# 3 Imputation

**Disclaimer**

The content of this chapter is now published as a paper (Huang et al. 2015). Text written in this chapter might overlap substantially with text in the published paper. In this chapter, I use “I” for the work that was mainly done by myself alone, while indicate clearly for work done by others. Bryan Howie, the co-first author of the published paper, implemented the IMPUTE2 software for merging reference panels and for using a new metric to sample haplotypes.

## 3.1 Introduction

### 3.1.1 How imputation works

Imputation is a statistical inference of missing genotypes, where genotyped markers from SNP arrays are used to infer unobserved genotypes from haplotype panels. Although there are quite a few different software for running imputation, the common underlying method is based on a hidden Markov model (HMM) that treats a sample haplotype as a mosaic of a pool of reference haplotypes and uses haplotype patterns in a reference panel to predict unobserved genotypes in a study dataset (Li and Stephens 2003, Scheet and Stephens 2006, Marchini et al. 2007, Browning and Browning 2009, Li et al. 2009). Imputation using large reference panels such as 1000GP has been made computationally efficient by pre-phasing of GWAS samples (Howie et al. 2012) and approximations that select a subset of reference haplotypes (Howie et al. 2011).

### 3.1.2 Use of imputation in GWAS

Imputation has been instrumental to the discovery of thousands of complex trait loci in genome-wide association studies (GWAS) (Howie et al. 2009). Imputation not only boosts genetic data through a most cost-effective approach and therefore increases statistical power, but also generates datasets with common list of SNPs that facilitate broad collaboration. By imputing individual SNP array dataset with customized content to the common set of variants in HapMap (International HapMap et al. 2007, International HapMap et al. 2010), the international society has been able to look at a common set of ~3 million variants across different cohorts and projects.

### 3.1.3 Imputation with WGS reference panels

Those variants in the HapMap reference panel are mainly common across populations, defined as MAF >5%. Although WGS provides near-complete characterization of genetic variation, it is still prohibitive for researchers to conduct WGS on large number of samples that are needed to study the effect on phenotypic variation by rare variants. Instead, using publically available WGS data as reference panels to impute existing datasets with genome-wide SNP array data would be a most cost-effective alternative. Built upon from the HapMap project, the 1000GP provides phased haplotypes for more than a thousand samples from diverse worldwide populations, thereby boosting variant coverage and imputation quality, particularly for variants with MAF of 1-5% (Abecasis et al. 2012). In my early work, I showed that imputations using the 1000GP data could identify novel genetic variants that were not identified in SNP arrays or through HapMap based imputation (Huang et al. 2012).

The 1000GP imputation reference panel currently widely used (Phase 1 version 3) includes a total of 1092 samples, 381 of which are European. In contrast, the UK10K project conducted WGS for 3,781 European samples with higher depth (~7X), and is powered to detect and impute variants with MAF down to 0.1% (The UK10K Consortium 2015). Using the UK10K panel or using the combination of UK10K and 1000GP are expected to provide more accurate imputation for low frequency and rare variants, which are a most effective approach for increasing statistical power along with a large sample size. Here I evaluate the utility of the UK10K WGS dataset as an imputation reference panel, above and beyond the WGS data from 1000GP.

### 3.1.4 Aims of this study

As the imputation reference panel includes thousands of reference haplotypes and tens of millions of variants, for each of the thousands of samples on the target panel to be imputed, the ideal scenario is that the best matched haplotype exists in the reference haplotype pool while the imputation program does not need to scan all haplotypes in order to use it for imputation. Combining multiple reference panels could improve the representativeness of the reference haplotype pool, while designing an algorithm to quickly narrow down the best matched haplotypes would substantially save computation time and cost. Therefore, the evaluation steps aims to find a preferred imputation strategy that maximizes haplotype representativeness and minimizes computational resources.

**Evaluation on performance of WGS reference panels**

Recently, a new option in the IMPUTE2 software (Howie et al. 2009, Howie et al. 2011) allowed two sets of haplotypes to be combined to form a single set of haplotypes at the union set of sites. Imputation into GWAS samples can then be carried out using this combined panel. This method can be used to combine two sets of haplotypes from two distinct population cohorts, such as UK10K and 1000GP, as described in this chapter. The results from my evaluation of UK10K and 1000GP should also help investigators who wish to use their own WGS data instead of UK10K to merge with 1000GP data.

The main difficulty in combining reference panels is that some sites will only have data in one or other of the panels. This could be due to population specific alleles, low-coverage of the non-reference allele, or cohort specific site filtering that removed the site from consideration. The new option in IMPUTE2 software uses HMM to impute the unobserved alleles in each panel while the other panel is used as reference. Once the two reference panels are imputed up to the union of their variants, the best-guess haplotypes are used to impute a GWAS cohort in the same way as using only one reference panel. IMPUTE2 could output the haplotypes of the merged reference panel so that they are used for future imputation without repeating this merging step. This new functionality is available in IMPUTE2 v2.3.0 or newer version (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html).

## Evaluation on approximation of haplotype sampling

Genotype imputation in GWAS has always been a computationally intensive task. Recent developments like pre-phasing have greatly reduced the computational cost of imputation, but growing reference panels continue to challenge existing methods. Previously, IMPUTE2 chose a different subset of $k_{hap}$ reference haplotypes (by default, 500) for each GWAS haplotype. The matching was based on an approximation of hamming distance metric. When this subset includes the most informative reference haplotypes, it can speed up the imputation calculations without sacrificing much accuracy. The cost of imputation with pre-phased GWAS data scales linearly with the number of reference haplotypes $N$, so the speedup expected from this approximation is roughly $N$ / $k_{hap}$ after accounting for the overhead of reading in a large data set. This speed-up would matter significantly since there are around ~10,000 haplotypes in the combined UK10K and 1000GP reference panel.

## 3.2 Methods

The various evaluations to be conducted aim to address the two key questions stated above: 1. Does UK10K reference panel perform better than 1000GP, or combining these two panels together would perform even better? 2. Is there a cost-effective approach for sampling only some of the reference haplotypes for imputing each sample? For the testing evaluations described in this chapter, the reference panels are UK10K WGS and 1000GP WGS, and the target panels are two pseudo-GWAS where some genetic variants are masked out to mimic the content of a SNP array panel. The masked out variants would then be used as “true” data to compare with the imputed data for the same sites and same sample. The evaluation was done sequentially. Once a preferred metric or design is identified in one round, the less preferred metrics or designs will not be evaluated again in the following rounds.

### 3.2.1 WGS Reference Haplotypes

**UK10K WGS**

The UK10K WGS data included 3,781 samples and contained over 42 million SNV and ~3.5 million insertion/deletion polymorphisms. To assess the quality of genotype data from low-depth sequencing, the UK10K study compared the variant sites and genotypes of 61 TwinsUK individuals with high-coverage exome data. A high level of concordance was observed (**Table 3.1**). Originally, the UK10K WGS panel was phased by Beagle during the genotype refinement step. In 2013, it was reported that re-phasing the 1000GP WGS panel using SHAPEIT v2 led to improved imputation quality (Delaneau et al. 2013), I therefore used SHAPEIT v2 for re-phasing the UK10K reference haplotypes. Per the recommendation of this software, the mean size of the windows in which conditioning haplotypes are defined is set to 0.5MB, instead of 2MB used for pre-phasing GWAS. Due to the significantly higher number of variants in the WGS data, the re-phasing was conducted by 3MB chunk with 250kb buffering regions, rather than by whole chromosomes as for the pseudo-GWAS. Imputation was carried out on the same chunks with the same flanking regions. To re-phase the UK10K final release sequencing data, I first converted the VCF files into PLINK binary format, each chromosomes split into 3MB chunks with +/-250kb flanking regions. I then used SHAPEIT v2 to re-phrase the haplotypes for each 3MB chunks with +/-250kb flanking regions. Although the chunk files could be used as reference panels directly, I also created whole chromosome files based on these re-phased chunks. To do that, phasing information from the SHAPEIT output was copied back to the original VCF files, by using the vcf-phased-join program from the VCFTOOLS package (Danecek et al.).

To merge UK10K reference panel with 1000GP reference panel for creating a combined reference panel, I first identified sites that need to be excluded. For UK10K, the following sites were excluded: 18,180,633 singletons that do not exist in 1000GP, 1,064,168 multi-allelic sites and 214,631mis-matched alleles sites. For 1000GP, the following sites were excluded: 7,053,246 singletons that do not exist in UK10K, 23,932 sites with a SNP and an INDEL at the same position and 443 within large structural deletions (**Table 3.2**). To identify these variants, I first used VCF-QUERY to get the summary statistics of the two sets of VCF files, including chromosome, position, reference and alternative alleles, and then compare the two summary statistics files against each other. I then used VCFTOOLS to exclude those sites to create a new set of VCF files. Finally, I used VCF-QUERY to convert the new VCF files into phased haplotypes and legend files that could be fed directly to IMPUTE2 for running imputation.

### 1000GP WGS

The 1000GP Phase I integrated variant set release (v3) for low-coverage whole-genomes in NCBI build 37 (hg19) coordinates was downloaded from 1000GP FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/, 23 Nov 2010 data freezes). This callset includes phased haplotypes for 1,092 individuals and 39,527,072 variants (22 autosome and chromosome X). The haplotypes were inferred from a combination of low-coverage genome sequence data, and they contain SNPs, short INDELs, and large deletions. As mentioned above, the following sites were excluded: 7,053,246 singletons that do not exist in UK10K, 23,932 sites with a SNP and an INDEL at the same position and 443 within large structural deletions. The final reference panel included all 1,092 samples and 32,449,428 sites.

### Merging two WGS reference panels

The following 3 steps were used to merge two WGS reference panels using IMPUTE2 (version 2.3 and later):

1. Impute the variants that are specific to panel 1 (1000GP) into panel 2 (UK10K).
2. Impute the variants that are specific to panel 2 (UK10K) into panel 1 (1000GP).
3. Treat the imputed haplotypes in both panels (with the union of variants from both) as known (i.e., take the best-guess haplotypes) and impute the GWAS cohort in the usual way.

**Data access**

UK10K reference haplotypes are available from the European Genome-phenome archive (EGA study: EGAS00001000713, EGA dataset: EGAD00001000776) under managed access conditions (see http://www.uk10k.org/data_access).

### 3.2.2 Test GWAS datasets

**UK10K Pseudo-GWAS**

A random set of 500 samples passing QC filters were chosen from the TwinsUK (N=1,854) and ALSPAC (N=1,927) WGS datasets. Genotypes for a total of 13,413 sites (corresponding to the content of the Illumina HumanHap610 SNP-array) on chromosome 20 were extracted from the UK10K WGS data in these 1,000 samples.

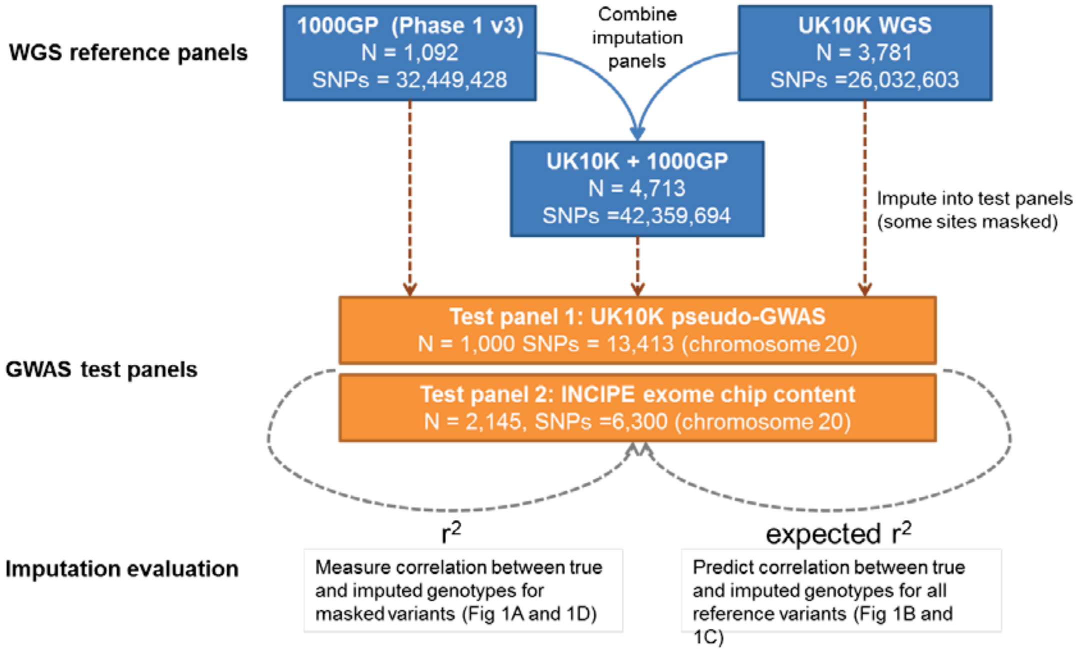
**INCIPE Pseudo-GWAS**

For the INCIPE study, 6,200 Caucasian participants were randomly chosen from the lists of registered patients of 62 randomly selected general practitioners based in four geographical areas in the Veneto region, north-eastern Italy (Gambaro et al. 2010). A total of total of 2,258 samples were genotyped with the HumanCoreExome-12v1-1 platform and were subject to further quality control (QC) evaluation as follows to determine sample and SNP quality. The details of QC for this dataset is presented elsewhere (Huang et al.). At the end, there are a total of 346,941 polymorphic variants on autosomes and 8,822 of those on chromosome 20 were retained for analysis. For the imputation evaluation, 2,522 exonic variants (i.e. those corresponding to the exome selected part of the array) on chromosome 20 were masked out. The remaining 6,300 SNPs were retained as a pseudo-GWAS imputation panel.

### 3.2.3 Running imputation

Prior to imputation, the two pseudo-GWAS datasets were pre-phased using SHAPEIT v2 (Delaneau et al. 2013) to increase phasing accuracy. The UK10K pseudo-GWAS panel was phased jointly with those samples in UK10K WGS. The INCIPE pseudo-GWAS of 2,145 participants was pre-phased separately. Imputation of genotypes from the three phased reference panels (UK10K, 1000GP and UK10K+1000GP) into the two test panels was carried out on chromosome 20, split in 3MB chunks with 250kb buffer regions. Imputation was performed using standard parameters with IMPUTE2. The accuracy of imputed variants was calculated as the Pearson correlation coefficient ($r^2$) between imputed genotype dosages in [0-2] and masked sequence genotypes in (0,1,2). The results were stratified into non-overlapping MAF bins for plotting. The overall flow of imputation evaluation is shown in **Figure 3.1**.

**Figure 3.1** imputation evaluation workflow



WGS reference panels
1000GP (Phase 1 v3) N = 1,092 SNPs = 32,449,428
Combine imputation panels
UK10K WGS N = 3,781 SNPs =26,032,603
UK10K + 1000GP N = 4,713 SNPs =42,359,694
Impute into test panels (some sites masked)
GWAS test panels
Test panel 1: UK10K pseudo-GWAS N = 1,000 SNPs = 13,413 (chromosome 20)
Test panel 2: INCIPE exome chip content N = 2,145, SNPs =6,300 (chromosome 20)
Imputation evaluation
r2
Measure correlation between true and imputed genotypes for masked variants (Fig 1A and 1D)
expected r2
Predict correlation between true and imputed genotypes for all reference variants (Fig 1B and 1C)

## 3.3 Results

### 3.3.1 Characteristics of UK10K WGS panel

The UK10K Cohorts Project (http://www.uk10k.org/studies/cohorts.html) includes two population samples from the UK. The TwinsUK registry comprises unselected, mostly female volunteers ascertained from the general population through national media campaigns in the UK (Moayyeri et al. 2012). The Avon Longitudinal Study of Parents and Children (ALSPAC) is a population-based birth cohort study that recruited more than 13,000 pregnant women resident in Bristol (formerly Avon) UK (Golding et al. 2001). A total of 1,990 individuals from TwinsUK and 2,040 individuals from ALSPAC were consented for sequencing. Variant sites and genotype likelihoods were called using SAMtools (Li et al. 2009), and genotypes were refined and phased using Beagle (Browning and Browning 2009), following similar procedures to the 1000GP (Abecasis et al. 2012). After quality control, 45,492,035 variant sites were retained (**Table 3.2**) in 1,854 and 1,927 individuals in the TwinsUK and ALSPAC panels, respectively. I downloaded the phased haplotypes of 1000GP (Phase 1 integrated v3), which include a total of 39,527,072 sites. For imputation, I removed multi-allelic sites and further excluded variants seen only once in the combined 1000GP+UK10K dataset. A total of 26,032,603 sites were retained for the imputation reference panel of UK10K panel, and 32,449,428 sites for the imputation reference panel of 1000GP. Given that 16,122,337 exist in both panels, combining the two reference panels results in a total of 42,359,694 sites (**Table 3.2**).

**Table 3.1** Sequence quality and variation metrics for UK10K Cohorts

This table was adopted from the UK10K study. The numbers in the table was provided by Klaudia Walter. For 61 overlapping TwinsUK individuals, the UK10K study compared the variant sites and genotypes of the low-coverage sequences with high-coverage exome data by non-overlapping AF bins (WGS versus Exomes). It considered 74,621 shared sites in non-overlapping AF bins, and calculated (i) the fraction of concordant over total sites, (ii) Non-Ref genotypes, (NRD, %) = number of non-reference genotypes and non-reference genotype discordance (NRD, in %) between WGS and Exomes; (iii) False discovery rate (FDR = FP=(FP + TP)), where it considered the exomes as the truth set; (iv) number of false positives (FP) and FDR for sites that are or not shared with the 1000 Genomes Project, PhaseI (1000GP); (v) false negative rate (FNR = FN=(FN + TP)), where AF bins were defined based on the 61 exomes. Furthermore, it compared 22 monozygotic (MZ) twin pairs at 880,280 bi-allelic SNV sites on chromosome 20, reporting (i) the percentage of concordant genotypes, non-reference genotypes and NRD. AF are from the set of 3,621 samples, which contains at most one of the two MZ twins from each pair. The discrepancies can be caused by errors in either twin, so the expected NRD to the truth would be half the NRD value given.

| | **WGS vs. Exomes** | | | | | | **MZ Twins** | |
|---|---|---|---|---|---|---|---|---|
| **AF** | **Total sites (concordant, %)** | **Non-Ref genotypes (NRD, %)** | **FP (FDR, %)** | **FP in 1000GP (FDR, %)** | **FP not in 1000GP (FDR, %)** | **FNR (%)** | **Total sites (concordant, %)** | **Non-Ref genotypes (NRD, %)** |
| **AC=1** | 2,963 (99.999) | 2,965 (0.1) | 125 (4.0) | 11 (3.8) | 114 (4.1) | n.a. | 411,583 (99.995) | 3,534 (12.7) |
| **AC=2** | 1,566 (99.998) | 1,577 (0.1) | 147 (8.6) | 25 (7.9) | 122 (8.7) | n.a. | 101,116 (99.989) | 1,594 (15.1) |
| **0:03-1%** | 16,303 (99.928) | 21,114 (3.3) | 1,160 (6.6) | 766 (5.5) | 394 (11.3) | 27.2 | 193,531 (99.954) | 19,034 (10.2) |
| **1-5%** | 16,356 (99.829) | 53,165 (3.2) | 1,038 (6.0) | 980 (5.7) | 58 (68.2) | 6.4 | 50,360 (99.776) | 56,554 (4.4) |
| **>5%** | 37,433 (99.688) | 1,151,178 (0.6) | 2,668 (6.7) | 2,653 (6.6) | 15 (46.9) | 7.3 | 123,690 (99.574) | 1,382,934 (0.8) |

**Table 3.2** Descriptive for imputation reference panels

For UK10K, the following sites were excluded: 18,180,633singletons that do not exist in 1000GP, 1,064,168 multi-allelic sites and 214,631mis-matched alleles sites. For 1000GP, the following sites were excluded: 7,053,246 singletons that do not exist in UK10K, 23,932 sites with a SNP and an INDEL at the same position and 443 within large structural deletions.

| | **UK10K** | **1000GP**<br>**(Phase 1 v3)** | **Combined** | **Overlap** |
|---|---|---|---|---|
| N samples (% European) | 3,781 (100%) | 1,092 (34.7%) | 4,873 | -- |
| N total sites in final release | 45,492,035 | 39,527,072 | -- | |
| N total sites after filtering | 26,032,603 | 32,449,428 | 42,359,694 | 16,122,337 |
| Autosome SNPs | 23,411,635 | 29,797,220 | 38,238,102 | 14,970,753 |
| Autosome INDELs | 1,698,262 | 1,370,819 | 2,407,858 | 661,223 |
| Chr X SNPs | 858,380 | 1,223,328 | 1,612,230 | 469,478 |
| Chr X INDELs | 64,326 | 58,061 | 101,504 | 20,883 |

### 3.3.2 Imputation evaluation on UK10K vs. 1000GP reference panels

As a first assessment of the UK10K reference panel, I performed a leave-one-out cross-validation on a sub-sample of 1,000 individuals from the UK10K WGS dataset (500 from TwinsUK and 500 from ALSPAC). I removed each sample from the reference panel in turn, selected 13,413 sites on chromosome 20 from the Illumina 610k bead chip, and imputed all other sites on this chromosome from a given reference panel. The imputation was conducted with three haplotype reference panels: the 1000GP panel, the “original” UK10K panel produced by initial genotype refinement and haplotyping with BEAGLE, and a “re-phased” UK10K panel that was generated by using SHAPEIT2 to estimate haplotypes from the BEAGLE genotypes. The accuracy of imputed variants was calculated as the Pearson correlation coefficient ($r^2$) between imputed genotype dosages in [0-2] and masked sequence genotypes in [0,1,2]. The results were stratified into non-overlapping MAF bins for plotting.

The results of this experiment are shown in **Figure 3.2A**, which focuses on variants with MAF<5%. Both UK10K reference panels (blue dotted and solid lines) produced higher accuracy than the 1000GP panel (black line), with greater gains at lower frequencies. These trends were expected due to the larger sample size and better ancestry matching of the UK10K reference panel to the pseudo-GWAS data. Notably, the UK10K reference panel yielded much higher imputation accuracy after re-phasing with SHAPEIT2 (solid vs. dotted blue lines): the mean $r^2$ at low frequencies increased by more than 0.1 (20%) after re-phasing, which implies a substantial boost in the power to detect associations. A large imputation panel is a resource that can inform a variety of association studies, so these results suggest that taking the time to improve a WGS panel’s haplotype quality could have substantial downstream benefits. Most recently, I evaluated the added value of using UK10K WGS reference panel on top of the latest 1000GP reference panel (phase 3), based on a US population (FHS samples). I observed significant improvement when adding the UK10K panel on top of 1000GP. At the MAF of 0.002, 0.01, 0.1, the mean r2 value increased from 0.438, 0.522, 0.844 to 0.532, 0.621, 0.876 respectively. This evaluation was based on pseudo-GWAS of 320 FHS WGS samples.

### 3.3.3 Evaluation of metrics for choosing reference haplotypes

I noticed that some rare variants were imputed much better when using the entire UK10K reference panel to drive imputation, yet poorly when using IMPUTE2's $k_{hap}$ approximation. This approximation reduces the computational cost of imputation by using a region-wide (e.g., across a 3MB imputation chunk) Hamming distance metric to reduce the number of reference haplotypes used by a given GWAS haplotype. The investigation of these variants led to the development of a new approximation that uses local (rather than region-wide) haplotype sharing to choose a subset of reference haplotypes. This was done by Bryan Howie. This approximation delivers the same substantial speed boost as the existing $k_{hap}$ approximation, but it does not sacrifice imputation accuracy at rare and low-frequency variants. For example, **Figure 3.2 B** shows the results of imputing the INCIPE pseudo-GWAS data with the UK10K reference panel. The full UK10K panel produced the highest accuracy (solid blue line), while the $k_{hap}$ approximation based on Hamming distance (solid orange line) was less accurate for SNPs with MAF<5%. By contrast, the new approximation based on haplotype tract sharing (dashed orange line) was nearly as accurate as the full reference panel, at ~10% of the computing time. Further speed improvements are possible for a modest price in accuracy. The evaluation of different K_hap (500 vs. 7562) and different sampling algorithm (tract sharing vs. hamming distance) was only run using the Italian isolates data. This is because imputing the UK10K pseudo-GWAS would need the leave-one-out approach, which would add an extra layer of complexity to the evaluation. Of note, the INCIPE pseudo-GWAS was generated from a SNP array data, not from WGS. Therefore, the number of variants masked out is much smaller and that in the UK10K pseudo-GWAS, and the r2 value between the two plots should be compared with this in mind.

The goal behind this new approximation is to ensure that each site in a study haplotype has the opportunity to copy the reference haplotype with the longest shared tract of allelic identity. The algorithm works as follows, from the point of view of a single GWAS haplotype:

1. For each reference haplotype, identify sets of contiguous sites that show no allele mismatches with the study haplotype; store these shared haplotype tracts for each reference haplotype.
2. At each site, generate a hash table whose keys are shared tract lengths (in genetic map units) and whose values are indices of the corresponding reference haplotypes. A given key can map to multiple values.

3. At each site, use the hash table created in the previous step to generate a list of reference haplotype indices ranked in descending order of shared tract length. Ties are broken at random.
4. Add the top-ranked haplotype index at each site to a list of unique reference haplotype indices; these states are marked for copying by the current study haplotype.
5. Go to the next-ranked haplotype index (“level”) and repeat Step 4 until $k_{hap}$ distinct reference haplotypes have been identified. If the number of selected haplotypes exceeds $k_{hap}$ at a particular level, choose a random subset of the reference indices at that level such that the total number of selected haplotypes is $k_{hap}$.

The advantage of the newly proposed tract sharing metric was illustrated in **Figure 3.3**. The computational cost of imputing a study haplotype with the Hamming distance approximation is *O(MN)*, where *M* is the number of sites shared between the study and reference panels and *N* is the number of reference haplotypes. By comparison, the cost of this new tract length approximation is roughly *O(4MN)* – the factor of four appears because this approximation scans the sites in a region multiple times. While the tract sharing approximation requires more calculations, it is still linear in *M* and *N*, and the Hamming distance approximation accounts for less than 0.2% of a typical imputation run (as determined by profiling the IMPUTE2 C++ code when imputing the INCIPE pseudo-GWAS with the UK10K reference panel). In summary, the new tract sharing approximation has a similar computational cost to the Hamming distance approximation of (Howie et al. 2011), but it is better at maintaining imputation accuracy for low-frequency and rare SNPs. This will be a useful approach as imputation reference panels continue to grow.

**Figure 3.2** Imputation performance for different reference panels and strategies

**(A)** Imputation accuracy in the UK10K pseudo-GWAS test panel using reference panels from 1000GP (black) and UK10K (blue). (**B**). Imputation accuracy in the INCIPE pseudo-GWAS panel using the UK10K reference panel and different imputation approximations.

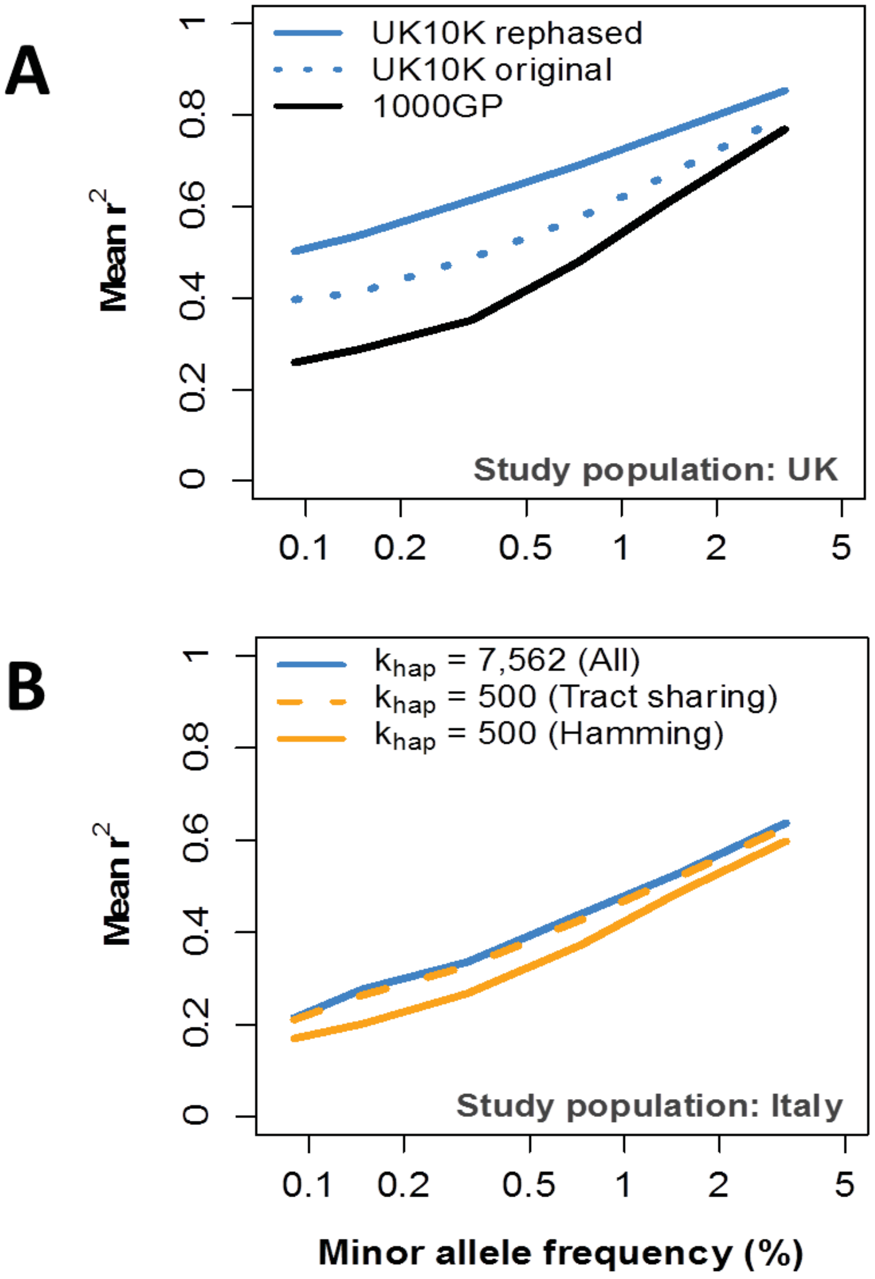
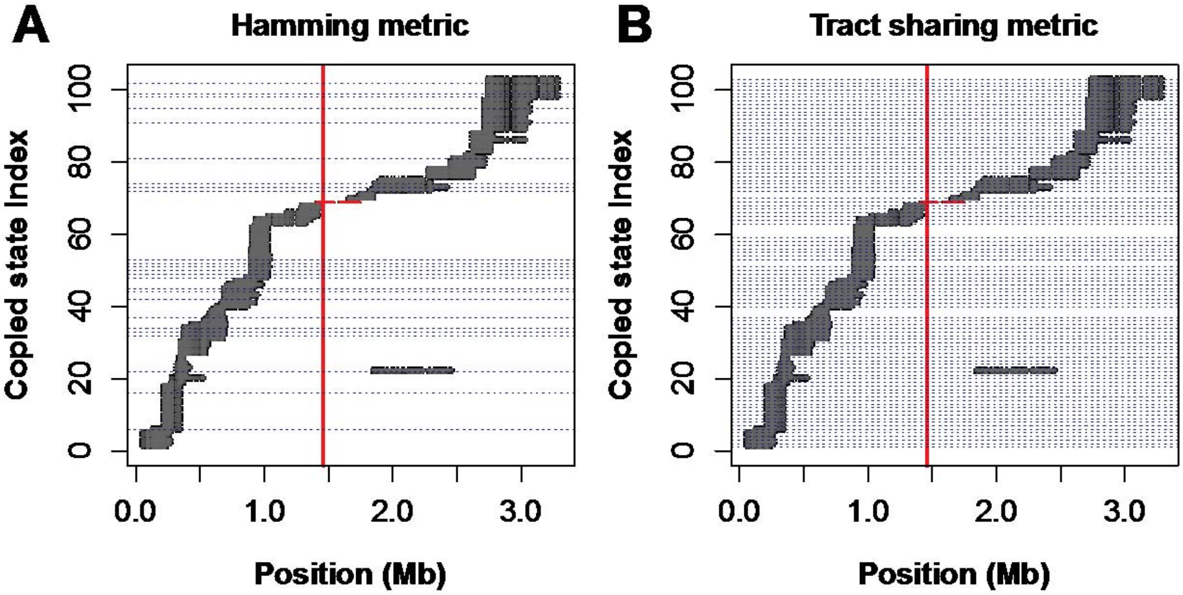


**Figure 3.3** Illustration of reference states (haplotypes) copied by IMPUTE2

This figure is based on imputing one INCIPE pseudo-GWAS haplotype from the UK10K reference panel in a 3Mb region on chromosome 20. Points at each position on the chromosome (x-axis) represent reference haplotypes that were copied with marginal (per-site) posterior probabilities of at least 0.01 when using the full UK10K reference panel (7,562 haplotypes). Copied reference haplotypes are ordered on the y-axis by the position at which they first surpassed this threshold. The location of the SNP examined is marked by a vertical red line, and points belonging to the haplotype that carries this variant are also coloured red. Subsets of reference states selected by different approximations are marked by dotted blue lines. **(A)** Reference states selected with $k_{hap}$=500 under a Hamming distance approximation. Of the 103 copied states in this plot, 25 (24%) were chosen under this approximation. **(B)** Reference states selected with $k_{hap}$=500 under a tract sharing approximation. Of the 103 copied states in this plot, 96 (93%) were chosen under this approximation.

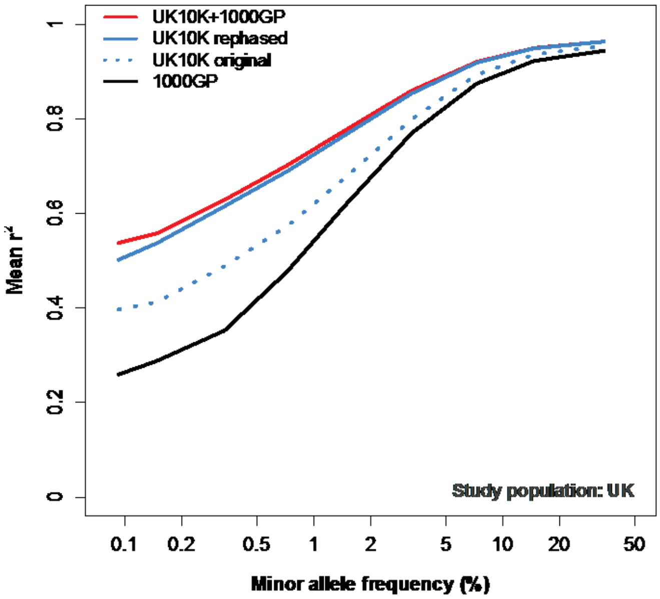


### 3.3.4 Evaluation of combining two reference panels

**Figure 3.4** shows how a combined 1000GP+UK10K panel (red) produced by this method performed against each panel separately (1000GP, black; UK10K, blue) when imputing a pseudo-GWAS of UK ancestry. The combined and UK10K panels produced very similar numbers of high-confidence (predicted $r^2$>0.8) variants at MAFs of 0.5% and higher, implying that the combined panel is neither helpful nor harmful for imputing common and low-frequency variants when a large, population-specific panel is available. On chromosome 20, the combined panel added 2,263 high-confidence rare variants that were not captured by the UK10K panel (MAF<0.5%; 4% increase), which could reflect mutations that have drifted to very low frequencies in the UK but persist on the same haplotype background elsewhere in Europe (Howie et al. 2011, Jewett et al. 2012). A similar result was observed when the imputation was run for a population in northern Italy (INCIPE cohort). The INCIPE cohort was newly genotyped in this study, using Illumina HumanCoreExome-12v1-1 arrays. After stringent quality control, the genotype data of chromosome 20 was split into an imputation panel (containing 6,300 SNPs genotyped in 2,145 study participants) and a test panel, corresponding to the exome content of the array (2,522 SNPs, all with MAF≤5%). In this dataset the UK10K reference panel outperformed the 1000GP panel in all frequency bins, despite the fact that the 1000GP includes a panel (TSI, or "Toscani in Italia") that is genetically more similar to the study population. As before, the combined 1000GP+UK10K panel yielded a larger number of high-confidence imputed variants than the UK10K panel alone – here, the combined panel added 3,729 well-imputed variants with MAF<0.5%, for a 20% increase in rare variants over the UK10K panel. These results suggest that it can be especially useful to combine the strengths of multiple panels when a large, population-specific reference set is not available for a particular GWAS population.

**Figure 3.4** Performance of combining UK10K and 1000GP panels

Imputation accuracy in the UK10K pseudo-GWAS test panel using reference panels from 1000GP (black), UK10K (blue), and UK10K+1000GP (red) across all MAFs. The rephased UK10K panel was combined with the 1000GP panel to produce the UK10K+1000GP panel.



## 3.4 Conclusion & Discussion

As WGS becomes a standard tool for population and disease genetics, there will be many questions about how to design sequencing studies, how to process the data, how to combine data across studies, and how to limit the computational costs of downstream analysis. With data from one of the most ambitious population sequencing studies to date, the above evaluations have demonstrated the value of a large, UK-specific reference panel for imputation in British cohorts and in other European populations. I showed that the UK10K reference panel greatly increases accuracy and coverage of low-frequency variants relative to a panel of 1,092 individuals from the 1000GP. The results show that state-of-the-art phasing methods like SHAPEIT v2 are essential for creating high-quality haplotype panels. Combining WGS data across studies is a desirable goal, which is now available in IMPUTE2 that can integrate sets of phased haplotypes to produce a unified reference panel. The combined panel is much larger than the 28.6 million imputable sites in the UK10K panel or 32.5 million imputable sites in the 1000GP panel. Finally, due to observations from my evaluation, a new approximation in IMPUTE2 was implemented that helps reduce the trade-off between imputation speed and accuracy as reference panels continue to grow.

As shown in chapter 2, sizable reductions in the magnitude of the effect sizes can be identified at any sample size through the use of the UK10K reference panel and the improved imputation quality. For instance, for a variant of MAF = 0.3% we have equivalent power when imputing from UK10K+1000GP into a 3,621 sample as we have when using the 1000GP imputation panel alone with 10,000 samples (**Figure 2.4a**). Similar, although weaker, increases in power were seen for region-based tests of rare variants. Although absolute power in **Figure 2.4b** is generally poor, there is demonstrable power improvements when data are better imputed or are directly sequenced (**Figure 2.4c**). The benefits of combining two reference panels in improving imputation for rare variants, as demonstrated in this study, could provide a good reference to future efforts that aim to combine a lot more WGS datasets. For example, the Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org/) so far combined WGS from 20 cohorts with more than 30,000 whole genomes. This is expected to significantly improve imputation especially for samples whose ancestries are not as well represented in the 1000GP or in UK10K.

In summary, my recommendation for future WGS based imputation would include the following: 1. pre-phase WGS panel with SHAPEIT; 2. combining two reference panels; 3. if computation cost is not an issue, use all haplotypes, otherwise, using the new IMPUTE2 to pick the top haplotypes; 4. run evaluations and check output data to confirm that the best strategy was adopted and the desirable imputation performance was achieved.