Accurate genotype imputation in multiparental populations from low-coverage sequence

Zheng, C., Boer, M. P., & van Eeuwijk, F. A.

This is a "Post-Print" accepted manuscript, which has been published in "Genetics"

This version is distributed under a non-commercial no derivatives Creative Commons  (CC-BY-NC-ND) user license, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited and not used for commercial purposes. Further, the restriction applies that if you remix, transform, or build upon the material, you may not distribute the modified material.

Please cite this publication as follows:

Zheng, C., Boer, M. P., & van Eeuwijk, F. A. (2018). Accurate genotype imputation in multiparental populations from low-coverage sequence. Genetics, 210(1), 71-82. DOI: 10.1534/genetics.118.300885

You can download the published version at:

https://doi.org/10.1534/genetics.118.300885

Accurate genotype imputation in multiparental populations from low-coverage sequence

Chaozhi Zheng*, Martin P. Boer, and Fred A. van Eeuwijk

Biometris, Wageningen University and Research, Wageningen, The Netherlands

July 24, 2018

Running head: Genotype imputation

Key words: Genotype imputation, genotyping by sequencing, hidden Markov model, cross pollinated (CP) population, multiparent advanced generation intercross (MAGIC)

***Corresponding author**:

Chaozhi Zheng

Biometris

Wageningen University and Research

PO Box 16, 6700AA Wageningen

The Netherlands

Email: chaozhi.zheng@wur.nl

Abstract

Many different types of multiparental populations have recently been produced to increase genetic diversity and resolution in quantitative trait loci (QTL) mapping. Low coverage genotyping by sequencing (GBS) technology has become a cost effective tool in these populations, despite large amounts of missing data in offspring and founders. In this work, we present a general statistical framework for genotype imputation in such experimental crosses from low coverage GBS data. Generalizing a previously developed hidden Markov model for calculating ancestral origins of offspring DNA, we present an imputation algorithm that doesn't require parental data and that is applicable to bi- and multiparental populations. Our imputation algorithm allows heterozygosity of parents and offspring as well as error correction in observed genotypes. Further, our approach can combine imputation and genotype calling from sequencing reads, and it also applies to called genotypes from single nucleotide polymorphism (SNP) array data. We evaluate our imputation algorithm by simulated and real datasets in four different types of populations: the F2, the advanced intercross recombinant inbred lines (AI-RIL), the multiparent advanced generation intercross (MAGIC), and the cross pollinated (CP) population. Because our approach uses marker data and population design information efficiently, the comparisons with previous approaches show that our imputation is accurate at even very low ($< 1\times$) sequencing depth, in addition to having accurate genotype phasing and error detection.

Introduction

Genotype imputation describes the process of imputing missing genotypes in study individuals, most often using a high density reference panel of genotypes. For human populations, HapMap (FRAZER *et al.* 2007) and the 1000 genome project (ALTSHULER *et al.* 2012) provide reference panels including millions of SNPs. Genotype imputation has become a key step in the genome wide association studies of human populations to increase the power of QTL detection and to facilitate meta-analyses of studies at different sets of SNPs (LI and FREUDENBERG 2009; MARCHINI and HOWIE 2010).

Genotype imputation leverages haplotype sharing between study individuals and reference panels. Along chromosomes, the pattern of haplotype sharing changes due to historical recombination. A crucial component of most genotype imputation methods is to infer the local haplotype clustering and the ancestral haplotypes from reference panels and study individuals (HOWIE *et al.* 2009; LI *et al.* 2010; BROWNING and BROWNING 2016). The accuracy of imputation depends on how well reference panels match study individuals in terms of ancestral haplotypes (PEI *et al.* 2008; ROSHYARA *et al.* 2016).

Next-generation sequencing technology has become an attractive and cost effective tool for QTL mapping in non-human populations (SPINDEL *et al.* 2013; HEFFELFINGER *et al.* 2014; KIM *et al.* 2016), and genotype imputation is essential for low coverage sequencing. The focus of this paper is on experimentally designed populations, particularly for plants, where study individuals are produced by multi-generation crossing from two or more founders. Many such multiparental populations have recently been created (e.g. KOVER *et al.* 2009; BANDILLO *et al.* 2013; MACKAY *et al.* 2014; SANNEMANN *et al.* 2015), aiming at increasing genetic diversity due to many founders and QTL mapping resolution due to accumulated recombination breakpoints over multiple generations.

The founders of multiparental populations are naturally used as the reference panel for genotype imputation. However, there are typically many missing founder genotypes particularly when both founders and offspring are genotyped by low coverage sequencing, and some of the founders may even be missing completely (THEPOT *et al.* 2015). In such cases, the population-based imputation methods (HOWIE *et al.* 2009; LI *et al.* 2010; BROWNING and BROWNING

2016) are not optimal. Alternatively, pedigree-based genotype imputation methods (ABECASIS *et al.* 2002; CHEUNG *et al.* 2013) are computationally intensive if not impossible, because of the large breeding pedigree being often partially or wholly unavailable, and most or all genotypes being missing in intermediate generations.

Recently, several imputation methods were proposed for experimental crosses. XIE *et al.* (2010) described a parent-independent genotyping method for two-way recombinant inbred lines (RILs), where parental genotypes were obtained using a maximum parsimony of recombination. SWARTS *et al.* (2014) described a Full-Sib Family Haplotype Imputation (FSFHap) method for biparental populations, where parental haplotypes were identified by a custom clustering method over non-overlapping windows with a window size of 50 loci along chromosomes. FRAGOSO *et al.* (2016) described a Low-Coverage Biallelic Impute (LB-Impute) algorithm for biparental populations, where parental genotypes were imputed only after offspring genotypes were imputed using a modified Viterbi algorithm over a sliding window (of size 7 loci) along chromosomes. See also HICKEY *et al.* (2015) for genotype imputation in biparental populations in plant breeding.

In experimental crosses, genotype imputation methods have mainly focused on biparental populations. There remain challenges for more complicated experimental designs. HUANG *et al.* (2014) described a genotype imputation method called mpimpute, which is however restricted to the funnel scheme 4- or 8-way RILs. In the funnel scheme, the founders of each line are randomly permuted. In this paper, we present a general statistical framework of genotype imputation from low coverage GBS data, applicable to many scenarios in experimental crosses. First, it applies to both bi- and multiparental populations. Second, it is parent-independent so that it applies even if some founders' genotypes are not available. Third, it integrates with parental phasing and thus applies to mapping populations with outbred founders. Last but not least, it integrates with genotype calling to account for the uncertainties in identifying heterozygous genotypes due to low read numbers.

Our imputation algorithm is called magicImpute, building on a hidden Markov model (HMM) framework that extends our previous work (ZHENG *et al.* 2014; ZHENG 2015; ZHENG *et al.* 2015, 2018). We first evaluate magicImpute with simulated data in four populations: the F2,

the AI-RIL, the funnel scheme 8-way RILs, and the CP. Then we analyze four sets of real data: the maize F2 (ELSHIRE *et al.* 2011), the maize AI-RIL (HEFFELFINGER *et al.* 2014), the rice MAGIC (BANDILLO *et al.* 2013), and the apple CP (GARDNER *et al.* 2014). The term MAGIC has been used for many different types of breeding designs, and the rice MAGIC is essentially a set of funnel scheme 8-way RILs (BANDILLO *et al.* 2013). In the evaluations by simulation and real data, we perform comparisons among magicImpute, Beagle v4.1 (BROWNING and BROWNING 2016), LB-impute (FRAGOSO *et al.* 2016) and mpimpute (HUANG *et al.* 2014), investigating, among other things, how imputation quality depends on amount of missing data, level of homozygosity and coverage of sequencing.

Methods

Overview of model

Consider a mapping population derived from a number $n_F \geq 2$ of founders. We assume that linkage groups (chromosomes) are independent, and thus consider only one group. The genotypic data matrix of sampled offspring is denoted by $\boldsymbol{y}^O = \{y_{ti}\}_{t=1...T,i=1...N}$, with element y_{ti} representing the genotype at locus t in offspring i . The founder genotype matrix is denoted by $\boldsymbol{y}^F = \{y_t^F\}_{t=1...T}$, with element y_t^F being the genotypes at locus t in all founders. We consider only bi-allelic markers, and denote the two alleles by 1 and 2. We model either the called genotypes from SNP array or GBS data, or the allelic depths of GBS data. The called unphased genotype at a locus can take one of six possible values: 11, 12, 22, $1U$, $2U$, or UU, where U denotes an uncertainty allele. For allelic depth data, the genotype is measured by read counts for each of two alleles. The ordering and genetic locations of markers are assumed to be known.

We build an integrated hidden Markov model for the genotypic data $\boldsymbol{y}^O$ and $\boldsymbol{y}^F$, but impute missing founder genotypes and missing offspring genotypes separately. The imputation diagram and the overview of the HMM are shown in Figure 1. Here the hidden founder haplotype matrix $\boldsymbol{h}^F = \{h_t^F\}_{t=1...T}$, where element h_t^F is similar to y_t^F except that it contains information on missing genotypes and genotype phases at locus t in founders. See an example in the following section on the genotype model. Conditional on estimated $\hat{\boldsymbol{h}}^F$, the genotypic data for each

offspring are analyzed independently by a sub-HMM, with x_{ti} being the hidden ancestral origin state at locus t in offspring i. The hidden Markov model will be further explained in the process model. See Table 1 for a list of symbols and their brief explanations.

The genotype model

Called genotype: The genotype model corresponds to the vertical relationships (arrows) in the directed acyclic graph of the HMM (Figure 1). Since the genotypes are independent conditional on the hidden states, we consider a single locus t. We first model the prior probability $P(h_t^F|y_t^F)$, which is assumed to follow a discrete uniform distribution over all possible combinations under the constraint of called parental genotypes y_t^F. Consider an example of four inbred founders with genotypes at locus t denoted by 11, 22, UU, and UU, respectively. We use $12UU$ as a shorthand for the four homozygous genotypes. Then h_t^F can take one of four possible values 1211, 1212, 1221, 1222 with equal probability. Consider the second example of a cross pollinated population, and the genotypes of two outbred parents are denoted by 12 and UU. Then h_t^F can take one of eight possible values 1211, 1212, 1221, 1222, 2111, 2112, 2121, and 2122, where the last four values account for the alternative phase of the first parent's genotype. The founder haplotype matrix $\boldsymbol{h}^F$ is known if all parental genotypes are observed and phased.

The hidden founder haplotype h_t^F is not the true founder haplotype, and it accounts for unknown phasing and missing values of called founder genotypes, but not allelic errors. The errors in called genotypes can be accounted for in the likelihood $l_{ti} = P(y_{ti}|h_t^F, x_{ti}, \epsilon_O, \epsilon_F)$ at locus t in offspring i, where ϵ_O and ϵ_F are the allelic error probabilities for offspring and founders, respectively. The calculation of likelihood l_{ti} has been described in detail in ZHENG *et al.* (2015). We describe it briefly as follows. We calculate l_{ti} by summing over the hidden true genotype z_{ti}, and it holds that

$$l_{ti} = \sum_{\boldsymbol{z_{ti}}} P(y_{ti}|z_{ti}, \epsilon_O) P(z_{ti}|d_{ti}, x_{ti}, \epsilon_F),$$

$$P(z_{ti}|d_{ti}, x_{ti}, \epsilon_F) \propto P(d_{ti}|z_{ti}, x_{ti}, \epsilon_F) P(z_{ti}|x_{ti}),$$

where d_{ti} denotes the derived genotype that is obtained from x_{ti} and h_t^F in a deterministic way. We assign an uninformative prior to $P(z_{ti}|x_{ti})$, and calculate $P(y_{ti}|z_{ti}, \epsilon_O)$ and $P(d_{ti}|z_{ti}, x_{ti}, \epsilon_F)$, assuming that typing errors occur independently and the observed allele is the alternative one if an error occurs with probability ϵ_O or ϵ_F. Here the derived genotype d_{ti} is the same as true genotype z_{ti} if there are no errors in observed founder genotypes ($\epsilon_F = 0$).

Allelic depth: We next consider the case that genotypes are represented by allelic depths of GBS data. We calculate prior probability $P(h_t^F|y_t^F)$ with y_t^F being called from founder allelic depths, where the genotype calling will be described in the next section. For likelihood l_{ti} at locus t in offspring i, only the calculation of $P(y_{ti}|z_{ti}, \epsilon_O)$ is different from the case of called genotypes. We introduce ε as the sequencing error probability that is given by $\varepsilon = 10^{-phred/10}$, where $phred$ is Phred quality score. The genotype y_{ti} is represented by (r_1, r_2), the number of reads for alleles 1 and 2, respectively. It holds that

$$\begin{aligned} P((r_1, r_2)|z' = 11, \varepsilon) &\propto (1-\varepsilon)^{r_1}\varepsilon^{r_2}, \\ P((r_1, r_2)|z' = 12, \varepsilon) &\propto (1/2)^{r_1+r_2}, \\ P((r_1, r_2)|z' = 21, \varepsilon) &\propto (1/2)^{r_1+r_2}, \\ P((r_1, r_2)|z' = 22, \varepsilon) &\propto \varepsilon^{r_1}(1-\varepsilon)^{r_2}, \end{aligned} \tag{1}$$

conditional on hidden genotype z' (XIE *et al.* 2010).

We interpret ϵ_O as a depth-independence allelic error probability, for example, due to the mis-assignment of reads to the reference genome. And we assume that z' results from the true genotype z_{ti} with error probability ϵ_O. Thus, $P(y_{ti}|z_{ti}, \epsilon_O, \varepsilon)$ can be calculated by summing over z' as follows

$$P(y_{ti} = (r_1, r_2)|z_{ti}, \epsilon_O, \varepsilon) = \sum_{z'} P((r_1, r_2)|z', \varepsilon)P(z'|z_{ti}, \epsilon_O)$$

where $P(z'|z_{ti}, \epsilon_O)$ is similar to $P(y_{ti}|z_{ti}, \epsilon_O)$ in the case of called genotypes, except that z' is phased. Specifically for $z_{ti} = 11$, we have $P(z'|z_{ti} = 11, \epsilon_O) = (1-\epsilon_O)^2$, $(1-\epsilon_O)\epsilon_O$, $\epsilon_O(1-\epsilon_O)$, and ϵ_O^2 for $z' = 11$, 12, 21, and 22, respectively. And similarly for $z_{ti} = 12$, 21, and 22. When there are no ambiguities, we suppress the dependence of ε for allelic depth data

in the description of the imputation algorithm.

Single genotype calling: We perform single genotype calling for founder allelic depths of GBS data before imputation, and for detecting potential erroneous genotypes among offspring during the last stage of imputation. For single genotype calling from allelic depths, we do not consider depth-independence errors. The calling is based on the following posterior probability

$$P(z_{ti}|y_{ti} = (r_1, r_2), \varepsilon) \propto P(y_{ti}|z_{ti}, \varepsilon)P(z_{ti}),$$

where $P(y_{ti}|z_{ti}, \varepsilon)$ is given by Equation 1 and $P(z_{ti}) = 1/4$, $1/2$, and $1/4$ for $z_{ti} = 11$, 12, and 22, respectively. Note that z_{ti} is unphased only in case of single genotype calling, and it is phased elsewhere. The genotype with posterior probability being greater than threshold P_{call} is called. If no genotype is called, we calculate the posterior probability

$$\begin{aligned} P(z_{ti} = 1U|y_{ti}, \varepsilon) =& P(z_{ti} = 11|y_{ti}, \varepsilon) + P(z_{ti} = 12|y_{ti}, \varepsilon), \\ P(z_{ti} = 2U|y_{ti}, \varepsilon) =& P(z_{ti} = 22|y_{ti}, \varepsilon) + P(z_{ti} = 12|y_{ti}, \varepsilon). \end{aligned}$$

The genotype $1U$ is called if $P(z_{ti} = 1U|y_{ti}, \varepsilon) > P_{call}$ and $P(z_{ti} = 1U|y_{ti}, \varepsilon) > P(z_{ti} = 2U|y_{ti}, \varepsilon)$, and similarly for genotype $2U$. The genotype is set to UU if no calling occurs.

The process model

The process model corresponds to the horizontal relationships (arrows) in the directed acyclic graph of the HMM (Figure 1). It has been described in detail (ZHENG *et al.* 2014; ZHENG 2015; ZHENG *et al.* 2015), and we give a brief summary in the following. The process $\{x_{ti}\}_{t=1}^{T}$ for offspring i describes how the ancestral origins change along chromosomes. At a locus t, let $x_{ti} = (x_{ti}^m, x_{ti}^p)$ be the ancestral origins on the maternally (m) and paternally (p) derived chromosomes. If offspring i is fully inbred, we have $x_{ti}^m = x_{ti}^p$ so that the ancestral origin process along the maternally derived chromosome is the same as the process along the paternally derived chromosome, and it is thus termed "depModel". On the other hand, if offspring i is completely outbred, the ancestral origin process along the maternally derived chromosome $\{x_{ti}^m\}_{t=1}^{T}$ is independent of the process $\{x_{ti}^p\}_{t=1}^{T}$ along the paternally derived chromosome, and

it is therefore termed "indepModel". In the general model called "jointModel", x_{ti}^m and x_{ti}^p are modeled jointly. We have kept the model terms (e.g. "jointModel") consistent with ZHENG *et al.* (2015).

In all three models, the ancestral origin process along two chromosomes is assumed to follow a Markov process, so that the ancestral origins x_{ti} at locus t depends only on $x_{t-1,i}$ at locus $t-1$ but not on the previous $\{x_{t',i}\}_{t'=1}^{t-2}$. Thus, the joint prior distribution of $\{x_{ti}\}_{t=1}^{T}$ can be specified by the initial distribution $\pi(x_{1i})$ and the transition probability $P(x_{ti}|x_{t-1,i})$ at $t = 2, ..., T$. The initial distribution $\pi(x_{1i})$ is specified by the stationary distribution of the Markov process, so that the prior process model does not depend on the direction of chromosomes. The initial distribution $\pi(x_{1i})$ and transition probability $P(x_{ti}|x_{t-1,i})$ can be specified from the breeding design of a mapping population, that is, how the sampled offspring is produced from the founders; the transition probability also depends on inter-marker distances. See ZHENG *et al.* (2014), ZHENG (2015), and ZHENG *et al.* (2018) for the details of calculating $\pi(x_{1i})$ and $P(x_{ti}|x_{t-1,i})$ under various breeding designs.

Founder imputation

Because the state space of the HMM exponentially increases with the number N of sampled offspring, the exact inference of the founder haplotype matrix $\boldsymbol{h}^F$ is computationally intractable, even using the forward-backward algorithm (RABINER 1989). In the following, we describe an approximate forward-backward procedure for maximum likelihood estimation of $\boldsymbol{h}^F$. Our forward algorithm calculates recursively the posterior probabilities $\gamma(h_t^F)$ and $\alpha(x_{ti}|h_t^F)$ for offspring $i = 1, ..., N$, conditional on genotypic data up to locus t. It proceeds as follows:

A0 Initialize at $t = 1$

$$\begin{aligned}
\tilde{\alpha}(x_{1i}|h_1^F) =& P(y_{1i}|h_1^F, x_{1i}, \epsilon_O, \epsilon_F)\pi(x_{1i}),\\
\gamma(h_1^F) \propto& P(h_1^F|y_1^F)\prod_{i=1}^{N}\sum_{x_{1i}}\tilde{\alpha}(x_{1i}|h_1^F),\\
\alpha(x_{1i}|h_1^F) =& \tilde{\alpha}(x_{1i}|h_1^F)/\sum_{x_{1i}}\tilde{\alpha}(x_{1i}|h_1^F).
\end{aligned}$$

A1 For $t = 2, ..., T$

$$\begin{aligned}
\tilde{\alpha}(x_{ti}|h_t^F) =& P(y_{ti}|h_t^F, x_{ti}, \epsilon_O, \epsilon_F) \sum_{x_{t-1,i}} P(x_{ti}|x_{t-1,i}) \sum_{h_{t-1}^F} \gamma(h_{t-1}^F)\alpha(x_{t-1,i}|h_{t-1}^F), \\
\gamma(h_t^F) \propto& P(h_t^F|y_t^F) \prod_{i=1}^{N} \sum_{x_{ti}} \tilde{\alpha}(x_{ti}|h_t^F), \\
\alpha(x_{ti}|h_t^F) =& \tilde{\alpha}(x_{ti}|h_t^F) / \sum_{x_{ti}} \tilde{\alpha}(x_{ti}|h_t^F),
\end{aligned}$$

where $\tilde{\alpha}(x_{ti}|h_t^F)$ is an unnormalized probability, and the normalization constant for $\gamma(h_t^F)$ is not shown. The key approximation comes from the independence of offspring in the calculation of $\gamma(h_t^F)$. ZHENG *et al.* (2016) have described a similar forward algorithm for haplotype reconstruction in tetraploid populations.

The maximum likelihood estimation of founder haplotypes is based on the posterior probabilities $\alpha(x_{ti}|h_t^F)$ and $\gamma(h_t^F)$ from algorithm A. The maximization proceeds backwardly as follows:

B0 Initialize at $t = T$: $\hat{h}_T^F = \texttt{argmax}\ \gamma(h_T^F)$ and $\hat{x}_{T,i} = \texttt{argmax}\ \alpha(x_{T,i}|h_T^F)$ for $i = 1, ..., N$.

B1 For $t = T - 1, ..., 1$

$$\begin{aligned}
\beta(x_{ti}|h_t^F) =& \alpha(x_{ti}|h_t^F) P(\hat{x}_{t+1,i}|x_{ti}), \\
\hat{h}_t^F =& \texttt{argmax}\ \gamma(h_t^F) \prod_{i=1}^{N} \sum_{x_{ti}} \beta(x_{ti}|h_t^F), \\
\hat{x}_{ti} =& \texttt{argmax}\ \beta(x_{ti}|\hat{h}_t^F).
\end{aligned}$$

It is possible that multiple argument values correspond to the same maximum. If such ties occur, we randomly choose one of these values. FRIEL and RUE (2007) have described a similar backward maximization algorithm for general factorisable models.

Preliminary simulations showed that our forward-backward procedure is occasionally less accurate on the left end of chromosomes in case of sparse data. We overcome this problem by two rounds of maximization. Specifically, we fix the founder haplotypes on the right-half chromosomes ($t > T/2$) after the first round of maximization, and then perform the second

round with reversed chromosome direction.

Offspring imputation

Conditional on the imputed founder haplotype matrix $\hat{\boldsymbol{h}}^F$, all the offspring are independent. For each offspring, we first perform the posterior decoding algorithm to calculate the posterior probabilities of ancestral origins at all loci (RABINER 1989; ZHENG *et al.* 2015). Then we calculate the posterior probabilities of true genotypes, from which missing genotypes can be imputed.

We obtain $P(z_{ti}|\boldsymbol{y}^O, \hat{\boldsymbol{h}}^F, \epsilon_O, \epsilon_F)$ by marginalizing the following joint posterior probability

$$P(z_{ti}, x_{ti}|\boldsymbol{y}^O, \hat{\boldsymbol{h}}^F, \epsilon_O, \epsilon_F) = P(z_{ti}|d_{ti}, x_{ti}, \epsilon_F)P(x_{ti}|\boldsymbol{y}^O, \hat{\boldsymbol{h}}^F, \epsilon_O, \epsilon_F),$$

where the posterior probability $P(x_{ti}|\boldsymbol{y}^O, \hat{\boldsymbol{h}}^F, \epsilon_O, \epsilon_F)$ can be calculated by the function magicReconstruct in the RABBIT software (ZHENG *et al.* 2015), which has been extended to analyze allelic depths of GBS data. Here the derived genotype d_{ti} is completely determined by x_{ti} and $\hat{h}_t^F$, and the calculation of $P(z_{ti}|d_{ti}, x_{ti}, \epsilon_F)$ has been described in the genotype model.

From the marginal posterior probability $P(z_{ti}|\boldsymbol{y}^O, \hat{\boldsymbol{h}}^F, \epsilon_O, \epsilon_F)$, we perform both imputation and error detection for offspring i. For imputation, the missing genotype in offspring i at locus t is imputed to be $\hat{z}_{ti}$ if its marginal posterior probability is larger than a given threshold P_{impute}. For error detection, the observed called genotype y_{ti} is corrected if the most probable genotype is different from y_{ti} and the maximal marginal posterior probability is larger than a given threshold P_{detect}.

Data simulation

We simulate sequence data, mimicking real data in the following mapping populations: the AI-RIL, the F2, the MAGIC(funnel scheme 8-way RIL), and the CP. These populations differ in the number of founders and the heterozygosity level of founders and offspring (Table 2). For each type of mapping population, we simulate independently three sample sizes: 100, 200, and 500, that is, the number of sampled offspring in the last generation. Independently for each type

of population with a given sample size, we first simulate the breeding pedigree according to the corresponding real data. The AI-RIL consists of five generations of random mating starting from the F1 generation and six generations of selfing; the size of the random mating population is set to 1000. For each offspring of the MAGIC, the founders are randomly permuted so that the number of funnels equals the sample size.

Given a breeding pedigree for each mapping population, we assign a unique founder genome label (FGL) to each inbred founder or to the haploid gamete of each outbred founder. We simulate only one linkage group. Each offspring gamete is a random mosaic of FGL blocks determined by chromosomal crossovers between two parental chromosomes. The number of crossovers in a gamete follows a Poisson distribution with mean being the chromosome length in Morgan, and the positions of crossovers are uniformly distributed across the chromosome. We set true founder haplotypes based on the founders imputed from the available real data (see Table 2), and obtain the true offspring genotypes by replacing FGLs with the true founder haplotypes. We apply the same error model to the true founder haplotypes with $\epsilon_F = 0.005$ and to the true offspring genotypes with $\epsilon_O = 0.005$.

We simulate read count data for each obtained founder or offspring genotype. Independently for each allele of a genotype, the number of reads is assumed to follow an exponential distribution with mean being $\lambda/2$, where we set $\lambda = 8$; the number of erroneous reads follow a binomial distribution with probability $\varepsilon = 0.001$, and the erroneous read corresponds to the alternative allele. The allelic depths of genotypes are obtained by combining reads of the two alleles. The allelic depths of founder and offspring genotypes are re-set to be missing with probabilities 0.25 and 0.15, respectively. We obtain 12 full datasets, 3 population sizes for each of the four mapping populations, with average offspring read depth 6.8. To study the dependence of sequencing coverage, we retain the same founder reads and randomly sample offspring reads with probability 2^{-i} for $i = 0, 1, ..., 10$, resulting in a total of 132 test datasets.

Real data

Table 2 shows a summary of real data after filtering. For the maize AI-RIL (HEFFELFINGER *et al.* 2014) and the maize F2 (ELSHIRE *et al.* 2011), we use the GBS data that have been

prepared by FRAGOSO *et al.* (2016) as the input data of LB-Impute. For the rice MAGIC (BANDILLO *et al.* 2013), we use the called genotypes that have been prepared by HUANG *et al.* (2014) for mpimpute. For the apple CP (GARDNER *et al.* 2014), we filter the original allelic depth data by removing markers with the missing fraction of called genotypes larger than 50%, and removing markers with segregation distortion at significant level 0.01. During the filtering process, a single genotype is called with threshold $P_{call} = 0.99$ and 0.95 for founders and offspring, respectively, as described in the previous section on single genotype calling. And the quality score is set to $phred = 30$ so that the sequencing error probability $\varepsilon = 10^{-phred/10} = 0.001$.

To calculate imputation accuracy, we mask a subset of high-confidence genotypes and use them as the pseudo-true genotypes. For the GBS data, the genotypes are first called with a very large threshold $P_{call} = 0.9999$ and the quality scores being 30 and 40 for apple and maize, respectively. The called genotypes (excluding UU, $1U$ and $2U$) are masked with probability being 0.25 and 0.05 for founders and offspring, respectively. After masking, the fractions of founder genotypes without reads are 0.23, 0.24, and 0.19 for the maize AI-RIL, the maize F2, and the apple CP, respectively. And the fractions of offspring genotypes without reads are 0.77, 0.16, and 0.095. For each of three masked full datasets, we retain the same founder reads and randomly sample offspring reads with probability 2^{-i} for $i = 0, 1, ..., 10$, resulting in 33 real sequencing datasets. For the called genotypes of the rice MAGIC, the missing fraction of founder genotypes after masking is 0.3. From this masked dataset, five datasets are produced independently by masking called offspring genotypes to give missing fractions from 0.5 to 0.9 at step size 0.1.

Algorithm evaluation

To set up the algorithm magicImpute, we perform sensitivity analysis of P_{impute}, P_{detect}, and ε_O. For each mapping population with size 200 and read depth 0.85, we impute the simulated dataset with the input data being called genotypes and the first two founders' genotypes being not available. By default, we set $\varepsilon_F = 0.005$, and the input genotypes are called from allelic depths with threshold $P_{call} = 0.99$ and 0.95 for founders and offspring, respectively. Figures

S1&S2 show that the accuracies of imputation and error detection increase slightly with P_{impute} from 0.6 to 0.95, while the fractions of imputation and error detection decrease slightly. Figures S1&S2 also show that the performances of imputation and error detection often become a bit worse when ε_O increases by a factor of 10. The effects of these parameters are marginal in general. Thus we set somewhat arbitrarily $P_{impute} = 0.9$, $P_{detect} = 0.9$, and $\varepsilon_O = 0.005$ in the following evaluations. The algorithm magicImpute also outputs the posterior probabilities of all possible genotypes for all offspring at all markers, from which we can perform imputation and error detection with different P_{impute} and P_{detect}.

We evaluate magicImpute by both simulated and real data in the four types of mapping populations. For each of the simulated datasets and the real GBS datasets, we run magicImpute in the four combinations: the first two founders' genotypes are available or not, and the input data are allelic depths or called genotypes. Here the quality scores are 30 for the simulated data and the real maize GBS data, and 40 for the real apple GBS data. For the real rice data, we run magicImpute in the two combinations: the first two founders' genotypes are available or not. Results of magicImpute are compared with those of Beagle v4.1 in all populations. We run Beagle v4.1 for the called genotypes in two ways: without reference panels and use the founder haplotypes imputed by magicImpute as the reference panels. Additionally, we run LB-Impute for the biparental populations AI-RIL and F2 with the input data being allelic depths, and run mpimpute for the MAGIC population with the input data being called genotypes. LB-Impute and mpimpute do not work if some founders' genotypes are not available. The running settings of magicImpute, Beagle v4.1, LB-Impute, and mpimpute are described in Supporting Information, File S1. See SWARTS *et al.* (2014) and FRAGOSO *et al.* (2016) for comparisons of FSFHap with Beagle and LB-Impute.

Data availability

The algorithm magicImpute is implemented in Mathematica 11.0 (WOLFRAM RESEARCH 2016), and it has been included as a function in the RABBIT software. RABBIT is available at `https://github.com/chaozhi/RABBIT.git`, and it is offered under the GNU Affero general public license, version 3 (AGPL-3.0). Example scripts for simulating genotypic

data are included. The real maize AI-RIL and F2 data have been described by HEFFELFINGER *et al.* (2014) and ELSHIRE *et al.* (2011), respectively, and they have been prepared by FRAGOSO *et al.* (2016) for LB-Impute. The rice MAGIC data have been described by BANDILLO *et al.* (2013), and they have been prepared by HUANG *et al.* (2014) for mpimpute. The apple CP data are available from GARDNER *et al.* (2014).

Results

Simulation evaluation

Figures 2-4 and Figures S3-S7 show the comparisons among magicImpute, Beagle, LB-Impute, and mpimpute in terms of imputation accuracy, error detection, and genotype phasing. All results are obtained from the simulated populations of size 200, except Figure S4 that shows the effects of population size.

Imputation accuracy: Figures 2 and S3 show the comparisons of imputation accuracy. One of the most striking patterns is that there exist break points for magicImpute and Beagle but not for LB-Impute and mpimpute. As shown in Figure 2 for the imputation accuracy of offspring genotypes, the break points of magicImpute are 0.053, 0.11, 0.21, and 0.21 read depth for the AI-RIL, the F2, the MAGIC, and the CP, respectively, much lower than the break points of 0.42, 3.4, 0.85, and 3.4 read depth for Beagle. As shown in the left panels of Figure S3, the break points of magicImpute for founder imputation are the same as those for offspring imputation; Beagle does not impute founder genotypes.

As for mpimpute and LB-Impute, they perform slightly worse than magicImpute. The imputation accuracy of mpimpute is $\sim 1.7\%$ lower than that of magicImpute when read depth $>$ 0.21 (Figure 2C). The imputation accuracies of LB-Impute at the highest read depth are similar to those of magicImpute, but they decrease gradually with decreasing read depth. In addition, the imputation fractions of LB-Impute at the highest read depth are around 0.8, much smaller than those of magicImpute (Figure S3B&D).

The unavailability of the first two founders' genotypes has no noticeable effects on the performance of magicImpute for the AI-RIL, the F2, and the MAGIC, as long as read depth is

higher than the break point. However for the CP, the availability of the two outbred founders' genotypes results in $\sim 2\%$ lower accuracy of imputing founder genotypes (Figure S3G), due to the calling errors in the available founder genotypes. As a result, the imputation accuracy of offspring genotypes is $\sim 4\%$ lower (Figure 2D).

Whether the input data are allelic depths or called genotypes has little influence on the performance of magicImpute. However for the almost homozygous populations AI-RIL and MAGIC, the ceiling limit of imputation accuracy decreases with increasing read depth instead of leveling off (Figure 2A&C). This is due to the assumption of homozygosity during the prior genotype calling, and the information on residual heterozygosity is lost after transforming allelic depths into called genotypes. The percentage of heterozygotes among missing genotypes increases with increasing read depth, and they are always missing and wrongly imputed.

Figure S4 shows that the main effect of population size is shifting the break points of the imputation accuracy obtained by magicImpute and Beagle.

Error detection: We evaluate the error detection of magicImpute in the case of the input data being called genotypes. A suspicious genotype error is detected by magicImpute when the most probable true genotype is different from the input called genotype and the maximum posterior probability is larger than the default threshold $P_{detect} = 0.9$. As shown in Figures 3 and S5, the unavailability of the first two founders' genotypes greatly improve the error detections for the F2, the CP, and the AI-RIL, but it has little effects on the MAGIC with multiple founders. This indicates that the errors in the available founder genotypes adversely affect the detection of offspring genotypes.

Figures 3 and S5 show that the error detection in the almost homozygous populations AI-RIL and the MAGIC is much worse than in the F2 and the CP. This is due to the homozygosity assumption under which the input genotypes are being called for the AI-RIL and the MAGIC; most offspring genotype errors are heterozygous and they cannot be detected and corrected when the heterozygosity information is lost during the prior genotype calling. Figure S6 shows that the error detection in the AI-RIL and the MAGIC is much better when homozygosity is not assumed.

Genotype phasing: We evaluate the phasing accuracy for the heterozygous populations F2 and

CP obtained by magicImpute and Beagle; mpimpute and LB-impute do not perform phasing. The phasing accuracy is measured in two ways: the switch accuracy is defined as one minus the number of switches divided by the number of opportunities for switch error, and the heterozygous accuracy denotes the percentage of correctly phased heterozygous genotypes. A switch error occurs if the heterozygous genotype at a site has phase switched related to that of the previous heterozygous site.

As shown in Figures 4 and S7, the phasing accuracy has similar patterns and the same break points as those of the imputation accuracy (Figure 2) for magicImpute and Beagle, so that the phasing of magicImpute is more robust to missing data. For the CP, the switch accuracy and the heterozygous accuracy of magicImpute are close to 1 when read depth is higher than the break point, whereas the heterozygous accuracy of Beagle is less than 0.8. The difference between switch and heterozygous accuracy indicates that the wrongly phased heterozygous genotypes occur in blocks and they could be corrected by a few switches between the two haplotypes within an offspring.

Figures 4 and S7 show that the availability of the two founders' genotypes are unimportant to genotype phasing. The phasing accuracy of Beagle increases slightly when read depth is higher than the break point. However for magicImpute in the CP, the ceiling limit of phasing accuracy decreases a bit, consistent with the decrease of ceiling imputation accuracy because of the errors in the available founder genotypes.

Evaluation by real data

Figures 5 and S8 show the results of genotype imputation obtained from the real data in the four mapping populations. Error detection and genotype phasing cannot be evaluated since true genotypes and phases are not available; the imputation accuracy is calculated based on masked genotypes. Figure 5 shows the patterns similar to those of the simulation evaluation. The break points for magicImpute are at much lower read depths or larger missing fractions than those of Beagle. The magicImpute accuracy is slightly larger than that of mpimpute, and it is always high until the break point. In contrast to that, the LB-Impute accuracy decreases gradually with read depth.

Maize AI-RIL and F2: Figure 5A&B and Figure S8A-D show the results of genotype imputation in the real biparental populations AI-RIL and F2. For magicImpute, the offspring imputation accuracies at the highest read depth are higher than 0.980 in the AI-RIL and 0.987 in the F2. The corresponding accuracies are 0.970 and 0.986 for Beagle, whereas they are 0.917 and 0.986 for LB-Impute. The imputation fractions at the highest read depth for both magicImpute and Beagle are larger than 0.960, whereas for LB-Impute they are 0.720 in the AI-RIL and 0.906 in the F2.

FRAGOSO *et al.* (2016) obtained the imputation accuracies 0.970 for the AI-RIL and 0.946 for the F2, and the differences may be due to the masking of founder genotypes and the usage of a small genotype error probability for magicImpute.

Rice MAGIC: Figure 5C shows that the imputation accuracies of magicImpute and mpimpute are almost independent of missing fraction of the input offspring genotypes in the range from 0.5 to 0.9. On average, the offspring imputation accuracy of magicImpute is higher than that of mpimpute by 2.5%. The Beagle imputation accuracy is comparable to that of magicImpute when the missing fraction is no greater than the break point of 0.7.

Figure S8E shows that the founder imputation accuracies are around 0.94 and 0.89 for mpimpute and magicImpute, respectively, whereas they are close to 1 in the simulation evaluation. The imputation fraction of founder genotypes for mpimpute gradually decreases from 0.947 to 0.922 with increasing missing fraction (Figure S8E); magicImpute imputes all missing founder genotypes. As a result, the offspring imputation fraction of mpimpute decreases rapidly from 0.92 to 0.6, whereas it is always around 0.96 for magicImpute (Figure S8F).

Apple CP: Figure 5D shows the results of offspring imputation accuracy obtained from the real apple data. The imputation accuracy of magicImpute decreases from 0.94 to 0.88 when read depth decreases from 15 to 0.46, in comparison with the almost constant accuracy of 0.96 in the simulated results in Figure 2D. The Beagle imputation accuracy is comparable to that of magicImpute, when read depth is no less than the break point of 3.7.

As shown in Figure S8G, the founder imputation accuracy of magicImpute at the highest read depth is around 0.96 when the two founders' genotypes are available, whereas it decreases to 0.75 when the two founders' genotypes are missing. The low accuracy is very likely because

of the mix up of the imputed genotypes between the two founders.

Running time: The running times for the four real datasets at the highest read depths or the smallest missing fractions are given in Table 1. Beagle is fastest in all populations. For the biparental populations, LB-Impute is much slower than magicImpute. And for the rice MAGIC, mpimpute is similar to Beagle, and faster than magicImpute.

The main computational load of magicImpute is the first two steps for founder imputation and phasing (Figure 1). The founder imputation of mpimpute and LB-impute is based on the decoding algorithm of the sub-HMM for each offspring, corresponding to the third step of magicImpute.

Discussion

We have implemented an HMM framework magicImpute for genotype imputation from low coverage sequence or SNP array data. The evaluations by simulation and real data in the four types of mapping populations demonstrate that magicImpute is accurate and flexible, despite the population being multiparental, founders being missing, founders being heterozygous, offspring being heterozygous, or sequencing coverage being low. The simulation evaluations also demonstrate the good performance of magicImpute for error detection and genotype phasing.

Although the dependence of imputation accuracy on sequence coverage varies with population size, marker density, and distribution of reads, magicImpute performs much better than Beagle, LB-Impute, and mpimpute at very low coverage. Beagle breaks down at much higher read depth in heterozygous populations than in almost homozygous populations, probably because of unsuccessful pre-phasing of Beagle imputation for heterozygous populations. Alternative pre-phasing methods might increase the follow-up imputation accuracy (WHALEN *et al.* 2017). The LB-Impute accuracy in biparental populations decreases with decreasing read depth, probably because the number of markers in the Markov trellis window is only 7 by default (large window size would result in dramatic increases in running time). The lower LB-Impute accuracy in the real AI-RIL than in the simulated AI-RIL may be due to the heavy tailed distribution of read depth in the real data and its inability of borrowing distant marker information.

Low coverage sequencing can be represented as allelic depths or called genotypes for the

input of magicImpute. The simulation and real evaluations show that the prior transformation of allelic depths into called genotypes has no appreciable effects, if homozygosity is not assumed for the transformation in almost homozygous populations. It indicates that little information is lost in the prior transformation, where the two half called genotypes ($1U$ and $2U$) keep sequence read information efficiently. Genotype likelihoods, a probabilistic representation of low coverage sequencing, have been alternatively used in many imputation methods such as Beagle v4.1.

It is implicitly assumed by magicImpute that sequencing reads are too short to cover more than two polymorphic sites, and the phasing information of long reads is ignored. Thus magicImpute would not rely on long reads. For very low coverage sequencing, the distances between detected neighbor polymorphic sites are expected to be too long, and very long reads are thus required to keep the phasing information. On the other hand, our HMM imputation framework provides a solid step for the extension to utilize phasing information.

One key assumption of magicImpute is no segregation distortion, when incorporating breeding design information into the HMM. The assumption is not expected to be a problem for biparental populations with only two inbred founders, as confirmed in our real data evaluation. For the MAGIC and the CP, the founder imputation accuracies in the real data evaluations are lower than simulation results, probably because of segregation distortion in the real data. For real MAGIC, magicImpute has higher offspring imputation accuracy and lower founder imputation accuracy than mpimpute, indicating that the offspring imputation is not affected by the possible segregation distortion.

Secondly, magicImpute assumes that the input genetic map is correct, as do Beagle, LB-Impute, and mpimpute. The assumption contributes to the differences of ceiling offspring imputation accuracy between simulation and real data evaluations. For the real apple CP, GARDNER *et al.* (2014) estimated the proportion of markers that are inconsistent with the physical grouping is as high as 18.3%, which might explain why the accuracy is relatively low (from 0.88 to 0.94) when read depth is no less than the break point (Figure 5D). See for example MONEY *et al.* (2015) and RUTKOSKI *et al.* (2013) for map-independent imputations in association panels.

Another assumption of magicImpute is on the conditional independence of offspring. In the

approximate forward algorithm for founder imputation, offspring are assumed to be independent given the posterior probabilities up to the current time. This approximation is well validated by the very accurate founder imputation in the simulation evaluations. Conditional on the imputed founder haplotypes, offspring are assumed to be independent, which is not always true because these offspring share parents in the intermediate generations. The algorithm magicImpute partly accounts for this relationship by the pre-calculated HMM parameters based on available breeding pedigrees, and thus the offspring imputation utilizes the marker information of the others indirectly via the founder imputation.

In conclusion, we have demonstrated that magicImpute is more accurate and robust to low sequencing depth than the current methods, because magicImpute can incorporate experimental design and utilize marker data efficiently. Furthermore, magicImpute is not restricted to specific experimental designs, and it can perform parental imputation and phasing in situations where most current methods are incapable.

Acknowledgments

The authors thank Emma Huang for helps on mpimpute, Cris Wijnen for valuable discussion on sequencing technology, and three anonymous reviewers for their constructive comments. This research was supported by the Stichting Technische Wetenschappen (STW) - Technology Foundation, which is part of the Nederlandse Organisatie voor Wetenschappelijk Onderzoek - Netherlands Organisation for Scientific Research, and which is partly funded by the Ministry of Economic Affairs. The specific grant number was STW-Rijk Zwaan project 12425.

Author contributions

CZ designed the study, created the model, developed the software and algorithm, and wrote the first draft of the manuscript. MPB and FAE provided critical feedback, helped shape the manuscript, and acquitted financial support. FAE supervised the project. All authors read and approved the final manuscript.

Literature Cited

ABECASIS, G. R., S. S. CHERNY, W. O. COOKSON, and L. R. CARDON, 2002 Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics **30**: 97–101.

ALTSHULER, D. M., R. M. DURBIN, G. R. ABECASIS, D. R. BENTLEY, A. CHAKRAVARTI, *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature **491**: 56–65.

BANDILLO, N., C. RAGHAVAN, P. A. MUYCO, M. A. L. SEVILLA, I. T. LOBINA, *et al.*, 2013 Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. RICE **6**: 11.

BROWNING, B. L., and S. R. BROWNING, 2016 Genotype imputation with millions of reference samples. American Journal of Human Genetics **98**: 116–126.

CHEUNG, C. Y. K., E. A. THOMPSON, and E. M. WIJSMAN, 2013 GIGI: an approach to effective imputation of dense genotypes on large pedigrees. American Journal of Human Genetics **92**: 504–516.

ELSHIRE, R. J., J. C. GLAUBITZ, Q. SUN, J. A. POLAND, K. KAWAMOTO, *et al.*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLOS ONE **6**: e19379.

FRAGOSO, C. A., C. HEFFELFINGER, H. Y. ZHAO, and S. L. DELLAPORTA, 2016 Imputing genotypes in biallelic populations from low-coverage sequence data. Genetics **202**: 487–495.

FRAZER, K. A., D. G. BALLINGER, D. R. COX, D. A. HINDS, L. L. STUVE, *et al.*, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature **449**: 851–U3.

FRIEL, N., and H. RUE, 2007 Recursive computing and simulation-free inference for general factorizable models. Biometrika **94**: 661–672.

GARDNER, K. M., P. BROWN, T. F. COOKE, S. CANN, F. COSTA, *et al.*, 2014 Fast and cost-effective genetic mapping in apple using next-generation sequencing. G3 (Bethesda) **4**: 1681–1687.

554 HEFFELFINGER, C., C. A. FRAGOSO, M. A. MORENO, J. D. OVERTON, J. P. MOTTINGER,
555 *et al.*, 2014 Flexible and scalable genotyping-by-sequencing strategies for population studies.
556 BMC Genomics **15**: 979.

557 HICKEY, J. M., G. GORJANC, R. K. VARSHNEY, and C. NETTELBLAD, 2015 Imputation of
558 single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations
559 with a hidden markov model. Crop Science **55**: 1934–1946.

560 HOWIE, B. N., P. DONNELLY, and J. MARCHINI, 2009 A flexible and accurate genotype im-
561 putation method for the next generation of genome-wide association studies. PLoS Genetics
562 **5**: e1000529.

563 HUANG, B. E., C. RAGHAVAN, R. MAULEON, K. W. BROMAN, and H. LEUNG, 2014 Ef-
564 ficient imputation of missing markers in low-coverage genotyping-by-sequencing data from
565 multiparental crosses. Genetics **197**: 401–404.

566 KIM, C., H. GUO, W. KONG, R. CHANDNANI, L.-S. SHUANG, *et al.*, 2016 Application of
567 genotyping by sequencing technology to a variety of crop breeding programs. Plant Science
568 **242**: 14–22.

569 KOVER, P. X., W. VALDAR, J. TRAKALO, N. SCARCELLI, I. M. EHRENREICH, *et al.*, 2009
570 A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis*
571 *thaliana*. PLoS Genetics **5**: e1000551.

572 LI, W., and J. FREUDENBERG, 2009 Two-parameter characterization of chromosome-scale
573 recombination rate. Genome Research **19**: 2300–2307.

574 LI, Y., C. J. WILLER, J. DING, P. SCHEET, and G. R. ABECASIS, 2010 MaCH: Using se-
575 quence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epi-
576 demiology **34**: 816–834.

577 MACKAY, I. J., P. BANSEPT-BASLER, T. BARBER, A. R. BENTLEY, J. COCKRAM, *et al.*,
578 2014 An eight-parent multiparent advanced generation inter-cross population for winter-sown
579 wheat: creation, properties, and validation. G3 (Bethesda) **4**: 1603–1610.

MARCHINI, J., and B. HOWIE, 2010 Genotype imputation for genome-wide association studies. Nature Reviews. Genetics **11**: 499–511.

MONEY, D., K. GARDNER, Z. MIGICOVSKY, H. SCHWANINGER, G. Y. ZHONG, *et al.*, 2015 LinkImpute: fast and accurate genotype imputation for nonmodel organisms. G3 (Bethesda) **5**: 2383–2390.

PEI, Y. F., J. LI, L. ZHANG, C. J. PAPASIAN, and H. W. DENG, 2008 Analyses and comparison of accuracy of different genotype imputation methods. PLOS ONE **3**: e3551.

RABINER, L., 1989 A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**: 257–286.

ROSHYARA, N. R., K. HORN, H. KIRSTEN, P. AHNERT, and M. SCHOLZ, 2016 Comparing performance of modern genotype imputation methods in different ethnicities. Scientific Reports **6**: 34386.

RUTKOSKI, J. E., J. POLAND, J. L. JANNINK, and M. E. SORRELLS, 2013 Imputation of unordered markers and the impact on genomic selection accuracy. G3 (Bethesda) **3**: 427–439.

SANNEMANN, W., B. E. HUANG, B. MATHEW, and J. LEON, 2015 Multi-parent advanced generation inter-cross in barley: high-resolution quantitative trait locus mapping for flowering time as a proof of concept. Molecular Breeding **35**: 86.

SPINDEL, J., M. WRIGHT, C. CHEN, J. COBB, J. GAGE, *et al.*, 2013 Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theoretical and Applied Genetics **126**: 2699–2716.

SWARTS, K., H. H. LI, J. A. R. NAVARRO, D. AN, M. C. ROMAY, *et al.*, 2014 Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome **7**: 1–12.

THEPOT, S., G. RESTOUX, I. GOLDRINGER, F. HOSPITAL, D. GOUACHE, *et al.*, 2015 Efficiently tracking selection in a multiparental population: the case of earliness in wheat. Genetics **199**: 609–621.

WHALEN, A., G. GORJANC, R. ROS-FREIXEDES, and J. M. HICKEY, 2017 Assessment of the performance of different hidden markov models for imputation in animal breeding. bioRxiv : 227157.

WOLFRAM RESEARCH, I., 2016 *Mathematica*. Wolfram Research, Inc., Champaign, Illinois, version 11.0 edition.

XIE, W. B., Q. FENG, H. H. YU, X. H. HUANG, Q. A. ZHAO, *et al.*, 2010 Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proceedings of the National academy of Sciences of the United States of America **107**: 10578–10583.

ZHENG, C., 2015 Modeling X-linked-linked ancestral origins in multiparental populations. G3 (Bethesda) **5**: 777–801.

ZHENG, C., M. P. BOER, and F. A. VAN EEUWIJK, 2014 A general modeling framework for genome ancestral origins in multiparental populations. Genetics **198**: 87–101.

ZHENG, C., M. P. BOER, and F. A. VAN EEUWIJK, 2015 Reconstruction of genome ancestry blocks in multiparental populations. Genetics **200**: 1073–1087.

ZHENG, C., M. P. BOER, and F. A. VAN EEUWIJK, 2018 Recursive algorithms for modeling genome blocks in a fixed pedigree. Submitted to G3 (Bethesda) .

ZHENG, C., R. E. VOORRIPS, J. JANSEN, C. A. HACKETT, J. HO, *et al.*, 2016 Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. Genetics **203**: 119–131.

Figures

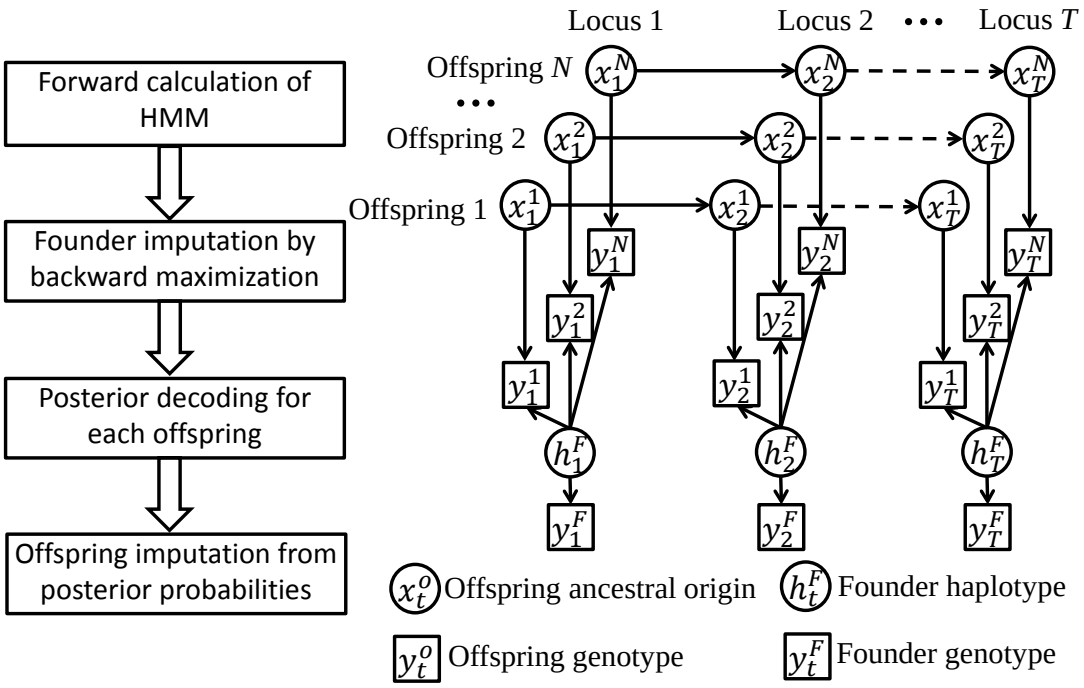

Figure 1

Figure 1: Overview of the imputation algorithm. The left panel shows the diagram of magicImpute. The right panel shows the directed acyclic graph of the HMM for N offspring at T loci, where the arrows denote probabilistic relationships that are described in the method section. See Table 1 for the symbols in the right panel. In the left panel, the second step of founder imputation results in the estimate of h_t^F and the third step of posterior decoding results in the posterior probability of x_t^o, conditional on genotypic data y_t^o and y_t^F for $t = 1...T$ and $o = 1...N$.

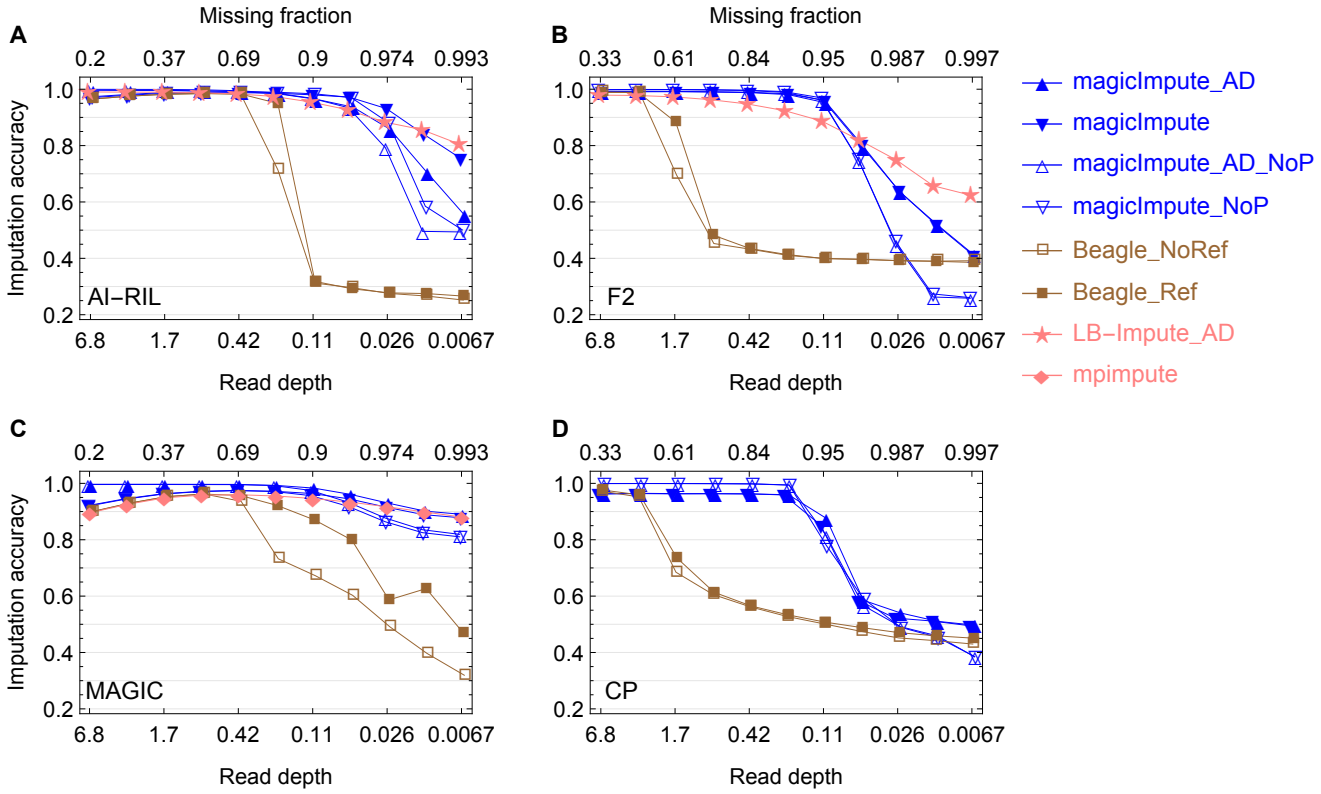

Figure 2

Figure 2: Simulation evaluation on the accuracy of imputing offspring genotypes. Panels A-D show the results for the AI-RIL, the F2, the MAGIC, and the CP, respectively. In the figure legend on the right side, "_AD" denotes that the input data are allelic depths rather than called genotypes, "_NoP" denotes that the first two founders' genotypes are not available, and "_Ref" and "_NoRef" denotes whether Beagle uses founder haplotypes as reference panels or not. When the input data are called genotypes, complete homozygosity is assumed for the AI-RIL and the MAGIC, and thus their missing fractions on the top axes are smaller than those of the F2 and the CP at the same depths.

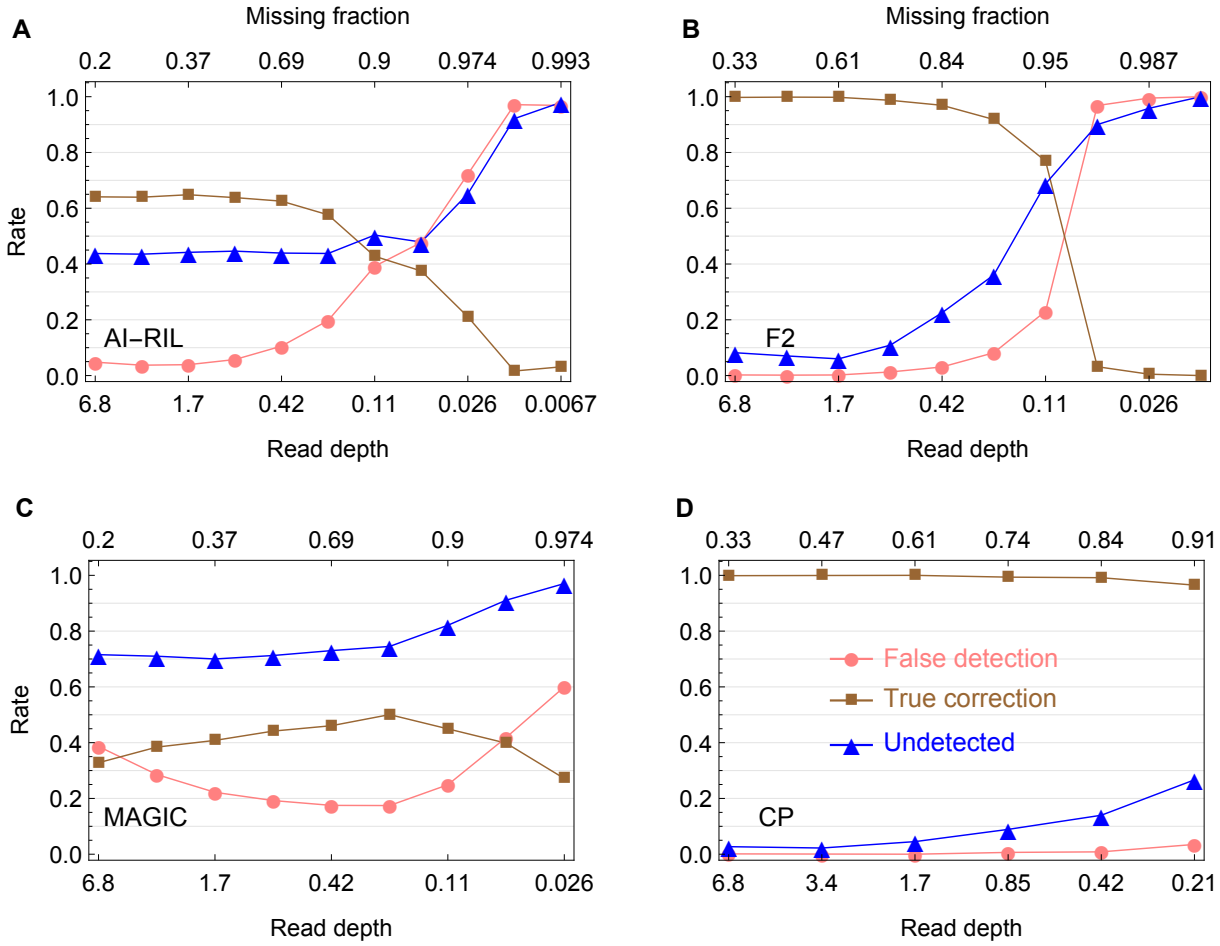

Figure 3

Figure 3: Simulation evaluation on the error detection in offspring genotypes. Panels A-D show the results for the AI-RIL, the F2, the MAGIC, and the CP, respectively, which are obtained by magicImpute with the first two founders' genotypes being unavailable and the input data being called genotypes. The false detection rate (●) denotes the percentage of estimated suspicious genotype errors being not true errors, the true correction rate (■) denotes the percentage of estimated suspicious genotype errors being true and being corrected into the true genotypes, and the undetected rate (▲) denotes the percentage of true genotype errors being not detected.

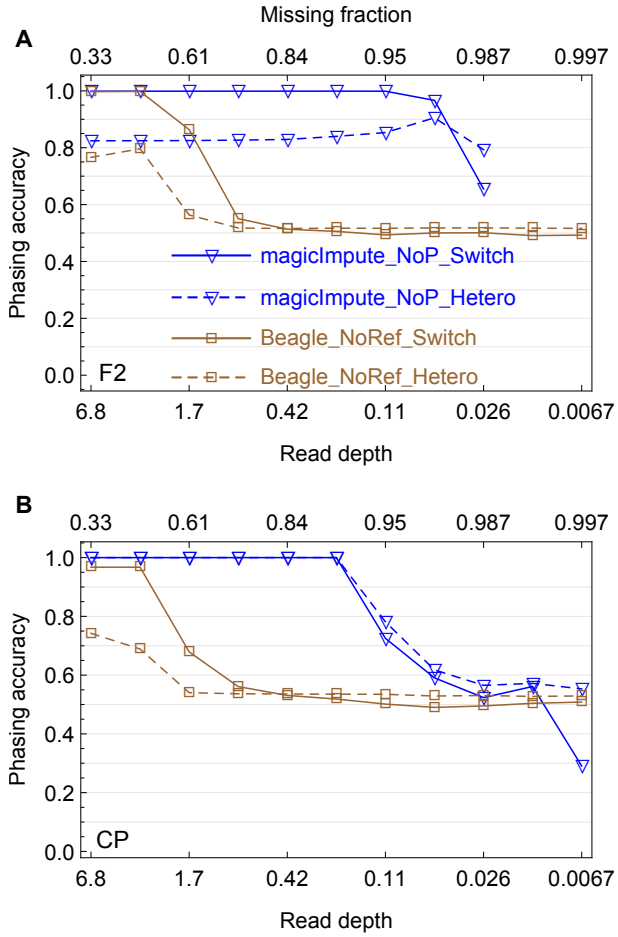

Figure 4

Figure 4: Simulation evaluation on the offspring genotype phasing. Panels A and B show the results obtained by magicImpute and Beagle for the F2 and the CP, respectively. For magicImpute, the first two founders' genotypes are unavailable ("_NoP"), and for Beagle there are no reference panels ("_NoRef"). The solid lines denote the switch accuracy ("_Switch"), one minus the percentage of switch errors to obtain the true haplotype phase; the dashed lines denote the percentage of correctly phased heterozygous genotypes ("_Hetero").

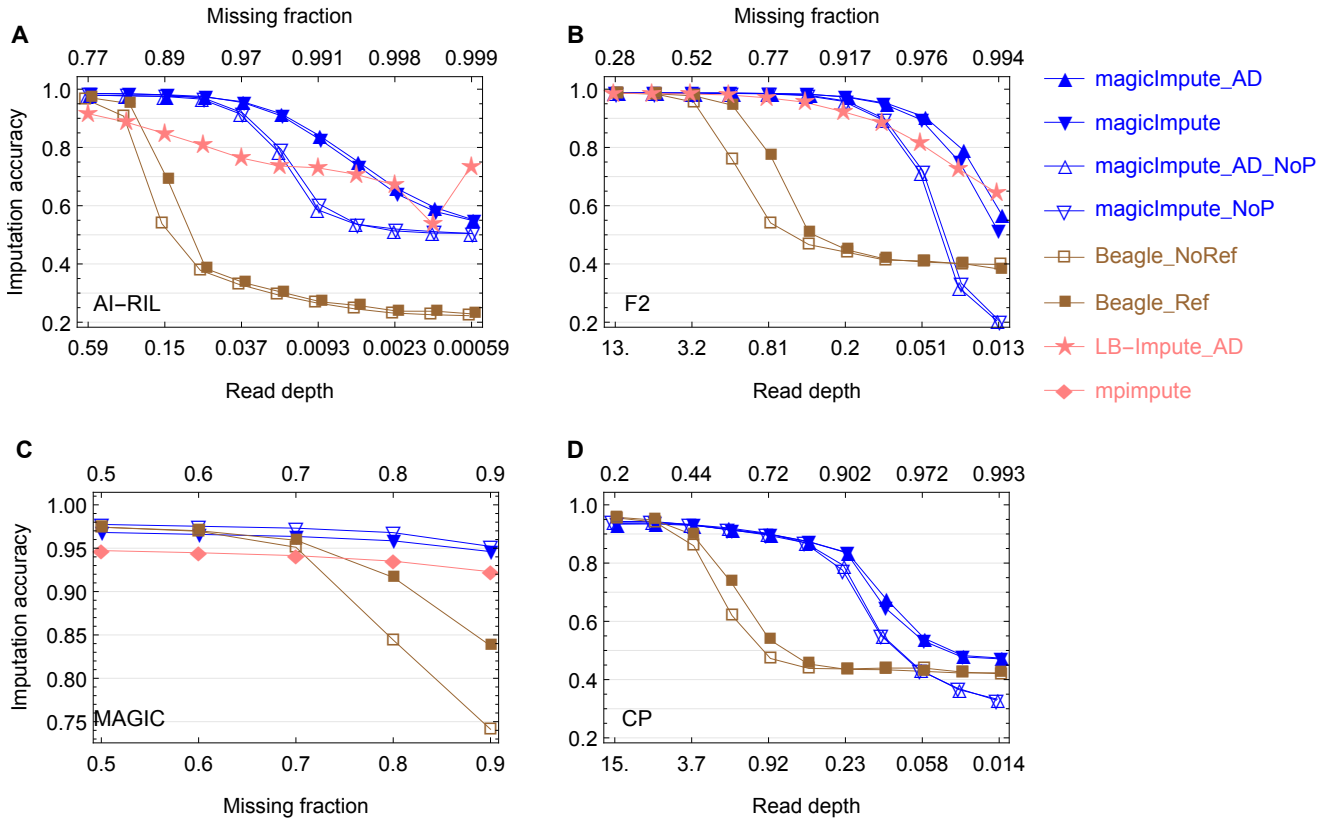

Figure 5

Figure 5: The accuracy of imputing offspring genotypes from real data. Panels A-D show the results for the AI-RIL, the F2, the MAGIC, and the CP, respectively. The figure legend on the right side is the same as that of Figure 2. Allelic depth data are not available for the MAGIC. The extreme large missing fraction or low read depth shows how genotype imputation approaches random imputation with decreasing amount of the input data. In panel A, the large variation of imputation accuracy of LB-Impute at low read depths is due to the corresponding imputation fraction being close to 0 (Figure S8B).

Table 1: List of symbols and their brief descriptions

Symbol	Description
n_F	Number of founders
N	Number of offspring
T	Number of markers (loci)
h_t^F	Hidden founder haplotype at locus t
$\boldsymbol{h}^F$	Hidden founder haplotype matrix $\boldsymbol{h}^F = \{h_t^F\}_{t=1...T}$
x_{ti}	Hidden ancestral origins at locus t in offspring i
x_{ti}^m, x_{ti}^p	$x_{ti} = (x_{ti}^m, x_{ti}^p)$ on maternally (m) or paternally (p) derived chromosome
d_{ti}	Genotype at locus t in offspring i that is completely determined by x_{ti} and h_t^F
z_{ti}	Hidden true genotype at locus t in offspring i
y_{ti}	Observed genotype at locus t in offspring i
$\boldsymbol{y}^O$	Observed offspring genotype matrix $\boldsymbol{y}^O = \{y_{ti}\}_{t=1...T,i=1...N}$
y_t^F	Observed genotypes for all founders at locus t
$\boldsymbol{y}^F$	Observed founder genotype matrix $\boldsymbol{y}^F = \{y_t^F\}_{t=1...T}$
$1U, 2U, UU$	Genotypes containing uncertain allele U
r_1, r_2	Number of reads for alleles 1 or 2
ϵ_O	Allelic error probability for offspring, independent of read depths
ϵ_F	Allelic error probability for founders, independent of read depths
$phred$	Phred quality score
ε	Sequencing error probability $\varepsilon = 10^{-phred/10}$
$\pi(x_{1i})$	Prior probability of x_{1i} at locus 1 in offspring i
$P(x_{ti}\|x_{t-1,i})$	Prior transition probability from $x_{t-1,i}$ to x_{ti}
l_{ti}	$l_{ti} = P(y_{ti}\|h_t^F, x_{ti}, \epsilon_O, \epsilon_F, \varepsilon)$ likelihood at locus t in offspring i
$\alpha(x_{ti}\|h_t^F)$	Posterior probability of x_{ti} conditional on h_t^F and genotypic data from loci 1 to t
$\tilde{\alpha}(x_{ti}\|h_t^F)$	Unnormalized conditional posterior probability of x_{ti}
$\gamma(h_t^F)$	Posterior probability of h_t^F conditional on genotypic data from loci 1 to t
$\hat{h}_t^F, \hat{x}_{ti}, \hat{z}_{ti}$	Hats denote maximum likelihood estimates
P_{call}	Single genotype call if probability of most probable genotype $>$ threshold P_{call}
P_{impute}	Impute if probability of most probable genotype is $>$ threshold P_{impute}
P_{detect}	Correct if probability of most probable genotype $>$ threshold P_{detect}

Table 2: The running time (in seconds) of genotype imputaton for the four real datasets.

Population	Maize AI-RIL	Maize F2	Rice MAGIC	Apple CP
Number of SNPs	13,912	127,059	37,240	13,493
Founder type	inbred	inbred	inbred	outbred
Offspring type	inbred	outbred	inbred	outbred
Number of founders	2	2	8	2
Number of offspring	275	87	178	87
magicImpute	784	212	3170	627
Beagle v4.1	178	31	445	39
LB-Impute	3698	3579	NA	NA
mpimpute	NA	NA	406	NA

File S1

Running setups of imputation packages

magicImpute

The Mathematica command line of magicImpute is given by

```
magicImpute[inputfile, model, popdesign, options]
```

where `inputfile` specifies the input genotypic data. Here `model` is set to be `{"depModel", "jointModel"}` for the population types AI-RIL and MAGIC, so that `"depModel"` is used for parental imputation, and `"jointModel"` is used for offspring imputation. And it is set to be `"jointModel"` for the population types F2 and CP, so that `"jointModel"` is used for both parental imputation and offspring imputation. See the online manual for details.

`popdesign` specifies the breeding design information that is used to compute the process parameter values of the HMM. For the F2, it is set to be `{"Pairing","Selfing"}`. For the AI-RIL, it is set to be `{"RM1-NE-1000","RM1-NE",...,"RM1-NE","Selfing", ...,"Selfing"}` where `"RM1-NE"` is repeated for 5 times, and `"Selfing"` is repeated for 6 times. For the MAGIC, it is set to be `{"Pairing", "Pairing", "Pairing", "Selfing", ..., "Selfing"}` where `"Selfing"` is repeated for 4 times. For the CP, it is specified in terms of a pedigree file.

There are many options for magicImpute. The option `imputingTarget -> All` so that we by default impute both founder and offspring. The options `founderAllelicError -> 0.005` and `offspringAllelicError -> 0.005` specify ε_F and ε_O, respectively. The option `isFounderInbred -> True` specifies that the founders are inbred for the F2, the AI-RIL, and the MAGIC, and `isFounderInbred -> False` is used for the CP. The option `imputingThreshold -> 0.9` specifies P_{impute}, The option `detectingThreshold -> 0.9` specifies P_{detect}. The option `minPhredQualScore -> 30` specifies that the quality score $phred$ so that $\varepsilon = 10^{-phred/10}$. The option `priorFounderCallThreshold -> 0.99` specifies the prior genotype calling threshold P_{call} when the input parental data are allelic depths.

Beagle v4.1

The command line used for Bealge v4.1 is given by

```
java -jar beagle.21Jan17.6cc.jar ne=100
```

where the effective population size is fixed to be 100. In addition, The gt option is used to specify input offspring genotype data, and the ref option is used to specify the imputed phased founder genotypes as the reference panel. We run Beagle with and without the reference panel.

LB-Impute

The command line used for LB-Impute is given by

```
java -jar LB-Impute.jar  -method impute -readerr 0.001
     -genotypeerr 0.01 -recombdist 10000000 -window 7
     -parentimpute -offspringimpute
```

Here the $-readerr$ option specifies the sequencing error, and it is set to be 0.001 corresponding to the quality score 30. The $-genotypeerr$ option specifies the genotype error to be 0.01, corresponding to the depth-independence allelic error probability of 0.005 in magicImpute. The two founder names are specified by the $-parents$ option, and the input and output files are specified by the options $-f$ and $-o$, respectively.

mpimpute

The R command line used for mpimpute is given by

```
mpimpute(object,what="both",threshold=0.5,calls="discrete")
```

Here the $what$ option is set so that we impute both founders and offspring, and input genotypic data and pedigree information are specified by the $object$.

Supplementary figures

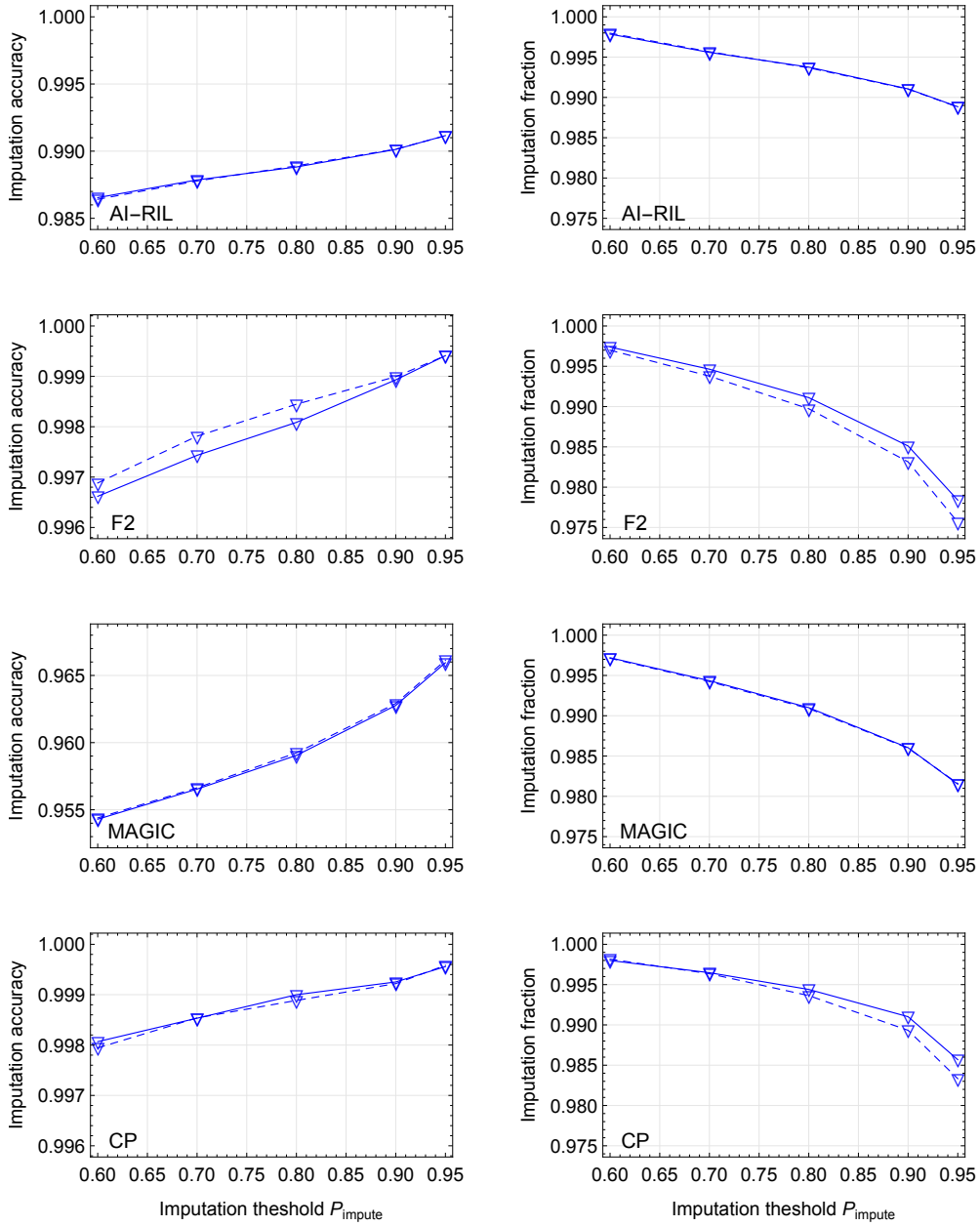

Figure S1: Sensitivity analysis of imputation threshold P_{impute} for the algorithm magicImpute. Panels from top to bottom denote the results for the AI-RIL, the F2, the MAGIC, and the CP, respectively. The solid and dashed lines denote the results corresponding to input parameter $\epsilon_O = 0.005$ and 0.05, respectively. The left and right panels denote the results for imputation accuracy and imputation fraction, respectively, which are obtained from the simulated datasets with the input data being called genotypes at read depth 0.85 and the first two founders' genotypes being not available.

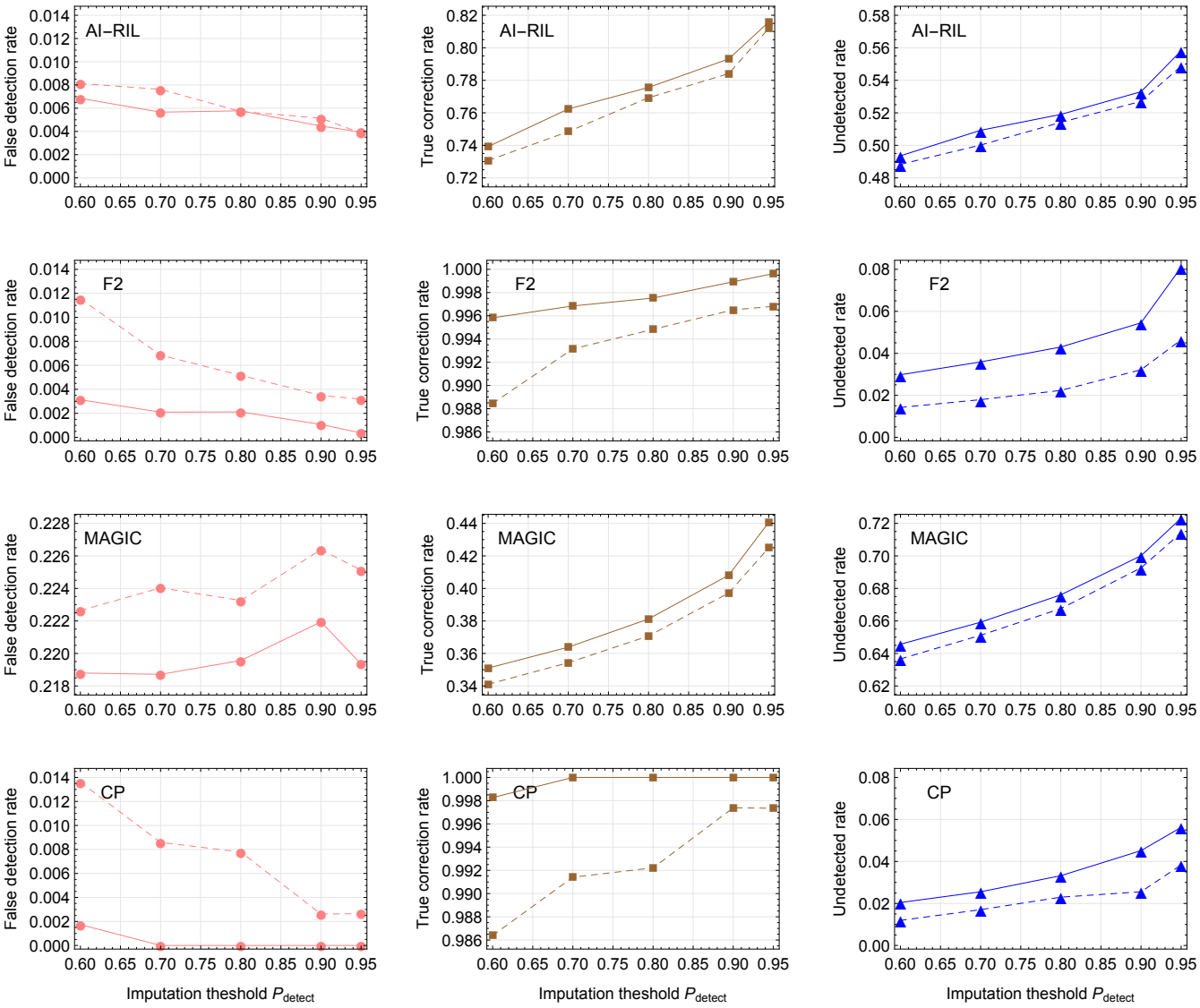

Figure S2: Sensitivity analysis of error detection threshold P_{detect} for the algorithm magicImpute. Panels from top to bottom denote the results for the AI-RIL, the F2, the MAGIC, and the CP, respectively. The solid and dashed lines denote the results corresponding to input parameter ϵ_O= 0.005 and 0.05, respectively. The left, middle and right panels denote false detection rate, true corretion rate, and undetected rate, respectively. For each simulated dataset with population size 200 and read depth 0.85, the results are obtained with the input data being called genotypes and the first two founders' genotypes being not available.

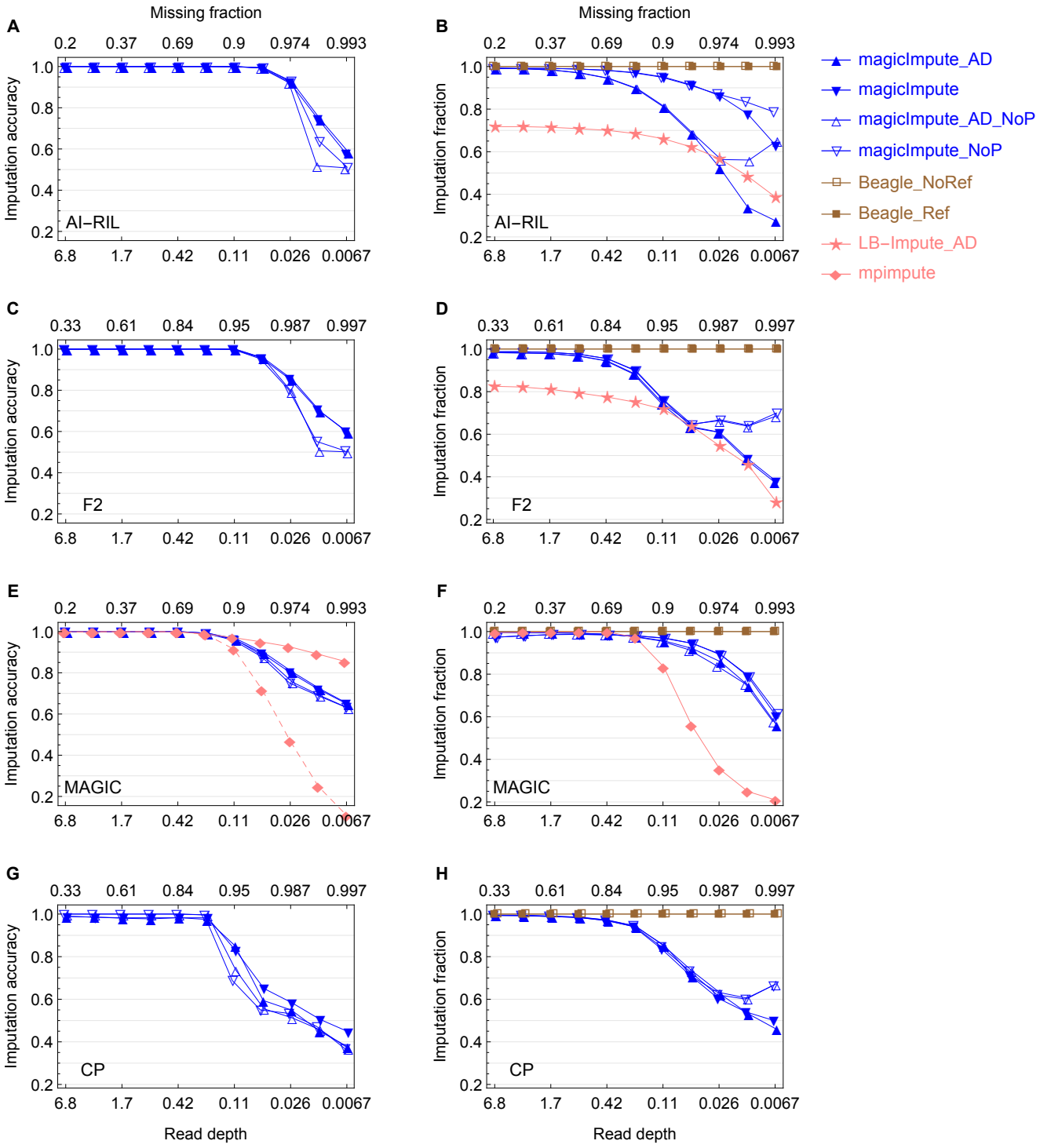

Figure S3: Simulation evaluation on the accuracy of imputing founder genotypes (left panels) and imputation fraction of offspring genotypes (right panels). Panels A&B, C&D, E&F, and G&H denote the results for the AI-RIL, the F2, the MAGIC, and the CP, respectively. The dashed lines in panel E denotes the mpimpute imputation fraction of founder genotypes. Beagle and LB-Impute do not impute founder genotypes, and magicImpute always imputes all the founder genotypes.

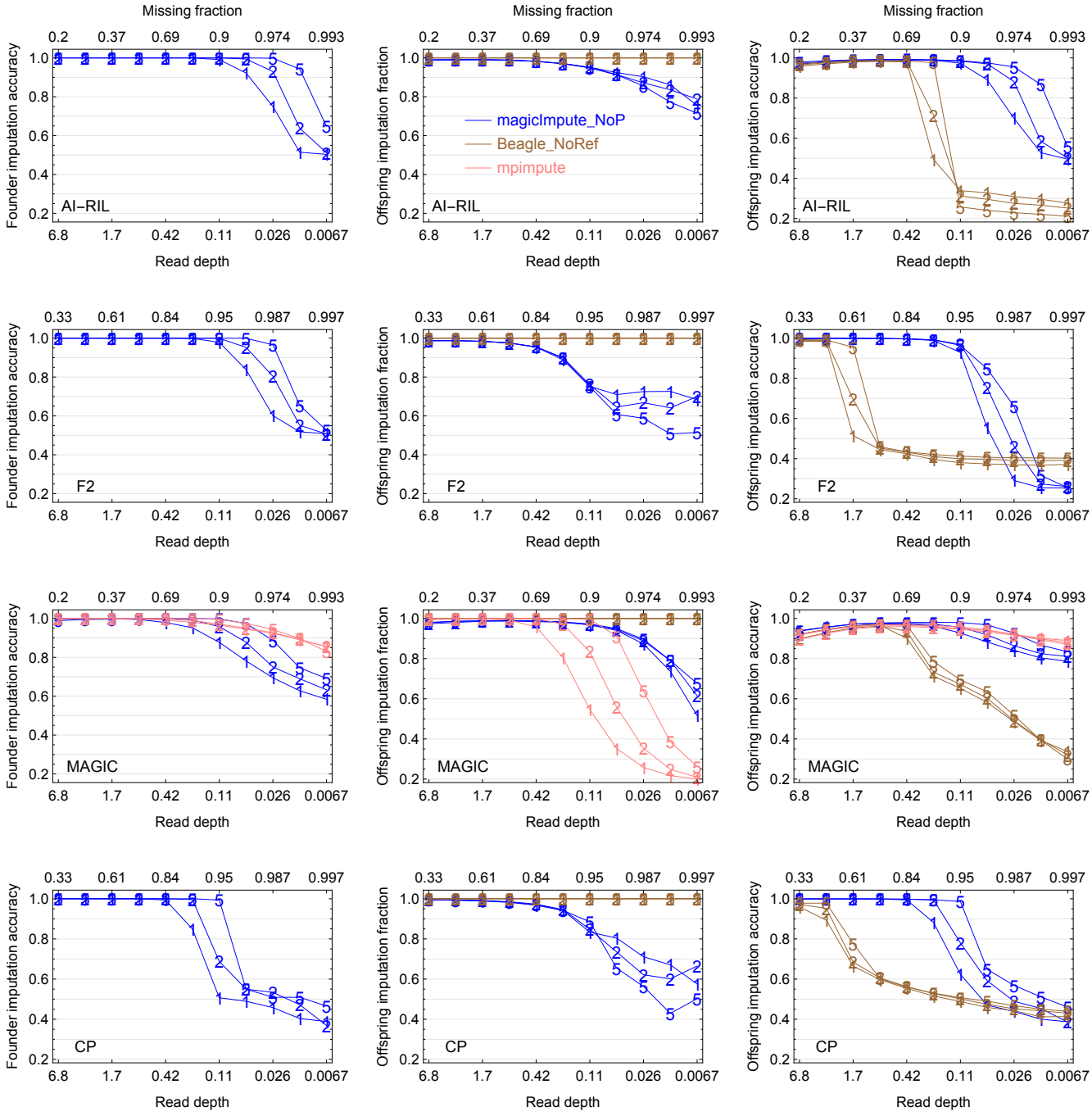

Figure S4: Dependencies of genotype imputation on population size. Panels from top to bottom denote the results for the AI-RIL, the F2, the MAGIC, and the CP, respectively. For magicImpute, the input data are called genotypes and the first two founders are missing; no reference panels for Beagle imputation. The plot markers "1", "2", and "5" denote population sizes 100, 200, and 500, respectively.

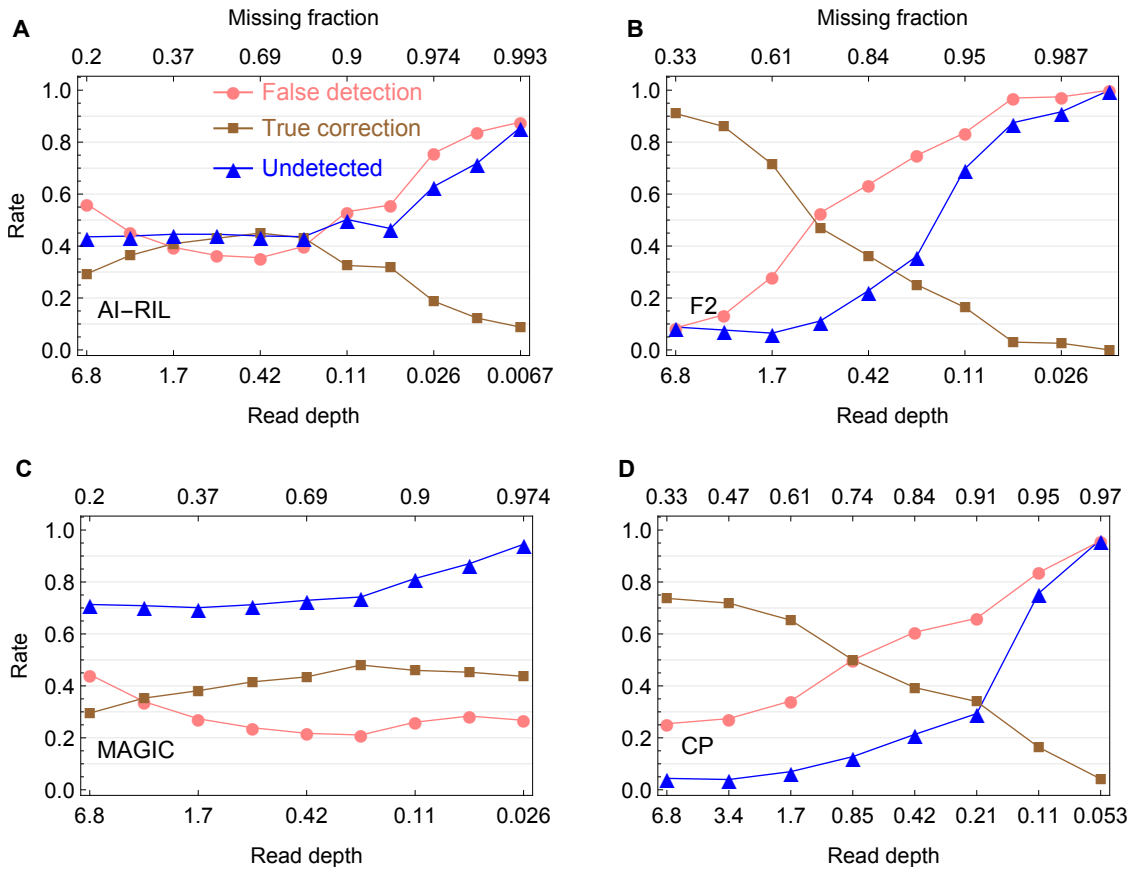

Figure S5: Similar to Figure 3 for the error detection by magicImpute but with the first two founders' genotypes being available.

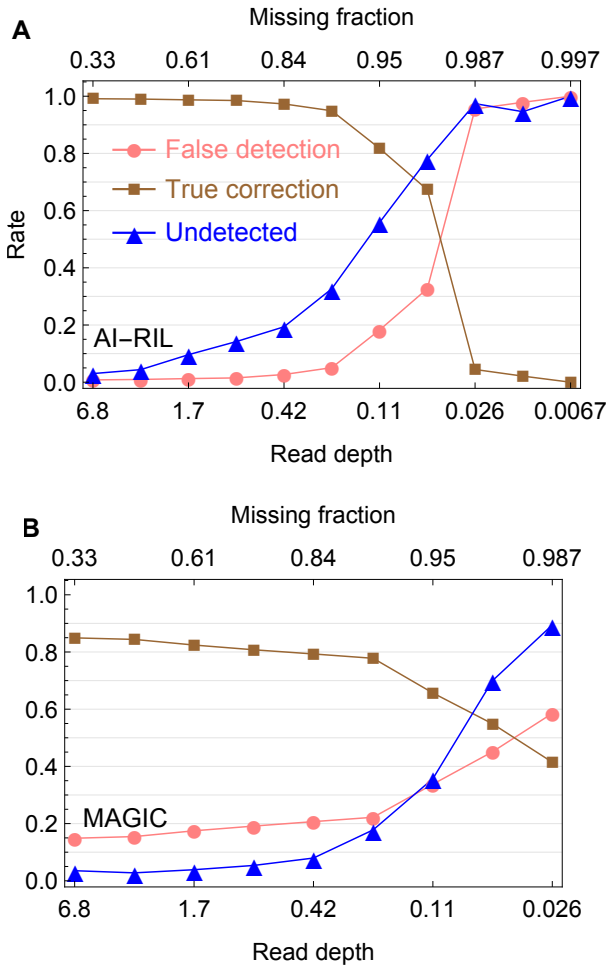

Figure S6: Similar to Figure 3 for the error detection by magicImpute but without assuming homozygosity for the almost homozygous populations AI-RIL (A) and MAGIC (B).

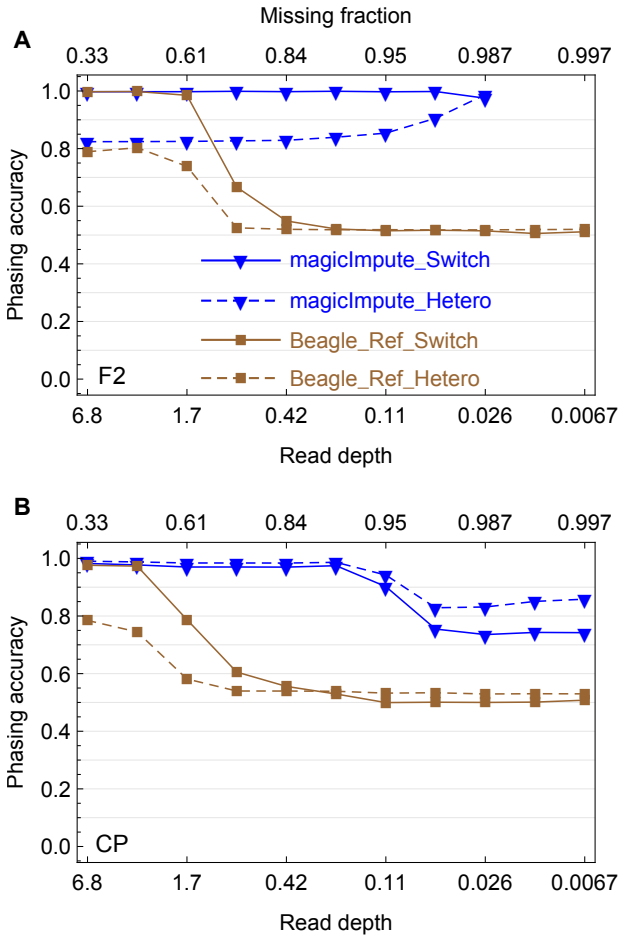

Figure S7: Similar to Figure 4 for the offspring phasing but for magicImpute with the first two founders' genotypes being available and for Beagle with the founder haplotypes (imputed by magicImpute) being the reference panels.

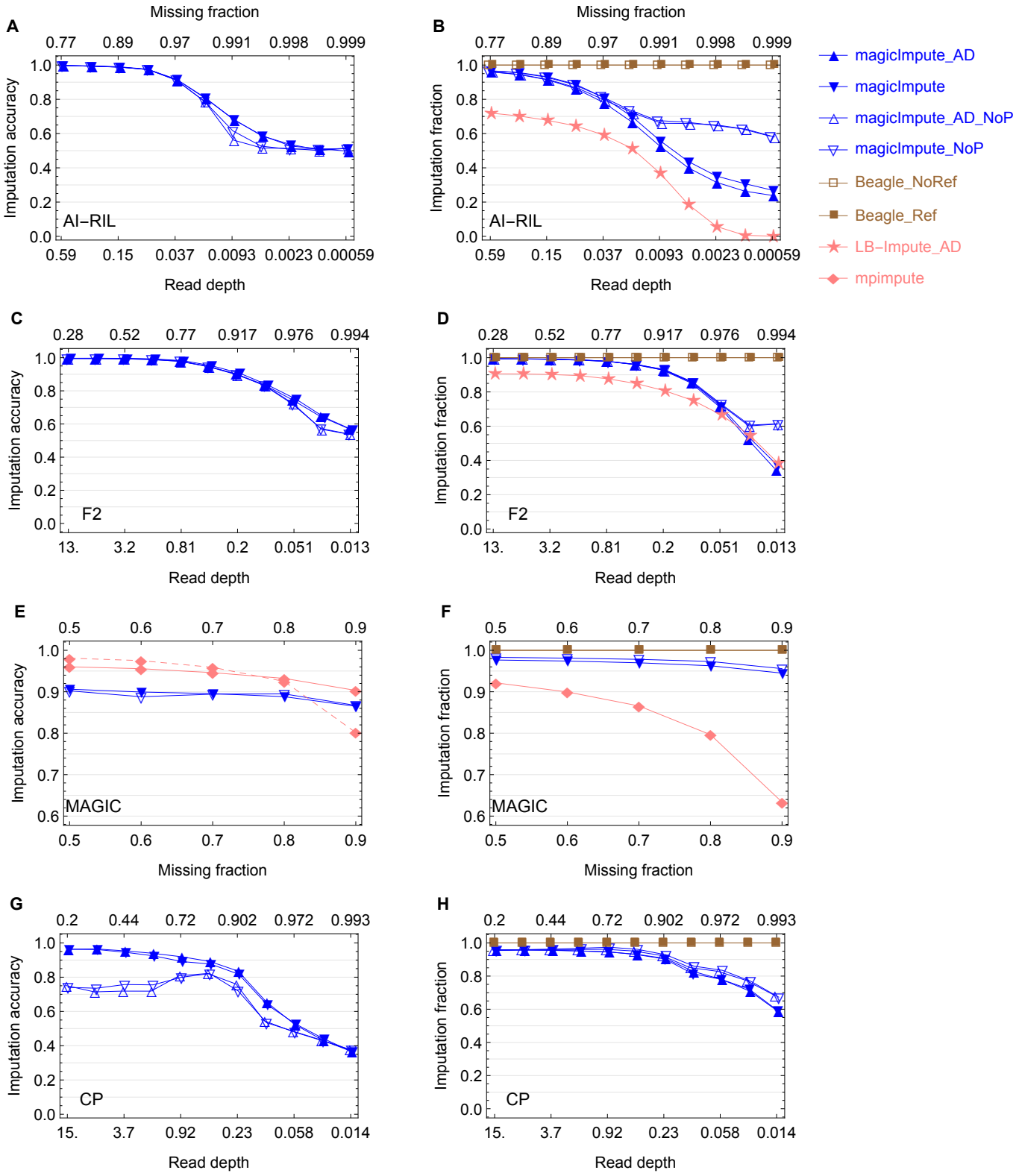

Figure S8: Evaluation on the accuracy of imputing founder genotypes and imputation fraction of offspring genotypes by real data. Panels A&B, C&D, E&F, and G&H denote the results for the AI-RIL, the F2, the MAGIC, and the CP, respectively. The dashed lines in panel E denotes the mpimpute imputation fraction of founder genotypes.