; Zheng, H.-F.; Cheng, C.-H.; Yan, H.; Wang, L.; Han, F.; Richards, J.B.; Karasik, D.; Kiel, D.P.; Hsu, Y.-H. A Combined Reference Panel from the 1000 Genomes and UK10K Projects Improved Rare Variant Imputation in European and Chinese Samples. *Scientific Reports* **2016**, *6*, 39313, doi:10.1038/srep39313.
20. Mitt, M.; Kals, M.; Pärn, K.; Gabriel, S.B.; Lander, E.S.; Palotie, A.; Ripatti, S.; Morris, A.P.; Metspalu, A.; Esko, T.; et al. Improved Imputation Accuracy of Rare and Low-Frequency Variants Using Population-Specific High-Coverage WGS-Based Imputation Reference Panel. *European Journal of Human Genetics* **2017**, *25*, 869–876, doi:10.1038/ejhg.2017.51.

21. Quick, C.; Anugu, P.; Musani, S.; Weiss, S.T.; Burchard, E.G.; White, M.J.; Keys, K.L.; Cucca, F.; Sidore, C.; Boehnke, M.; et al. Sequencing and Imputation in GWAS: Cost-Effective Strategies to Increase Power and Genomic Coverage across Diverse Populations. *Genetic epidemiology* **2020**, *44*, 537–549, doi:10.1002/gepi.22326.
22. Deelen, P.; Menelaou, A.; van Leeuwen, E.M.; Kanterakis, A.; van Dijk, F.; Medina-Gomez, C.; Francioli, L.C.; Hottenga, J.J.; Karssen, L.C.; Estrada, K.; et al. Improved Imputation Quality of Low-Frequency and Rare Variants in European Samples Using the "Genome of The Netherlands." *Eur J Hum Genet* **2014**, *22*, 1321–1326, doi:10.1038/ejhg.2014.19.
23. Herzig, A.F.; Velo-Suárez, L.; Frex Consortium; FranceGenRef Consortium; Dina, C.; Redon, R.; Deleuze, J.-F.; Génin, E. Can Imputation in a European Country Be Improved by Local Reference Panels? The Example of France. *bioRxiv* **2022**, 480829, doi:10.1101/2022.02.17.480829.
24. Yasuda, J.; Katsuoka, F.; Danjoh, I.; Kawai, Y.; Kojima, K.; Nagasaki, M.; Saito, S.; Yamaguchi-Kabata, Y.; Tadaka, S.; Motoike, I.N.; et al. Regional Genetic Differences among Japanese Populations and Performance of Genotype Imputation Using Whole-Genome Reference Panel of the Tohoku Medical Megabank Project. *BMC Genomics* **2018**, *19*, 551, doi:10.1186/s12864-018-4942-0.
25. Cocca, M.; Barbieri, C.; Concas, M.P.; Robino, A.; Brumat, M.; Gandin, I.; Trudu, M.; Sala, C.F.; Vuckovic, D.; Girotto, G.; et al. A Bird's-Eye View of Italian Genomic Variation through Whole-Genome Sequencing. *Eur J Hum Genet* **2020**, *28*, 435–444, doi:10.1038/s41431-019-0551-x.
26. Kals, M.; Nikopensius, T.; Läll, K.; Pärn, K.; Tõnis Sikka, T.; Suvisaari, J.; Salomaa, V.; Ripatti, S.; Palotie, A.; Metspalu, A.; et al. Advantages of Genotype Imputation with Ethnically Matched Reference Panel for Rare Variant Association Analyses. *bioRxiv* **2019**, 579201, doi:10.1101/579201.
27. McCarthy, S.; Das, S.; Kretzschmar, W.; Delaneau, O.; Wood, A.R.; Teumer, A.; Kang, H.M.; Fuchsberger, C.; Danecek, P.; Sharp, K.; et al. A Reference Panel of 64,976 Haplotypes for Genotype Imputation. *Nature Genetics* **2016**, *48*, 1279–1283, doi:10.1038/ng.3643.
28. Baum, L.E.; Petrie, T.; Soules, G.; Weiss, N. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics* **1970**, *41*, 164–171, doi:10.1214/aoms/1177697196.
29. Rabiner, L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE* **1989**, *77*, 257–286, doi:10.1109/5.18626.
30. The 1000 Genomes Project Consortium A Global Reference for Human Genetic Variation. *Nature* **2015**, *526*, 68–74, doi:10.1038/nature15393.
31. Lawson, D.J.; Hellenthal, G.; Myers, S.; Falush, D. Inference of Population Structure Using Dense Haplotype Data. *PLOS Genetics* **2012**, *8*, e1002453, doi:10.1371/journal.pgen.1002453.
32. Durbin, R. Efficient Haplotype Matching and Storage Using the Positional Burrows-Wheeler Transform (PBWT). *Bioinformatics* **2014**, *30*, 1266–1272, doi:10.1093/bioinformatics/btu014.
33. Alexander, D.H.; Novembre, J.; Lange, K. Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Res* **2009**, *19*, 1655–1664, doi:10.1101/gr.094052.109.
34. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.W.; Daly, M.J.; et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **2007**, *81*, 559–575.
35. Chacón-Duque, J.-C.; Adhikari, K.; Fuentes-Guajardo, M.; Mendoza-Revilla, J.; Acuña-Alonzo, V.; Barquera, R.; Quinto-Sánchez, M.; Gómez-Valdés, J.; Everardo Martínez, P.; Villamil-Ramírez, H.; et al. Latin Americans Show Wide-Spread Converso Ancestry and Imprint of Local Native Ancestry on Physical Appearance. *Nat Commun* **2018**, *9*, 5388, doi:10.1038/s41467-018-07748-z.
36. Perdry, H.; Dandine-Rolland, C.; Banddyopadhyay, D.; Kettner, L. Gaston: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models. *CRAN* **2018**, https://cran.r-project.org/web/packages/gaston/index.html.
37. Delaneau, O.; Marchini, J.; Zagury, J.-F. A Linear Complexity Phasing Method for Thousands of Genomes. *Nat Methods* **2011**, *9*, 179–181, doi:10.1038/nmeth.1785.
38. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve Years of SAMtools and BCFtools. *Gigascience* **2021**, *10*, giab008, doi:10.1093/gigascience/giab008.
39. Vince, N.; Douillard, V.; Geffard, E.; Meyer, D.; Castelli, E.C.; Mack, S.J.; Limou, S.; Gourraud, P. SNP-HLA Reference Consortium (SHLARC): HLA and SNP Data Sharing for Promoting MHC-centric Analyses in Genomics. *Genet Epidemiol* **2020**, *44*, 733–740, doi:10.1002/gepi.22334.
40. Yu, K.; Das, S.; LeFaive, J.; Kwong, A.; Pleiness, J.; Forer, L.; Schönherr, S.; Fuchsberger, C.; Smith, A.V.; Abecasis, G.R. Meta-Imputation: An Efficient Method to Combine Genotype Data after Imputation with Multiple Reference Panels. *Am J Hum Genet* **2022**, *109*, 1007–1015, doi:10.1016/j.ajhg.2022.04.002.