Strategies for Phasing and Imputation in a Population Isolate

Anthony Herzig, Teresa Nutile, Marie-Claude Babron, Marina Ciullo, Céline Bellenguez, Anne-Louise Leutenegger

Strategies for Phasing and Imputation in a Population Isolate

Anthony Francis Herzig (1,2)

Teresa Nutile (3)

Marie-Claude Babron (1,2)

Marina Ciullo (3,4)

Céline Bellenguez (5,6,7,8)

Anne-Louise Leutenegger (1,2,8)

(1) Université Paris-Diderot, Sorbonne Paris Cité, U946, F-75010 Paris, France

(2) Inserm, U946, Genetic variation and Human diseases, F-75010 Paris, France

(3) Institute of Genetics and Biophysics A. Buzzati-Traverso - CNR, Naples, Italy

(4) IRCCS Neuromed, Pozzilli, Isernia, Italy

(5) Inserm, U1167, RID-AGE - Risk factors and molecular determinants of aging-related diseases, F-59000 Lille, France

(6) Institut Pasteur de Lille, F-59000 Lille, France

(7) Université de Lille, U1167 - Excellence Laboratory LabEx DISTALZ, F-59000 Lille, France

(8) These authors contributed equally to this study

Correspondence:

Anthony Francis Herzig

INSERM UMR 946

27 Rue Juliette Dodu

75010, PARIS, FRANCE.

anthony.herzig@inserm.fr

+33172639313

Abstract

In the search for genetic associations with complex traits, population isolates offer the advantage of reduced genetic and environmental heterogeneity. In addition, cost-efficient next-generation association approaches have been proposed in these populations where only a sub-sample of representative individuals is sequenced and then genotypes are imputed into the rest of the population. Gene mapping in such populations thus requires high quality genetic imputation and preliminary phasing. To identify an effective study-design, we compare by simulation a range of phasing and imputation software and strategies.

We simulated 1,115,604 variants on chromosome 10 for 477 members of the large complex pedigree of Campora, a village within the established isolate of Cilento in southern Italy. We assessed the phasing performance of IBD-based software ALPHAPHASE and SLRP, LD-based software SHAPEIT2, SHAPEIT3, and BEAGLE, and new software EAGLE which combines both methodologies. For imputation we compared IMPUTE2, IMPUTE4, MINIMAC3, BEAGLE, and new software PBWT. Genotyping errors and missing genotypes were simulated to observe their effects on the performance of each software.

Highly accurate phased data were achieved by all software with SHAPEIT2, SHAPEIT3, and EAGLE2 providing the most accurate results. MINIMAC3, IMPUTE4, and IMPUTE2 all performed strongly as imputation software and our study highlights the considerable gain in imputation accuracy provided by a genome sequenced reference panel specific to the population isolate.

Key Words: Founder Effect, Genotyping Errors, Identity By Descent, Linkage Disequilibrium, Study Specific Panel.

Introduction

For many complex traits, attention has turned to the search for associations with low-frequency or rare variants. This follows the success of genome-wide association studies (GWAS) in identifying associations with many common variants but without yet gaining a satisfactorily complete description of the genetic heritability for various complex traits. The large sample sizes required to achieve sufficient power to detect associations with rare variants (particularly if effect size is modest), combined with the sequencing cost, limit the opportunities for finding such associations.

Population isolates have inherent characteristics beneficial to the study of complex traits, namely reduced environmental and genetic heterogeneity (Bourgain & Génin, 2005; Hatzikotoulas, Gilly, & Zeggini, 2014). Because of the bottleneck at the founding of the population followed by generations of genetic drift, some mutations which would be described as 'rare' in general populations can occur with greater frequency in the population isolate. Fewer individuals are hence required to achieve sufficient power for analyses. Also, unique patterns of linkage disequilibrium (LD) are expected within such populations and long haplotypes will be identical by descent (IBD) among members of the population even when not closely related.

To take advantage of the prevalence of shared IBD regions, a subset of the study population can be whole-genome sequenced (WGS) and then made available as a Study Specific Panel (SSP) for genetic imputation on to the remainder of the genotyped sample (Asimit & Zeggini, 2012; Holm et al., 2011; Zeggini, 2011). Alternatively, public reference panels could be employed for imputation: for example the 1000 Genomes Project (1000G) (The 1000 Genomes Project Consortium, 2015) or the Haplotype Reference Consortium (HRC) (McCarthy et al., 2016). All study designs require efficient phasing and imputation, and a range of software has been developed to this end.

Methods for phasing can be classified as either LD-based (Browning & Browning, 2016; Delaneau, Zagury, & Marchini, 2013; O'Connell et al., 2016) or IBD-based (Glodzik et al., 2013; Hickey et al., 2011; Livne et al., 2015; Palin, Campbell, Wright, Wilson, & Durbin, 2011). O'Connell et al. (2014) found that despite the prevalence of IBD regions in an isolate, LD-based methods outperformed the IBD-based method proposed by Palin et al. (2011) when tested in several population isolates. Recently a new method was proposed to combine both LD-based and IBD-based approaches and was shown to achieve increased phasing accuracy over LD-based methods in a large outbred population (Loh, Danecek, et al., 2016; Loh, Palamara, & Price, 2016). However, this new approach is yet to be evaluated in a population isolate.

Several studies investigating imputation strategies have shown that using an imputation panel specific to the population under study increases imputation accuracy compared to using larger multi-ethnic public reference panels. This has been observed in population isolates (Joshi et al., 2013; Pistis et al., 2015; Surakka et al., 2010) and in outbred populations (Deelen et al., 2014; Mitt et al., 2017; Roshyara & Scholz, 2015). However, no study has compared imputation software and imputation strategies together in a population isolate since the recent releases of updated software versions (Browning & Browning, 2016; Bycroft et al., 2017; Das et al., 2016), new methods (Durbin, 2014), and larger and denser reference panels (McCarthy et al., 2016; The 1000 Genomes Project Consortium, 2015).

In population isolates, genealogical data may be available. There exist many methods for phasing and imputation using in part or solely pedigree data (Abecasis, Cherny, Cookson, & Cardon, 2002; Chen & Schaid, 2014; Cheung, Thompson, & Wijsman, 2013; Hickey et al., 2011; Livne et al., 2015). The size and complexity of the pedigrees typical to isolates precludes the application of some methods which use only pedigree data. However, methods that combine IBD inference from both genetic and pedigree information should be well adapted for population isolates (Hickey et al., 2011; Livne et al., 2015).

Here we provide an updated evaluation of state-of-the-art phasing and imputation methods in the context of a population isolate. We test the latest versions of existing software as well as recently released software on simulated data with the structure of the population isolate of Campora in southern Italy. The effects of errors and missingness on the performance of each software were also assessed. The design of our study also gives the opportunity to observe in detail the effects of isolate characteristics on phasing and imputation software in order to provide recommendations for future studies of population isolates.

Methods

Campora - Pedigree and genetic data for Campora have previously been gathered as part of the Vallo di Diano Project. The pedigree contains 2,894 members, including 495 founders and spans the 16th century to the present day (Colonna et al., 2007). The pedigree of Campora was reconstructed from parish records (Supplementary Figure 1). Whilst the pedigree captures many loops and connections that result in a high level of relatedness, it falls short of reaching back to the founding event of Campora. Previous analysis of sex chromosomes and mitochondrial DNA in Campora concluded that around 96.7% of the genetic variability was explained by 17 female and 20 male lineages. Hence, whilst the recorded pedigree contains 495 founders, the true founding event in Campora likely involved closer to 37 founders (Colonna et al., 2007).

Of the present day individuals, 477 have high quality genotypes, all of whom have been genotyped on an Illumina 370K SNP-chip array (ARRAY). A subset of 93 individuals has whole exome sequencing (WES) data and another subset of 18 individuals has whole-genome sequencing (WGS) data. The WES subset was selected to serve as an SSP using the method described in Uricchio, Chong, Ross, Ober, and Nicolae (2012) but with genetic kinship in the place of genealogical kinship. This way we selected a subset with a high level of relatedness to the remaining unselected individuals whilst avoiding high levels of relatedness among the selected individuals. This resulted in a selection of 93 individuals spread across the bottom four generations of the Campora pedigree with a higher proportion coming from the bottom two generations. The set of 93 individuals does not contain multiple members of any single nuclear family.

Simulation - Genetic data were simulated with similar characteristics to those observed in the real genetic data from Campora (Supplementary Figure 2). Gene-dropping of chromosome 10 (chr10) was performed on the entire pedigree using the MORGAN package Genedrop (Wijsman, Rothstein, & Thompson, 2006). For time efficiency, Genedrop was only provided with a coarse genetic map, we then sampled precise location of recombination events on the far denser genetic map used in our study as in Gazal et al. (2014).

We considered two approaches to generate the founder haplotypes, both enlisting the haplotypes of the UK10K panel (UK10K) (The UK10K Consortium, 2015) (see URLs). The UK10K contains member of the TwinsUk cohort; for the purposes of the simulation one member from each pair of monozygotic and dizygotic twins was removed leading to a pool of 7,500 haplotypes. In a first simulation strategy we sampled the 990 pedigree founder haplotypes without replacement from the pool of UK10K haplotypes. In a second simulation strategy we first sampled 80 haplotypes from UK10K to approximate the founding event of roughly 37 founders in Campora and then used HapGen2 (Su, Marchini, & Donnelly, 2011) to simulate recombination events and mutations to create a pool of mosaic haplotype from which the 990 founder haplotypes of the pedigree were sampled without replacement. From hence we refer to these two simulation strategies as 'Pedigree' and 'HapGen+Pedigree' respectively. Further details on HapGen2 parameters are given in Supplementary Materials. Each strategy was independently replicated 100 times with independent draws for the 990 and 80 haplotypes respectively. In each replicate we simulated variants at ARRAY positions for all 477 individuals and WGS positions for the 93 SSP individuals. We observed that the HapGen+Pedigree simulation produced simulated data with a mean pairwise genetic kinship (estimated on ARRAY genotypes) closer to the mean observed in

Campora (Supplementary Figure 3) suggesting the HapGen+Pedigree simulation better mimicked the data of Campora.

Error models - Errors and sporadic missingness were simulated in the data. Both were introduced independently in the two simulated platforms (ARRAY and WGS).

Missing genotypes observed in the ARRAY data in Campora were set to missing in the simulated data. Errors on the ARRAY data were simulated with a simple un-directed error model where one allele from a genotype can change to the other available allele (major or minor) at that position with an error rate of 0.001.

For the WGS data, we simulated multiple reads for each genotype (including erroneous reads), from which genotype likelihoods and genotype quality scores were estimated using a similar methodology to previous studies involving next generation sequencing data simulation (Kim et al., 2011; Vieira, Albrechtsen, & Nielsen, 2016). Genotypes which emerged with a quality score less than 20 were set to missing, otherwise the genotype of greatest likelihood was kept. Our error model was tuned to produce missingness rates close to the observed missingness rate in Campora (between 0.01 and 0.02) and error rates similar to those expected on the sequencing platform used in Campora (between 0.003 and 0.004). Full details of our WGS data simulation and the error model are given in Supplementary Materials and specific nucleotide error rates in Supplementary Table 1.

To assess the effect of genotyping errors and missingness on the performance of each phasing and imputation algorithm, we completed the same phasing and imputation steps using simulated data with both genotype errors and missingness (Imperfect data) but also without any such imperfections (Perfect data).

Quality Control – No Quality control was performed on individuals. For imperfect data, all genotypes in the nuclear family were set to missing each time a Mendelian error was introduced by our error models. In all files, variants were removed for low Minor Allele Frequency (MAF), significant deviation from Hardy-Weinberg equilibrium and for high missingness in the case of imperfect data (Supplementary Materials).

Phasing - Phasing algorithms can be separated into two main methodological classes:

LD-based methods which rely on Hidden Markov Models (HMM) are employed by phasing algorithms SHAPEIT2 (Delaneau, Zagury, et al., 2013) and BEAGLE (Browning & Browning, 2016). Phase is estimated with respect to LD patterns and haplotype similarity and is built for each individual as a mosaic of current haplotype estimations of all other sample individuals as well as external reference haplotypes if they are made available to the algorithm. For SHAPEIT2 we considered the use of the 'duohmm' option (O'Connell et al., 2014) which harnesses parent-offspring or duo information for phasing. We also tested SHAPEIT3 (O'Connell et al., 2016), a new version of SHAPEIT2 designed for large sample sizes.

In IBD-based methods, long stretches of IBD can be directly sought between pairs of individuals in order to phase directly each individual in turn in an approach named Long Range Phasing (Kong et al., 2008). We tested two software that employ Long Range Phasing: SLRP (Palin et al., 2011) and ALPHAPHASE (Hickey et al., 2011). ALPHAPHASE was developed for livestock populations and is able to use pedigree information in addition to genotypes. SLRP, which was specifically designed for population isolates, uses only the genotypes.

Two releases of a new method which combines LD-based and IBD-based methods were also tested: EAGLE version 1 (EAGLE1) (Loh, Palamara, et al., 2016) and version 2 (EAGLE2) (Loh, Danecek, et al., 2016). EAGLE1 was aimed at general populations and was developed to phase data with very large sample sizes It employs Long Range Phasing followed by an HMM in a second step. EAGLE2 focuses on harnessing an external reference panel. It no longer uses Long Range Phasing and instead is based on the positional Burrows-Wheeler transform (Durbin, 2014) and an HMM. Yet if EAGLE2 is used without a reference panel it adds the Long Range Phasing algorithm of EAGLE1 as an initial step.

BEAGLE, SHAPEIT2, SHAPEIT3, and EAGLE2 can make inference from an external reference panel when phasing. We tested all software without an external panel and SHAPEIT2 and EAGLE2 with the 1000G panel.

Switch Error Rate (SER) is the standard measure to assess the accuracy of an estimation of genetic phase. A switch error is observable between two consecutive heterozygous sites and occurs if phase at the second heterozygous site is incorrect with respect to that of the first. The SER is the fraction of pairs of heterozygous sites where a switch error has occurred out of the total number of possible pairs. A description of SER calculation in the presence of known genotype errors is given in the Supplementary Materials. We calculated SERs on the entirety of chr10: globally over all individuals and variants, for each individual, and for each variant. We compared the SER per variant to MAF calculated naively on the simulated ARRAY genotypes and the mean SER of each individual to the individual’s mean genetic kinship with all other sample members. Kinship was estimated from the simulated ARRAY genotypes using the R package ‘Gaston’ (see URLs).

Imputation – LD-based imputation methods IMPUTE2 (Howie, Donnelly, & Marchini, 2009), IMPUTE4 (Bycroft et al., 2017), BEAGLE v4.1 (Browning & Browning, 2016), and MINIMAC3 (Das et al., 2016) were compared when using the 1000G as a reference panel. We included all 2,504 individuals from all populations of the 1000G for imputation as this has been shown to be the best approach (Howie, Marchini, & Stephens, 2011). We also used the HRC panel but only for MINIMAC3 due to the computational burden associated with this panel. The HRC panel used was the version made available for download through the European Genome-phenome Archive, which contains 27,165 individuals, including all samples from the 1000G. As our simulations were based on the UK10K, we removed all UK10K haplotypes, leading to 23,450 individuals. We also tested the PBWT software (Durbin, 2014) on 20 of our replicates through use of the Wellcome Trust's Sanger Imputation Service and again using the 1000G as a reference panel. We did not test PBWT with the HRC panel as we could not remove the UK10K haplotypes from the panel when using this imputation service. To restrict to 20 replicates per simulation strategy was a pragmatic decision based on the time required to upload data to the server.

The benefits of imputation using an SSP (either alone or combined with a public reference panel) were investigated. In each simulation replicate, we first created an SSP: WGS and ARRAY data for the 93 SSP individuals were combined (setting discordant genotypes created by our error models to missing in the case of Imperfect data) and then phased. Imputation was performed with IMPUTE2 with a combination of this SSP and the 1000G panel, using the software option which allows the combination of two reference panels through cross imputation. We also tested MINIMAC3 with a combination of the SSP and the HRC panel. As MINIMAC3 does not offer an option for cross imputation, the two panels were first restricted to the set of variants in common between them and then merged. We denote a phasing or imputation strategy by the name of the software added to the panels employed, for example: EAGLE2+1000G, IMPUTE2+1000G, or MINIMAC3+HRC+SSP.

Imputation accuracy of software was assessed in each replicate by the squared Pearson’s correlation between imputed genotype dosages and original simulated genotypes for each biallelic SNP polymorphic in the simulated data and present in the output of every imputation software. Imputation was restricted to the telomeric region of the short arm of chr10 (20Mb in length). As imputation scenarios involving the SSP of 93 individuals were tested, imputation accuracy was measured for all scenarios on the complimenting set of 384 non-SSP individuals. Mean imputation accuracy was calculated over distinct partitions of the observed range of MAF by averaging across all variants in each MAF bin considered. MAF was estimated naively on all 7,500 UK10K haplotypes. All imputation software were run on pre-phased data arising from the best phased data found when comparing phasing software. For general populations, it is possible that pre-phasing could lead to a loss of imputation accuracy (Roshyara, Horn, Kirsten, Ahnert, & Scholz, 2016) but this is unlikely to be significant in population isolates where highly accurate phased data is achievable (Howie, Fuchsberger, Stephens, Marchini, & Abecasis, 2012).

All imputation software provided imputation quality scores per variant; the calculation of such scores varies between imputation software but the scores have been shown to be highly correlated to each other (Marchini & Howie, 2010). We investigated the consequences of post imputation quality control based on imputation quality scores in a separate analysis.

Speed - Since we only concentrate on a single chromosome with a moderate number of individuals, computation time was not an issue for our simulation. However, many of the algorithms considered were designed with speed and low memory usage in mind. Indeed, EAGLE1, EAGLE2, BEAGLE, MINIMAC3, PBWT, IMPUTE4 and SHAPEIT3 are all geared towards performance when analysing very large numbers of individuals or when leveraging very large external reference panels. We measured real and computational time elapsed during a single replicate of the HapGen+Pedigree simulation. All phasing and imputation executions were completed on a 2×6 core, 2×12 thread 2.66GHz Intel Xeon Processor X5650 with 96Gb of random access memory.

The options used for phasing and imputation software are discussed in the Supplementary Materials and the software versions used are detailed in the URLs.

Results

LD-based Phasing - For analyses of phasing performance, we present results from only the HapGen+Pedigree simulation unless otherwise indicated as the patterns of results were very similar between the two simulation strategies. Imperfect ARRAY data initially spanned 13,599 variants on chr10 and following quality control an average of 13,262 variants remained on the HapGen+Pedigree simulation strategy. Totalling over the 477 individuals and across the entirety of chr10, phasing algorithms were required to phase an average of 2,150,627 heterozygous sites in each simulation replicate. All LD-based phasing algorithms considered were able to phase the ARRAY data to a high degree of accuracy with global SERs below 0.002 (Figure 1). EAGLE2 delivered improved SER compared to EAGLE1 (Supplementary Figure 4) and so we only present detailed results for EAGLE2. SHAPEIT2 provided the most accurately phased data and the additions of the 'duohmm' option and the 1000G as an external reference panel further improved its performance. SHAPEIT3 performed similarly to SHAPEIT2 and for subsequent analysis we will only present results for SHAPEIT2+duohmm+1000G. SHAPEIT2+duohmm+1000G achieved a mean SER of 1.9×10^{-4} whilst EAGLE2 achieved 3.2×10^{-4}. The mean global SERs for all phasing strategies considered are given in Supplementary Table 2.

IBD-based Phasing - We note that EAGLE2 outperformed EAGLE2+1000G; conversely to what was observed for SHAPEIT2 (Figure 1). This result can be interpreted as evidence of the utility of the EAGLE2 Long Range Phasing routine for population isolates as this routine is irrevocably omitted from the algorithm when using an external reference panel.

ALPHAPHASE and SLRP both provided added complications because they only phase sites that were found IBD between individuals. SLRP outperformed ALPHAPHASE in terms of SER even though ALPHAPHASE had access to the pedigree information (Supplementary Figure 5). ALPHAPHASE however phased more heterozygous sites than SLRP which may explain some of the difference in SER between the two. We chose to compare only SLRP to other software (Figure 2) as SLRP was clearly stronger than ALPHAPHASE. Owing to the sites left unphased by SLRP, a separate calculation of SER restricted to the set of sites phased by SLRP in each replicate was carried out. SLRP produced higher SERs than SHAPEIT2+duohmm+1000G and EAGLE2 and reducing the analysis to these sites resulted in lower SERs for all other phasing software (when compared to Figure 1). On these sites, SHAPEIT2+duohmm+1000G achieved a mean SER of 1.4×10^{-4} whilst EAGLE2 achieved 2.7×10^{-4} and so a considerable proportion of the switch errors observed in Figure 1 occurred on the small percentage (1.6% on average) of heterozygous sites left unphased by SLRP. This suggests that the sites left unphased by SLRP, which are by definition in areas where SLRP was unable to identify IBD between individuals, are precisely those sites that other software frequently phased incorrectly.

Factors which impact Phasing Performance - To further explore the performance of phasing software, we performed a series of sub-analyses to identify patterns in the distributions of switch errors on chr10.

Variants with low MAF had demonstrably higher SERs, whether using LD-based software or EAGLE2 (Supplementary Figure 6).

The levels of IBD in the simulated populations clearly affected phasing performance as all software had improved phasing accuracy in the presence of the elevated IBD in the HapGen+Pedigree simulation as compared to the Pedigree simulation strategy (Supplementary Figure 7). Similarly, SLRP and ALPHAPHASE both phased many more sites on the HapGen+Pedigree simulation (Supplementary Figures 8a-b). At the individual level, all phasing algorithms had lower performance for the individuals with the lowest mean pairwise genetic kinship to the rest of the sample (Supplementary Figures 9a-c).

Phasing software returned slightly higher SERs when phasing data with errors and missingness (Supplementary Figure 10) and ALPHAPHASE and SLRP phased significantly less sites when errors and missingness were present (Supplementary Figures 8a-b). The effect of imperfections within the data was noticed particularly on the Long Range Phasing algorithms (ALPHAPHASE, SLRP, and EAGLE2).

We specifically investigated the IBD status at switch errors sites in the Pedigree simulation strategy for EAGLE2 and SHAPEIT2+duohmm+1000G (Supplementary Materials and Supplementary Figure 11) as in only this simulation strategy, true IBD sharing was accessible from Genedrop. For both phasing approaches, there were a lower number of true IBD haplotypes at switch errors sites (6 IBD haplotypes on average) compared to correctly phased sites (17 IBD haplotypes on average). These true IBD haplotypes are the haplotypes that the software can use as phase informative. Hence the performance of the LD-based method SHAPEIT2 was implicitly linked to the prevalence of IBD.

Accuracy of Imputation Software - Results pertain to imputation of phased Imperfect ARRAY data from both simulations strategies unless otherwise stated. Following the results from the phasing software evaluation, we phased ARRAY and WGS data with SHAPEIT2+duohmm+1000G. This phasing strategy was also found to be the most accurate for WGS data (Supplementary Figure 12).

In each replicate, mean imputation accuracy was calculated across all polymorphic SNPs found within the output of every software. On average this entailed a selection of 40,989 SNPs for the Pedigree simulation and 40,407 SNPs for the HapGen+Pedigree simulation. This difference is ascribed to the presence of more monomorphic variants in the HapGen+Pedigree simulation.

When using 1000G as the reference panel, MINIMAC3 provided the best imputation accuracy in both simulation strategies followed closely by IMPUTE4 and then IMPUTE2 (Figure 3). Variants with low MAF were universally harder to impute. BEAGLE and PBWT consistently delivered lower imputation accuracy than IMPUTE2, IMPUTE4, and MINIMAC3. Whilst IMPUTE4 marginally outperformed IMPUTE2, it currently does not offer the option to combine reference panels necessary for subsequent analyses in which we hence compare IMPUTE2 and MINIMAC3.

Genotype errors and missingness on the ARRAY data had minimal impact on imputation accuracy but such imperfections simulated on the WGS SSP had slightly more effect (Supplementary Figures 13 & 14).

Impact of Reference Panel Choice - By comparing the two simulation strategies, we were able to identify the consequences of reference panel choice in a population isolate. When the 1000G was chosen as the external reference panel, imputation accuracy was significantly lower in the HapGen+Pedigree simulation strategy than in the Pedigree one (Figure 3). This difference in imputation accuracy may be due to differences in MAF between the simulated data and the 1000G reference panel (Supplementary Materials and Supplementary Figure 15). MAFs on the HapGen+Pedigree simulation had drifted further away from the 1000G reference panel and the variants with the highest differences in MAF to the 1000G reference panel were imputed with lower accuracy than random selections of similar variants (Supplementary Figure 16a).

Imputation with the SSP was an improvement upon imputation with the 1000G for both IMPUTE2 and MINIMAC3 (Figures 4 and 5). When using the SSP, the simulation strategy with the highest imputation accuracy was the HapGen+Pedigree simulation, contrary to when using only the 1000G (Figure 3). This can be ascribed the higher levels of IBD between the 93 SSP members and the 384 other individuals in this simulation strategy. Indeed, the most accurately imputed individuals were consistently those with higher values of mean pairwise kinship to the set of SSP individuals (Supplementary Figure 17).

For MINIMAC3, imputation accuracy was clearly improved by using the HRC over the 1000G (Figure 5). Imputation which included the SSP again produced more accurate results than imputation with only public reference panels on the HapGen+Pedigree simulation strategy. Rare variants were however imputed more accurately by MINIMAC3+HRC than by MINIMAC3+SSP on the Pedigree simulation. The results of Figures 3, 4, and 5 are summarised in Supplementary Table 3.

The founding event in an isolate will result in higher MAFs for certain variants as compared to general populations. Variants with a high difference in MAF compared to the 1000G were imputed as well as the random selections of comparable variants under IMPUTE2+SSP, but with lower accuracy under IMPUTE2+1000G (Supplementary Figure 16a). When changing reference panel from the 1000G to the SSP, we observed that imputation accuracy increased the most for variants with a MAF higher in the sample than the 1000G (Supplementary Materials and Supplementary Figure 16b). Another consequence of using solely the 1000G as a reference panel was the fact that some variants which were monomorphic in the sample were imputed with dosages compatible with being heterozygous for many individuals, i.e. polymorphic in the sample (Supplementary Figures 16c-d).

Imputation Quality Scores - Finally, we analysed the effect of applying various thresholds of the 'info' score for IMPUTE2 and the 'RSQ' score for MINIMAC3. Each successive threshold improved imputation accuracy for both IMPUTE2 and MINIMAC3 with the latter still providing higher accuracy in each MAF bin (Supplementary Materials and Supplementary Figure 18a-b). The ‘RSQ’ measure gave a better indication of imputation accuracy than ‘info’ and we also found that higher thresholds than the standard ones were arguably preferable for both rare and common variants in both simulation strategies (Supplementary Materials and Supplementary Table 4).

Speed - For phasing, BEAGLE, EAGLE1 and EAGLE2 were the fastest because they allow for multiple threading. SHAPEIT2 required more computation time than other algorithms. For imputation, the quickest software were BEAGLE and IMPUTE4. MINIMAC3+1000G was quicker than IMPUTE2+1000G. We observed the additional complexity encountered by IMPUTE2 when performing cross imputation. The full list of times is given in Supplementary Table 5.

Discussion

Using simulated genetic data, we have rigorously tested the performance of a range of phasing and imputation software in a population isolate. EAGLE2 (without a reference panel) and SHAPEIT2 were the strongest performing phasing software with SHAPEIT2+duohmm+1000G giving the most accurately phased data. MINIMAC3, IMPUTE4, and IMPUTE2 all performed well and we observed a slight advantage for MINIMAC3. MINIMAC3 imputation was more accurate with the HRC as an external reference panel rather than the 1000G. The use of an SSP proved to be a very successful strategy, when used alone, but even more so when combined with a large external reference panel. MINIMAC3+HRC+SSP proved the most effective imputation strategy. Genotype errors and missingness were shown to have only a small effect on the performance of all phasing and imputation software considered.

If we compare our phasing results to published results for outbred populations, it is clear that all methods performed with greater accuracy (SERs at least one order of magnitude smaller) on our simulated data. Indeed, for outbred populations, very large sample sizes have been required to achieve the high level of phasing accuracy observed in our population isolate study. For examples, see Bycroft et al. (2017), Loh, Danecek, et al. (2016), O'Connell et al. (2016), and Mitt et al. (2017).

IBD-based phasing methods did not prove as effective as the LD-based software SHAPEIT2 which appeared itself to directly profit from IBD in the sample. O'Connell et al. (2014) also observed SHAPEIT2 benefiting from IBD. Indeed, the performance of IBD-based and LD-based software followed a similar pattern: all were less accurate when less IBD was present and all had difficulty when phasing the likely non-IBD regions of the genome and when phasing individuals with a low average kinship to the rest of the sample. IBD-based methods were the most affected by imperfections in the data.

EAGLE was expected to perform strongly on population isolate data as it should combine the appeal of Long Range Phasing and the strengths of LD-based methods such as SHAPEIT2. Though the combination of IBD-based and LD-based approaches in EAGLE1 and EAGLE2 is a clear improvement over previous Long Range Phasing software, it does not provide more accurate phasing that the LD-based approach implemented in SHAPEIT2. This is in accord with the results of Mitt et al. (2017) in a cohort of intermediate size but not with those of Loh, Danecek, et al. (2016) in much larger cohorts. EAGLE2 was developed with the aim of handling large sample sizes but as gene-mapping studies in population isolates will remain by nature small-scale, SHAPEIT2 remains the optimum choice for phasing.

Published results for SHAPEIT3 in outbred populations suggest that it may return less accurate phased data compared to SHAPEIT2 (O'Connell et al., 2016). Of the two, SHAPEIT2 is recommended for sample sizes less than 20,000 which would encompass the realm of population isolates. In our study, SHAPEIT2 and SHAPEIT3 performed very similarly.

Our comparisons on imputation strategies agree with recent literature (Deelen et al., 2014; Mitt et al., 2017; Pistis et al., 2015) in terms of the improvement in accuracy brought by a reference panel specific to the population under study. Mitt et al. (2017) concluded that for certain outbred populations, such a panel can outperform an order of magnitude larger and more diverse reference panel (the HRC). We show that for a population isolate, an SSP can be far smaller and still outperform the HRC. As discussed in Asimit and Zeggini (2012), the appropriate size of the SSP will depend on the diversity of the isolate.

The HapGen+Pedigree simulation strategy gave the best representation of a true isolate with a strong founder effect producing large disparities to general populations represented in public databases. Of the two simulation strategies, imputation accuracy was significantly lower on this simulation when using only a public reference panel. This suggests that for a population isolate with a very small set of founders and high relatedness between individuals, using public reference panels alone is not a completely appropriate strategy for imputation. A better solution is to sequence a subset of the isolate to serve as an SSP. Even with a very large external reference panel, such as the HRC (here 23,450 individuals), imputation accuracy could not match the level reached by an SSP of 93 individuals. Using an SSP was particularly effective when imputing variants with MAFs higher in the sample than in an external reference panel. As such variants are precisely those which motivate the study of population isolates, this strengthens the argument for using an SSP in a population isolate.

We observed that the best results came from combining an external reference panel and our SSP together for imputation. IMPUTE2 facilitates cross-imputation of two reference panels with variants at non-identical sets of positions. This is an attractive strategy for isolates as all positions from both panels can be imputed including variants specific to the isolate.

The accuracy of imputation can be directly linked to the statistical power of subsequent association tests (Browning & Browning, 2009; Huang, Wang, & Rosenberg, 2009; Li, Willer, Ding, Scheet, & Abecasis, 2010; Surakka et al., 2010). Indeed, if N is the number of individuals in a study and a variant is imputed with an imputation accuracy of $r^2 = \alpha$, then the statistical power of an association test using the imputed dosages is equivalent to that of a test performed on observed genotypes for αN samples. This is the intended interpretation of imputation quality scores which are estimates of the true r^2 statistics (Marchini & Howie, 2010). To give an example, we have observed differences in imputation accuracy of around 0.2 for rare variants (MAF ≤ 0.05) and 0.1 for common variants (MAF > 0.05) between MNIMAC3+1000G and MINIMAC3+HRC+SSP on the HapGen+Pedigree simulation (Supplementary Table 3). Imputation accuracy was measured on a sample of size N = 384 (non-SSP individuals), hence the observed differences in imputation accuracy would correspond to losses of power equivalent to removing around 77 or 38 of these individuals from subsequent analyses respectively. Studies in isolates typically involve unavoidably modest sample sizes. Hence, there is great importance in attaining the highest imputation accuracy possible in such studies in order to preserve power.

One possible option for SHAPEIT2 that we did not consider is the PIR option which harnesses phase informative reads (Delaneau, Howie, Cox, Zagury, & Marchini, 2013). To include this in our simulation would have required the creation of the original read data which was judged to be too great a computational burden for our study. This option was tested in Mitt et al. (2017) and did not significantly improve the global performance of SHAPEIT2. Another version of SHAPEIT2, SHAPEITR (Sharp, Kretzschmar, Delaneau, & Marchini, 2016), sets out to improve phasing by concentrating on rare variants. However, as it is so far only available through the Oxford Statistics Phasing Server (see URLs), it is not suitable for an in-house simulation.

One software in particular which we have not tested is PRIMAL which uses Long Range Phasing and is designed for phasing and imputation in population isolates (Livne et al., 2015). PRIMAL specifically requires pedigree information for phasing and an SSP for imputation. We were unable to successfully setup and run PRIMAL on our simulated datasets and we have been advised by the authors to wait for a new version which is soon to be released.

In this study, we have strived to create realistic isolate data to thoroughly test a range of phasing and imputation software and strategies. Our study design allowed us to observe how phasing and imputation algorithms are impacted by certain characteristics of isolate data, namely IBD between sample members and characteristics arising from isolation such as divergent MAFs compared to reference populations. We found that the best strategy for phasing in a population isolate was to use SHAPEIT2 with the 'duohmm' option and with an external reference panel. For imputation, if no SSP is sequenced in the isolate, it is desirable to use the largest public reference panel available which would lead to the use of MINIMAC3 or IMPUTE4 as these software can handle very large reference panels. If an SSP is available in the isolate it should be used and the option in IMPUTE2 that combines reference panels through cross imputation makes it an attractive choice of imputation software. In this case the largest available public reference panel compatible with IMPUTE2 should be used with the SSP. At the time of publication, IMPUTE4 and MINIMAC3 do not offer the option of combining two reference panels, but, if such options do become available, then a strategy which both combines the HRC and an SSP by cross imputation would likely be both fast and highly accurate in a population isolate.

Acknowledgements: We address special thanks to the people of Campora for their participation in the study. We kindly thank the European Genome-phenome Archive at the European Bioinformatics Institute for making available the UK10K imputation panel (EGAD00001000776) and HRC imputation panel (EGAD00001002729) for the use in our simulation study. We also thank the two anonymous reviewers for their comments which greatly improved the manuscript.

Funding: ESGI - The research leading to these results has received funding from the Seventh Framework Programme [FP7/2007-2013] under grant agreement n° 262055.

A.H. was funded by an international Ph.D. fellowship from Sorbonne Paris Cité (convention HERZI15RDXMTSPC1LIETUE).

Conflict of Interest: None Declared

URLs:

1. ALPHAPHASE (v1.2), http://www.alphagenes.roslin.ed.ac.uk/alphasuite-softwares/alphaphase/.

2. BEAGLE (v4.1), http://faculty.washington.edu/browning/beagle/beagle.html.

3. EAGLE2 (v2.3.2) & EAGLE1 (v1.0), http://www.hsph.harvard.edu/alkes-price/software/.

4. SHAPEIT2 (v2.837), http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html.

5. SHAPEIT3 (v1.0), https://jmarchini.org/shapeit3/.

6. SLRP (v1.0), https://github.com/kpalin/SLRP.

7. IMPUTE2 (v2.3.2), https://mathgen.stats.ox.ac.uk/impute/impute_v2.html.

8. IMPUTE4 (v1.0), https://jmarchini.org/impute-4/.

9. MINIMAC3 (v.2.0.1), http://genome.sph.umich.edu/wiki/Minimac3.

10. 1000 Genomes data set (Phase 3) , http://www.1000genomes.org/.

11. Haplotype Reference Consortium, http://www.haplotype-reference-consortium.org/.

12. UK10K Project, https://www.uk10k.org/.

13. Sanger Imputation Service, https://imputation.sanger.ac.uk/.

14. Michigan Imputation Server, https://imputationserver.sph.umich.edu/.

15. Oxford Statistics Phasing Server, https://phasingserver.stats.ox.ac.uk/.

16. R-package 'Gaston', https://cran.r-project.org/web/packages/gaston/index.html.

17. European Genome-phenome Archive, https://www.ebi.ac.uk/ega/home.

References

Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet, 30(1), 97-101. doi: 10.1038/ng786

Asimit, J. L., & Zeggini, E. (2012). Imputation of rare variants in next generation association studies. Human heredity, 74(0), 196-204. doi: 10.1159/000345602

Bourgain, C., & Génin, E. (2005). Complex trait mapping in isolated populations: Are specific statistical methods required? Eur J Hum Genet, 13(6), 698-706.

Browning, Brian L., & Browning, Sharon R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet, 84(2), 210-223. doi: 10.1016/j.ajhg.2009.01.005

Browning, Brian L., & Browning, Sharon R. (2016). Genotype Imputation with Millions of Reference Samples. Am J Hum Genet, 98(1), 116-126. doi: 10.1016/j.ajhg.2015.11.020

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . Marchini, J. (2017). Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv. doi: 10.1101/166298

Chen, W., & Schaid, D. J. (2014). PedBLIMP: extending linear predictors to impute genotypes in pedigrees. Genet Epidemiol, 38(6), 531-541. doi: 10.1002/gepi.21838

Cheung, C. Y., Thompson, E. A., & Wijsman, E. M. (2013). GIGI: an approach to effective imputation of dense genotypes on large pedigrees. Am J Hum Genet, 92(4), 504-516. doi: 10.1016/j.ajhg.2013.02.011

Colonna, V., Nutile, T., Astore, M., Guardiola, O., Antoniol, G., Ciullo, M., & Persico, M. G. (2007). Campora: A Young Genetic Isolate in South Italy. Human heredity, 64(2), 123-135. doi: 10.1159/000101964

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., . . . Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. Nat Genet, 48(10), 1284-1287. doi: 10.1038/ng.3656

Deelen, P., Menelaou, A., van Leeuwen, E. M., Kanterakis, A., van Dijk, F., Medina-Gomez, C., . . . Swertz, M. A. (2014). Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. Eur J Hum Genet, 22(11), 1321-1326. doi: 10.1038/ejhg.2014.19

Delaneau, O., Howie, B., Cox, Anthony J., Zagury, J.-F., & Marchini, J. (2013). Haplotype Estimation Using Sequencing Reads. Am J Hum Genet, 93(4), 687-696. doi: 10.1016/j.ajhg.2013.09.002

Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat Meth, 10(1), 5-6. doi: 10.1038/nmeth.2307

Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). Bioinformatics, 30(9), 1266-1272. doi: 10.1093/bioinformatics/btu014

Gazal, S., Sahbatou, M., Perdry, H., Letort, S., Génin, E., & Leutenegger, A. L. (2014). Inbreeding Coefficient Estimation with Dense SNP Data: Comparison of Strategies and Application to HapMap III. Human heredity, 77(1-4), 49-62.

Glodzik, D., Navarro, P., Vitart, V., Hayward, C., McQuillan, R., Wild, S. H., . . . McKeigue, P. (2013). Inference of identity by descent in population isolates and optimal sequencing studies. Eur J Hum Genet, 21(10), 1140-1145. doi: 10.1038/ejhg.2012.307

Hatzikotoulas, K., Gilly, A., & Zeggini, E. (2014). Using population isolates in genetic association studies. Briefings in Functional Genomics, 13(5), 371-377. doi: 10.1093/bfgp/elu022

Hickey, J. M., Kinghorn, B. P., Tier, B., Wilson, J. F., Dunstan, N., & van der Werf, J. H. J. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genetics, Selection, Evolution : GSE, 43(1), 12-12. doi: 10.1186/1297-9686-43-12

Holm, H., Gudbjartsson, D. F., Sulem, P., Masson, G., Helgadottir, H. T., Zanon, C., . . . Stefansson, K. (2011). A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nat Genet, 43(4), 316-320. doi: 10.1038/ng.781

Howie, B., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet, 5(6), e1000529. doi: 10.1371/journal.pgen.1000529

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet, 44(8), 955-959. doi: 10.1038/ng.2354

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. G3 (Bethesda), 1(6), 457-470. doi: 10.1534/g3.111.001198

Huang, L., Wang, C., & Rosenberg, N. A. (2009). The relationship between imputation error and statistical power in genetic association studies in diverse populations. Am J Hum Genet, 85(5), 692-698. doi: 10.1016/j.ajhg.2009.09.017

Joshi, P. K., Prendergast, J., Fraser, R. M., Huffman, J. E., Vitart, V., Hayward, C., . . . Navarro, P. (2013). Local Exome Sequences Facilitate Imputation of Less Common Variants and Increase Power of Genome Wide Association Studies. PLOS ONE, 8(7), e68604. doi: 10.1371/journal.pone.0068604

Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., . . . Nielsen, R. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinformatics, 12, 231-231. doi: 10.1186/1471-2105-12-231

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., . . . Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet, 40(9), 1068-1075. doi: 10.1038/ng.216

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. Genet Epidemiol, 34(8), 816-834. doi: 10.1002/gepi.20533

Livne, O. E., Han, L., Alkorta-Aranburu, G., Wentworth-Sheilds, W., Abney, M., Ober, C., & Nicolae, D. L. (2015). PRIMAL: Fast and Accurate Pedigree-based Imputation from Sequence Data in a Founder Population. PLoS Computational Biology, 11(3), e1004139. doi: 10.1371/journal.pcbi.1004139

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., . . . L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet, 48(11), 1443-1448. doi: 10.1038/ng.3679

Loh, P.-R., Palamara, P. F., & Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat Genet, 48(7), 811-816. doi: 10.1038/ng.3571

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat Rev Genet, 11(7), 499-511. doi: 10.1038/nrg2796

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., . . . The Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet, 48(10), 1279-1283. doi: 10.1038/ng.3643

Mitt, M., Kals, M., Parn, K., Gabriel, S. B., Lander, E. S., Palotie, A., . . . Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet. doi: 10.1038/ejhg.2017.51

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., . . . Marchini, J. (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genetics, 10(4), e1004234. doi: 10.1371/journal.pgen.1004234

O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., . . . Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. Nat Genet, 48(7), 817-820. doi: 10.1038/ng.3583

Palin, K., Campbell, H., Wright, A. F., Wilson, J. F., & Durbin, R. (2011). Identity-by-Descent-Based Phasing and Imputation in Founder Populations Using Graphical Models. Genet Epidemiol, 35(8), 853-860. doi: 10.1002/gepi.20635

Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., . . . Sanna, S. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. Eur J Hum Genet, 23(7), 975-983. doi: 10.1038/ejhg.2014.216

Roshyara, N. R., Horn, K., Kirsten, H., Ahnert, P., & Scholz, M. (2016). Comparing performance of modern genotype imputation methods in different ethnicities. Scientific Reports, 6, 34386. doi: 10.1038/srep34386

Roshyara, N. R., & Scholz, M. (2015). Impact of genetic similarity on imputation accuracy. BMC Genetics, 16, 90. doi: 10.1186/s12863-015-0248-2

Sharp, K., Kretzschmar, W., Delaneau, O., & Marchini, J. (2016). Phasing for medical sequencing using rare variants and large haplotype reference panels. Bioinformatics, 32(13), 1974-1980. doi: 10.1093/bioinformatics/btw065

Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics, 27(16), 2304-2305. doi: 10.1093/bioinformatics/btr341

Surakka, I., Kristiansson, K., Anttila, V., Inouye, M., Barnes, C., Moutsianas, L., . . . Ripatti, S. (2010). Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. Genome Res, 20(10), 1344-1351. doi: 10.1101/gr.106534.110

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526(7571), 68-74. doi: 10.1038/nature15393

The UK10K Consortium. (2015). The UK10K project identifies rare variants in health and disease. Nature, 526(7571), 82-90. doi: 10.1038/nature14962

Uricchio, L. H., Chong, J. X., Ross, K. D., Ober, C., & Nicolae, D. L. (2012). Accurate imputation of rare and common variants in a founder population from a small number of sequenced individuals. Genet Epidemiol, 36(4), 312-319. doi: 10.1002/gepi.21623

Vieira, F. G., Albrechtsen, A., & Nielsen, R. (2016). Estimating IBD tracts from low coverage NGS data. Bioinformatics, 32(14), 2096-2102. doi: 10.1093/bioinformatics/btw212

Wijsman, E. M., Rothstein, J. H., & Thompson, E. A. (2006). Multipoint Linkage Analysis with Many Multiallelic or Dense Diallelic Markers: Markov Chain–Monte Carlo Provides Practical Approaches for Genome Scans on General Pedigrees. Am J Hum Genet, 79(5), 846-858.

Zeggini, E. (2011). Next-generation association studies for complex traits. Nat Genet, 43(4), 287-288. doi: 10.1038/ng0411-287

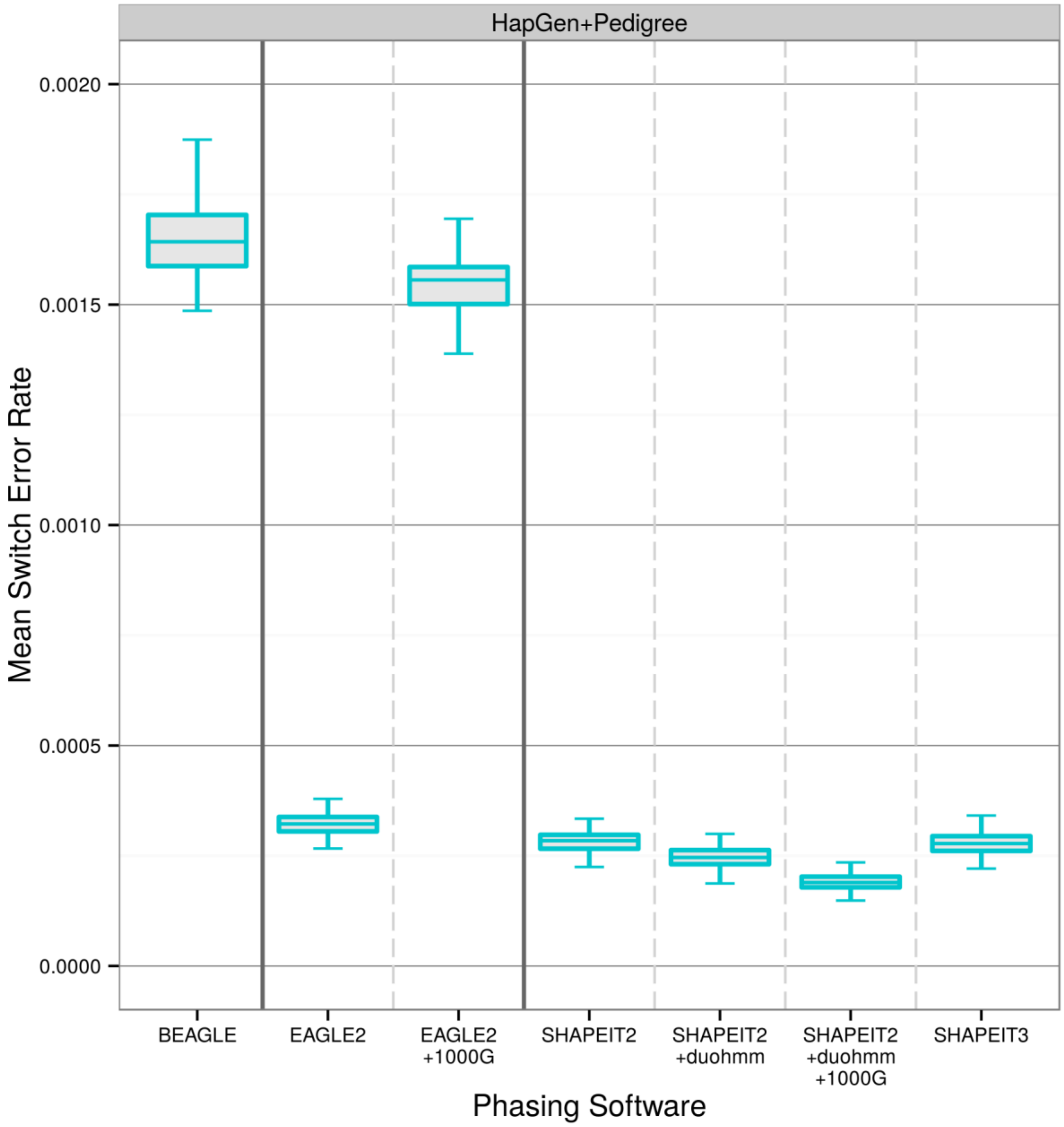

Figure 1. Global switch error rates for BEAGLE, EAGLE2, SHAPEIT2, and SHAPEIT3 for the HapGen+Pedigree simulation strategy.

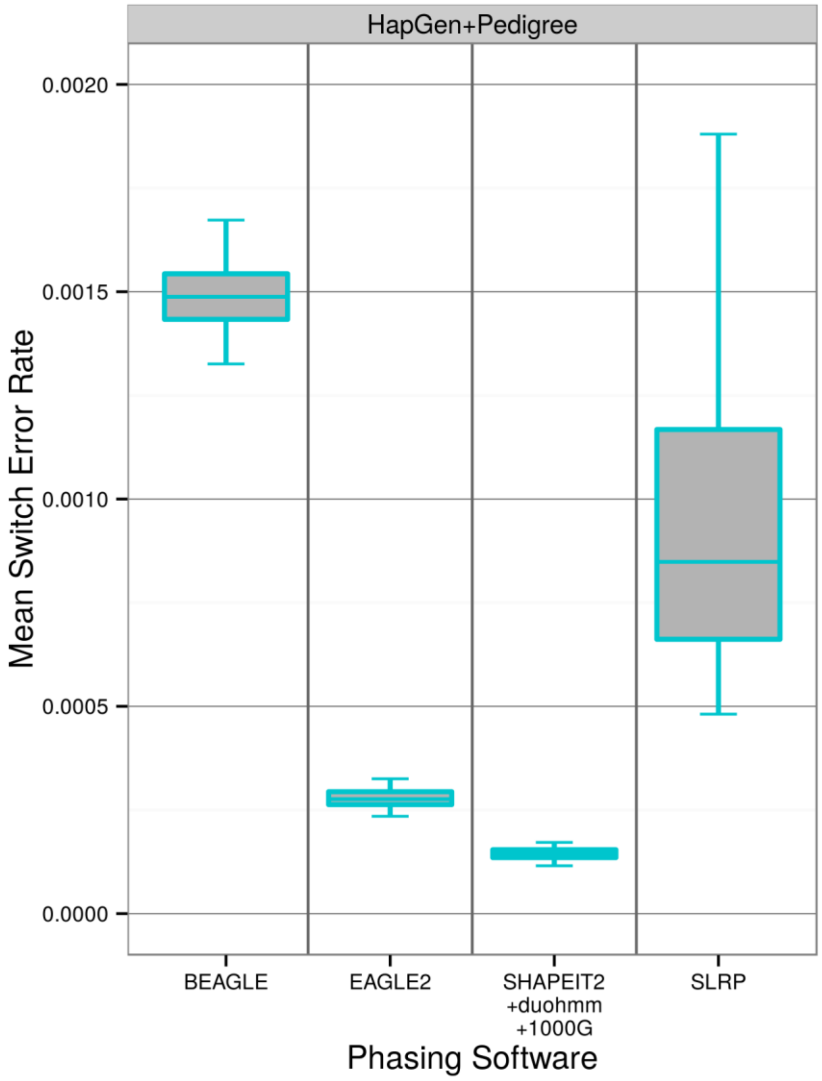

Figure 2. Global switch error rates for BEAGLE, SLRP, EAGLE2, and SHAPEIT2+duohmm+1000G for the HapGen+Pedigree simulation strategy on the set of variants successfully phased by SLRP in each replicate.

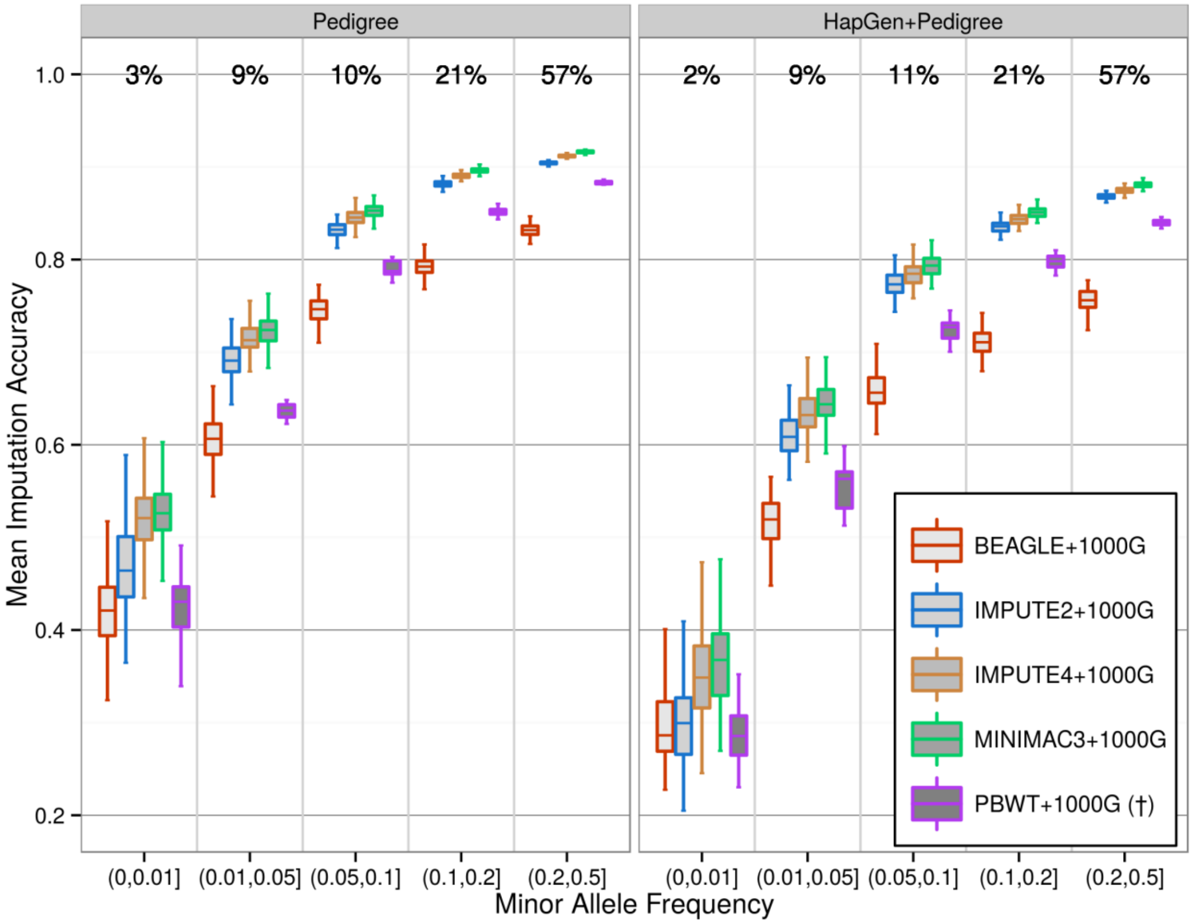

Figure 3. Software imputation accuracy with the 1000G as an external reference panel and for the Pedigree and HapGen+Pedigree simulation strategies. The percentages of variants in each MAF bin are displayed atop the figure. Total number of variants for each strategy: 40,989 (Pedigree) and 40,407 (HapGen+Pedigree). † PBWT was only run on 20 replicates of each simulation strategy.

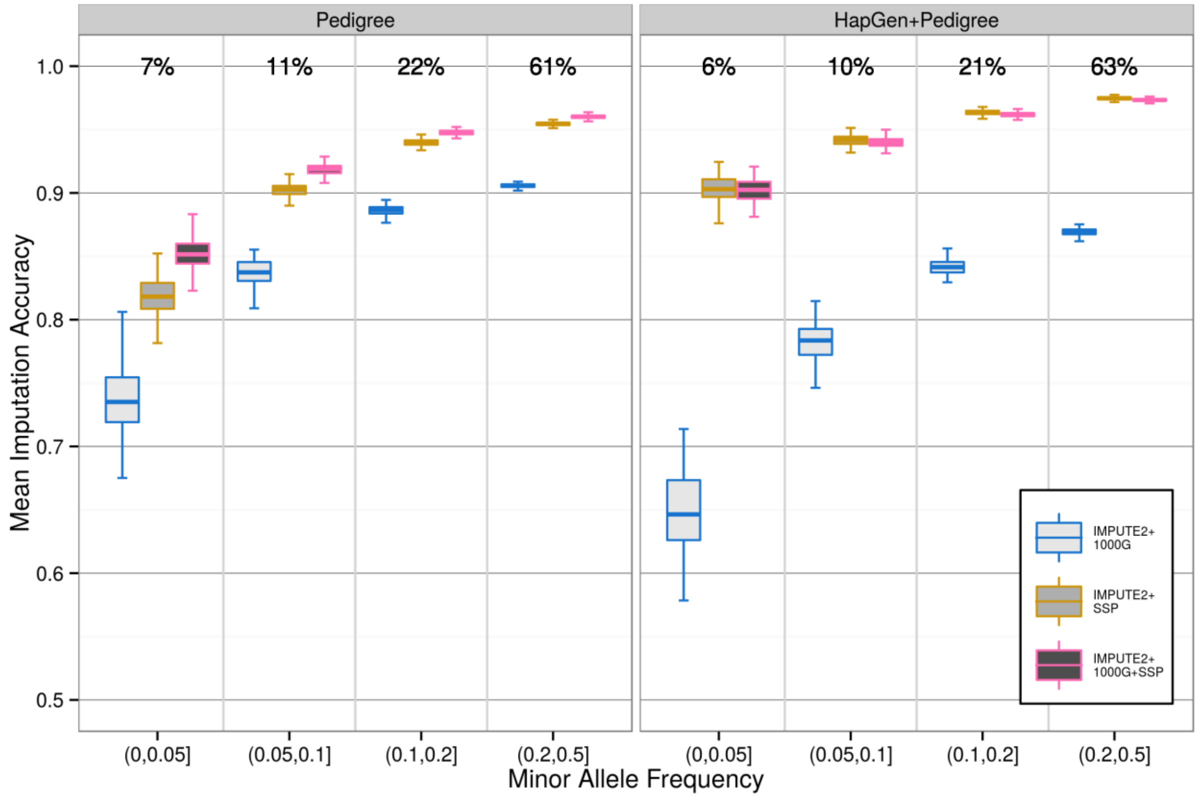

Figure 4. Imputation accuracy of IMPUTE2 when using various reference panels for the Pedigree and HapGen+Pedigree simulation strategies. The set of variants used for comparison is a reduction of the set used in Figure 3 because using only the SSP as a reference panel limits the set of possible variants to compare imputed dosages and true genotypes. This depleted the number of variants in the [0,0.01) MAF category, which was therefore merged with that of [0.01,0.05) MAF. Total number of variants for each strategy: 35,058 (Pedigree) and 34,065 (HapGen+Pedigree).

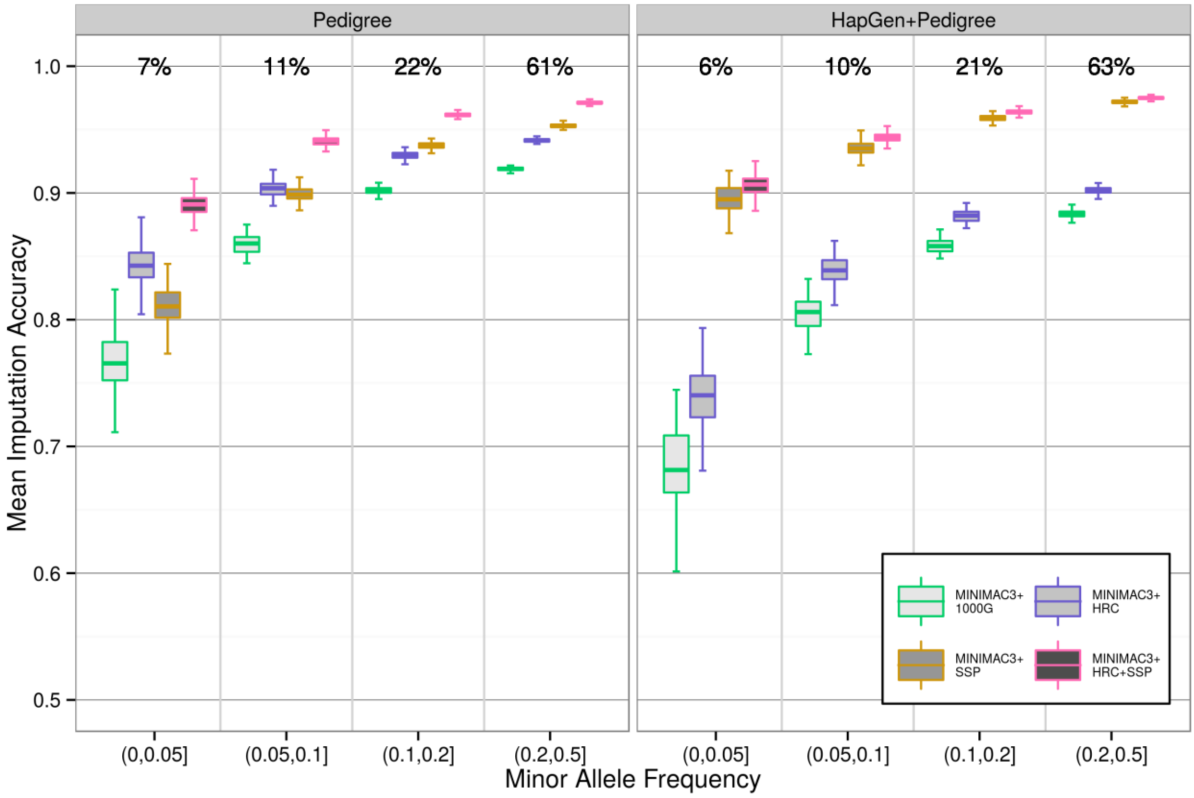

Figure 5. Imputation accuracy of MINIMAC3 with various reference panels on the same set of variants as used in Figure 4.