



HAL
open science

Strategies for Phasing and Imputation in a Population Isolate

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

► **To cite this version:**

Anthony Francis Herzig, Teresa Natile, Marie-Claude Babron, Marina Ciullo, Céline Bellenguez, et al.. Strategies for Phasing and Imputation in a Population Isolate. Genetic Epidemiology, 2017, Epub ahead of print. 10.1002/gepi.22109 . inserm-01645064

HAL Id: inserm-01645064

<https://inserm.hal.science/inserm-01645064>

Submitted on 22 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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.

1 Introduction

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

8 Population isolates have inherent characteristics beneficial to the study of complex traits, namely
9 reduced environmental and genetic heterogeneity (Bourgain & Génin, 2005; Hatzikotoulas, Gilly, & Zeggini,
10 2014). Because of the bottleneck at the founding of the population followed by generations of genetic drift,
11 some mutations which would be described as 'rare' in general populations can occur with greater frequency in
12 the population isolate. Fewer individuals are hence required to achieve sufficient power for analyses. Also,
13 unique patterns of linkage disequilibrium (LD) are expected within such populations and long haplotypes will be
14 identical by descent (IBD) among members of the population even when not closely related.

15 To take advantage of the prevalence of shared IBD regions, a subset of the study population can be
16 whole-genome sequenced (WGS) and then made available as a Study Specific Panel (SSP) for genetic
17 imputation on to the remainder of the genotyped sample (Asimit & Zeggini, 2012; Holm et al., 2011; Zeggini,
18 2011). Alternatively, public reference panels could be employed for imputation: for example the 1000 Genomes
19 Project (1000G) (The 1000 Genomes Project Consortium, 2015) or the Haplotype Reference Consortium (HRC)
20 (McCarthy et al., 2016). All study designs require efficient phasing and imputation, and a range of software has
21 been developed to this end.

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

30 Several studies investigating imputation strategies have shown that using an imputation panel specific
31 to the population under study increases imputation accuracy compared to using larger multi-ethnic public
32 reference panels. This has been observed in population isolates (Joshi et al., 2013; Pistis et al., 2015; Surakka et
33 al., 2010) and in outbred populations (Deelen et al., 2014; Mitt et al., 2017; Roshyara & Scholz, 2015).
34 However, no study has compared imputation software and imputation strategies together in a population isolate
35 since the recent releases of updated software versions (Browning & Browning, 2016; Bycroft et al., 2017; Das et
36 al., 2016), new methods (Durbin, 2014), and larger and denser reference panels (McCarthy et al., 2016; The
37 1000 Genomes Project Consortium, 2015).

38 In population isolates, genealogical data may be available. There exist many methods for phasing and
39 imputation using in part or solely pedigree data (Abecasis, Cherny, Cookson, & Cardon, 2002; Chen & Schaid,
40 2014; Cheung, Thompson, & Wijnsman, 2013; Hickey et al., 2011; Livne et al., 2015). The size and complexity
41 of the pedigrees typical to isolates precludes the application of some methods which use only pedigree data.
42 However, methods that combine IBD inference from both genetic and pedigree information should be well
43 adapted for population isolates (Hickey et al., 2011; Livne et al., 2015).

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

50 **Methods**

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

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

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

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

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

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

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

94 For the WGS data, we simulated multiple reads for each genotype (including erroneous reads), from
95 which genotype likelihoods and genotype quality scores were estimated using a similar methodology to previous
96 studies involving next generation sequencing data simulation (Kim et al., 2011; Vieira, Albrechtsen, & Nielsen,
97 2016). Genotypes which emerged with a quality score less than 20 were set to missing, otherwise the genotype
98 of greatest likelihood was kept. Our error model was tuned to produce missingness rates close to the observed
99 missingness rate in Campora (between 0.01 and 0.02) and error rates similar to those expected on the
100 sequencing platform used in Campora (between 0.003 and 0.004). Full details of our WGS data simulation and
101 the error model are given in Supplementary Materials and specific nucleotide error rates in Supplementary Table
102 1.

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

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

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

111 LD-based methods which rely on Hidden Markov Models (HMM) are employed by phasing algorithms
112 SHAPEIT2 (Delaneau, Zagury, et al., 2013) and BEAGLE (Browning & Browning, 2016). Phase is estimated
113 with respect to LD patterns and haplotype similarity and is built for each individual as a mosaic of current
114 haplotype estimations of all other sample individuals as well as external reference haplotypes if they are made

115 available to the algorithm. For SHAPEIT2 we considered the use of the 'duohmm' option (O'Connell et al.,
116 2014) which harnesses parent-offspring or duo information for phasing. We also tested SHAPEIT3 (O'Connell
117 et al., 2016), a new version of SHAPEIT2 designed for large sample sizes.

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

124 Two releases of a new method which combines LD-based and IBD-based methods were also tested:
125 EAGLE version 1 (EAGLE1) (Loh, Palamara, et al., 2016) and version 2 (EAGLE2) (Loh, Danecek, et al.,
126 2016). EAGLE1 was aimed at general populations and was developed to phase data with very large sample sizes
127 It employs Long Range Phasing followed by an HMM in a second step. EAGLE2 focuses on harnessing an
128 external reference panel. It no longer uses Long Range Phasing and instead is based on the positional Burrows-
129 Wheeler transform (Durbin, 2014) and an HMM. Yet if EAGLE2 is used without a reference panel it adds the
130 Long Range Phasing algorithm of EAGLE1 as an initial step.

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

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

143 **Imputation** – LD-based imputation methods IMPUTE2 (Howie, Donnelly, & Marchini, 2009), IMPUTE4
144 (Bycroft et al., 2017), BEAGLE v4.1 (Browning & Browning, 2016), and MINIMAC3 (Das et al., 2016) were
145 compared when using the 1000G as a reference panel. We included all 2,504 individuals from all populations of
146 the 1000G for imputation as this has been shown to be the best approach (Howie, Marchini, & Stephens, 2011).
147 We also used the HRC panel but only for MINIMAC3 due to the computational burden associated with this
148 panel. The HRC panel used was the version made available for download through the European Genome-
149 phenome Archive, which contains 27,165 individuals, including all samples from the 1000G. As our simulations
150 were based on the UK10K, we removed all UK10K haplotypes, leading to 23,450 individuals. We also tested
151 the PBWT software (Durbin, 2014) on 20 of our replicates through use of the Wellcome Trust's Sanger
152 Imputation Service and again using the 1000G as a reference panel. We did not test PBWT with the HRC panel
153 as we could not remove the UK10K haplotypes from the panel when using this imputation service. To restrict to
154 20 replicates per simulation strategy was a pragmatic decision based on the time required to upload data to the
155 server.

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

166 Imputation accuracy of software was assessed in each replicate by the squared Pearson's correlation
167 between imputed genotype dosages and original simulated genotypes for each biallelic SNP polymorphic in the
168 simulated data and present in the output of every imputation software. Imputation was restricted to the telomeric
169 region of the short arm of chr10 (20Mb in length). As imputation scenarios involving the SSP of 93 individuals
170 were tested, imputation accuracy was measured for all scenarios on the complementing set of 384 non-SSP
171 individuals. Mean imputation accuracy was calculated over distinct partitions of the observed range of MAF by
172 averaging across all variants in each MAF bin considered. MAF was estimated naively on all 7,500 UK10K

173 haplotypes. All imputation software were run on pre-phased data arising from the best phased data found when
174 comparing phasing software. For general populations, it is possible that pre-phasing could lead to a loss of
175 imputation accuracy (Rosshyara, Horn, Kirsten, Ahnert, & Scholz, 2016) but this is unlikely to be significant in
176 population isolates where highly accurate phased data is achievable (Howie, Fuchsberger, Stephens, Marchini,
177 & Abecasis, 2012).

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

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

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

192 **Results**

193 **LD-based Phasing** - For analyses of phasing performance, we present results from only the HapGen+Pedigree
194 simulation unless otherwise indicated as the patterns of results were very similar between the two simulation
195 strategies. Imperfect ARRAY data initially spanned 13,599 variants on chr10 and following quality control an
196 average of 13,262 variants remained on the HapGen+Pedigree simulation strategy. Totalling over the 477
197 individuals and across the entirety of chr10, phasing algorithms were required to phase an average of 2,150,627
198 heterozygous sites in each simulation replicate. All LD-based phasing algorithms considered were able to phase
199 the ARRAY data to a high degree of accuracy with global SERs below 0.002 (Figure 1). EAGLE2 delivered
200 improved SER compared to EAGLE1 (Supplementary Figure 4) and so we only present detailed results for
201 EAGLE2. SHAPEIT2 provided the most accurately phased data and the additions of the 'duohmm' option and

202 the 1000G as an external reference panel further improved its performance. SHAPEIT3 performed similarly to
203 SHAPEIT2 and for subsequent analysis we will only present results for SHAPEIT2+duohmm+1000G.
204 SHAPEIT2+duohmm+1000G achieved a mean SER of 1.9×10^{-4} whilst EAGLE2 achieved 3.2×10^{-4} . The mean
205 global SERs for all phasing strategies considered are given in Supplementary Table 2.

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

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

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

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

228 The levels of IBD in the simulated populations clearly affected phasing performance as all software
229 had improved phasing accuracy in the presence of the elevated IBD in the HapGen+Pedigree simulation as
230 compared to the Pedigree simulation strategy (Supplementary Figure 7). Similarly, SLRP and ALPHAPHASE

231 both phased many more sites on the HapGen+Pedigree simulation (Supplementary Figures 8a-b). At the
232 individual level, all phasing algorithms had lower performance for the individuals with the lowest mean
233 pairwise genetic kinship to the rest of the sample (Supplementary Figures 9a-c).

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

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

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

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

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

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

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

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

275 For MINIMAC3, imputation accuracy was clearly improved by using the HRC over the 1000G (Figure
276 5). Imputation which included the SSP again produced more accurate results than imputation with only public
277 reference panels on the HapGen+Pedigree simulation strategy. Rare variants were however imputed more
278 accurately by MINIMAC3+HRC than by MINIMAC3+SSP on the Pedigree simulation. The results of Figures
279 3, 4, and 5 are summarised in Supplementary Table 3.

280 The founding event in an isolate will result in higher MAFs for certain variants as compared to general
281 populations. Variants with a high difference in MAF compared to the 1000G were imputed as well as the
282 random selections of comparable variants under IMPUTE2+SSP, but with lower accuracy under
283 IMPUTE2+1000G (Supplementary Figure 16a). When changing reference panel from the 1000G to the SSP, we
284 observed that imputation accuracy increased the most for variants with a MAF higher in the sample than the
285 1000G (Supplementary Materials and Supplementary Figure 16b). Another consequence of using solely the
286 1000G as a reference panel was the fact that some variants which were monomorphic in the sample were

287 imputed with dosages compatible with being heterozygous for many individuals, i.e. polymorphic in the sample
288 (Supplementary Figures 16c-d).

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

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

301 **Discussion**

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

311 If we compare our phasing results to published results for outbred populations, it is clear that all
312 methods performed with greater accuracy (SERs at least one order of magnitude smaller) on our simulated data.
313 Indeed, for outbred populations, very large sample sizes have been required to achieve the high level of phasing
314 accuracy observed in our population isolate study. For examples, see Bycroft et al. (2017), Loh, Danecek, et al.
315 (2016), O'Connell et al. (2016), and Mitt et al. (2017).

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

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

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

334 Our comparisons on imputation strategies agree with recent literature (Deelen et al., 2014; Mitt et al.,
335 2017; Pistis et al., 2015) in terms of the improvement in accuracy brought by a reference panel specific to the
336 population under study. Mitt et al. (2017) concluded that for certain outbred populations, such a panel can
337 outperform an order of magnitude larger and more diverse reference panel (the HRC). We show that for a
338 population isolate, an SSP can be far smaller and still outperform the HRC. As discussed in Asimit and Zeggini
339 (2012), the appropriate size of the SSP will depend on the diversity of the isolate.

340 The HapGen+Pedigree simulation strategy gave the best representation of a true isolate with a strong
341 founder effect producing large disparities to general populations represented in public databases. Of the two
342 simulation strategies, imputation accuracy was significantly lower on this simulation when using only a public
343 reference panel. This suggests that for a population isolate with a very small set of founders and high relatedness
344 between individuals, using public reference panels alone is not a completely appropriate strategy for imputation.
345 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 ($MAF \leq 0.05$) 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

376 pedigree information for phasing and an SSP for imputation. We were unable to successfully setup and run
377 PRIMAL on our simulated datasets and we have been advised by the authors to wait for a new version which is
378 soon to be released.

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

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

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

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

401 **Conflict of Interest:** None Declared

402 **URLs:**

- 403 1. ALPHAPHASE (v1.2), <http://www.alphagenes.roslin.ed.ac.uk/alphasuite-sofwares/alphaphase/>.
- 404 2. BEAGLE (v4.1), <http://faculty.washington.edu/browning/beagle/beagle.html>.

- 405 3. EAGLE2 (v2.3.2) & EAGLE1 (v1.0), <http://www.hsph.harvard.edu/alkes-price/software/>.
- 406 4. SHAPEIT2 (v2.837), http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html.
- 407 5. SHAPEIT3 (v1.0), <https://jmarchini.org/shapeit3/>.
- 408 6. SLRP (v1.0), <https://github.com/kpalin/SLRP>.
- 409 7. IMPUTE2 (v2.3.2), https://mathgen.stats.ox.ac.uk/impute/impute_v2.html.
- 410 8. IMPUTE4 (v1.0), <https://jmarchini.org/impute-4/>.
- 411 9. MINIMAC3 (v.2.0.1), <http://genome.sph.umich.edu/wiki/Minimac3>.
- 412 10. 1000 Genomes data set (Phase 3) , <http://www.1000genomes.org/>.
- 413 11. Haplotype Reference Consortium, <http://www.haplotype-reference-consortium.org/>.
- 414 12. UK10K Project, <https://www.uk10k.org/>.
- 415 13. Sanger Imputation Service, <https://imputation.sanger.ac.uk/>.
- 416 14. Michigan Imputation Server, <https://imputationserver.sph.umich.edu/>.
- 417 15. Oxford Statistics Phasing Server, <https://phasingserver.stats.ox.ac.uk/>.
- 418 16. R-package ‘Gaston’, <https://cran.r-project.org/web/packages/gaston/index.html>.
- 419 17. European Genome-phenome Archive, <https://www.ebi.ac.uk/ega/home>.
- 420

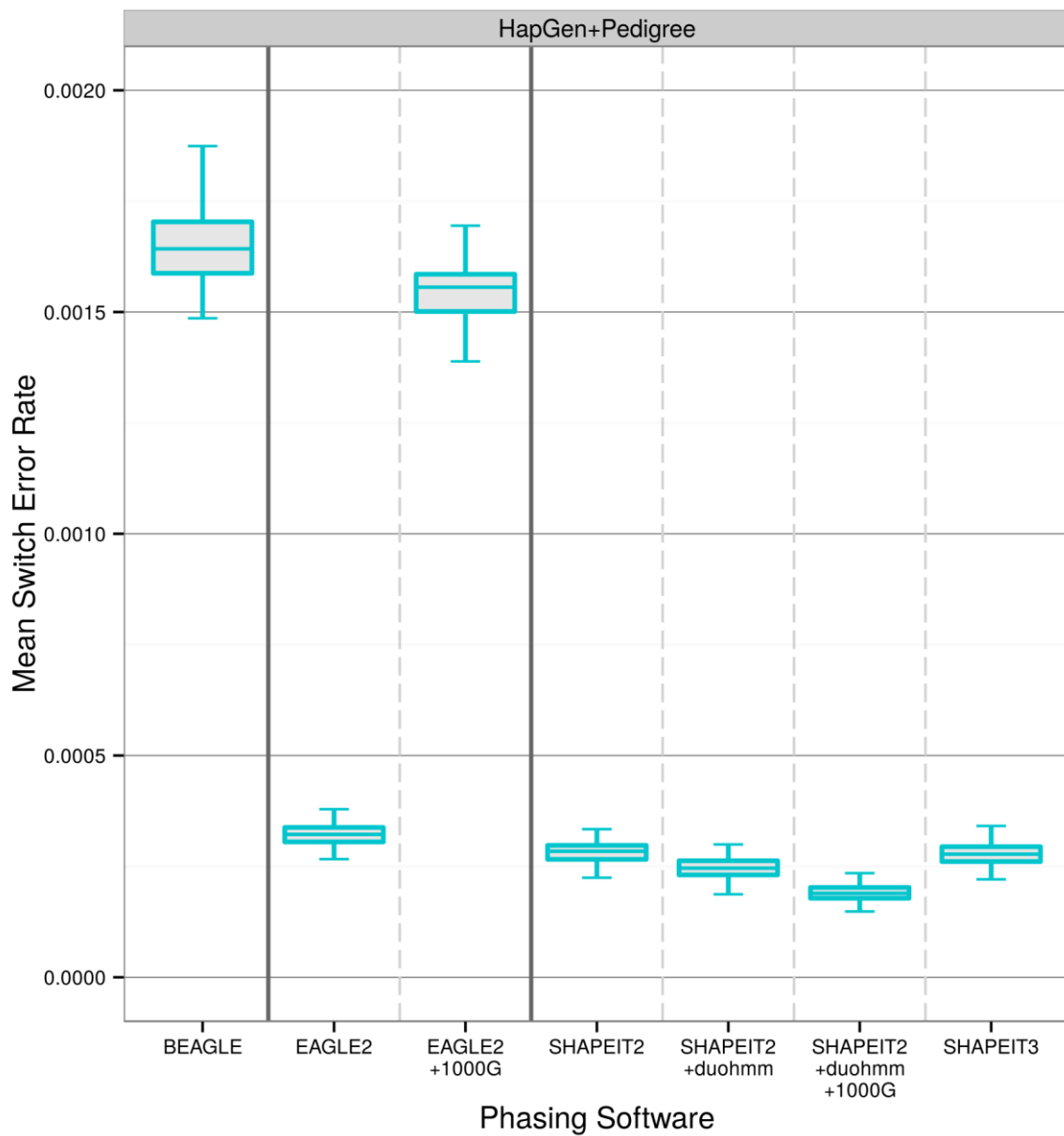
421 References

- 422 *Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin--rapid analysis of dense*
 423 *genetic maps using sparse gene flow trees. Nat Genet, 30(1), 97-101. doi: 10.1038/ng786*
- 424 *Asimit, J. L., & Zeggini, E. (2012). Imputation of rare variants in next generation association studies.*
 425 *Human heredity, 74(0), 196-204. doi: 10.1159/000345602*
- 426 *Bourgain, C., & Génin, E. (2005). Complex trait mapping in isolated populations: Are specific*
 427 *statistical methods required? Eur J Hum Genet, 13(6), 698-706.*
- 428 *Browning, Brian L., & Browning, Sharon R. (2009). A unified approach to genotype imputation and*
 429 *haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum*
 430 *Genet, 84(2), 210-223. doi: 10.1016/j.ajhg.2009.01.005*
- 431 *Browning, Brian L., & Browning, Sharon R. (2016). Genotype Imputation with Millions of Reference*
 432 *Samples. Am J Hum Genet, 98(1), 116-126. doi: 10.1016/j.ajhg.2015.11.020*
- 433 *Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . Marchini, J. (2017).*
 434 *Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv. doi:*
 435 *10.1101/166298*
- 436 *Chen, W., & Schaid, D. J. (2014). PedBLIMP: extending linear predictors to impute genotypes in*
 437 *pedigrees. Genet Epidemiol, 38(6), 531-541. doi: 10.1002/gepi.21838*
- 438 *Cheung, C. Y., Thompson, E. A., & Wijsman, E. M. (2013). GIGI: an approach to effective imputation of*
 439 *dense genotypes on large pedigrees. Am J Hum Genet, 92(4), 504-516. doi:*
 440 *10.1016/j.ajhg.2013.02.011*
- 441 *Colonna, V., Natile, T., Astore, M., Guardiola, O., Antoniol, G., Ciullo, M., & Persico, M. G. (2007).*
 442 *Campora: A Young Genetic Isolate in South Italy. Human heredity, 64(2), 123-135. doi:*
 443 *10.1159/000101964*

- 444 Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., . . . Fuchsberger, C. (2016). Next-
445 generation genotype imputation service and methods. *Nat Genet*, 48(10), 1284-1287. doi:
446 10.1038/ng.3656
- 447 Deelen, P., Menelaou, A., van Leeuwen, E. M., Kanterakis, A., van Dijk, F., Medina-Gomez, C., . . .
448 Swertz, M. A. (2014). Improved imputation quality of low-frequency and rare variants in
449 European samples using the 'Genome of The Netherlands'. *Eur J Hum Genet*, 22(11), 1321-
450 1326. doi: 10.1038/ejhg.2014.19
- 451 Delaneau, O., Howie, B., Cox, Anthony J., Zagury, J.-F., & Marchini, J. (2013). Haplotype Estimation
452 Using Sequencing Reads. *Am J Hum Genet*, 93(4), 687-696. doi: 10.1016/j.ajhg.2013.09.002
- 453 Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease
454 and population genetic studies. *Nat Meth*, 10(1), 5-6. doi: 10.1038/nmeth.2307
- 455 Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows–Wheeler
456 transform (PBWT). *Bioinformatics*, 30(9), 1266-1272. doi: 10.1093/bioinformatics/btu014
- 457 Gazal, S., Sahbatou, M., Perdry, H., Letort, S., Génin, E., & Leutenegger, A. L. (2014). Inbreeding
458 Coefficient Estimation with Dense SNP Data: Comparison of Strategies and Application to
459 HapMap III. *Human heredity*, 77(1-4), 49-62.
- 460 Glodzik, D., Navarro, P., Vitart, V., Hayward, C., McQuillan, R., Wild, S. H., . . . McKeigue, P. (2013).
461 Inference of identity by descent in population isolates and optimal sequencing studies. *Eur J
462 Hum Genet*, 21(10), 1140-1145. doi: 10.1038/ejhg.2012.307
- 463 Hatzikotoulas, K., Gilly, A., & Zeggini, E. (2014). Using population isolates in genetic association
464 studies. *Briefings in Functional Genomics*, 13(5), 371-377. doi: 10.1093/bfpgp/elu022
- 465 Hickey, J. M., Kinghorn, B. P., Tier, B., Wilson, J. F., Dunstan, N., & van der Werf, J. H. J. (2011). A
466 combined long-range phasing and long haplotype imputation method to impute phase for
467 SNP genotypes. *Genetics, Selection, Evolution : GSE*, 43(1), 12-12. doi: 10.1186/1297-9686-
468 43-12
- 469 Holm, H., Gudbjartsson, D. F., Sulem, P., Masson, G., Helgadóttir, H. T., Zanon, C., . . . Stefansson, K.
470 (2011). A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet*,
471 43(4), 316-320. doi: 10.1038/ng.781
- 472 Howie, B., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method
473 for the next generation of genome-wide association studies. *PLoS Genet*, 5(6), e1000529. doi:
474 10.1371/journal.pgen.1000529
- 475 Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate
476 genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*,
477 44(8), 955-959. doi: 10.1038/ng.2354
- 478 Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3
479 (Bethesda)*, 1(6), 457-470. doi: 10.1534/g3.111.001198
- 480 Huang, L., Wang, C., & Rosenberg, N. A. (2009). The relationship between imputation error and
481 statistical power in genetic association studies in diverse populations. *Am J Hum Genet*,
482 85(5), 692-698. doi: 10.1016/j.ajhg.2009.09.017
- 483 Joshi, P. K., Prendergast, J., Fraser, R. M., Huffman, J. E., Vitart, V., Hayward, C., . . . Navarro, P.
484 (2013). Local Exome Sequences Facilitate Imputation of Less Common Variants and Increase
485 Power of Genome Wide Association Studies. *PLOS ONE*, 8(7), e68604. doi:
486 10.1371/journal.pone.0068604
- 487 Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., . . . Nielsen, R. (2011).
488 Estimation of allele frequency and association mapping using next-generation sequencing
489 data. *BMC Bioinformatics*, 12, 231-231. doi: 10.1186/1471-2105-12-231
- 490 Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., . . . Stefansson, K.
491 (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat
492 Genet*, 40(9), 1068-1075. doi: 10.1038/ng.216

- 493 Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using Sequence and Genotype
494 Data to Estimate Haplotypes and Unobserved Genotypes. *Genet Epidemiol*, 34(8), 816-834.
495 doi: 10.1002/gepi.20533
- 496 Livne, O. E., Han, L., Alkorta-Aranburu, G., Wentworth-Sheilds, W., Abney, M., Ober, C., & Nicolae, D.
497 L. (2015). PRIMAL: Fast and Accurate Pedigree-based Imputation from Sequence Data in a
498 Founder Population. *PLoS Computational Biology*, 11(3), e1004139. doi:
499 10.1371/journal.pcbi.1004139
- 500 Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., . . . L Price, A.
501 (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nat*
502 *Genet*, 48(11), 1443-1448. doi: 10.1038/ng.3679
- 503 Loh, P.-R., Palamara, P. F., & Price, A. L. (2016). Fast and accurate long-range phasing in a UK
504 Biobank cohort. *Nat Genet*, 48(7), 811-816. doi: 10.1038/ng.3571
- 505 Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev*
506 *Genet*, 11(7), 499-511. doi: 10.1038/nrg2796
- 507 McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., . . . The Haplotype
508 Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype
509 imputation. *Nat Genet*, 48(10), 1279-1283. doi: 10.1038/ng.3643
- 510 Mitt, M., Kals, M., Parn, K., Gabriel, S. B., Lander, E. S., Palotie, A., . . . Palta, P. (2017). Improved
511 imputation accuracy of rare and low-frequency variants using population-specific high-
512 coverage WGS-based imputation reference panel. *Eur J Hum Genet*. doi:
513 10.1038/ejhg.2017.51
- 514 O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., . . . Marchini, J. (2014). A
515 General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS*
516 *Genetics*, 10(4), e1004234. doi: 10.1371/journal.pgen.1004234
- 517 O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., . . . Marchini, J. (2016). Haplotype
518 estimation for biobank-scale data sets. *Nat Genet*, 48(7), 817-820. doi: 10.1038/ng.3583
- 519 Palin, K., Campbell, H., Wright, A. F., Wilson, J. F., & Durbin, R. (2011). Identity-by-Descent-Based
520 Phasing and Imputation in Founder Populations Using Graphical Models. *Genet Epidemiol*,
521 35(8), 853-860. doi: 10.1002/gepi.20635
- 522 Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., . . . Sanna, S. (2015). Rare variant
523 genotype imputation with thousands of study-specific whole-genome sequences:
524 implications for cost-effective study designs. *Eur J Hum Genet*, 23(7), 975-983. doi:
525 10.1038/ejhg.2014.216
- 526 Roshyara, N. R., Horn, K., Kirsten, H., Ahnert, P., & Scholz, M. (2016). Comparing performance of
527 modern genotype imputation methods in different ethnicities. *Scientific Reports*, 6, 34386.
528 doi: 10.1038/srep34386
- 529 Roshyara, N. R., & Scholz, M. (2015). Impact of genetic similarity on imputation accuracy. *BMC*
530 *Genetics*, 16, 90. doi: 10.1186/s12863-015-0248-2
- 531 Sharp, K., Kretzschmar, W., Delaneau, O., & Marchini, J. (2016). Phasing for medical sequencing using
532 rare variants and large haplotype reference panels. *Bioinformatics*, 32(13), 1974-1980. doi:
533 10.1093/bioinformatics/btw065
- 534 Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs.
535 *Bioinformatics*, 27(16), 2304-2305. doi: 10.1093/bioinformatics/btr341
- 536 Surakka, I., Kristiansson, K., Anttila, V., Inouye, M., Barnes, C., Moutsianas, L., . . . Ripatti, S. (2010).
537 Founder population-specific HapMap panel increases power in GWA studies through
538 improved imputation accuracy and CNV tagging. *Genome Res*, 20(10), 1344-1351. doi:
539 10.1101/gr.106534.110
- 540 The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation.
541 *Nature*, 526(7571), 68-74. doi: 10.1038/nature15393

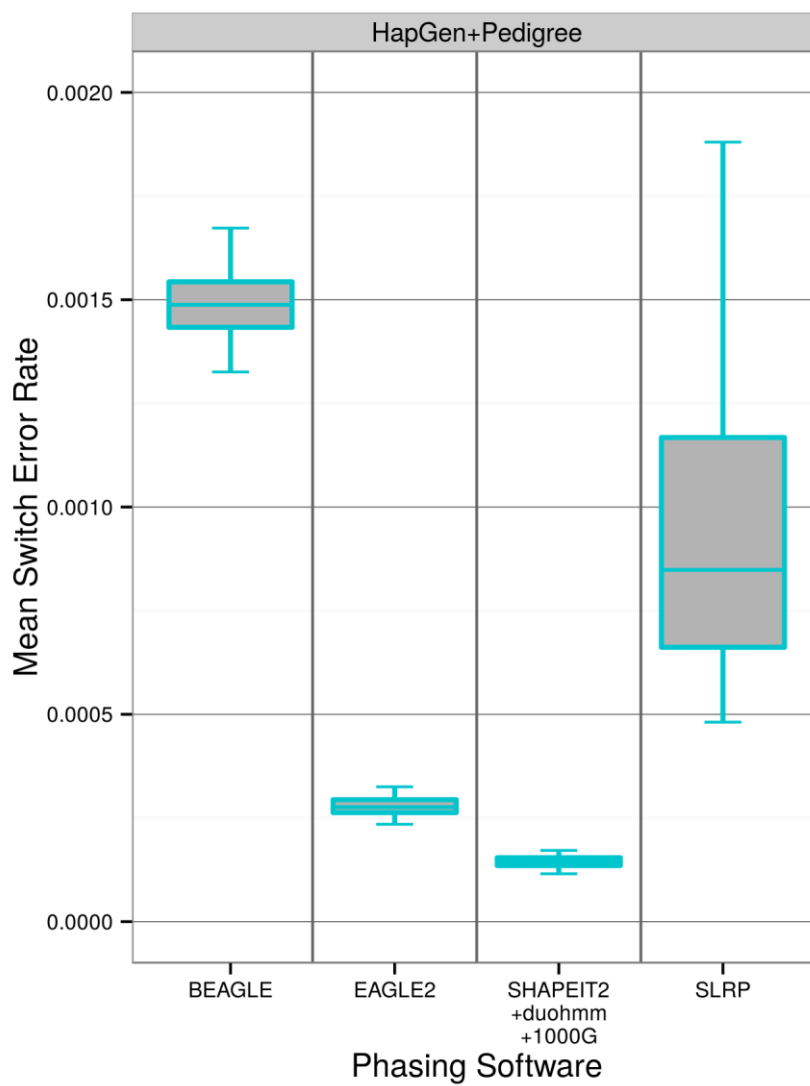
- 542 *The UK10K Consortium. (2015). The UK10K project identifies rare variants in health and disease.*
543 *Nature, 526(7571), 82-90. doi: 10.1038/nature14962*
- 544 *Uricchio, L. H., Chong, J. X., Ross, K. D., Ober, C., & Nicolae, D. L. (2012). Accurate imputation of rare*
545 *and common variants in a founder population from a small number of sequenced individuals.*
546 *Genet Epidemiol, 36(4), 312-319. doi: 10.1002/gepi.21623*
- 547 *Vieira, F. G., Albrechtsen, A., & Nielsen, R. (2016). Estimating IBD tracts from low coverage NGS data.*
548 *Bioinformatics, 32(14), 2096-2102. doi: 10.1093/bioinformatics/btw212*
- 549 *Wijsman, E. M., Rothstein, J. H., & Thompson, E. A. (2006). Multipoint Linkage Analysis with Many*
550 *Multiallelic or Dense Diallelic Markers: Markov Chain–Monte Carlo Provides Practical*
551 *Approaches for Genome Scans on General Pedigrees. Am J Hum Genet, 79(5), 846-858.*
- 552 *Zeggini, E. (2011). Next-generation association studies for complex traits. Nat Genet, 43(4), 287-288.*
553 *doi: 10.1038/ng0411-287*
- 554
- 555



556

557 **Figure 1.** Global switch error rates for BEAGLE, EAGLE2, SHAPEIT2, and SHAPEIT3 for the

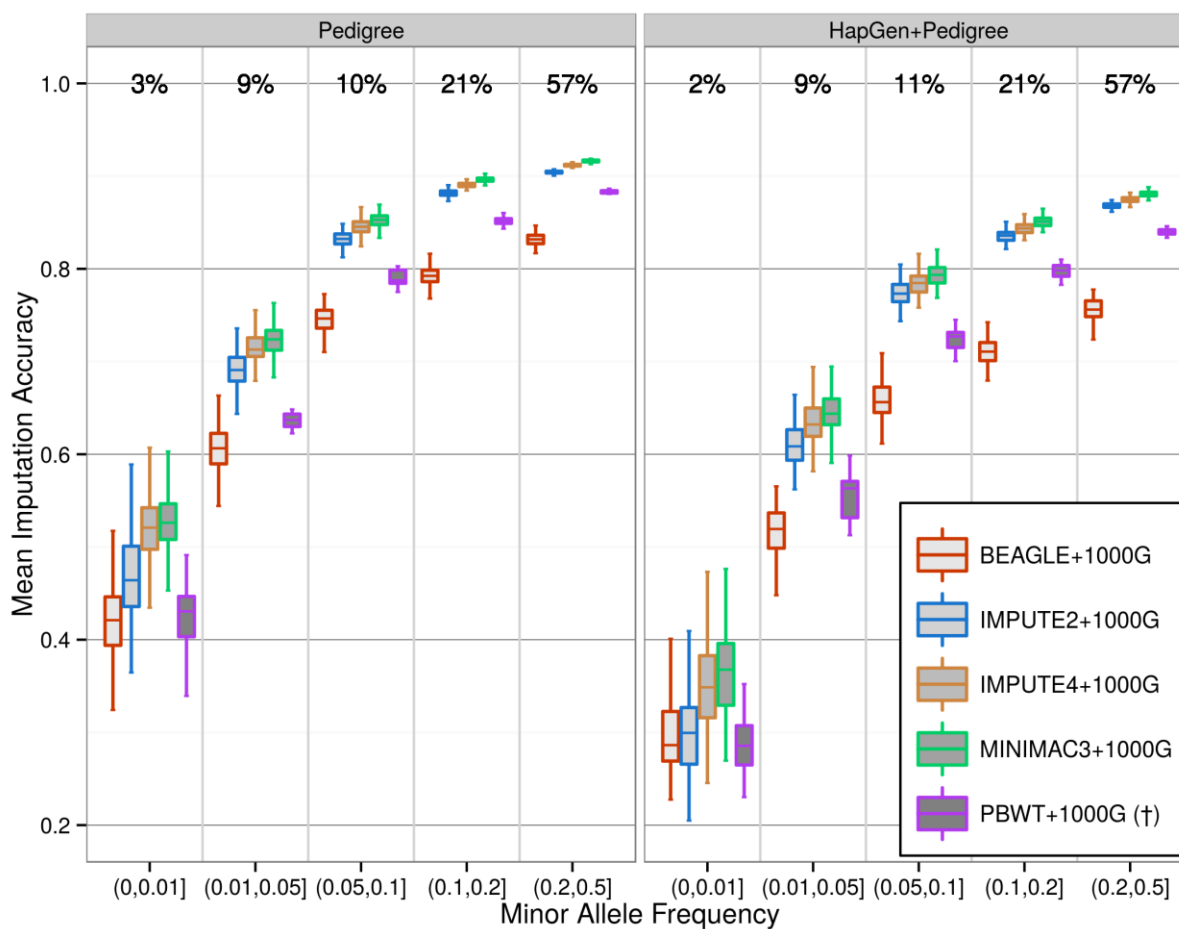
558 HapGen+Pedigree simulation strategy.



559

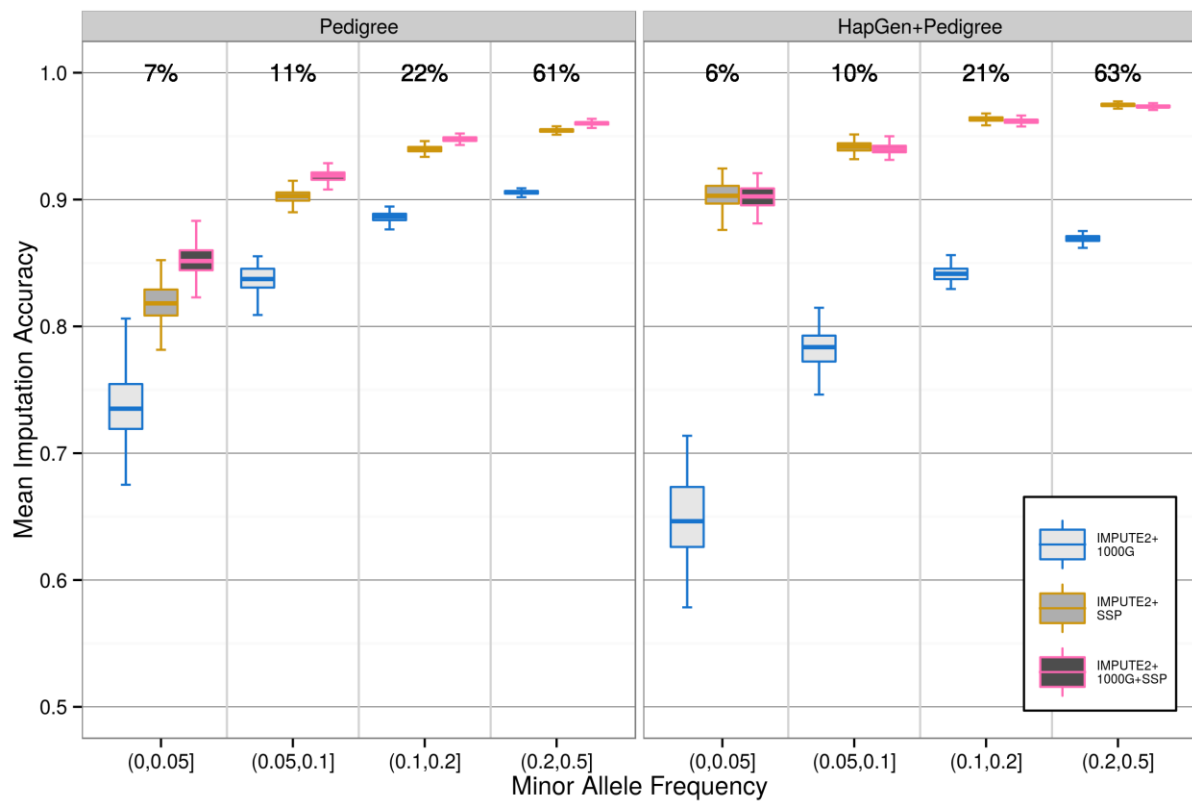
560 **Figure 2.** Global switch error rates for BEAGLE, SLRP, EAGLE2, and SHAPEIT2+duohmm+1000G for the

561 HapGen+Pedigree simulation strategy on the set of variants successfully phased by SLRP in each replicate.



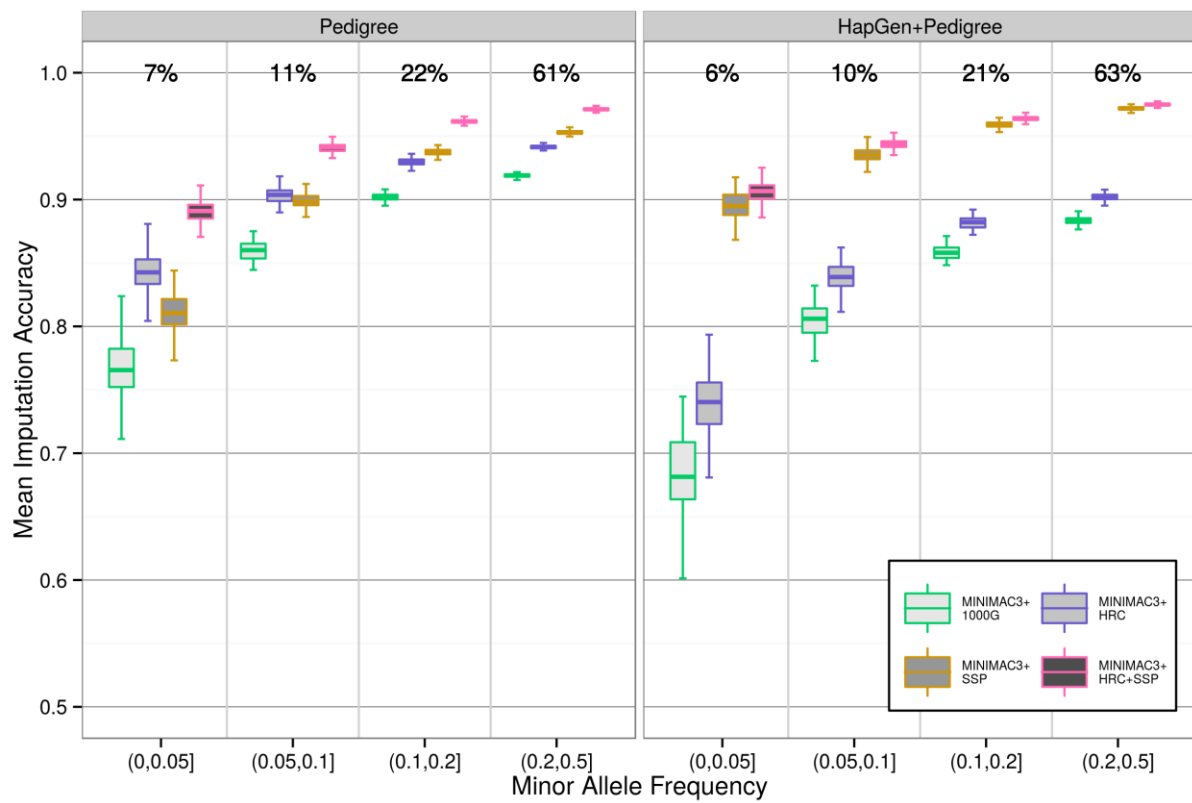
562

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



567

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



574

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