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

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. Kinpute uses identity by descent information—due to either recent, familial relatedness or distant, unknown ancestors—to alleviate this problem. Kinpute is designed to use a within-study reference panel and works in conjunction with other genotype imputation software to improve the quality of the imputed genotypes. In addition, Kinpute can select an optimal set of study individuals to sequence as the within-study reference panel.

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, a substantial fraction will typically be discarded due to low quality. This is particularly true when the study population is not well represented by available reference panels. For instance, indigenous and founder populations, or even isolated populations of European ancestry, can have significantly improved genotype imputation results when population-specific sequence data are included (Deelen *et al.*, 2014; Hou *et al.*, 2017; Mitt *et al.*, 2017; Pistis *et al.*, 2015; Sidore *et al.*, 2015; Zhou *et al.*, 2017).

In studies where there exists substantial identity-by-descent (IBD) between subjects, for instance due to founder effects or familial relatedness, IBD can increase imputation quality (McCarthy *et al.*, 2016). In the case of families, with a known pedigree, there have been several methods that make use of IBD to perform imputation (Burdick *et al.*, 2006; Chen and

Schaid, 2014; Cheung *et al.*, 2013; Livne *et al.*, 2015). Long-range phasing approaches (Kong *et al.*, 2008; Palin *et al.*, 2011; Uricchio *et al.*, 2012) make use of IBD for imputation, also. Our method, Kinpute, does not require pedigree information, can impute genotypes in IBD regions in both closely and distantly related individuals, does not require phase information (and, hence, is robust to phasing errors), and is novel in that it can leverage the accuracy of standard imputation software (Howie *et al.*, 2012; Fuchsberger *et al.*, 2015; Browning and Browning, 2016) by using their output as prior genotype probabilities. Kinpute requires a within-study sequenced reference panel and can select an optimal set of such individuals. If genotype probabilities from a standard imputation method are not available, Kinpute can use allele frequencies to generate prior genotype probabilities.

3 Methods

The software consists of two components, one which selects an optimal set of individuals to use as a reference panel while leaving the rest as the imputation panel, and another which performs genotype imputation on the imputation panel given sequence data on the reference panel. Both components require prior estimates of IBD sharing. Here, we present a brief description of the two methods. Full details are given in the Supplementary Information.

To select an optimal internal reference panel, we assume knowledge of either kinship coefficients, from a pedigree for instance, or 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. Our method, given a fixed size of the desired reference panel, seeks a division of the sample such that a global measure of the probability that the reference panel individuals are not IBD with the imputation sample is minimized.

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 reference panel.

Our approach to using IBD to enhance imputation is to use the framework set of markers to compute IBD probabilities, across the genome, for a non-sequenced individual with every reference panel individual. In addition, we use IBD probabilities within the reference panel individuals to inform the imputation. Letting G_u be the unknown genotype of a non-sequenced individual u and G_i the known genotype of the ith reference panel individual r_i, we wish to compute $\Pr(G_u|G_1, \ldots, G_N, \{S\})$, where $\{S\}$ is the set of all pairwise IBD states for the set of individuals $u, r_1, \ldots, r_N$. This is a highly computationally challenging problem, so instead we apply Bayes' theorem $\Pr(G_u|G_1, \ldots, G_N, \{S\}) = \Pr(G_1, \ldots, G_N|G_u, \{S\}) \Pr(G_u|\{S\}) / \sum_{G_u} \Pr(G_1, \ldots, G_N|G_u, \{S\}) \Pr(G_u|\{S\})$. This allows us to use the prior probabilities computed from LD methods for $\Pr(G_u|\{S\})$. Details of this computation are given in the Supplementary Information.

4 Results

We applied our method to a human founder population in which 98 individuals were sequenced. Details on the subjects are given in the Supplementary Information. Of the 98 individuals, 50 were selected as an internal reference panel while imputation was done in the remaining 48. The sequence data of the 48 were hidden from imputation and used as the "ground truth" against which to compare the imputed genotypes. IBD probabilities between all pairs of the 98 individuals was computed from genotype array data using the IBDLD software package (Han and Abney, 2011, 2013), and IMPUTE2 was used with the 1000 Genomes data combined with the internal reference panel to compute initial imputed genotype probabilities. These probabilities were used as prior probabilities for our Kinpute

method.

To assess the imputation improvement provided by Kinpute, we stratify the 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 population specific reference panel). Minor allele frequency was computed from the pooled reference panels for shared SNPs, or from the population specific reference panel for private SNPs. In addition, each 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 certainty (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 imputed dosage and the true genotype from both IMPUTE2 and Kinpute. The additional quality metrics, concordance, heterozygote sensitivity, and heterozygote positive predictive value, are given in the Supplementary Information.

In all but one stratification levels, 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 Conclusions

We have shown that in the presence of IBD, Kinpute can significantly improve the quality of imputed sequence data over imputation methods such as Impute2. Note that Kinpute should be viewed as a supplement, as opposed to an alternative, to standard imputation methods such as Impute2, Beagle or minimac. In the absence of IBD, Kinpute returns the supplied prior probabilities. When IBD is present, however, and particularly when the external reference panel does not represent the study population well, Kinpute provides an additional useful tool to maximize the use of a study's sequence data.

Info	SNP type	MAF	genotype certainty	% of genos	Impute2	Kinpute
> 0.4	shared	> 0.02	High	59.7	0.90	**0.93**
> 0.4	shared	> 0.02	Low	9.6	0.32	**0.68**
> 0.4	shared	≤ 0.02	High	7.9	0.90	**0.94**
> 0.4	shared	≤ 0.02	Low	0.3	0.25	**0.70**
> 0.4	private	> 0.02	High	4.1	0.89	**0.92**
> 0.4	private	> 0.02	Low	0.7	0.35	**0.66**
> 0.4	private	≤ 0.02	High	1.1	0.53	**0.66**
> 0.4	private	≤ 0.02	Low	0.1	0.07	**0.47**
< 0.4	shared	> 0.02	High	4.8	0.63	**0.80**
< 0.4	shared	> 0.02	Low	2.2	0.28	**0.73**
< 0.4	shared	≤ 0.02	High	5.8	0.61	**0.77**
< 0.4	shared	≤ 0.02	Low	0.2	**0.13**	0.0
< 0.4	private	> 0.02	High	0.6	0.63	**0.82**
< 0.4	private	> 0.02	Low	0.2	0.34	**0.76**
< 0.4	private	≤ 0.02	High	2.5	0.45	**0.52**
< 0.4	private	≤ 0.02	Low	0.1	0.06	**0.36**

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

6 Acknowledgements

We thank Dr. Carole Ober for allowing access to the Hutterite data set. This work was supported by the National Institutes of Health [HG02899].

References

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

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.

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.

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.

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.

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.

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.

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.