Kinpute: Using identity by descent to improve genotype imputation

Mark Abney* and Aisha El Sherbiny

Department of Human Genetics, University of Chicago,

Chicago, IL 60637, USA

*To whom correspondence should be addressed, abney@uchicago.edu

1 Abstract

Motivation Genotype imputation, though generally accurate, often results in many genotypes being poorly imputed, particularly in studies where the individuals are not well represented by standard reference panels. When individuals in the study share regions of the genome identical by descent (IBD), it is possible to use this information in combination with a study specific reference panel (SSRP) to improve the imputation results. Kinpute uses IBD information—due to either recent, familial relatedness or distant, unknown ancestors—in conjunction with the output from linkage disequilibrium (LD) based imputation methods to compute more accurate genotype probabilities. Kinpute uses a novel method for IBD imputation, which works even in the absence of a pedigree, and results in substantially improved imputation quality.

Results Given initial estimates of average IBD between subjects in the study sample, Kinpute uses a novel algorithm to select an optimal set of individuals to sequence and use as an SSRP. Kinpute is designed to use as input both this SSRP and the genotype probabilities output from other LD based imputation software, and uses a new method to combine the LD imputed genotype probabilities with IBD configurations to substantially improve imputation. We tested Kinpute on a human population isolate where 98 individuals have been sequenced. In half of this sample, whose sequence data was masked, we used Impute2 to perform LD based imputation and Kinpute was used to obtain higher accuracy genotype probabilities. Measures of imputation accuracy improved significantly, particularly for those genotypes that Impute2 imputed with low certainty.

Availability Kinpute is an open-source and freely available C++ software package that can be downloaded from `https://github.com/markabney/Kinpute/releases`.

2 Introduction

Genotype imputation methods have been a boon to researchers, allowing them to maximize available resources by allowing them to computationally infer the alleles at untyped variants for many individuals. Although a majority of genotypes may be imputed with high accuracy using standard reference panels, a substantial fraction of variants will typically be discarded due to low quality imputation. For study populations that are not well represented by available reference panels, this problem is exacerbated (Herzig *et al.*, 2018). When imputing genotypes in indigenous and founder populations, or even isolated populations of European ancestry, the results can be significantly improved when a subset of the sample are sequenced and included as a study specific reference panel (SSRP) (Deelen *et al.*, 2014; Hou *et al.*, 2017; Mitt *et al.*, 2017; Pistis *et al.*, 2015; Sidore *et al.*, 2015; Zhou *et al.*, 2017; Herzig *et al.*, 2018).

When identity by descent (IBD) from close relatives to an unsequenced individual exists, linkage disequilibrium (LD) based imputation approaches are expected to have high accuracy (McCarthy *et al.*, 2016). This is because LD-based approaches will, in principle, phase and match the haplotype of the unsequenced individual with the IBD segment from a sequenced individual. In practice, this phasing and matching are done imperfectly, allowing for the possibility that IBD based imputation methods might be an alternative. IBD based imputation methods generally are either ones that find and use extended IBD regions to both phase and impute genotypes (Kong *et al.*, 2008; Palin *et al.*, 2011; Uricchio *et al.*, 2012) or use a pedigree to inform imputations (Burdick *et al.*, 2006; Chen and Schaid, 2014; Cheung *et al.*, 2013; Livne *et al.*, 2015). Pedigree based imputation can outperform LD based methods, particularly at variants with low allele frequencies (Ullah *et al.*, 2019), but it is less clear whether IBD can improve on LD based imputation in the absence of a pedigree once an SSRP is included (Herzig *et al.*, 2018).

Here, we show that our method, Kinpute, which does not use pedigree data, can use IBD information to obtain substantially improved imputation accuracy. Kinpute uses a novel approach to imputation in that genotype probabilities output from LD imputation methods,

such as Impute2 (Howie *et al.*, 2012), minimac (Fuchsberger *et al.*, 2015; Das *et al.*, 2016), and Beagle (Browning *et al.*, 2018; Browning and Browning, 2016) are used as prior probabilities in a unified probabilistic model that includes IBD information. Other approaches to combine IBD and LD based imputation include (1) selecting either the pedigree imputed genotype, if a pedigree based method can be used, or the LD imputed genotype according to some criterion (Saad and Wijsman, 2014; Blue *et al.*, 2014; Livne *et al.*, 2015) or (2) by taking a linear combination of the two sets of imputed genotypes (Chen and Schaid, 2014). By using the LD imputed genotypes as prior probabilities, Kinpute provides a natural, probabilistic model to integrate the two sources of imputation information while allowing for a wide range of study designs (e.g. presence or absence of pedigree data). In addition, Kinpute does not require phase information, making it robust to phasing errors, but does require an SSRP, which may also be used when performing LD-based imputation, and implements a novel algorithm to select an optimal SSRP.

3 Methods

The software consists of two components, both of which require prior estimates of IBD sharing. The first component selects an optimal set of individuals to use as an SSRP while leaving the rest as the imputation panel. The second component performs genotype imputation on the imputation panel given sequence data on the SSRP, prior genotype probabilities, and the previously computed IBD information.

3.1 Optimal study specific reference panel

To select an optimal SSRP we assume knowledge of kinship coefficients, computed from a pedigree or from a genetic relationship matrix estimated from genotype data, for every pair of individuals in the study. These coefficients measure the probability of a pair of individuals being IBD at an arbitrary locus.

We seek a set of individuals of a prespecified size from the study sample that, when sequenced, will provide the most informative subsample (i.e. the SSRP) to impute the sequence data into the remaining individuals (i.e. the imputation sample). We use IBD sharing as the measure of informativeness for imputation. Consider a locus where IBD probabilities have been computed between all pairs in the sample. Let $\{\phi_{ij}\}$ be the set such that ϕ_{ij} is the probability that, at that locus, an allele drawn randomly from individual i is IBD with a randomly drawn allele from individual j. When computed from a pedigree in the absence of marker data, this definition is formally equivalent to the kinship coefficient. Also, let R be the set of sequenced individuals and U the set of genotyped but not sequenced individuals, and let $r \in R$ and $u \in U$. The quantity ϕ_{ru}, then, is a measure of how informative r is with respect to imputing u's genotype at that locus. For two sequenced individuals r_1, r_2 our information for imputing u's genotype is measured by $1 - (1 - \phi_{r_1 u})(1 - \phi_{r_2 u})$. This is the probability that at least one of the randomly drawn alleles from the sequenced individuals will be IBD with a random allele from u at the locus. As this probability approaches one we become increasingly certain that at least one of the sequenced individual will be IBD with u. We extend this pattern to form a score by taking the product over all $r \in R$ and summing over all $u \in U$ for a given division of subjects into R and U. The optimal set of individuals to sequence at this locus is obtained by minimizing the loss function $L = \sum_{u \in U} \prod_{r \in R} (1 - \phi_{ru})$ over all divisions of subjects for a given number of subjects in R. To obtain a genomewide score we let ϕ_{ij} be either the kinship coefficient or the ijth component of the genetic relatedness matrix computed from known genotype data. Alternatively, one may define a value L_m at each marker m and set $L = \sum_m L_m$, though we do not implement this latter approach.

We use a simple greedy algorithm to select which individuals in the sample to include in R. That is, the first individual from the sample is chosen such that L is minimized for that person in R and everyone else in U. Subsequent individuals are added to R by keeping the individuals already in R and finding which one of the remaining individuals in U minimizes L. This procedure is continued until the desired size of R is met. Note that this algorithm

will not, in general, result in the globally optimal solution for R. It does, however, provide a computationally tractable approach to obtaining a good, if not the best, solution.

Note that under the case where we are perfectly informed with respect to IBD sharing, the method proposed in (Gusev *et al.*, 2012) is a special case of our more general algorithm. In their approach, rather than basing the loss function on IBD probabilities it is based on hard calls of segments being IBD or not. IBD segment calling, however, is in practice far from perfect and can have significant false positive and/or false negative error rates. Using probabilities has the advantage of taking uncertainty into account and allowing moderate probabilities to combine for greater information. This is not possible if the probabilities are thresholded to a binary classification of IBD or not.

3.2 Imputation

The imputation method is designed to work in conjunction with the output of other genotype imputation methods, e.g. Impute2 (Howie *et al.*, 2012), minimac (Fuchsberger *et al.*, 2015) or Beagle (Browning and Browning, 2016), that output genotype probabilities at specific markers. These genotype probabilities are used as prior probabilities for our method. In the absence of genotype probabilities from other methods, Kinpute will compute prior probabilities from the allele frequencies in the SSRP.

Consider a study where the entire sample has been genotyped at a set of markers (i.e. the framework markers) and in which a subset of the sample of individuals have been sequenced. The sequenced individuals are the SSRP $R = \{r_i\}$, where r_i is the ith reference panel individual, $i = 1, \ldots, N$. Let G_i^m be the true, unphased sequence data at marker m, $m = 1, \ldots, M$ in the SSRP with $i = 1, \ldots, N$, and O_i^m be the corresponding observed, unphased sequence data. The imputation problem is to find $\Pr(G_u^m | O_1^*, \ldots, O_N^*, F_u)$, where F_u is the framework genotype data of u, G_u^m is the unknown sequence at marker m of sample individual u who has not been sequenced and the superscript asterisk indicates the set over all markers. In addition, we assume that from the framework genotype data we have estimates $\hat{S}_{ab}^*$ on

the IBD condensed identity states $S^l_{ab} = k$ where $k \in \{1, \ldots, 9\}$ indexes the identity state at framework marker l, $l = 1, \ldots, L < M$, and $a, b \in \{u, r_1, \ldots, r_N\}$. (Supplementary Figure 1 displays the nine condensed identity states.)

Standard imputation methods typically use LD information to first phase the observed genotype and sequence data then compute the probability $\Pr(G^m_u | O^*_1, \ldots, O^*_N, F_u)$. That is, a primary component of the imputed genotype at a marker is the phased observed alleles of u at nearby markers. Errors in phasing, or in the observed genotypes or sequences, can reduce the accuracy of the imputation. Kinpute on the other hand, uses IBD, which extends over a longer range than LD. Given this IBD information Kinpute seeks to compute the probability $\Pr(G^m_u | O^m_1, \ldots, O^m_N, \{\hat{S}^m_{ab}\})$. Note that in LD-based imputation the probability is computed given the genomewide sequence data of the reference panel and genomewide framework genotype data, while the conditional probability used by Kinpute apparently conditions only on the reference panel sequence data and IBD estimates at marker m. In fact, Kinpute implicitly conditions on the genomewide framework data because this is used in the computation of the IBD estimates $\hat{S}^m_{ab}$ and on the reference panel genomewide sequence through the use of a prior genotype probability that depends on this, as described below. Note also that though the probabilities are conditioned on genomewide data, in practice information local to m is most influential. Rather than compute this conditional probability directly under some probability model (often some form of a hidden Markov model), as is done with LD based imputation, we instead use Bayes' law,

$$\begin{aligned}\Pr(G_u | O_1, \ldots, O_N, \{\hat{S}_{ab}\}) &= \frac{\Pr(O_1, \ldots, O_N | G_u, \{\hat{S}_{ab}\}) \Pr(G_u | \{\hat{S}_{ab}\})}{\sum_{g=0}^{2} \Pr(O_1, \ldots, O_N | G_u = g, \{\hat{S}_{ab}\}) \Pr(G_u = g | \{\hat{S}_{ab}\})} \\ &\approx \frac{\Pr(O_1, \ldots, O_N | G_u, \{\hat{S}_{ab}\}) \Pr(G_u)}{\sum_{g=0}^{2} \Pr(O_1, \ldots, O_N | G_u = g, \{\hat{S}_{ab}\}) \Pr(G_u = g)},\end{aligned} \tag{1}$$

where we have dropped the marker m superscript for notational simplicity. We have used $\Pr(G_u = g | \{\hat{S}_{ab}\}) \approx \Pr(G_u = g)$, which is an approximation because, for instance, the presence of inbreeding can affect the probabilities of u's genotypes. Unless inbreeding is

substantial, however, the approximation is highly accurate. Furthermore, this formulation allows us to choose prior probabilities $\Pr(G_u)$ based on other sources of information. In particular, we use the imputed genotype probabilities of standard LD based imputation as prior probabilities. Below, we focus on our approach to computing $\Pr(O_1, \ldots, O_N | G_u, \{\hat{S}_{ab}\})$.

3.2.1 Error model

Throughout, we allow for errors to exist between the true genotype G_i and the observed genotype O_i. We further assume that the observed genotypes are conditionally independent of all other random variables given the true genotypes so that, for instance, $\Pr(O_1, O_2 | G_1, G_2) = \Pr(O_1 | G_1) \Pr(O_2 | G_2)$. We assume an allelic error rate of ϵ with errors of the two alleles in a genotype being independent and the probability of error at different markers also independent. (See Supplementary Table 1). This error model is certainly a simplification; however, it does not play a strong role in the final imputed genotype probabilities. Primarily, the error model prevents numerical failure of the algorithm that can occur from inconsistent observations, for instance, when the observed genotypes are inconsistent with the IBD state. In our tests small non-zero error rates, up to at least 5%, had only minor impact on the imputed probabilities. In the Kinpute software package we set $\epsilon = 0.005$.

3.2.2 Independence within the reference panel

At times, one simplifying assumption we use is that the genotypes of the reference panel individuals are conditionally independent given G_u. In this case we have

$$\begin{aligned} \Pr(O_1, \ldots, O_N | G_u, \{\hat{S}_{ab}\}) &= \prod_{i=1}^{N} \Pr(O_i | G_u, \hat{S}_{iu}) \\ &= \prod_{i=1}^{N} \sum_{G_i} \Pr(O_i | G_i) \Pr(G_i | G_u, \hat{S}_{iu}) \end{aligned} \tag{2}$$

The estimated IBD information $\hat{S}_{iu}$ consists of probabilities of each identity state given the framework set of markers, $\Pr(S_{iu} = k | F) = \hat{\Delta}_k^{iu}$, where $k = 1, \ldots, 9$ indexes the identity

state. That is, $\hat{\Delta}_k^{iu}$ is the estimated locus-specific probability of the identity state given the genotype data, and we assume this is already known. In practice, we compute these probabilities using the IBDLD (Han and Abney, 2011, 2013) software package. Furthermore, we have

$$\begin{aligned}\Pr(G_i|G_u, \hat{S}_{iu}) &= \sum_{k=1}^{9} \Pr(G_i|G_u, S_{iu} = k) \Pr(S_{iu} = k|G_u, \hat{S}_{iu}) \\ &= \sum_{k=1}^{9} \Pr(G_i|G_u, S_{iu} = k) \frac{\Pr(G_u|S_{iu} = k)\hat{\Delta}_k^{iu}}{\Pr(G_u|\hat{S}_{iu})}.\end{aligned} \tag{3}$$

In the above we use

$$\begin{aligned}\Pr(G_u = 0|\hat{S}_{iu}) &= \hat{f}_u(1-p) + (1-\hat{f}_u)(1-p)^2 \\ \Pr(G_u = 1|\hat{S}_{iu}) &= (1-\hat{f}_u)2p(1-p) \\ \Pr(G_u = 2|\hat{S}_{iu}) &= \hat{f}_u p + (1-\hat{f}_u)p^2\end{aligned} \tag{4}$$

where $\hat{f}_u$ is the estimated inbreeding coefficient of individual u at the current marker, and p is the frequency of the 1 allele. The probabilities $\Pr(G_u|S_{iu} = k)$ are given in Supplementary Table 2, and the probabilities $\Pr(G_i|G_u, S_{iu} = k)$ are given in Supplementary Table 3. The quantities in these two tables together with equation (3) allows us to compute the probability in equation (2).

3.2.3 IBD within the reference panel

It may often be the case that IBD exists between members of the reference panel at some markers. We can leverage this IBD to obtain more accurate estimates of the genotype probability of individual u at that marker. We do this by altering the factoring that occurs in Equation (2) to reflect this dependence. For instance, let the reference panel be composed of three individuals $1, 2, 3$ and there is an individual u in whom we are imputing the genotypes at a SNP. Assume that individuals $u, 1, 2$ are in an informative 3-way IBD state S_{u12} with

some probability Ψ_{u12}, while individual 3 is not IBD with either 1 or 2. We get,

$$\begin{aligned}
&\Pr(G_1, G_2, G_3 | G_u, \{\hat{S}_{ab}\}) \\
&\quad = \sum_{S_{u12}, S_{u3}} \Pr(G_1, G_2 | G_u, S_{u12}) \Pr(S_{u12} | G_u, \{\hat{S}_{iu}\}) \Pr(G_3 | G_u, S_{u3}) \Pr(S_{u3} | G_u, \hat{S}_{u3}) \\
&\quad = \sum_{S_{u12}} \Pr(G_1, G_2 | G_u, S_{u12}) \frac{\Pr(G_u | S_{u12}) \hat{\Psi}_{u12}}{\Pr(G_u | \{\hat{S}_{iu}\})} \sum_{S_{u3}} \Pr(G_3 | G_u, S_{u3}) \frac{\Pr(G_u | S_{u3}) \hat{\Delta}^{3u}_{S_{u3}}}{\Pr(G_u | \hat{S}_{u3})}.
\end{aligned} \tag{5}$$

The probabilities $\Pr(G_u | \hat{S}_{iu})$ are given by equations (4). There are many possible 3-way IBD states at a locus (Thompson, 1974), and computing them would be challenging. Instead, we use an approximation based on the pairwise IBD states at the locus and define $S_{u12} = (S_{u1}, S_{u2}, S_{12})$. Then, letting $s = (S_{u1} = k, S_{u2} = l, S_{12} = m)$ such that (k, l, m) define a legal 3-way IBD state, we define the probability of the 3-way IBD configuration as,

$$\hat{\Psi}_{u12} = \Pr(S_{u12} = s | F) = \frac{\Pr(S_{u1} = k | F) \Pr(S_{u2} = l | F) \Pr(S_{12} = m | F)}{\sum_{s \in \mathcal{S}} \Pr(S_{u1} = k | F) \Pr(S_{u2} = l | F) \Pr(S_{12} = m | F)}, \tag{6}$$

where $\mathcal{S}$ is the set of only the eight most probable values of s, for computational expediency.

In a collection of reference panel individuals, we need an algorithm to place them into 3-way configurations with u. Random, or arbitrary groupings are unlikely to be generally useful as only some 3-way configurations will be informative for imputation.

The algorithm proceeds by first picking an individual u into whom we wish to impute genotypes. At each marker, the reference panel individuals are grouped into 3-way configurations (as listed in Supplementary Table 4), if possible. Reference panel individuals who do not end up in a 3-way configuration are treated as independent. For purposes of classifying into configurations, the IBD state for a pair (e.g. S_{u1}) is the IBD state with the highest probability. If a reference panel individual belongs to more than one configuration, only one of them is chosen. For computational expediency we do not attempt to optimize the placement into configurations (some configurations may be more informative than others based on the

genotypes of the individuals), and instead simply place a reference panel individual into the first configuration found. Once reference panel individuals are placed into configurations the genotype probabilities are computed as given in Supplementary Tables 5 – 13, individuals who are not in a configuration have genotype probabilities given by Supplementary Table 3. These probabilities are combined, as in Equation (5), with 3-way configurations weighted according to their probability, as computed by Equation (6). We further apply our error model to obtain,

$$\begin{aligned}\Pr(O_1,\ldots,O_n|G_u,\{\hat{S}_{ab}\}) = &\prod_{ij\in\mathcal{C}}\sum_{G_i,G_j}\Pr(O_i|G_i)\Pr(O_j|G_j)\Pr(G_i,G_j|G_u,\{\hat{S}_{iu}\})\times \\ &\prod_{k\notin\mathcal{C}}\sum_{G_k}\Pr(O_k|G_k)\Pr(G_k|G_u,\hat{S}_{ku}),\end{aligned} \tag{7}$$

where $\mathcal{C}$ is the set of all reference panel individuals in a 3-way configuration. The final imputed genotype probabilities are, given by Equation (1).

3.3 Data and analysis

We applied our method to a human founder population. The study comprised 1,415 individuals from the Hutterite kindred. A framework set of markers, consisting of 271,486 SNPs that passed quality control and had a minor allele frequency of at least 0.05, was obtained from a combination of three Affymetrix arrays (500k, 5.0 and 6.0). The framework markers were not pruned for LD. From these individuals, 98 were sequenced by Complete Genomics, resulting in 6,617,627 SNPs that passed quality control. Details of the genotype and sequence data, including quality control is described in Livne *et al.* (2015). Although these individuals are linked by a known pedigree, all our analyses described below are done without knowledge of the pedigree.

From the 98 sequenced individuals, 50 were chosen to act as an SSRP while the other 48 were used as subjects for imputation given their framework genotypes. The sequence data of the 48 were hidden from imputation and used as the "ground truth" against which

to compare the imputed genotypes. To assess the degree to which Kinpute can use IBD information to improve genotype imputation we focus on the genotypes on chromosome 22 only. On chromosome 22 there are 3059 SNPs in the framework set and 111,369 SNPs in the 98 sequenced individuals. We performed an initial round of imputation on the 48 subjects using Impute2 (Howie *et al.*, 2012) with pre-phasing by ShapeIt2 (O'Connell *et al.*, 2014). We used a reference panel that consisted of the 1000 Genomes panel (1000 Genomes Project Consortium *et al.*, 2015) merged with the 50 SSRP individuals, with merging performed by Impute2. The resulting imputed genotype probabilities were used as prior probabilities for our Kinpute method, as described below.

To apply our Kinpute method we first used the IBDLD software package (Han and Abney, 2011) with the GIBDLD method (Han and Abney, 2013), which does not use a pedigree, to estimate probabilities of IBD sharing at each SNP in the framework set for every pair of the 98 individuals. Using these IBD probabilities, the SSRP, and the output from Impute2 as the genotype prior probabilities, we used Kinpute to obtain posterior probabilities for every genotype on chromosome 22 in the 48 subjects. We then compared the imputed genotypes, both those done solely by Impute2 and those done with the combination of Impute2 and Kinpute, with the sequenced genotypes in the 48 subjects to assess accuracy.

4 Results

To assess the imputation improvement provided by Kinpute, we stratify the imputed genotypes by genotype and SNP metrics. SNPs are stratified by Impute2 Info score (> 0.4 or < 0.4), minor allele frequency (≤ 0.02 or > 0.02), and whether the SNP was shared (i.e. in both reference panels), or private (i.e. only in the SSRP). Minor allele frequency was computed from the pooled reference panels for shared SNPs, or from the SSRP for private SNPs. In addition, each imputed genotype was stratified based on whether the Impute2 results had high certainty (i.e. the maximum probability for a genotype was at least 0.9), or low cer-

Info	SNP type	MAF	genotype certainty	Number of genotypes (%)	Impute2	Impute2 & Kinpute
> 0.4	shared	> 0.02	High	2911525 (59.7)	0.90	**0.93**
> 0.4	shared	> 0.02	Low	469987 (9.6)	0.32	**0.68**
> 0.4	shared	≤ 0.02	High	387932 (7.9)	0.90	**0.94**
> 0.4	shared	≤ 0.02	Low	15460 (0.3)	0.25	**0.70**
> 0.4	private	> 0.02	High	201672 (4.1)	0.89	**0.92**
> 0.4	private	> 0.02	Low	31896 (0.7)	0.35	**0.66**
> 0.4	private	≤ 0.02	High	54182 (1.1)	0.53	**0.66**
> 0.4	private	≤ 0.02	Low	3178 (0.1)	0.07	**0.47**
< 0.4	shared	> 0.02	High	235481 (4.8)	0.63	**0.80**
< 0.4	shared	> 0.02	Low	106855 (2.2)	0.28	**0.73**
< 0.4	shared	≤ 0.02	High	284529 (5.8)	0.61	**0.77**
< 0.4	shared	≤ 0.02	Low	8751 (0.2)	**0.13**	0.0
< 0.4	private	> 0.02	High	28684 (0.6)	0.63	**0.82**
< 0.4	private	> 0.02	Low	7988 (0.2)	0.34	**0.76**
< 0.4	private	≤ 0.02	High	123111 (2.5)	0.45	**0.52**
< 0.4	private	≤ 0.02	Low	5913 (0.1)	0.06	**0.36**

Table 1: Imputed genotype dosage R^2 with the true genotypes

tainty (i.e. the complement of high certainty). Within each stratification bin, we compute imputation quality metrics using all genotypes in that bin. In Table 1 we show the genotype dosage R^2 between the true genotype and the imputed dosages computed from Impute2 and from Kinpute with Impute2 priors. The additional quality metrics, heterozygote sensitivity, heterozygote positive predictive value, and concordance are given in Supplementary Tables 14–16.

In all but one stratification level, where neither method does well, Kinpute improves the imputation quality of Impute2. For genotypes that are of low certainty, in particular, Kinpute can substantially increase R^2, especially for SNPs that have low INFO score.

5 Discussion

We have shown that in the presence of IBD, even when pedigree data is absent, Kinpute can significantly improve the quality of imputed sequence data over using LD based imputation methods alone. Note that Kinpute should be viewed as a supplement, as opposed to an alternative, to LD-based imputation methods such as Impute2, Beagle or minimac. These LD-based methods provide useful information to Kinpute's IBD-based approach in the form of prior genotype probabilities. The IBD approach used by Kinpute is distinct from pedigree-based approaches, which also use IBD. Pedigree approaches (Burdick *et al.*, 2006; Chen and Schaid, 2014; Cheung *et al.*, 2013; Livne *et al.*, 2015) require the pedigree to estimate IBD, often via the construction of inheritance vectors. Thus only sequence data within the pedigree can be used for the IBD-based imputation. When a pedigree has only one or a few individuals who were sequenced, pedigree based imputation will be of limited use. For instance, in the Hutterite data set used here, we were able to assess the imputation accuracy by holding out 48 sequenced individuals while using 50 sequenced individuals as an SSRP. Even though a pedigree is available in the Hutterite population, the subpedigrees that would have been needed for the pedigree based methods (Burdick *et al.*, 2006; Chen and Schaid, 2014; Cheung *et al.*, 2013) would have left too few SSRP individuals in any subpedigree for the methods to have effectively imputed genotypes in the imputation panel. Kinpute, on the other hand, does not need a pedigree and can use IBD from all sequenced individuals, including those that are cryptically related. This is particularly useful in populations where many individuals may be related, but do not have a recorded pedigree.

Kinpute is advantageous in that it can use IBD even from distant or cryptically related individuals to improve LD-based imputation results. However, like any IBD based method, its utility is limited by the amount of IBD that is detectable in the sample. For samples that come from large, outbred populations, where IBD is sparse, every IBD-based method will be of limited use. When imputing a genotype where the information from IBD goes to zero, the genotype probabilities returned from Kinpute will converge to the provided

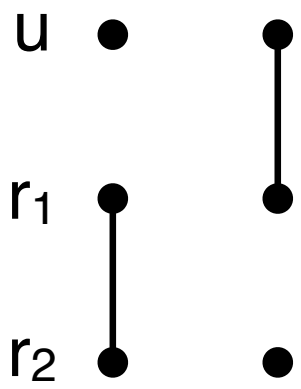

Figure 1: Highly informative IBD configuration. Individual u has IBD 1 with r_1 and IBD 0 with r_2 while r_1 and r_2 have IBD 1. When r_1 has genotype (0, 1) and r_2 has genotype (0, 0), we can immediately infer that u must have at least one copy of the 1 allele.

prior probabilities, typically the LD-based estimates. As samples increase in IBD, Kinpute's performance will improve. Samples from populations such as the Hutterites, where IBD is common, will gain the most. Similar types of gains in imputation can be expected from studies on population isolates, indigenous groups, and endogenous populations. Note that having as high a level of IBD as in the Hutterites is not required to gain from using Kinpute. Much of the imputation gain results from the novel use of IBD configurations. For instance, though an IBD value of 1 between an unknown genotype and a SSRP genotype provides only limited information for imputation, when this IBD occurs within certain configurations, the information for imputation can be dramatically higher. One instance where this arises is when the SSRP individual's genotype is heterozygous (i.e. (0, 1)) and is also IBD = 1 with a homozygous (e.g. (0, 0)) SSRP genotype. If the unknown genotype has an IBD value of 0 with this second SSRP genotype, we can immediately conclude that the unknown genotype must have at least a 1 allele (Figure 1). When the 1 allele is rare, LD based imputation often has trouble imputing the presence of this allele in the unknown genotype. In this situation, even though the amount of IBD is not high, IBD provides significant added value and results in both much higher heterozygote sensitivity (Table 14 in the Supplementary Information) and heterozygote positive predictive value (Table 15 in the Supplementary Information).

Kinpute relies on the results of other computational tools and, as a consequence, the

final results can be dependent on the tools used and their accuracy. When selecting an optimal sample to sequence, Kinpute requires kinship coefficients. These values may differ depending on whether the coefficients are computed from a pedigree or from genotype data, and, when computed from genotype data, what method was used. Different estimates of the kinship coefficient may lead to different samples chosen for the SSRP but do not directly affect the imputation. Insofar as different methods give similar estimates of the kinship coefficients, we expect the selected SSRPs will be similar and the imputation results to be largely unaffected. When performing imputation, Kinpute requires prior probabilities for each genotype. We recommend that the output of LD-based imputation be used, but any set of prior probabilities are allowed. If no prior probabilities are provided, Kinpute will use allele frequencies in the SSRP to compute prior genotype probabilities assuming Hardy-Weinberg equilibrium. Depending on how informative the IBD is at a genotype, the posterior imputed probabilities may be sensitive to the priors used.

As a general rule, the larger the SSRP, the more accurate LD based imputation will be. Similarly, when a sample comes from a population that is close to a population in the standard reference panels, LD based imputation will generally be of high accuracy. However, even in this case, with a reference panel in the tens of thousands, some genotypes, particularly at SNPs with a rare minor allele frequency, are not very well imputed (McCarthy *et al.*, 2016). Approaches that can measurably improve imputation quality, then, can be of great benefit. This is particularly true when budgets limit the number of individuals that can be sequenced and the study population has drifted significantly from those in standard reference panels. When relatedness exists in the sample, whether close, distant or cryptic, Kinpute provides an additional useful tool to maximize the use of the study's sequence data.

6 Funding

This work was supported by the National Institutes of Health [HG02899].

7 Acknowledgements

We thank Dr. Carole Ober for use of the Hutterite data.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74.

Blue, E. M., Cheung, C. Y., Glazner, C. G., Conomos, M. P., Lewis, S. M., Sverdlov, S., Thornton, T., and Wijsman, E. M. (2014). Identity-by-descent graphs offer a flexible framework for imputation and both linkage and association analyses. *BMC Proc*, **8**(Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo), S19.

Browning, B. L. and Browning, S. R. (2016). Genotype imputation with millions of reference samples. *Am J Hum Genet*, **98**(1), 116–26.

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am J Hum Genet*, **103**(3), 338–348.

Burdick, J. T., Chen, W.-M., Abecasis, G. R., and Cheung, V. G. (2006). In silico method for inferring genotypes in pedigrees. *Nat Genet*, **38**(9), 1002–4.

Chen, W. and Schaid, D. J. (2014). PedBLIMP: extending linear predictors to impute genotypes in pedigrees. *Genet Epidemiol*, **38**(6), 531–541.

Cheung, C. Y. K., Thompson, E. A., and Wijsman, E. M. (2013). GIGI: an approach to effective imputation of dense genotypes on large pedigrees. *Am J Hum Genet*, **92**(4), 504–516.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G.,

Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G. R., and Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nat Genet*, **48**(10), 1284–1287.

Deelen, P., Menelaou, A., van Leeuwen, E. M., Kanterakis, A., van Dijk, F., Medina-Gomez, C., Francioli, L. C., Hottenga, J. J., Karssen, L. C., Estrada, K., Kreiner-Møller, E., Rivadeneira, F., van Setten, J., Gutierrez-Achury, J., Westra, H.-J., Franke, L., van Enckevort, D., Dijkstra, M., Byelas, H., van Duijn, C. M., Genome of the Netherlands Consortium, de Bakker, P. I. W., Wijmenga, C., Swertz, M. A., and Genome of the Netherlands Consortium (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.

Fuchsberger, C., Abecasis, G. R., and Hinds, D. A. (2015). minimac2: faster genotype imputation. *Bioinformatics*, **31**(5), 782–784.

Gusev, A., Shah, M. J., Kenny, E. E., Ramachandran, A., Lowe, J. K., Salit, J., Lee, C. C., Levandowsky, E. C., Weaver, T. N., Doan, Q. C., Peckham, H. E., McLaughlin, S. F., Lyons, M. R., Sheth, V. N., Stoffel, M., De La Vega, F. M., Friedman, J. M., Breslow, J. L., and Pe'er, I. (2012). Low-pass genome-wide sequencing and variant inference using identity-by-descent in an isolated human population. *Genetics*, **190**(2), 679–689.

Han, L. and Abney, M. (2011). Identity by descent estimation with dense genome-wide genotype data. *Genet Epidemiol*, **35**(6), 557–567.

Han, L. and Abney, M. (2013). Using identity by descent estimation with dense genotype data to detect positive selection. *Eur J Hum Genet*, **21**, 205–211.

Herzig, A. F., Nutile, T., Babron, M.-C., Ciullo, M., Bellenguez, C., and Leutenegger, A.-L. (2018). Strategies for phasing and imputation in a population isolate. *Genet Epidemiol*, **42**(2), 201–213.

Hou, L., Kember, R. L., Roach, J. C., O'Connell, J. R., Craig, D. W., Bucan, M., Scott, W. K., Pericak-Vance, M., Haines, J. L., Crawford, M. H., Shuldiner, A. R., and McMahon, F. J. (2017). A population-specific reference panel empowers genetic studies of anabaptist populations. *Sci Rep*, **7**(1), 6079.

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

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D. F., Stefansson, H., and Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet*, **40**(9), 1068–1075.

Livne, O. E., Han, L., Alkorta-Aranburu, G., Wentworth-Sheilds, W., Abney, M., Ober, C., and Nicolae, D. L. (2015). PRIMAL: Fast and accurate pedigree-based imputation from sequence data in a founder population. *PLoS Comput Biol*, **11**(3), e1004139.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., Veldink, J., Peters, U., Pato, C., van Duijn, C. M., Gillies, C. E., Gandin, I., Mezzavilla, M., Gilly, A., Cocca, M., Traglia, M., Angius, A., Barrett, J. C., Boomsma, D., Branham, K., Breen, G., Brummett, C. M., Busonero, F., Campbell, H., Chan, A., Chen, S., Chew, E., Collins, F. S., Corbin, L. J., Smith, G. D., Dedoussis, G., Dorr, M., Farmaki, A.-E., Ferrucci, L., Forer, L., Fraser, R. M., Gabriel, S., Levy, S., Groop, L., Harrison, T., Hattersley, A., Holmen, O. L., Hveem, K., Kretzler, M., Lee, J. C., McGue, M., Meitinger, T., Melzer, D., Min, J. L., Mohlke, K. L., Vincent, J. B., Nauck, M., Nickerson, D., Palotie, A., Pato, M., Pirastu, N., McInnis, M., Richards, J. B., Sala, C., Salomaa, V., Schlessinger,

D., Schoenherr, S., Slagboom, P. E., Small, K., Spector, T., Stambolian, D., Tuke, M., Tuomilehto, J., Van den Berg, L. H., Van Rheenen, W., Volker, U., Wijmenga, C., Toniolo, D., Zeggini, E., Gasparini, P., Sampson, M. G., Wilson, J. F., Frayling, T., de Bakker, P. I. W., Swertz, M. A., McCarroll, S., Kooperberg, C., Dekker, A., Altshuler, D., Willer, C., Iacono, W., Ripatti, S., Soranzo, N., Walter, K., Swaroop, A., Cucca, F., Anderson, C. A., Myers, R. M., Boehnke, M., McCarthy, M. I., Durbin, R., and Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*, **48**(10), 1279–1283.

Mitt, M., Kals, M., Pärn, K., Gabriel, S. B., Lander, E. S., Palotie, A., Ripatti, S., Morris, A. P., Metspalu, A., Esko, T., Mägi, R., and 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*, **25**(7), 869–876.

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J. E., Rudan, I., McQuillan, R., Fraser, R. M., Campbell, H., Polasek, O., Asiki, G., Ekoru, K., Hayward, C., Wright, A. F., Vitart, V., Navarro, P., Zagury, J.-F., Wilson, J. F., Toniolo, D., Gasparini, P., Soranzo, N., Sandhu, M. S., and Marchini, J. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet*, **10**(4), e1004234.

Palin, K., Campbell, H., Wright, A. F., Wilson, J. F., and Durbin, R. (2011). Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genet Epidemiol*, **35**(8), 853–860.

Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., Busonero, F., Mulas, A., Zoledziewska, M., Maschio, A., Brennan, C., Lai, S., Miller, M. B., Marcelli, M., Urru, M. F., Pitzalis, M., Lyons, R. H., Kang, H. M., Jones, C. M., Angius, A., Iacono, W. G., Schlessinger, D., McGue, M., Cucca, F., Abecasis, G. R., and 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.

Saad, M. and Wijsman, E. M. (2014). Combining family- and population-based imputation data for association analysis of rare and common variants in large pedigrees. *Genet Epidemiol*, **38**(7), 579–590.

Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziewska, M., Mulas, A., Pistis, G., Steri, M., Danjou, F., Kwong, A., Ortega Del Vecchyo, V. D., Chiang, C. W. K., Bragg-Gresham, J., Pitzalis, M., Nagaraja, R., Tarrier, B., Brennan, C., Uzzau, S., Fuchsberger, C., Atzeni, R., Reinier, F., Berutti, R., Huang, J., Timpson, N. J., Toniolo, D., Gasparini, P., Malerba, G., Dedoussis, G., Zeggini, E., Soranzo, N., Jones, C., Lyons, R., Angius, A., Kang, H. M., Novembre, J., Sanna, S., Schlessinger, D., Cucca, F., and Abecasis, G. R. (2015). Genome sequencing elucidates sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet*, **47**(11), 1272–1281.

Thompson, E. A. (1974). Gene identities and multiple relationships. *Biometrics*, **30**(4), 667–680.

Ullah, E., Mall, R., Abbas, M. M., Kunji, K., Nato, Jr, A. Q., Bensmail, H., Wijsman, E. M., and Saad, M. (2019). Comparison and assessment of family- and population-based genotype imputation methods in large pedigrees. *Genome Res*, **29**(1), 125–134.

Uricchio, L. H., Chong, J. X., Ross, K. D., Ober, C., and 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), 312–9.

Zhou, W., Fritsche, L. G., Das, S., Zhang, H., Nielsen, J. B., Holmen, O. L., Chen, J., Lin, M., Elvestad, M. B., Hveem, K., Abecasis, G. R., Kang, H. M., and Willer, C. J.

(2017). Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels. *Genet Epidemiol*, **41**(8), 744–755.