Imputation of Single Nucleotide Polymorphism Genotypes in Biparental, Backcross, and Topcross Populations with a Hidden Markov Model

John M. Hickey,* Gregor Gorjanc, Rajeev K. Varshney, and Carl Nettelblad

ABSTRACT

Genomic selection offers great potential to increase the rate of genetic improvement in plant breeding programs. The ability to accurately impute missing genotypes for a large number of individuals, screened with low marker density, at low cost is crucial for achieving this. In this research an existing general algorithm for tracing allele inheritance in known pedigrees was modified to enable genotype imputation in specific crosses (biparental, backcross, and topcross) that are common in plant breeding. The extension was tested with a series of representative simulated examples of these crosses. The results show success of imputation is affected by many factors including the number of low-density markers per cM, level of inbreeding or intercrossing of the individuals to have genotypes imputed, level of inbreeding of the parents of a cross, and genome length; but not by the number of high-density markers or by the interaction between the genome length and the number of low-density markers. With as few as one or two markers per 20 cM genotype imputation was successful when parents were inbred. Therefore, genotyping strategies in which inbred parents of a cross are genotyped at high-density and their descendants are genotyped with 200 to 400 markers genome wide may be cost effective and useful in practical plant breeding programs that utilize genomic selection.

J.M. Hickey, The Roslin Institute and Royal (Dick) School of Veterinary Studies, Univ. of Edinburgh, Easter Bush Research Centre, Midlothian EH25 9RG, UK; G. Gorjanc, The Roslin Institute and Royal (Dick) School of Veterinary Studies, Univ. of Edinburgh, Easter Bush Research Centre, Midlothian EH25 9RG, UK; R.K. Varshney, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Centre of Excellence in Genomics (CEG), Patancheru 502324, A.P., India; C. Nettelblad, Lab. of Molecular Biophysics, Dep. of Cell and Molecular Biology, Uppsala Univ., Husargatan 3 (Box 596), SE-751 24 Uppsala, Sweden. Received 24 Sept. 2014. Accepted 17 Mar. 2015. *Corresponding author (john.hickey@roslin.ed.ac.uk).

Abbreviations: GS, genome selection; GWAS, genome wide association studies; HMM, hidden Markov models; SNP, single nucleotide polymorphism.

Genomic selection (GS) and genome wide association studies (GWAS) are valuable tools in plant breeding programs. Both are most powerful when many individuals are genotyped with high-density markers. Genotyping many individuals provides more accurate estimates of breeding values in GS and marker effects in GWAS and genotyping more selection candidates allows for greater selection intensity in GS and this enables greater response to selection. However, genotyping a large number of individuals with high-density marker arrays can be expensive and genotype imputation is a cost-effective way to achieve the practical equivalent of high-density genotype information for many individuals.

To generalize, genotype imputation involves genotyping some individuals with high-density markers, other individuals with low-density markers; and using this information to impute the untyped markers. The assumption underlying imputation is that

Published in Crop Sci. 55:1–13 (2015).
doi: 10.2135/cropsci2014.09.0648
Freely available online through the author-supported open-access option.

the haplotypes carried by the individuals that are genotyped at low-density are mosaics of the haplotypes carried by the individuals genotyped at high-density.

Biparental populations, and similar populations such as backcross and topcross, are widely used in plant breeding and plant genetics research. These populations are ideal for imputation for four reasons: (i) The favorable ratio between the numbers of individuals genotyped at low-density and high-density enables imputation to be a cost effective technique. For example, biparental populations have a small number of founder individuals (i.e., two) that need to be genotyped at high-density and a large number of descendants (e.g., hundreds of F_2 individuals and their descendants in the further generations), which could be genotyped at low-density. (ii) The high levels of inbreeding typical in parental lines simplifies the task of inferring the parental haplotypes, since homozygous loci in the parents are de-facto phased. (iii) There is a very small number of recombination events between the individuals that would be genotyped at high-density and those that would be genotyped at low-density. Therefore, long haplotype blocks are preserved, allowing very accurate imputation. (iv) These populations (bi-parental and similar) have very defined pedigree structures that could be explicitly utilized, through inheritance tracking, to empower imputation.

Many imputation algorithms have been developed for application in human genetics (e.g., Beagle–Browning and Browning, 2007; IMPUTE2–Howie et al., 2009; MaCH–Li et al., 2010) and animal breeding (e.g., Findhap–VanRaden et al., 2011; AlphaImpute–Hickey et al., 2011; FIMPUTE–Sargolzaei et al., 2011). While these algorithms can be applied to biparental populations in plant breeding, they do not optimally capitalize on the genetic structure of such populations. To our knowledge no publicly available algorithm has been explicitly designed for genotype imputation in plant biparental populations but there have been several algorithms designed for the tracking of inheritance in these populations (e.g., Haley and Knott, 1992; Broman et al., 2003; Nettelblad et al., 2009), which can be used for developing powerful algorithms for imputation in plant breeding.

The objective of this research was to adapt and evaluate an inheritance-tracking algorithm based on hidden Markov models (HMM) (Nettelblad et al., 2009; Nettelblad, 2012) for imputation of single nucleotide polymorphism (SNP) genotypes in plant breeding populations under different scenarios and breeding practices. The accuracy of imputation was quantified in detail using simulated data with different marker densities, levels of inbreeding in parents and their descendants, intercrossing, recombination rates, and population type. The results show that an HMM that tracks inheritance can be highly effective for imputation in many plant breeding scenarios.

MATERIALS AND METHODS

Description of the Imputation Algorithm

An HMM describes the generation of a sequence of observed symbols (Rabiner, 1989). The emission of those symbols is controlled by the model states. In the HMM, the states and transitions between states are not directly observable. There is, therefore, a hidden sequence of states, and the probability of observing a symbol at a specific position in the sequence depends only on the state at that specific position. In addition, the probability of a particular state at position $x+1$ depends only on the state at position x. These conditional independencies represent the so-called Markov property. When the sequence of observed symbols is known, or partially known, the posterior probabilities for the hidden states as well as unknown symbols can be computed using the Forward–Backward algorithm (Rabiner, 1989). It is this latter property of being able to infer unknown symbols that we are using for imputation.

In our genetic model the symbols are defined as the observed marker alleles in an individual and its immediate ancestors. The states are defined as the identity of marker alleles by descent, that is, which grandparent allele, carried by parents, was transmitted to the focal individual (Nettelblad, 2012). Transitions between states correspond to recombination events. By assuming the Haldane model of recombination probabilities, the Markov property between states holds. When the same founder/parent individual appears in multiple offspring pedigrees, the total information, in all those pedigrees, can be used to iteratively refine an estimate of the phase for each marker in the founder (Nettelblad, 2012).

In this work the implementation of the described genetic model was adapted to map the inferred states to possible emitted marker values for those markers that are not observed in the offspring, that is, genotype imputation. This implementation was named PlantImpute and is available on request from carl.nettelblad@icm.uu.se. In the application, the aim was to identify which of the four (for biparental and backcross populations) or six (for topcross populations involving a third parent) grandparental alleles were transmitted to an F_2 individual (i.e., F_2 descendant of the parents used to form the cross) at genotyped loci; and based on the inferred states for non-genotyped loci to compute the corresponding posterior genotype probabilities. In the case of fully inbred parents of a biparental population both (grandparental) haplotypes in each respective parent are identical, but this was not exploited to allow for variable homozygosity in parental lines as described later. This HMM formulation has 64 states for each locus, representing the phase identity of the haplotype transmitted in the gamete to the next generation from a total of two parents and four grandparents. The phase in each gamete is a separate Markov process with two states, and in joining them the state space thus grows to 2^n. While the method as presented in Nettelblad (2012) does allow for the inference of haplotypes through successive generations it could be used in the imputation and population scenarios presented here. Such support is not relevant when the low-density genotyping pattern is identical in all individuals, as in this work. Since the genotypes are lacking in all offspring individuals, there is no additional information available for phasing heterozygous markers in the founder individuals. With a varied pattern of genotyped vs. missing markers in the offspring, for

Table 1. Transition probabilities between superstates (fixed/non-fixed parental haplotypes in successive selfing) in terms of the probability p that one observable recombination occurs in any of d generations.

Previous	New	No haplotype fixed	Haplotype 1 fixed	Haplotype 2 fixed
No haplotype	fixed	$1-2p$	p	p
Haplotype 1	fixed	$2p/(2^d-1)$	$1-2p/(2^d-1)-p^2$	p^2
Haplotype 2	fixed	$2p/(2^d-1)$	p^2	$1-2p/(2^d-1)-p^2$

example as occurs with genotyping-by-sequencing data, then the support for the inference of haplotypes through successive generations would be useful.

The model was also extended to efficiently handle selfing for multiple generations. The selfing extension was accomplished by adding three super-states representing the state of both haplotypes transmitted from the original individual. The main model still represents the state of the original offspring, that was then selfed. The super-states cover how this original haplotype configuration was then transmitted through successive generations of selfing, with the following possibilities: (i) both haplotypes being transmitted from the original individual, (ii) haplotype 1 transmitted in two copies from the original individual; and (iii) haplotype 2 transmitted in two copies from the original individual. The super-state does not describe at what generation of selfing the fixation happened. Transitions in main states and super-states are independent and the total transition probabilities between a composite state (main state and super-state) is thus a simple product. For main states the transitions are in turn, as in Nettelblad (2012), the product of the recombination probabilities for those gametes where the transition indicates recombination taking place and the complement probabilities for those gametes where it did not. The Haldane mapping function is used, but sex-dependent (and even individual-dependent) recombination rates are supported. Since these transition probabilities change the asymptotic state distribution, the initial state probabilities are also adjusted accordingly. For the super-state of no fixation to hold, each successive generation of selfing must result in no additional fixation happening. Thus, the probabilities in Table 1 contain both the simple probability for fixation p (computed using the Haldane mapping function and the number of generations d), and an explicit compensation by d for the "loss of fixation" probability, that is, the transition from a fixed super-state to the non-fixed one. In the end, these distributions correctly reflect the loss of heterozygosity over any number of generations of selfing, while not modeling the gametes in each generation explicitly.

The approach with super-states was chosen instead of the general pedigree approach described above due to computational tractability. Using the general pedigree approach, while still taking into account that there is only a single (duplicated) parent with selfing, would result in multiplying the total number of states by 4 for each successive generation of selfing; or 4194,304 states in total for selfed F_{10} individuals. With the introduction of selfing-specific super-states, any number of generations of selfing can be adequately represented by a total state space of 192 states (3 new super states combined with the existing state space of 64 states) and transition probabilities that reflect the number of repeated selfings, to accurately model the increased number of recombinations and the depression in terms of the number of heterozygotes. The importance of this can vary depending on the layout of the marker maps.

Data Simulation

Data was simulated in two steps using the Markovian coalescent simulation (MaCS–Chen et al., 2009) and gene drop simulation (AlphaDrop–Hickey and Gorjanc, 2012). Initially, 100 founder chromosomal haplotypes were simulated for one chromosome using the coalescent. Chromosomal haplotypes comprised 10^8 base pairs and were simulated using a per site mutation rate of 1.0×10^{-8}, a recombination rate per site that varied over scenarios, and an effective population size (N_e) that varied over time. The different recombination rates simulated were 0.25×10^{-8}, 0.5×10^{-8}, 1.0×10^{-8}, 1.5×10^{-8}, 2.0×10^{-8}, 3.0×10^{-8}, and 4.0×10^{-8} that resulted in the genetic lengths of chromosome being 25, 50, 100, 150, 200, 300, and 400 cM, respectively. Changes in the effective population size over time roughly mimicked the historical changes of N_e in a crop such as maize (*Zea mays* L.). The population size was set to $N_e = 100$ at the final generation of simulation, to $N_e = 1000$ at 100 generations ago, and to $N_e = 10{,}000$ at 2000 generations ago with linear changes in between. The resulting haplotypes had approximately 80,000 segregating sites varying between simulations with different genetic lengths of a chromosome.

After the simulation of founder haplotypes, a pedigree of 11,266 individuals was constructed (Fig. 1), where individuals had their chromosomes simulated by dropping founder chromosomal haplotypes through the pedigree with recombination (crossovers occurred with 1% probability per cM and were uniformly distributed along the chromosomes). Several pedigrees with the same structure were simulated with varying genetic lengths of the chromosomes. The simulated pedigrees were designed to quantify the accuracy of imputation with varying levels of inbreeding in parents and their descendants, rounds of inter-crossing, and population type. In detail, the pedigree (Fig. 1) was constructed by initiating three biparental populations each generated from two outbred founders. These biparental populations were used to generate individuals that were selfed to different levels (i.e., resulting in different levels of inbreeding) that then served as parents of subsequent populations of individuals (descendants) where imputation was performed. A series of biparental populations were generated with varying levels of inbreeding in the parents (F_1, F_2, F_4, F_6, F_8, F_{10}, and F_{20}), while backcross and topcross populations were generated only with inbred parents from F_{20} (Fig. 1). Each population of descendants had 100 individuals per generation from F_1 to F_{10}. To properly propagate the residual heterozygosity in the parents, each population of descendants was initiated by generating 100 pairs of F_1 individuals, that were in turn mated to generate 100 F_2 individuals, that were in turn selfed to generate 100 F_3 individuals, etc. (Fig. 1). Backcross and topcross populations had an intermediate step of generating the backcross (B_1) and topcross (T_1) individuals (Fig. 1). Intercrossing was simulated only for the biparental population with inbred parents from F_{20} with random mating of the F_2 individuals for several rounds.

Finally, several high-density and low-density marker arrays were constructed and each individual was genotyped with all

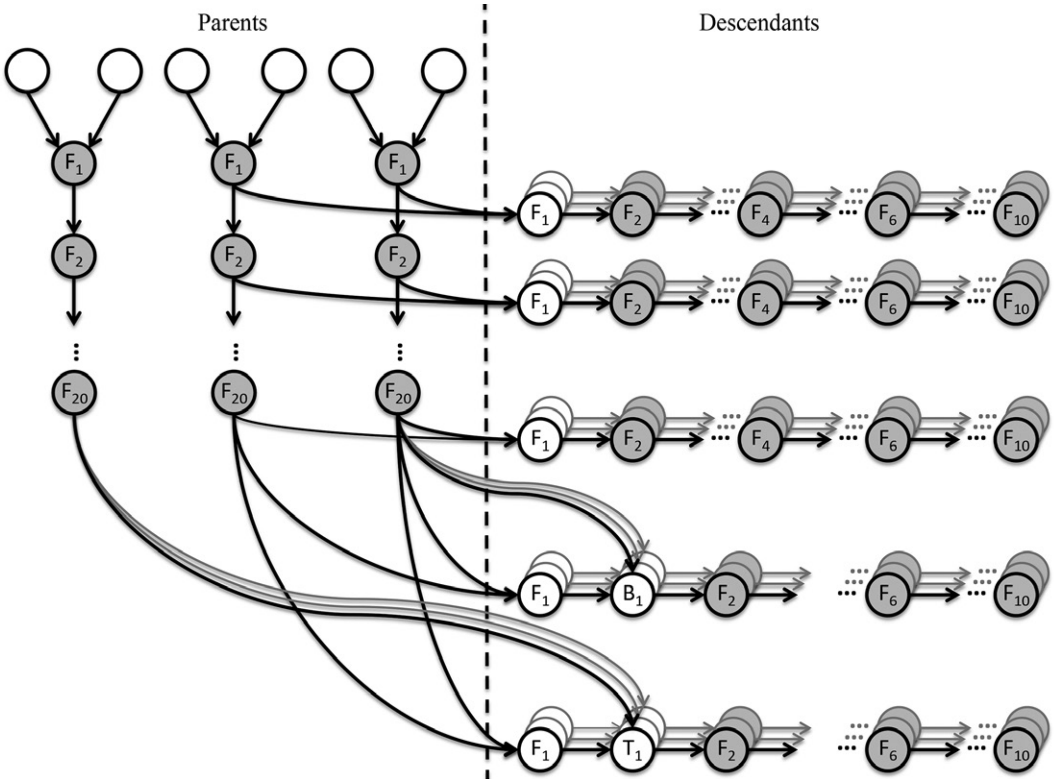

Figure 1. Pedigree design with different levels of inbreeding of the parents and their descendants and the population type (shaded circles denote individuals with genotype data)

the arrays for different analyses as described below. High-density arrays had 1000, 5000, or 25,000 markers per chromosome, while low-density arrays had 3, 5, 10, 20, 50, 100, 200, or 400 markers per chromosome. All marker arrays were nested, that is, the 3 markers from the smallest array were present on the array with 5 markers that were in turn present on the array with 10 markers etc. all up to the largest array with 25,000 markers, and all markers were assumed to be genotyped without error. Arrays were constructed by aiming to select a set of loci that segregated in the F_{20} generation of parents (Fig. 1) to obtain the same number of informative (segregating) markers in all populations. Out of approximately 80,000 segregating loci in the founder chromosomes 24,999 loci segregated in the F_{20} generation of parents and these loci constituded the high-density marker array. Two regions (at approximately 40 and 60 cM) were fixed *n* the F_{20} generation of parents, of which one harbored the fixed marker on an array.

Scenarios

To quantify the effect of different factors on the success of imputation, the generated data were analysed in six different scenarios. In all scenarios imputation to the high-density array with 1000 markers was performed separately for each low-density marker array but imputation to the two other high-density arrays (i.e., 5000 and 25,000) was only performed in the first scenario. Scenarios were:

- The first scenario (Sc1) was used to quantify the effect of imputing to different marker densities (1000, 5000, or 25,000 markers) in a biparental population where parents were fully inbred (from the F_{20} generation), descendants were from the F_2 generation, and chromosomes were 100 cM in length.
- The second scenario (Sc2) was used to quantify the effect of the level of selfing in the descendants (F_2, F_4, F_6, or F_{10}) in a biparental population where parents were fully inbred (from the F_{20} generation) and chromosomes were 100 cM in length. In this scenario the described selfing extension of the imputation algorithm was also tested.
- The third scenario (Sc3) was used to quantify the effect of the rounds of intercrossing (from 1 to 10 rounds) in a biparental population where parents were fully inbred (from the F_{20} generation) and chromosomes were 100 cM in length.
- The fourth scenario (Sc4) was used to quantify the effect of the level of inbreeding in the parents (i.e., parents were F_1, F_2, F_4, F_6, F_8, F_{10}, or F_{20}) of a biparental population where descendants were from the F_2 generation and chromosomes were 100 cM in length.
- The fifth scenario (Sc5) was used to quantify the effect of recombination rate (25, 50, 100, 150, 200, 300, or 400 cM) in a biparental population where parents were fully inbred (from the F_{20} generation) and descendants were from the F_2 generation.
- The sixth scenario (Sc6) was used to quantify the effect of the population type (biparental, backcross, or topcross) where parents were fully inbred (from the F_{20} generation), descendants were from the F_2 generation (Fig. 1), and chromosomes were 100 cM in length.

Analysis

Imputation was performed using the pedigree, as well as, high and low-density genotype information. In each case two or three parents had high-density genotypes and the 100 descendants had low-density genotypes. Accuracy of imputation was measured individually and summarized over all individuals within each scenario. Measures of the accuracy of imputation were the genotype and allele concordance. Genotype concordance was defined as posterior genotype probability $Pr\left(g_{i,j,k} \mid \text{data}\right)$ assigned to the true genotype, where $g_{i,j,k}$ is the *k*th genotype of the *j*th marker of the *i*th individual. For example, if posterior genotype probabilities were 0.05, 0.20, 0.75 for the respective Genotypes 0/0, 0/1, and 1/1 and the true genotype was 1/1, then genotype concordance was 0.75. If the true genotype would be 0/1, then genotype concordance would be 0.2. Allele concordance was defined as the weighted sum of posterior genotype probabilities, assigned to the genotypes that have alleles in common to the true genotype, with weights according to the number of alleles in common. For example, if posterior genotype probabilities are 0.05, 0.20, 0.75 for the respective Genotypes 0/0, 0/1, and 1/1 and the true genotype is 1/1, then allele concordance is $1/2 \times 0.20+1 \times 0.75 = 0.85$. If the true genotype would be 0/1, then allele concordance would be 0.6. Both genotype and allele concordance were obtained for each marker of each individual and then averaged over all markers of an individual to obtain individual specific measures.

The described definitions of genotype and allele concordance do not take into account that success of imputation depends, to a large extent, on allele frequency. For example, if the frequency of allele 1 is $p = 0.80$, then the expected (prior) genotype frequencies according to the Hardy–Weinberg equilibrium are $Pr(g = 0/0) = (1 - p)^2 = 0.04$, $Pr(g = 0/1) = 2(1 - p)\ p = 0.32$, and $Pr(g = 1/1) = p^2 = 0.64$. Therefore, imputing Genotype 1/1 gives a high chance of imputing the correct genotype (in 64% of cases) and the correct alleles (in 80% of cases) due to a high frequency of allele 1. To avoid overstating the success of imputation the metrics referred to as the "gain in genotype concordance" and "gain in allele concordance" were computed as the differences between genotype and allele concordance evaluated under the posterior and prior genotype probabilities. For example, if allele frequency was 0.8, posterior genotype probabilities were 0.05, 0.20, 0.75 for the respective Genotypes 0/0, 0/1, and 1/1, and the true genotype was 1/1, then gain in genotype concordance was $0.75 - 0.64 = 0.11$; and gain in allele concordance was $0.85 - 0.80 = 0.05$. If the true genotype would be 0/1, then gain in genotype concordance would be $0.20 - 0.32 = -0.12$; and gain in allele concordance would be $0.60 - 0.66 = -0.06$. Prior genotype probabilities were computed according to the Hardy–Weinberg equilibrium with allele frequencies computed from the high-density genotype data in parental lines according to the population type. Specifically, allele frequency for the *j*th marker was computed as $p_j = ½p_{A,j} + ½p_{B,j}$ for the biparental population, as $p_j = ¼p_{A,j} + ¾p_{B,j}$ for the backcross population, and as $p_j = ¼p_{A,j} + ¼p_{B,j} + ½p_{C,j}$ for the topcross population, where $p_{i,j}$ is allele frequency in the *i*th parental line at the *j*th marker.

RESULTS

The success of imputation was measured using the genotype and allele concordance at each marker, of each individual, and summarized per individual. To illustrate, an example of genotype and allele concordance, along the chromosome for three F_2 individuals, imputed from 3 low-density markers up to 1000 high-density markers, is shown in Fig. 2. The first individual inherited non-recombined chromosomes from inbred parents (Fig. 2, top). All markers except one were fixed for the opposing allele in the inbred parents, which implies that allele frequencies were 0.5 in the inbred parents. Imputing genotypes in the first individual based solely on the prior information would therefore give an a priori genotype concordance of 50%. Observing three low-density markers and performing imputation resulted in a posterior genotype concordance of 90%, a gain of 40%. Genotype concordance was high in the regions close to the observed marker positions and tended towards the prior genotype concordance in the regions between the observed markers. Gain in genotype concordance, due to imputation, was therefore at maximum 50% and tended toward 0% in the regions between the observed markers. For the fixed marker both the prior and posterior genotype concordances were 100% with gains of 0%. For segregating markers the allele concordance was higher (prior 75%, posterior 95%) than the genotype concordance as it is easier to impute at least one allele correctly than both alleles. However, the gain in allele concordance due to imputation (20%) was smaller than for genotype concordance (40%). The second individual inherited recombined chromosomes from both inbred parents, which lead to lower genotype concordance (81%) and allele concordance (91%) after imputation (Fig. 2, middle) than in the first individual that did not undergo recombination. However, in the second individual the recombination caused the imputation of genotypes based solely on the before be worse in comparison to the first individual. Therefore the gains in genotype concordance (46%) and allele concordance (30%) were higher for the second individual than for the first individual. The third individual inherited recombined chromosomes from outbred parents (Fig. 2, bottom). This situation provided very limited information for imputation, with genotype concordance at 43% and allele concordance at 66%. The gain in genotype concordance was only 2 and 0% in allele concordance. In comparison to the inbred parent example, there was as set of distinct values of genotype and allele concordances present in this outbred parent example (seen as "horizontal" lines; Fig. 2, bottom). These distinct values arise due to variable allele frequencies in outbred parents from locus to locus (0.00, 0.25, 0.50, 0.75, or 1.00) and variation in inherited alleles in progeny. For example, for the last marker the outbreed parents had Genotypes 0/0 and 0/1. This gives an allele frequency of 0.25 and

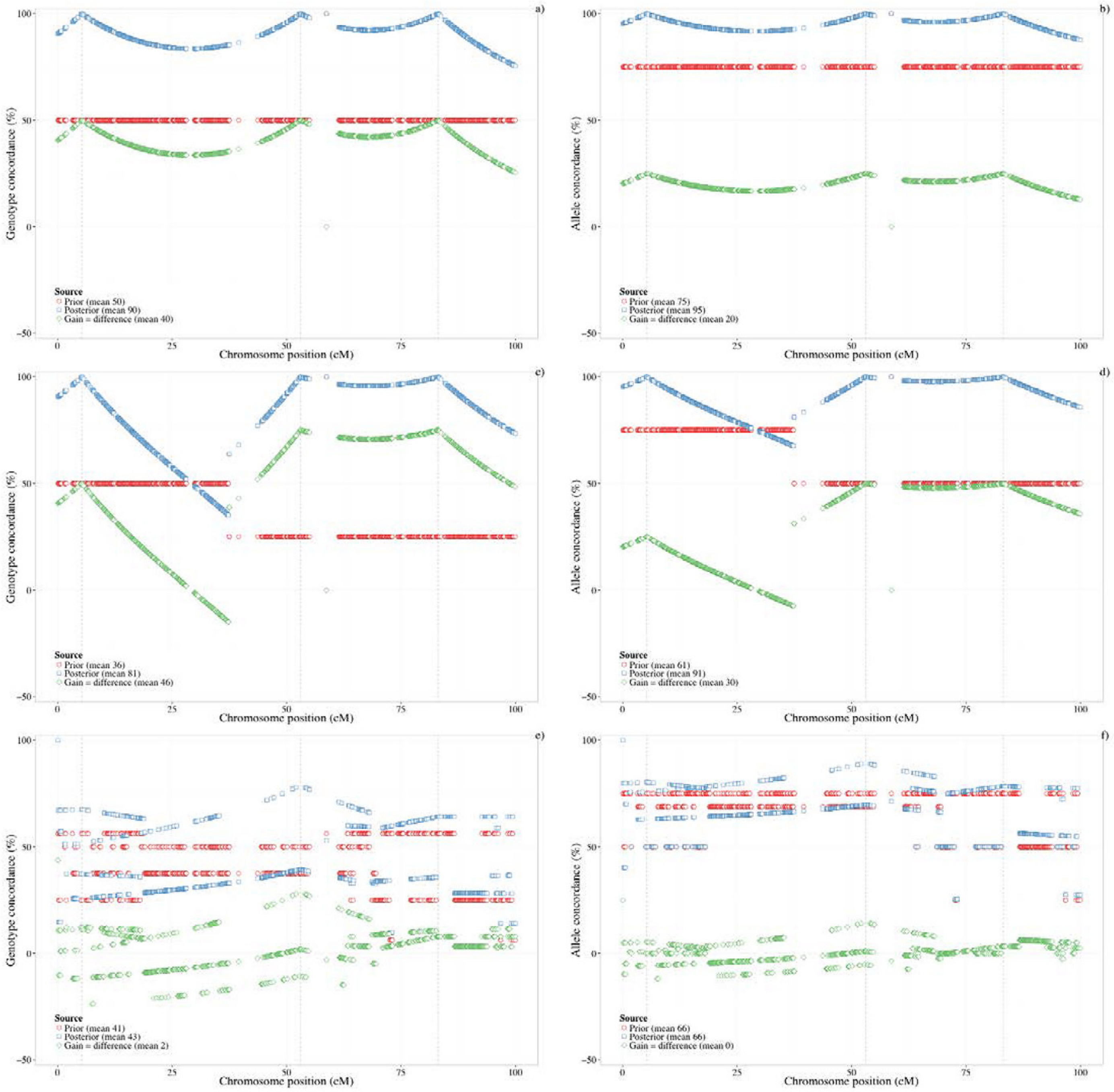

Figure 2. Accuracy of imputation measured with prior (circle), posterior (square), and gain (diamond) in genotype concordance (left–a, c, e) and allele concordance (right–b, d, f) along the chromosome for three F_2 individuals (top–inheriting non-recombined chromosomes from inbred parents, middle–inheriting recombined chromosomes from inbred parents, and bottom–inheriting recombined chromosomes from outbred parents); dashed vertical lines mark positions of observed low-density markers.

prior allele concordances of 0.75, 0.6875, and 0.25 for the respective Genotypes 0/0, 0/1, and 1/1. The progeny had 0/1 genotype so the prior allele concordance for this locus is 0.25. Other allele frequencies give other prior allele concordance values, but because there are only few possible allele frequencies when considering two parents, the number of different prior allele concordance values is limited (0.00, 0.25, 0.50, 0.6875, 0.75, and 1.00).

The results for Sc1 showed that the number of high-density markers to which the low-density genotypes were imputed did not affect the success of imputation. The allele concordance and gain in allele concordance when imputing to 1000, 5000, or 25,000 high-density markers were equal (Fig. 3), as were the values for genotype concordance and gain in genotype concordance (Supplemental Fig. S1). This is in line with the theory behind the statistical model.

The results for Sc2 showed that the level of inbreeding, in individuals who were to have genotypes imputed, affected the success of imputation. There was an interaction between this factor and the number of low-density

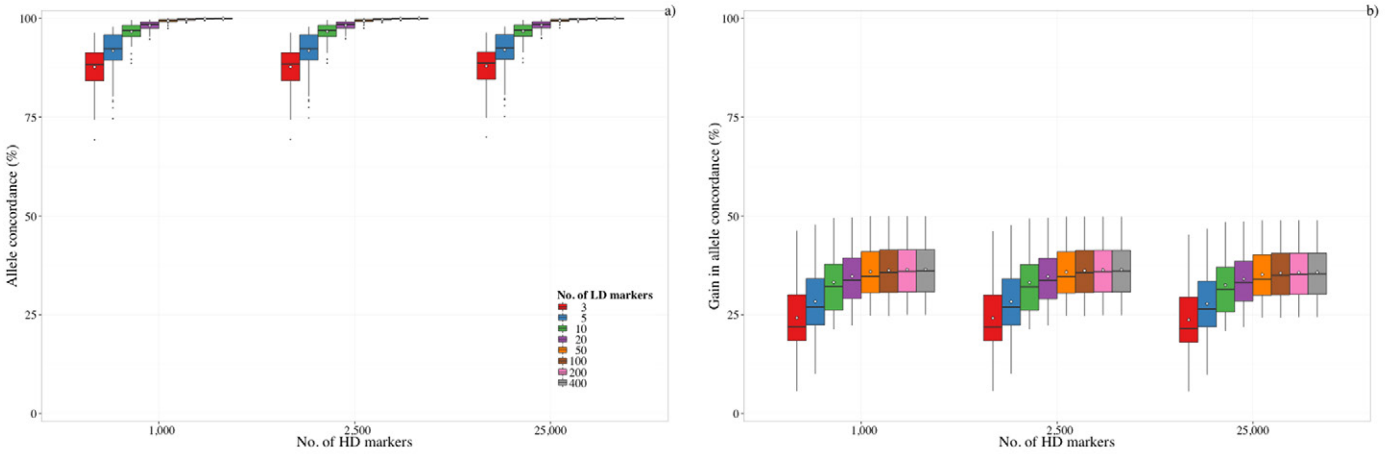

Figure 3. (a) Allele concordance per individual and (b) gain in allele concordance per individual when imputing to different numbers of high-density (HD) markers using differing numbers of low-density (LD) markers (Sc1).

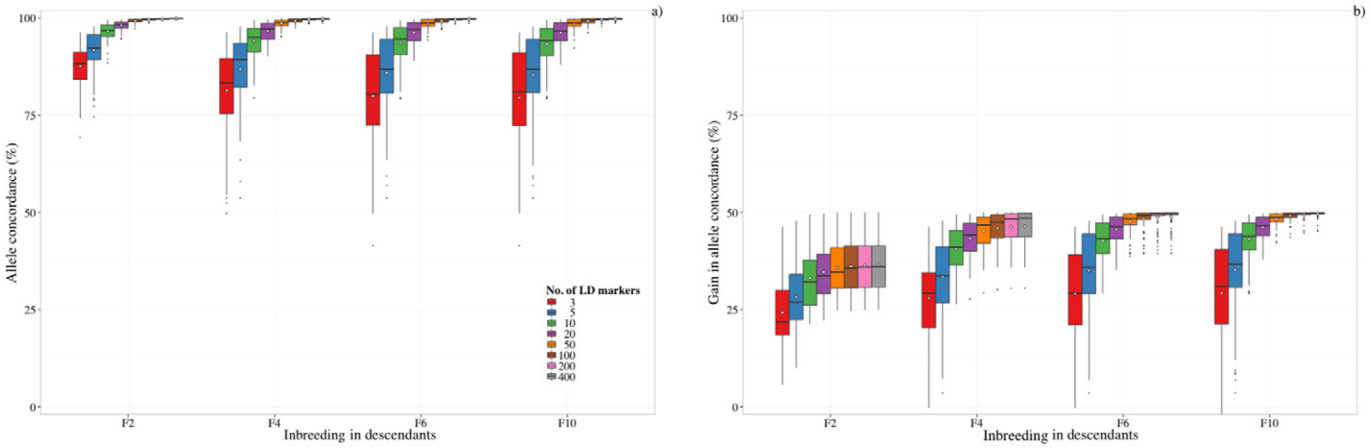

Figure 4. (a) Allele concordance per individual and (b) gain in allele concordance per individual when imputing individuals with different levels of inbreeding using differing numbers of low-density (LD) markers (Sc2).

markers (Fig. 4 and Supplemental Fig. S2). With many low-density markers (i.e., ≥100) almost perfect allele concordance could be obtained for F_2, F_4, F_6, or F_{10} individuals. With smaller numbers of low-density markers the imputation was less successful for more advanced generations than it was for early generations. For example, when imputing with a three marker array the median allele concordance dropped from 88% for F_2 to 81% for F_{10} individuals. The variance in the success of imputation was also greater for advanced generations and for lower numbers of low-density markers. When imputing with three marker arrays the standard deviation of allele concordance was 6% for F_2 individuals and 11% for F_{10} individuals. Within F_2 individuals the standard deviation of allele concordance dropped from 6% when imputing with three marker arrays to 0% when imputing with 100 or more marker array. Gain in allele concordance was at maximum 50% since parents were fully inbred. Interaction between the level of inbreeding and marker density was observed also for gain in allele concordance. Gain in allele concordance was lower and more variable in the early generations and increased with the more advanced generations. In the advanced generations the rate of increase in allele concordance and the rate of decrease of its variability was higher with higher marker densities. In the F_{10} individuals the gain in allele concordance was exactly the same as allele concordance (but decreased by a constrant of 50% owing to the way in which the statistic was calculated), while it was always more than 50% lower in the earlier generations.

The proposed selfing option in this study improved concordance of imputed genotypes in the F_4, F_6, and F_{10} individuals (Fig. 5). This effect was more evident for cases with sparse low-density marker maps, and especially so in the more advanced generations. In some cases use of the selfing option reduced genotype concordance. There was no improvement in allele concordance in general (Fig. 5), except for the F_4 and F_6 individuals, where the use of the selfing option decreased allele concordance, when very small numbers of low-density markers were used (Fig. 5, Supplemental Fig. S3).

The results for Sc3 showed that the level of intercrossing affected the success of imputation and that there was

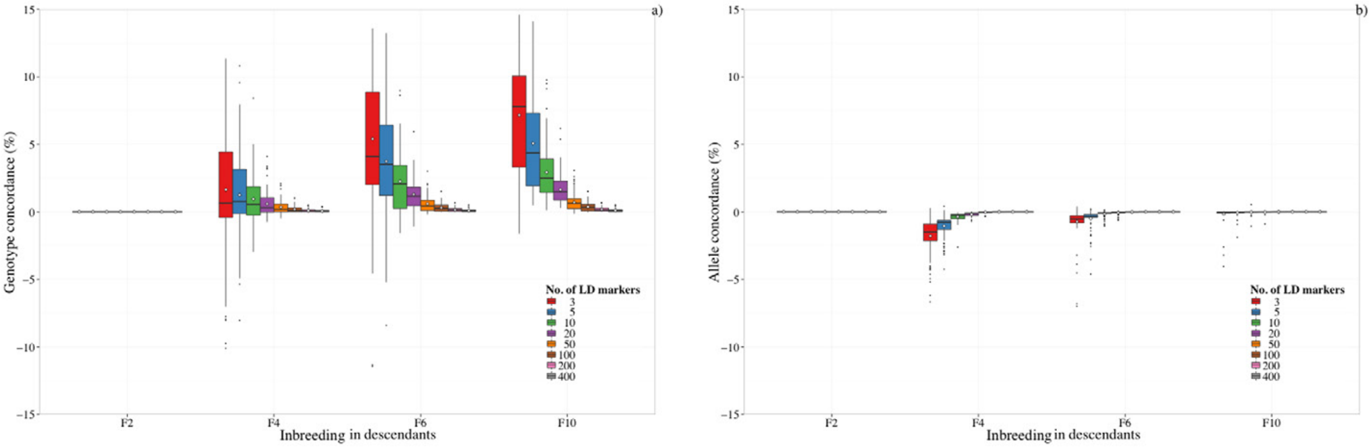

Figure 5. Improvement in (a) genotype concordance and (b) allele concordance per individual with explicit modeling of selfing in comparison to no modeling of selfing when imputing individuals with different levels of inbreeding using differing numbers of low-density (LD) markers (Sc2).

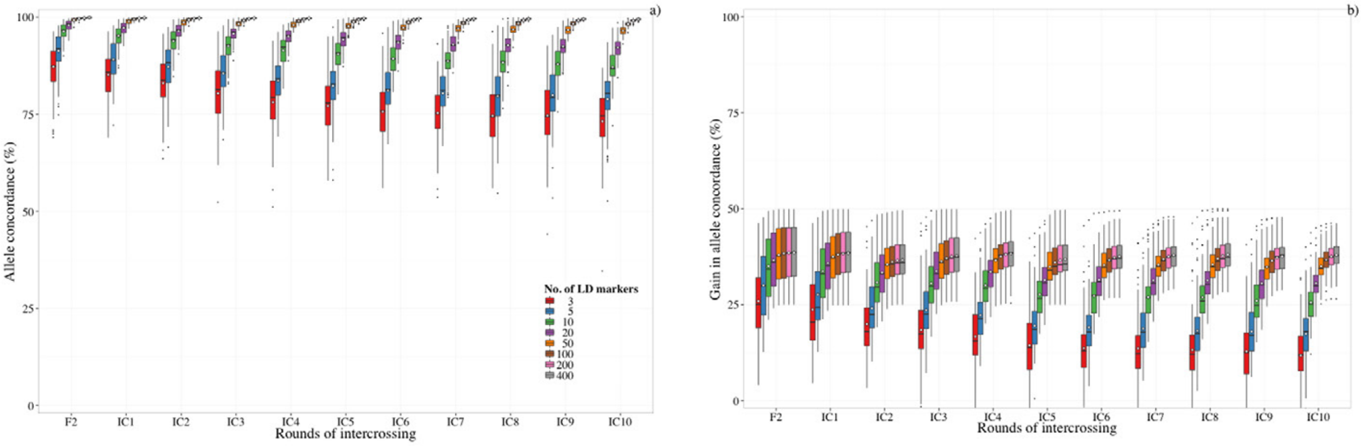

Figure 6. (a) Allele concordance per individual and (b) gain in allele concordance per individual when imputing individuals from different rounds of intercrossing using differing numbers of low-density (LD) markers (Sc3).

an interaction between the level of intercrossing and the number of low-density markers (Fig. 6 and Supplemental Fig. S4). The reduction in imputation success was almost linear with increasing rounds of intercrossing; the greatest affect being for the most sparse low-density marker scenarios. After 10 rounds of intercrossing the median allele concordance reduced to 75% when imputing with three markers. With 100 or more low-density markers the reduction in imputation success was minimal, even after 10 rounds of intercrossing. Compared to selfing, the reduction in imputation success with intercrossing was greater. When imputing with three markers the allele concordance for F_{10} individuals was 81%, whereas for 10 rounds of intercrossing it was only 75%. The variance of imputation success was lower for intercrossing than it was for selfing. When imputing with three markers the standard deviation of allele concordance for selfing at F_{10} was 13%, whereas for 10 rounds of intercrossing it was 8%. Gain in allele concordance generally followed the same pattern with larger variability than for allele concordance.

The results for Sc4 showed that as the level of inbreeding in the parents increased so did the success of imputation in their F_2 descendants and that there was an interaction between the number of low-density markers and this (Fig. 7 and Supplemental Fig. S5). With parents from inbreeding generation F_6 or later the specific degree of inbreeding had little impact on the success of imputation. When parents were F_1, F_2, or F_4 the imputation success was much lower. For example, when imputing with three markers the median allele concordance was 68% when the parents were F_1 but was 88% when parents were F_6 or more. Even with larger numbers of low-density markers it was not possible to achieve almost perfect allele concordance when parents were F_1, F_2, or F_4. With 400 markers median allele concordance was 77, 83, and 96% when parents were F_1, F_2, or F_4, respectively. In comparison median allele concordance with 400 markers was 100% when parents were F_6 or greater. Gain in allele concordance followed the same pattern as allele concordance with smaller values, but larger variability as already shown for the F_2 progeny of F_{20} parents in Sc2.

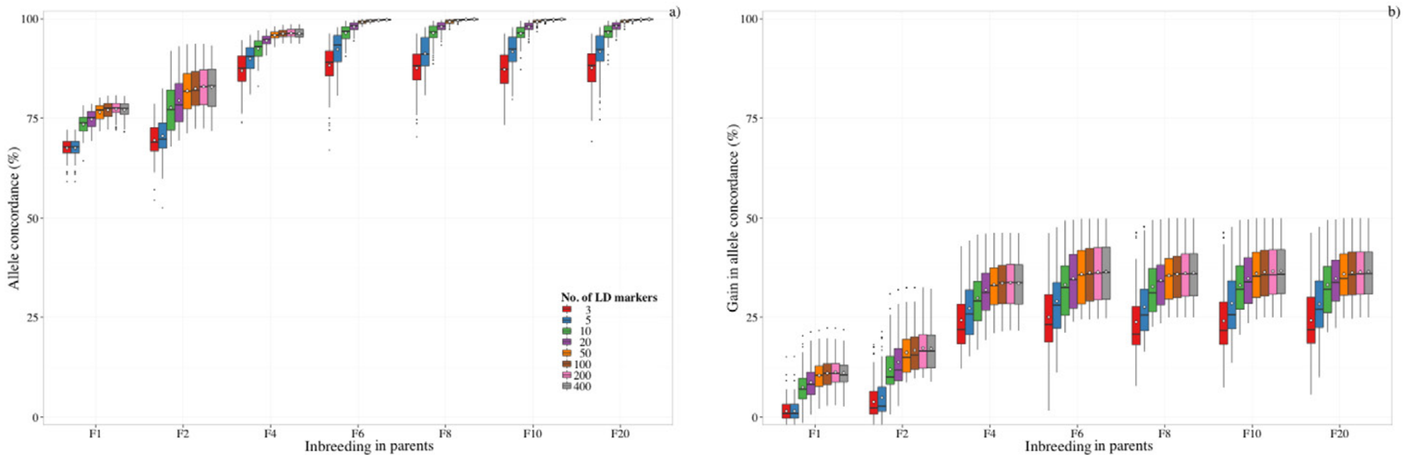

Figure 7. (a) Allele concordance per individual and (b) gain in allele concordance per individual when imputing individuals whose parents have different levels of inbreeding using differing numbers of low-density (LD) markers (Sc4).

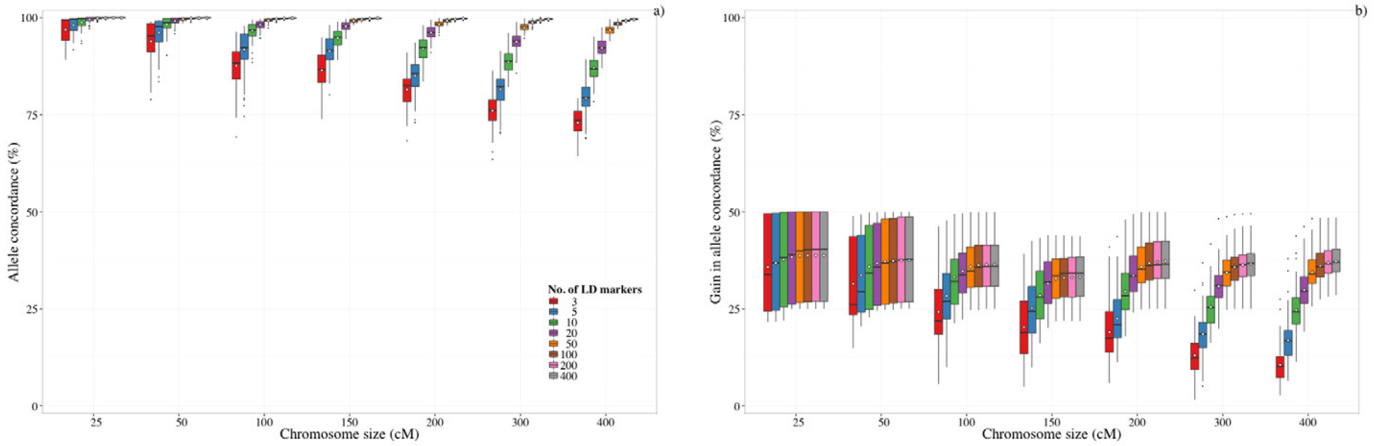

Figure 8. (a) Allele concordance per individual and (b) gain in allele concordance per individual when imputing chromosomes with different recombination rate using differing numbers of low-density (LD) markers (Sc5).

The results for Sc5 showed that as chromosome length increased more markers were required to achieve high levels of imputation success (Fig. 8 and Supplemental Fig. S6). However, the ratio of the number of low-density markers to the length of chromosome did not affect the level of imputation success. Long, intermediate, and short chromosomes required the same ratio to achieve any given level of imputation success. Gain in allele concordance was more variable than allele concordance, especially with shorter chromosomes. The number of low-density markers had little effect on this metric when chromosomes were short and the affect became stronger with increasing length of chromosomes.

The results for Sc6 showed that the population type (biparental, backcross, or topcross) affected the imputation success of descendants to a small degree when the number of low-density markers was low (Fig. 9 and Supplemental Fig. S7). Biparentals had the highest median allele concordance, followed by backcrosses, and finally topcrosses. The differences between biparentals and backcrosses, although consistent, were small, while the differences between these and topcrosses were noticeably greater. The variance in imputation success was greatest for backcrosses.

DISCUSSION

Using simulation, the performance of PlantImpute was evaluated across a wide range of scenarios and the results showed that imputation can be accurate in the types of populations that are typically made by plant breeders. The number of low-density markers contained in an array, level of inbreeding in the individuals to be imputed, level of inbreeding in their parents, genome length, and population type, affected imputation accuracy. The number of high-density markers to be imputed and the interaction between the number of low-density markers and chromosome length did not affect imputation accuracy.

The observation that marker density and genome length affected imputation accuracy was not surprising. However, it was surprising that the interaction between these two factors did not affect imputation accuracy in

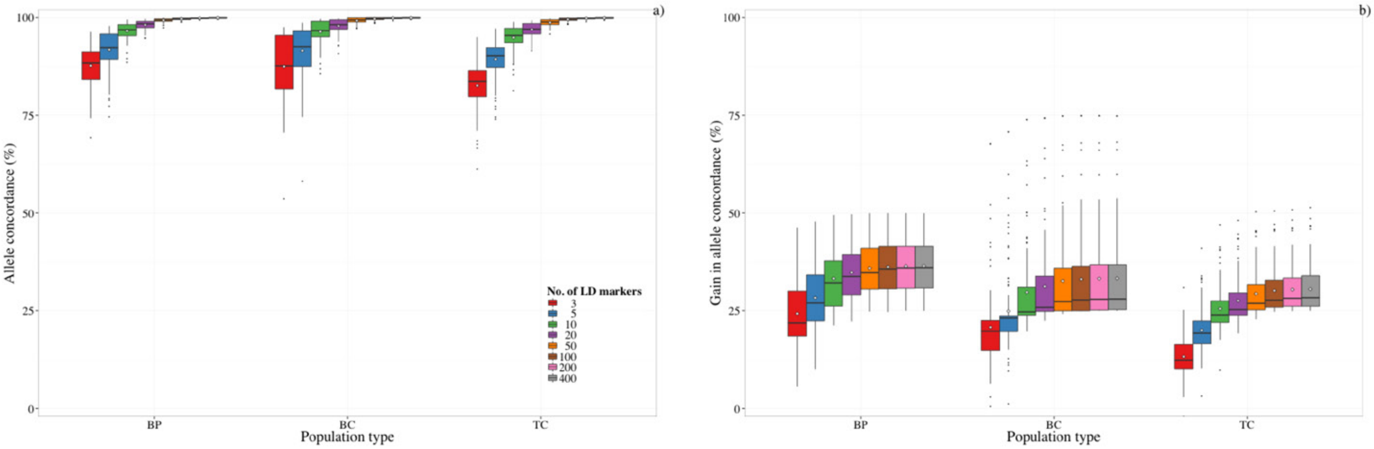

Figure 9. (a) Allele concordance per individual and (b) gain in allele concordance per individual when imputing individuals from biparental (BP), backcross (BC), and topcross (TC) populations, using differing numbers of low-density (LD) markers (Sc6).

the types of populations examined. We expected that these factors would interact to affect imputation accuracy because any imputation process can be expected to require a set of markers that serve as anchor points to broadly determine the combination of parental gametes on each gamete of an individual. Based on this expectation we envisaged that the markers positioned towards each end of a chromosome would be more important as anchor points than those in the middle, regardless of the chromosome length, thus meaning chromosomes that are of greater length would require proportionately less markers than shorter chromosomes. Perhaps an interaction between marker density and genome length was not observed because small chromosomes (i.e., 25 or 50 cM) have very few recombinations and therefore always have high imputation accuracy regardless of marker density. Such an interaction may be observed with unrealistically long chromosomes (e.g., >1000 cM).

The level of inbreeding in parents affected the imputation accuracy is in line with expectations. Scenarios in which the parents had lower levels of inbreeding showed lower imputation accuracy because the imputation process used in PlantImpute requires that the parents' haplotypes can be determined. PlantImpute looks strictly at the flow of alleles within the known pedigree, and thus cannot determine haplotype structure for markers only genotyped in individuals genotyped at high density. Inbred parents predominantly carry homozygous loci and thus the underlying haplotypes were phased de facto. In cases with lower inbreeding of founders, adding high-density genotyping, of a limited subset of offspring, could substantially improve overall imputation quality. Accurate phasing is central to imputation because once haplotypes of the high-density genotyped individuals are resolved, the process of imputation reduces to the tracking of which combinations of these haplotypes are carried by the individuals that are genotyped at low density (e.g., Hickey et al., 2011; Nettelblad, 2012). For the types of crosses used in this study the tracking of the combinations of these haplotypes is a relatively easy task because there are a small number of haplotype alleles (i.e., in a biparental population there are only two) at any locus, and there are a relatively small number of generations (i.e., opportunity for recombination) separating the individuals that are genotyped at high density and those genotyped at low density. That imputation success was lower with increasing rounds of selfing of the individual to be imputed and increasing rounds of intercrossing of the individual to be imputed is consistent with the increasing number of generations that separate the individuals genotyped at high density and at low density measured both with genotype or allele concordance. With the increasing number of generations there are more recombinations that break the parental haplotypes and thus larger numbers of low-density markers are required to achieve accurate imputation. Interestingly, although the genotype and allele concordance of imputation decreased with the increasing rounds of selfing of the individuals to be imputed, the gain in concordance due to imputation increased as the value of prior information (allele frequency in parents) diminished with each round of selfing. In other words, the value of imputation beyond naïve imputation based on allele frequencies in parents is increases with the increased rounds of selfing as drift and potentially also selection change allele frequency in individuals to be imputed. The same did not hold for the increasing rounds of intercrossing, because random selection and mating used in the simulation do not change allele frequency too much and the individuals to be imputed were quite heterogenous in comparison to the repeated rounds of selfing, which limits the power of imputation.

The pattern of change in imputation accuracy due to the type of population (i.e., biparental, backcross, and topcross) was consistent with expectations. Backcrosses had the highest imputation accuracy and topcrosses the lowest. Backcrosses have the lowest level of genetic variation among the descendants while topcrosses have the highest.

Imputation is expected to work better when the levels of genetic variation in a population are lower because there are fewer parental haplotypes and combinations of parental haplotypes present.

Imputation is an essential component of a cost effective plant breeding program that utilizes genomic information. The results of this study suggest that in those population types studied, accurate imputation can be achieved with a small number of markers, that is, one or two markers per 20 cM on average. These values translate to between 100 and 200 markers for a genome of 2000 cM in length and 200 and 400 markers for a genome of 4000 cM in length. In a practical plant breeding program imputation can be used to lower the cost of genotyping large numbers of individuals in a training set for genomic selection or to lower the cost of genotyping large numbers of individuals in the genomic selection prediction set, that is, the selection candidates. The results of this study suggest that one such strategy for achieving both, in practical breeding programs, could be to genotype all parents of crosses with high-density markers and any of their descendants that are to be training individuals or selection candidates at low density. Imputation can then be used to resolve high-density genotypes for all the individuals. Large training sets can then be constructed by accumulating training individuals from many crosses. Such an approach would facilitate genomic prediction across families (Bernardo and Yu, 2007; Heffner et al., 2011; Riedelsheimer et al., 2013; Hickey et al., 2014). The conclusion that sufficiently accurate imputation can be obtained with 100 and 200 markers for a genome of 2000 cM in length and 200 and 400 markers for a genome of 4000 cM in length is context dependent. In a commercial pig breeding, for example, levels of imputation accuracy similar to what were obtained in this study (Cleveland and Hickey, 2013) result in acceptable levels of genomic selection accuracies. Determining the value of particular levels of imputation accuracy in the context of genomic selection depends on where imputation is being used in the breeding program, the total available financial resources for that breeding program (including potential genotyping budget), the time horizon of the objectives, the overall design of the breeding program and its components, and the quantitative genetics of the traits being selected on. For example, the level of imputation accuracy may have different impacts when imputation is used in the training set or the prediction set and these impacts may different depending on the design of the training set (e.g., close vs. distant relatives or large vs. small training sets (Hickey et al., 2014)). Genotyping cost per individual and total genotyping budget interact to affect the total number of possible selection candidates and the total number of selection candidates affects the intensity of selection and therefore response to selection. Through these interactions a breeder has a choice in some cases between a higher selection intensity vs. a lower selection accuracy and determining the optimum point on this curve is not trivial (Gorjanc et al., 2015). Simple traits controlled by a small number of quantitative trait loci (QTL) with larger effects may have different properties to complex traits and imputation accuracy may have different impacts on the short- and long-term response to selection and rates of utilisation of genetic variance. Due to these complexities a complete evaluation of the impact of imputation accuracy on the outcome of genomic selection is beyond the scope of this study.

The specific nature of the studied population types allowed the derivation of a novel way to describe the success of imputation, which is composed of two components. Family average contributes to the success of imputation, where imputing genotype or allele probabilities based on allele frequencies in parents gives substantial imputation success with no genotyping of descendants; and, therefore, at virtually no extra cost, for example, a priori genotype and allele concordance in inbred parents that have the opposing alleles fixed are 50 and 75%, respectively. However, such imputation has no value for making within-family selection decisions, as segregation is not captured. It does, however, allow for better estimation of family mean and therefore comparison of families.

The second contribution to the success of imputation comes from capturing segregation beyond the expectation; which PlantImpute achieves by tracking the inheritance of parental haplotypes based on the observed low-density markers. The difference observed between the overall concordance and a priori concordance (gain in concordance) shows the accuracy with which imputation algorithms resolve the segregation of genotypes within families. This measure has a direct relationship to the concordance of imputed data and accuracy of evaluated Mendelian sampling terms for genomic prediction, that is, if the gain in allele concordance is 0% then accuracy of genomic prediction is only due to capturing family structure and no segregation within family is captured.

More work is required to evaluate the effect of different components of the success of imputation on the accuracy of genomic prediction. It should also be noted that PlantImpute does not attempt to optimize allele or genotype concordance, but rather the maximum likelihood of the true sequence given the posterior probabilities. This means that the selfing extension introduced into the model sometimes gives worse performance than the non-extended model according to the metrics used here, since it models more recombinations and therefore penalizes deviations from equilibrium in gaps between markers more strongly. If the goal was to only maximize genotype concordance with no probabilistic interpretation, the single genotype with the highest probability should always be chosen. If this is done, the selfing extension performs favorably. However, over a full population, consistently choosing the imputed genotype in such a manner could introduce undesirable biases and discontinuities. A simple example would be the case where all individuals are actually matching homozygotes at two adjacent markers in the map. If the distance between them is long enough, there would be a sharp transition at a specific distance in the interspersing region where the imputed genotype for all of them would suddenly turn to the heterozygote, due to the model asymptotically approaching Hardy–Weinberg equilibrium (or the relevant equilibrium at the chosen level of selfing), when lacking other data, making the heterozygote the most likely per-individual genotype. Thus, the imputed genotypes taken together would not correctly reflect that equilibrium, but rather a uniform set of exclusively heterozygotes, possibly impairing any downstream analysis and seriously disrupting statistics such as allele frequencies. If only maximum per-individual accuracy is desired, the imputed genotype should be chosen to be the most likely one.

The current version of PlantImpute does not perform well when the parents of a cross are not inbred, and only if parents have been genotyped with high density. The haplotype inference support in PlantImpute is only based on inference of recombination within the pedigree, and thus cannot exploit general shorter patterns of shared haplotypes within a population. With the marker densities simulated here, such patterns are expected to exist. For breeding programs that involve crossing outbred parents (e.g., apples [*Malus* spp.] or a genomic selection enabled rapid-cycling breeding program for cereals) several options exist. If some densely genotyped offspring were to be added, inference of parental haplotypes is possible. A pre-phasing of parents could also be done with other, population-based tools. The PlantImpute algorithm could be given those haplotypes both as a fixed truth, or as an initial value, to be refined iteratively, augmented by the haplotype inference that is made possible by a limited subset of high-density offspring. If genotyping by sequencing approaches are used, for example, the linkage information made available from markers present within the same read could also be employed by this method.

The current implementation of PlantImpute is computationally intensive. For the basic scenario of imputing up to 1000 markers the imputation took ~3 h during which memory consumption was ~4 GB. Increasing the size of the HD array to 25,000 markers increased the memory consumption to more than 100 GB.

CONCLUSIONS

This study shows that the success of imputation is affected by many factors including the number of low-density markers per cM, level of inbreeding or intercrossing of the individuals to have genotypes imputed, level of inbreeding of the parents of a cross, and genome length; but not by the number of high-density markers or by the interaction between the genome length and the number of low-density markers. Effective imputation was possible with as few as one or two markers per 20cM genotype imputation when parents were inbred. Therefore, the results of this study show that genotyping strategies in which inbred parents of a cross are genotyped at high-density and their descendants are genotyped with 200 to 400 markers genome wide may be cost effective and useful in practical plant breeding programs that utilize genomic selection.

Supplemental Information Available

Supplemental information is available with the online version of this manuscript.

Acknowledgments

J.M. Hickey and G. Gorjanc received support from a BBSRC strategic grant to The Roslin Institute, from The Seeds of Discovery, and from ICRISAT. C. Nettelblad and computational infrastructure work was supported by the Swedish Research Council, the Knut and Alice Wallenberg Foundation, the European Research Council, the Röntgen-Ångström Cluster, and Stiftelsen Olle Engkvist Byggmästare. Some development work done by C. Nettelblad was also performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project snic2014-3-30. Editing and review of the manuscript was completed by S.M. Cleveland.

References

Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47:1082. doi:10.2135/cropsci2006.11.0690

Broman, K.W., H. Wu, Ś. Sen, and G.A. Churchill. 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890. doi:10.1093/bioinformatics/btg112

Browning, S.R., and B.L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097. doi:10.1086/521987

Chen, G.K., P. Marjoram, and J.D. Wall. 2009. Fast and flexible simulation of DNA sequence data. Genome Res. 19:136–142. doi:10.1101/gr.083634.108

Cleveland, M.A., and J.M. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J. Anim. Sci. 91:3583–3592. doi:10.2527/jas.2013-6270

Gorjanc, G., M.A. Cleveland, R.D. Houston, and J.M. Hickey. 2015. Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet. Sel. Evol.

Haley, C.S., and S.A. Knott. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324. doi:10.1038/hdy.1992.131

Heffner, E.L., J.-L. Jannink, and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Gen. 4:65–75. doi:10.3835/plantgenome.2010.12.0029

Hickey, J.M., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu, B.M. Prasanna et al. 2014. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci. 54:1476–1488.

Hickey, J.M., and G. Gorjanc. 2012. Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. G3:Genes, Genomes, Genetics 2:425–427.

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

Howie, B.N., P. Donnelly, and J. Marchini. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5(6):E1000529. doi:10.1371/journal.pgen.1000529

Li, Y., C.J. Willer, J. Ding, P. Scheet, and G.R. Abecasis. 2010. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34:816–834. doi:10.1002/gepi.20533

Nettelblad, C. 2012. Inferring haplotypes and parental genotypes in larger full sib-ships and other pedigrees with missing or erroneous genotype data. BMC Genet. 13:85. doi:10.1186/1471-2156-13-85

Nettelblad, C., S. Holmgren, L. Crooks, and Ö. Carlborg. 2009. cnF2freq: Efficient determination of genotype and haplotype probabilities in outbred populations using Markov models. In: S. Rajasekaran, editor, Bioinformatics and computational biology. Lecture Notes in Computer Science. Springer, Berlin. p. 307–319.

Rabiner, L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77:257–286. doi:10.1109/5.18626

Riedelsheimer, C., J.B. Endelman, M. Stange, M.E. Sorrells, J.-L. Jannink, and A.E. Melchinger. 2013. Genomic predictability of interconnected biparental maize populations. Genetics 194:493–503. doi:10.1534/genetics.113.150227

Sargolzaei, M., J.P. Chesnais, and F.S. Schenkel. 2011. FImpute-An efficient imputation algorithm for dairy cattle populations. J. Dairy Sci. 94 (E-Suppl. 1): 421.

VanRaden, P.M., J.R. O'Connell, G.R. Wiggans, and K.A. Weigel. 2011. Genomic evaluations with many more genotypes. Genet. Sel. Evol. 43:10. doi:10.1186/1297-9686-43-10