Molecular Breeding manuscript No.
(will be inserted by the editor)

Marker imputation efficiency for Genotyping-By-Sequencing data in rice (*Oryza sativa*) and alfalfa (*Medicago sativa*)

Nelson Nazzicari · Filippo Biscarini · Paolo Cozzi · E. Charles Brummer · Paolo Annicchiarico

Received: / Accepted:

Abstract Genotyping-by-sequencing (GBS) is a rapid and cost-effective genome-wide genotyping technique applicable whether a reference genome is available or not. Due to the cost-coverage trade-off, however, GBS typically produces large amounts of missing marker genotypes, whose imputation becomes therefore both challenging and critical for later analyses.

In this work, the performance of four general imputation methods (K-Nearest Neighbors, Random Forest, Singular Value Decomposition, and Mean Value) and two genotype-specific methods ("Beagle" and FILLIN) were measured on GBS data from alfalfa (*Medicago sativa L.*, autotetraploid, heterozygous, without reference genome) and rice (*Oryza sativa L.*, diploid, 100% homozygous, with reference genome). Alfalfa SNP were aligned on the genome of the closely

Nelson Nazzicari and Filippo Biscarini contributed equally to the work
Corresponding author: Nelson Nazzicari E-mail: nelson.nazzicari@crea.gov.it

N. Nazzicari, P. Annicchiarico
Council for Agricultural Research and Economics (CREA) Research Centre for Fodder Crops and Dairy Productions, Lodi (Italy)

F. Biscarini, P. Cozzi
Dipartimento di Bioinformatica, Fondazione Parco Tecnologico Padano, Lodi (Italy)

E.C. Brummer
University of California, Plant Sciences Department, Davis, CA (USA)

related species *Medicago truncatula L.*. Benchmarks consisted in progressive data filtering for marker call-rate (up to 70%) and increasing proportions (up to 20%) of known genotypes masked for imputation. The relative performance was measured as the total proportion of correctly imputed genotypes, globally and within each genotype class (two homozygotes in rice, two homozygotes and one heterozygote in alfalfa).

We found that imputation accuracy was robust to increasing missing rates, and consistently higher in rice than in alfalfa. Accuracy was as high as 90-100% for the major (most frequent) homozygous genotype, but dropped to 80-90% (rice) and below 30% (alfalfa) in the minor homozygous genotype. Beagle was the best performing method, both accuracy- and time-wise, in rice. In alfalfa, KNNI and RFI gave the highest accuracies, but KNNI was much faster.

Keywords SNP, Genotyping by Sequencing (GBS), K-Nearest Neighbors Imputation (KNNI), Random Forest Imputation (RFI), Singular Value Decomposition Imputation (SVDI), Beagle, FILLIN, alfalfa, rice, imputation, reference genome

1 Introduction

Imputation of missing alleles and genotypes is a preliminary step for a wide range of genetic analyses. In fact most models and software for population genetics, genomic selection (GS) and genome-wide association studies (GWAS) can not easily handle missing data and require complete datasets (e.g. [13, 2,11,29]). SNP-array represent one of the leading genotyping technologies and produce datasets that, after low call-rate filtering, usually still contain a small proportion of uncalled genotypes (e.g., <5%) randomly distributed along the genome. Genotyping-by-sequencing (GBS) is a relatively recent technique which is considered a viable alternative to SNP array-based genotyping to

produce SNP genotype data [10]. GBS is platform-independent and is conveniently used in species that lack commercial SNP-chips or even lack a reference genome sequence. GBS data typically present a much larger proportion of sporadic missing genotypes: e.g., $\sim$52% and $\sim$ 58% average in two maize experimental populations [9], or "up to 80% missing data per marker" in wheat [31]. This is due to both intrinsic properties of the technology and to its application to species without a mature genome assembly [12].

There are several methods that are routinely used for the imputation of missing genotypes. Some rely solely on genotypic information, some make use also of genealogies, most of them are based on reconstructing haplotypes to be used in predictive models [27]. Most methods for genotype imputation have been applied to SNP array data, and typically yield very high imputation accuracy. For instance, >95% correctly imputed genotypes were reported in maize [15] and cattle [43]. The same imputation methods can also be applied to GBS data (see [16] and [39] for applications in rice and maize). When a reference genome is available, SNP loci can be aligned against its sequence and are therefore ordered, thus allowing the exploitation of local linkage disequilibrium (LD). However, GBS data may be generated also for species lacking a reference genome assembly. In this case the GBS output is a series of "floating" loci not linked to any chromosome: alignment is not possible and SNPs are therefore considered unordered. Unordered data add additional complexity to genotype imputation. Exploitation of linkage disequilibrium and haplotype reconstruction are less straightforward, consecutive loci may lie on different chromosomes or scaffolds, and overall data are less homogeneous. Rutkoski et al. [34] considered the application of general data imputation methods to wheat, maize and barley populations, obtaining best case scenarios accuracies as high as 0.84-0.94 (measured as R^2 between true and imputed genotypes). Imputation of missing GBS data without alignment to the reference genome

has been reported in alfalfa (*Medicago sativa L.*, [33]) and red raspberry (*Rubus idaeus L.*, [44]). These studies however do not focus on measuring imputation accuracy and do not compare alternative imputation methods.

Compared to SNP-array data, GBS is more challenging: many more missing genotypes, high variability between runs, intrinsically noisy data, reads from different loci that can overlap. For all these reasons, imputation of missing genotypes with GBS data is still a relatively immature technique, not yet amenable to produce highly standardised and repeatable results. There is therefore scientific and practical interest in gaining further insights into how genotype imputation with GBS data works, and in developing the best methods and strategies to accurately and efficiently impute such missing data. This would be relevant not only for the scientific community but also for the breeding industry, because of its direct consequences on genomic applications.

In this paper, we imputed missing genotypes from GBS data in two agronomically important crop species: the cereal crop rice (*Oryza sativa L.*) whose reference genome has been assembled in 2005 [17], and updated in 2013 [18], and the forage crop alfalfa (*Medicago sativa L.*), for which a reference genome is not available yet. Alfalfa genotypes were mapped on the close relative *Medicago truncatula L.*, a model species for legumes, whose genome is available [45]. The species were chosen as representative of very different scenarios, alfalfa being autotetraploid with high heterozygosity and rice being diploid with essentially 100% homozygozity. The present study thus encompasses a wide span of real application cases.

We compared the imputation accuracy of four general imputation methods (Mean Value Imputation, K-Nearest Neighbor Imputation, Singular Value Decomposition Imputation and Random Forest Imputation) with the performance of two methods specific for genotype imputation (localized haplotype clustering imputation, implemented in the software package Beagle, and the

method of based on haplotype reconstruction around recombination sites implemented in the FILLIN algorithm as part of the Tassel suite). The four general algorithms were chosen as well known imputation strategies implemented in several freely available software libraries. The two genotype-specific methods represent the state of the art for genotypes lacking pedigree information.

The relative efficiency of the different imputation methods was assessed in terms of accuracy and computation time. Accuracy was measured as fraction of correctly imputed genotypes, both as a total and within genotype classes (major/minor homozygotes, and heterozygotes, when present).

2 Materials and methods

2.1 Plant material

Samples from two autotetraploid alfalfa (*Medicago sativa L.*) and one diploid rice (*Oryza sativa L.*) populations were available (see Table 1).

Alfalfa populations included elite germplasms from the Po Valley (Alfalfa-PV) and Mediterranean-climate environments (Alfalfa-Med). For Alfalfa-PV, 124 parent genotypes were chosen by stratified mass selection for dry matter yield over three harvests. For Alfalfa-Med, 154 parent genotypes were derived from two cycles of free intercrossing among three outstanding populations in a multi-environment study. For further details see [1].

The Rice dataset included 437 plants belonging to 391 accessions (46 duplicates) from the Rice Germplasm Collection maintained at CRA-Rice Research Unit (Vercelli, Italy). The sampled collection comprised accessions from the five main sub-populations of *O. sativa* (274 temperate japonica, 108 tropical japonica, 28 indica, 16 aus and 11 aromatic). All accessions were purified

through single seed descent, and were genotypically essentially 100% homozygous.

2.2 Genotyping by sequencing

The 715 alfalfa (Alfalfa-Med and Alfalfa-PV) and rice (Rice) samples were genotyped using the genotyping-by-sequencing (GBS) approach [10]. Different GBS protocols were used to genotype alfalfa and rice, since genotyping was carried out under different projects.

In alfalfa, DNA was isolated from fresh leaf tissues by the Wizard ® Genomic DNA Purification Kit (Promega, A1125) and quantified with a Quant-iT PicoGreen dsDNA assay kit (Life Technologies, P7589). Two libraries were constructed for the two populations, where 100 ng of each DNA was digested with ApeKI (NEB, R0643L) and then ligated to a unique barcoded adapter and a common adapter. For the reference population Alfalfa-Med, 5 nmoles each of the primers and NEB 2X Taq Master Mix (NEB Cat # M0270S) were included in the PCR reaction according to [10] original protocol, while for Alfalfa-PV the KAPA library amplification readymix (Kapa Biosystems Cat # KK2611) was used instead. Each library was sequenced in two lanes on Illumina HiSeq 2000 at the Genomic Sequencing and Analysis Facility at the University of Texas at Austin, Texas, USA. For both populations post sequencing analysis and SNP calling was carried out using Tassel UNEAK pipeline [24].

Alfalfa is an autotetraploid plant species and can therefore present three different heterozygous genotypes: AAAB, AABB and ABBB. Sequencing depth in this study was not sufficient for accurate tetraploid allelic dosage, but following [23] and [22] reliable genotype calls based on diploid allelic dosage were obtained considering diploid heterozygotes (i.e. AB), while the two ho-

mozygous genotypes (AAAA and BBBB) were considered diploid homozygotes (i.e., AA or BB). A further quality filter, implemented through ad-hoc Python scripts, removed heterozygous loci with less than 4 and homozygous loci with less than 11 aligned reads. A similar filtering was performed in [33] using less restrictive thresholds.

In rice, DNA was isolated from three-weeks old leaves using the DNeasy Plant Mini Kit (QIAGEN) with a TECAN Freedom EVO150 liquid handling robot (TECAN Group Ltd, Switzerland). DNA digestion was performed on 100 ng samples in 96-well plates using ApeKI, which was shown to cut every 1 kb on average in a in-silico digestion of the Nipponbare reference genome, and 96-plex libraries constructed according to the GBS protocol. The libraries were loaded into Genome Analyzer II (Illumina, Inc., San Diego, CA) for sequencing. Raw sequence data filtering, sequence alignment to the rice reference genome (Os-Nipponbare-Reference-IRGSP-1.0, [18]), and SNP calling from GBS genotyping were carried out using the Tassel GBS pipeline [12].
In total, 32 706, 40 734 and 166 418 SNP markers were called by GBS in Alfalfa-PV, Alfalfa-Med and Rice, respectively (see Table 1). In all datasets SNP variants were renamed so that AA represented the most common homozygote, and BB the least common homozygote.

2.3 Imputation methods

We considered six imputation methods: Mean Value Imputation (MNI), K-Nearest Neighbors Imputation (KNNI), Random Forest Imputation (RFI), Singular Value Decomposition Imputation (SVDI), Localized Haplotype Clustering Imputation ("Beagle") and haplotype reconstruction around the recombination sites (FILLIN from Tassel suite). For all algorithms we imputed a $n \times m$ matrix of n individuals and m markers whose data points, defined

in the set {0,1,2,NA}, represented the three possible genotypes (AA, AB, and BB) and the missing value, respectively.

MNI : in Mean Value Imputation each missing data point was replaced by the mean of the non-missing values for that marker, then discretized to the closest value in {0,1,2}.

KNNI : in K-Nearest Neighbors Imputation missing data points were imputed based on the weighted average of the K closest markers [41] defined by the Simple Matching Coefficient distance function [35], specifically designed for categorical data. $K = 4$ was used in KNNI, and neighbors values were weighted by the reciprocal of their distance from the data point to be imputed.

SVDI : Singular Value Decomposition (SVD) imputation was based on the following factorization of the genotype matrix M (n individuals, m markers):

$$M = U\Sigma V^T \tag{1}$$

where U is a $n \times n$ unitary matrix (i.e. $UU^T = I$), Σ is a $n \times m$ rectangular diagonal matrix of singular values and V^T is the $m \times m$ conjugate transpose of the unitary matrix V. The columns of matrix U and matrix V contain the eigenvectors of $MM^T_{(n\times n)}$ and $M^T M_{(m\times m)}$, respectively, and the corresponding non-zero singular values in Σ are equivalent to the square-root of the non-zero eigenvalues of MM^T and $M^T M$. The first k eigenvectors in U - ordered by decreasing eigenvalue (from Σ) - capture most of the information in the marker genotype matrix M, and were used to generate linear combinations (principal components) of the original m markers for the imputation of missing data points. The imputation procedure comprised: (i) initial imputation of missing genotypes using MNI, since SVD can only be performed on

complete matrices; (ii) SVD to select the most informative k eigenvectors of the marker genotype matrix; (iii) these k eigenvectors were used as predictors in a linear regression model for marker genotypes:

$$Y = U^*\beta + \varepsilon \tag{2}$$

with Y as vector of n genotypes at marker j; U^* the $n \times k$ matrix of the first k eigenvectors; β the vector of k regression coefficients; ε the random error terms; (iv) all eigenvectors (matrix U^*) and β were used to estimate missing values at marker j. The procedure was repeated (steps (ii) to (iv)) until convergence. The final imputed data points were then discretized to the closer genotype in {0,1,2}. A value of $k = 4$ for the eigenvectors to be selected in U^* was used in our implementation of SVDI, based on empirical results in the range 3–20 (data not shown). Additional details on SVDI can be found in [41]

RFI : in Random Forest (RF) imputation, missing genotypes at marker j were imputed by means of RF multiple regression trees [5] where all markers other than j were used for the prediction. At each marker j, 100 RF regression trees $\Theta_1 \dots \Theta_{100}$ were grown from a bootstrapped sample of the individuals in Y and a random subset x of $\sqrt{m-1}$ markers. Missing genotypes were imputed averaging predictions over the 100 RF trees:

$$\hat{Y} = \frac{1}{100}\sum_{i=1}^{100} h(x, \Theta_i) \tag{3}$$

where $h(x, \Theta_i)$ is a function of the i_{th} tree and subset of markers. RFI was repeated until convergence or for maximum 10 iterations. After regression, the imputed data were then discretized to the closer genotype in {0,1,2}.

Beagle : "localized haplotype clustering imputation" is a method implemented in the software "Beagle" [7]. Originally developed for human genetics, Beagle has since found wide application also in animal and plant genetics. Beagle has become so popular that the name of the software is commonly used to refer metonymically to the method it implements, making the two hardly distinguishable. Beagle infers haplotypes and imputes sporadic missing alleles both with known and unknown phase, using a localized haplotype cluster model. This is a class of directed acyclic graphs which empirically models haplotype frequencies on a local scale and therefore adapts to local structure in the data. Beagle makes use of a hidden Markov model to find the most likely haplotype pair for each individual, given the genotype data for that individual and the graphical haplotype frequency model. The method works iteratively using an expectation-maximization (EM) approach. The imputed missing data, probabilities of missing genotypes and inferred haplotypes are calculated from the model that is fitted in the last iteration. For the imputation of missing genotypes in the completely homozygous rice accessions, a likelihood file (with the prior likelihood of each of the three possible SNP genotypes) was defined to specifically exclude the imputation of non-possible heterozygous genotypes (details in [8]).

FILLIN The "Fast Inbred Line Library ImputatioN" ("FILLIN") is an imputation method optimised for inbred populations implemented in the "Tassel" programming suite [39]. FILLIN is based on haplotype reconstruction around recombination break points. Haplotypes are clustered per genotype similarity using the Hamming distance function. This information is eventually used to impute the target locus in an iterative approach that attemps, through a Markov Chain MonteCarlo (MCMC) process, to maximise the likelihood of the observed SNP calls given the unobserved imputed genotypes. If convergence

criteria are not met genotypes are left uninputed. Among the considered algorithms, FILLIN is the only one that can thus produce partially imputed results.

2.4 Imputation procedure

In order to assess the imputation performance of the different methods, we introduced increasing proportions of artificial missing genotypes in the data sets. These were then imputed with the various algorithms, measuring the resulting imputation accuracy and computation time. From the overall data, we extracted four datasets that had a maximum of 10%, 20%, 40% or 70% missing data per marker. For each of these datasets, we randomly introduced an additional 1%, 5%, 10% or 20% missing genotypes, on which imputation accuracy could be measured. Table 1 reports the number of markers and average missing rate (before introducing artificial missing genotypes) for the four call-rate thresholds. This procedure produced 16 combinations for each population: 4 call rates $\times$ 4 levels of artificial missing markers. On each combination, we applied the 6 described imputation methods. Rice data were analysed as a whole and by each of the 12 chromosomes separately. In absence of the alfalfa genome, the *M. Truncatula* genome was used as reference for those algorithms (Beagle and FILLIN) requiring marker position.

To further investigate the effect of presence (or absence) of genome information we generated "reshuffled" datasets where marker positions were randomly assigned inside each chromosome. This way, the linkage disequilibrium between markers is broken and the relative performance decrease can be directly assessed. The reshuffled datasets were tested only on Beagle and FILLIN.

2.5 Imputation accuracy and efficiency assessment

We used the artificial missing genotypes to measure the performance of the six imputation methods. The rice dataset contained only two genotype classes: AA (homozygous for the major allele) and BB (homozygous for the minor allele). Both alfalfa datasets contained also the third genotype classes AB (heterozygous, pooling the three possible heterozygotes AAAB, AABB, ABBB).

For each experiment we measured the overall imputation accuracy and the imputation accuracy within each genotype class. The imputation accuracy was computed as the number of missing data correctly imputed divided by the total number of artificially missing data (proportion of correct imputations):

$$accuracy = \frac{1}{n}\sum_{i=1}^{n} I(g_i = \hat{g}_i) \tag{4}$$

where $I()$ is an indicator function that returns 1 if the imputed ($\hat{g}$) and true (g) genotypes are equal, 0 otherwise. We obtained four accuracy measures for alfalfa and three for rice.

For each imputation method, the total computation time needed to complete the analyses was measured as an indicator of their relative performance. The measured time equals the total dedicated CPU time in case of single thread execution, and corresponds to a fraction of it in case of algorithm supporting multi thread execution. In our experiments only RFI implemented parallel execution, and it was configured to use 10 CPUs: thus, all RFI reported times are to be considered as time elapsed while employing ten times more computational resources compared to the other algorithms.

To ensure consistency in the experimental conditions, all analysis were run on the same computing platform at PTP Science Park (www.ptp.it): a multi-

node cluster with 64 CPUs AMD Opteron(tm) 2400 MHz and 256GB RAM for each node.

2.6 Software

Data handling, editing and preparation, summary of results and plots were all performed using the open-source environment for statistical programming R [32]. All imputation methods except Beagle and FILLIN were implemented in R: MNI directly in base R; KNNI using the function *knncatimpute()* from the Scrime package [36]; SVDI with the R package "bcv" [30]; RFI using the "MissForest" [37] R package (with parameters: ntree=100, maxiter=10, parallelize='variables'). The "Beagle" localized haplotype clustering imputation method was implemented with the Beagle software version 3.3.2 [7]. The FILLIN algorithm [39] is implemented using the Tassel CLI plugin *FILLINFindHaplotypesPlugin* followed by *FILLINImputationPlugin*. The latter allowed accuracy measurements through *-accuracy* and *-propSitesMask* options. The Bowtie 2 tool [20] was used to query the consensus sequence of each alfalfa tag pair containing a SNP against the *M. truncatula* reference genome Version 4.0 using the *-verysensitivelocal* preset. The BWA alignment tool [21] was used to align rice tags on Rice Genome Annotation Project Release 7.

3 Results

3.1 Genotypes

GBS detected 32 706, 40 734 and 166 418 SNP loci in Alfalfa-PV, Alfalfa-Med and Rice, respectively (Table 1). The overall missing rate was 66.6%, 59.6% and 53.4% in the three datasets. Supplementary Figure S1 shows the density distribution of missing rate per marker in the complete datasets. The amount

of missing marker genotypes varied with the threshold of maximum per-marker missing-rate allowed in the data (10%, 20%, 40% and 70%) and the proportion of artificial missing genotypes that were introduced (1%, 5%, 10% and 20%). Missing data point counts ranged from a minimum of 3 878 in Alfalfa-PV with 10% allowed and 1% artificial missing genotypes, to a maximum of 5 580 616 in Rice with 70% allowed and 20% artificial missing genotypes. Over all datasets and allowed/artificial missing genotypes thresholds, the amount of missing data points to be imputed averaged 249 405.

Minor Allele Frequency (MAF) was 0.172, 0.170 and 0.140 in Alfalfa-PV, Alfalfa-Med and Rice, respectively. The three datasets were therefore unbalanced with respect to genotype classes: Rice had the lowest MAF, while Alfalfa-PV and Alfalfa-Med had the smallest minor classes (minor homozygotes 3.5% and 3.2% of the total genotypes). Figure 1 reports the proportion of genotypes in each class in the complete datasets (100%) and as a function of the maximum allowed missing rate per marker. The relative proportion of genotype classes remained approximately stable in rice. In contrast, the heterozygous class in alfalfa tended to become smaller with increasing proportion of allowed genotypes. This reflected the different GBS calling criteria for homozygous and heterozygous SNP: heterozygous SNPs required fewer overlapping reads to be called, while the implemented quality filters on alfalfa required a larger number of reads to call a homozygous locus. Therefore homozygous loci tended to have higher missing rates, and to be included progressively more frequently with more relaxed thresholds on allowed missing level. The average missing-rate and MAF in the 12 rice chromosomes were comparable to the whole-rice dataset.

Alignment on reference genomes assigned a position to 88.66%, 57.54% and 57.86% of Rice, Alfalfa-Med and Alfalfa-PV markers, respectively, reflecting

the difference between having a reference genome available (rice) or using that of a closely related species (alfalfa).

Accuracies are reported, averaged, in Supplementary Table 1. In rice, the average imputation accuracy over all genotype classes was above 80% for all methods. No significant difference was found between chromosomes. Thus average accuracy over all chromosomes is therefore presented (Figure 2).
The overall accuracy for alfalfa, averaged across the two datasets, was 12-25% lower than rice, ranging between 66% and 82%. No significant differences were found between the two datasets. Thus average accuracy over the Alfalfa-Med and Alfalfa-PV is presented (Figure 3).

The imputation accuracy varied across the different genotype classes, with the highest accuracy in the most common genotype (averaging 92.27% and 96.47% in alfalfa and rice, respectively), while the least common genotype class showed much lower accuracy (averaging 5.79% and 69.88% in alfalfa and rice, respectively). Alfalfa datasets included heterozygous SNP loci, accounting for an average of 27% of individuals per locus (Table 1). Heterozygotes had an intermediate imputation accuracy ranging from 0.5% with Beagle to 66.03% with KNNI. Results for each individual alfalfa population and the 12 individual rice chromosomes are similar to those discussed here (see Supplementary Figures S2 to S15).
Missing rate did not affect the imputation accuracy substantially, with most algorithms showing a flat or almost flat response to increased missing rate. KNNI showed a decreasing imputation accuracy with increasing missing rate, both in alfalfa and rice.

Beagle outperformed all other imputation methods in rice, making efficient use of marker position (Figure 2). When markers were randomly reshuffled (thus offsetting position information), though, general imputation methods (other than MNI) showed a higher imputation accuracy than Beagle. The ac-

curacy loss with shuffled markers is exacerbated in alfalfa datasets, and in particular in heterozygous and minor homozygous genotypes.

Similarly, FILLIN resulted in a 5-8% accuracy drop when marker positions were shuffled. FILLIN performances were in general lower than other methods due to the fraction of unimputed genotypes, here accounted as imputation errors. On average, FILLIN did not impute 14.6% of missing data in Rice and practically the entirety of Alfalfa data. Even trying to relax FILLIN parameters, we couldn't obtain meaningful results in alfalfa. Thus, FILLIN accuracy results are not reported in Figure 3.

When examining general imputation methods, resulting accuracies highlighted RFI and KNNI as the best ones. The two methods had comparable performances over all experiments. SVDI followed, and MNI ranked last.

In Rice, for 70% allowed missing rate, the imputation accuracy of RFI was still higher than 90%, while it dropped towards 85% for KNNI and SVDI. As expected, MNI had an imputation accuracy of 100% in the major homozygous class, but dropped to 0% in the minor homozygous class.

In alfalfa KNNI showed the highest imputation accuracy over all classes (averaging 82.20%) and in the heterozygous (66.03%) and minor homozygous (14.68%) genotype classes (Figure 3). RFI and SVDI were second ranking. They did not differ much from MNI in the overall accuracy (MNI: 78.38%, RFI: 79.50%, SVDI: 77.36%) and in the major (MNI: 88.69%, RFI: 89.05%, SVDI 85.81%) homozygous class, but gave higher imputation accuracy in the heterozygous (MNI: 58.88%, RFI: 61.75%, SVDI 61.67%). In the minor homozygous class only SVDI resulted in some sporadic correct imputation, while RFI and MNI did not produce any sensible result (MNI: 0%, RFI: 0.01%, SVDI 4.47%) classes.

3.2 Computation time

The amount of time required to complete the imputation process was recorded for each method. Only the implementation of RFI could leverage a multi-core/multi-thread environment, so that RFI computation times should be evaluated considering 10 CPUs used in parallel, while all other algorithms used only one CPU at a time.

Imputation efficiency (Figure 4) was assessed with respect to the gross dimension of the data set (i.e., number of markers $\times$ number of samples). An alternative analysis relating computation times to the number of missing genotypes brought to the same results (data not shown).

RFI required by far the longest computation times (in spite of parallelization), which grew approximately exponentially ($O(e^N)$) with problem size, and easily required tens of hours for individual rice chromosomes, and hundred of hours (up to 937 hours) for the complete rice dataset. The second slowest algorithm was KNNI, with computation times growing approximately quadratically ($O(N^2)$) with problem complexity (the chosen KNNI implementation contains no heuristic method to prune the all vs. all distance calculation). KNNI was however significantly faster than RFI (on average 16.7 times faster), requiring a time ranging from 20.68 seconds (rice chromosome 7, 10% allowed, 1% artificial) to about 28 hours (rice complete dataset, 70% allowed, 20% artificial) to complete the imputation task.

All other algorithms were faster, with computation times growing linearly or logarithmically with problem size. Beagle, FILLIN and SVDI resulted in similar execution times. MNI was by far the fastest imputation algorithm, being scarcely affected by the size of the problem (computation times ranging from 0.23 to 31.29 seconds).

4 Discussion

4.1 Minor Allele Frequency and data (un)balancedness

GBS data pose a greater challenge than SNP-array data to imputation algorithms, mainly as a consequence of the much larger quantity of missing genotypes they contain. In our data sets we found that missing rates varied from 53% to 67%. With larger amounts of missing data the complexity of the imputation problem increases. Imputation errors can have a negative impact on successive analyses (e.g. genomic predictions [34, 1]).

The imputation of missing SNP genotypes is a special case of the broader family of classification problems: three-class missing genotypes (AA, AB and BB) at any given SNP locus are classified based on known genotypes at all remaining data points. Classification problems are known to be harder when data are unbalanced, i.e. the classes appear at different frequencies in the datasets, with typically one over-represented class (see [19] and [38] for a review). In SNP genotype imputation, data balancedness is directly related to the minor allele frequency (MAF). In the classification of unbalanced observations, it is important to look not only at the total classification accuracy, but also at the per-class accuracy. The total classification accuracy may be misleading, being "dominated" by the majority class [14]. Indeed, we found that for most methods, even when the total classification accuracy was very high, relatively large error rates were present in the minority classes. The dependency of imputation results on MAF has already been acknowledged (e.g., [15] in maize; [25] in cattle; [28] in humans). Our per-class dissection of results allowed a deeper insight into the imputation process. Indeed, all imputation algorithms performed considerably better in the majority class rather than in the less frequent classes. In alfalfa, KNNI was easily the best imputation method in the

heterozygous and minor homozygous classes, while SVDI and RFI performed only slightly better than MNI. In rice, Beagle gave the best imputation results, with accuracy close to 100% both in the major and minor homozygous classes.

4.2 Missing rate and the "curse of dimensionality"

In general, imputation methods were relatively robust to increasing missing rates. Only KNNI, and to a lesser extent RFI, showed slowly degrading performances at very high missing rates. The divergent response of the different imputation methods to missing rate became apparent only in the most challenging scenarios, when markers with up to 40-70% missing genotypes were allowed in the dataset.

KNNI's decrease in imputation accuracy with increasing missing rates can be interpreted in relation to the phenomenon known as the "curse of dimensionality" [3,26]: with the same amount of data, the increasing number of parameters - the "dimensions" of the problem - increases and complicates the identification of k neighbors which are close enough to the data point to be classified/imputed. On average, the side l of the hypercube to include k neighbors is a function of k, n (sample size) and p (number of parameters): $l = \left(\frac{k}{n}\right)^{1/p}$. With n constant, the hypercube in which the k neighbors lie gets bigger as p increases. This holds especially for local learning methods (e.g., K-Nearest Neighbors methods, local regression) which rely heavily on the the information content of the neighborhood. RFI - which only partially relies on local structure of the data - suffered marginally from high missingness. SVD and MNI, which build on more general properties of the data, and Beagle - whose learning process is fundamentally different and specific for genotype imputation - were scarcely affected by missing rates.

4.3 Imputation efficiency: differences in rice vs. alfalfa

Imputation results were significantly different in rice and alfalfa: all imputation algorithms performed consistently better in rice, where Beagle, RFI and KNNI achieved performances comparable to what is reported in literature for SNP-chip data (e.g., [42] for bovine data; [40] for human data). On the other hand, imputation accuracy in alfalfa was much lower. This can be ascribed to four main factors:

i) the imputation problem was simpler in rice than in alfalfa, since the rice dataset comprised only two genotype classes (AA and BB) instead of alfalfa's three (AA, AB, BB);

ii) rice is natively diploid while alfalfa, autotetraploid, has been rendered diploid "in silico" during SNP calling. This simplification step made alfalfa data less adherent to the underlying biology;

iii) rice data have higher marker density (2.4 Kbp/SNP, compared to 24.5 Kbp/SNP for Alfalfa-PV and 19.6 Kbp/SNP for Alfalfa-Med) due to both a higher number of markers and a smaller genome (400 Mb for rice, 800 Mb for Alfalfa). This allowed for higher average values of linkage disequilibrium (LD) among SNP markers, thus facilitating the imputation process;

iv) rice markers were aligned on their native genome, while alfalfa markers were aligned on the genome of a different species.

There was no population structure in alfalfa [1], while the rice dataset was intrinsically stratified —being a collection of five subpopulations (*indica*, *temperate japonica tropical japonica*, *aus* and *aromatic*). Most imputation methods implicitly exploit population structure (e.g. KNNI computes distances based on genetic relatedness; Beagle reconstructs haplotypes based on genetic similarities), without explicitly modelling it. The imputation of missing genotypes, however, is not an inferential problem such as GWAS (genome-wide

association study), where the significance and estimate of SNP effects are known to be inflated if population structure is not included in the model, and the impact on the accuracy of imputation is likely to be small. In cattle from Scandinavian countries, Brondum et al [6] actually found higher imputation accuracy when combining populations (cattle breeds) in Beagle (without explicitly modelling the population structure), most likely as a result of the larger sample size. On the other hand, population stratification may help explain why FILLIN performed poorly in our rice dataset, which is a collection of subpopulations, while this algorithm is optimised for homogeneous inbred populations. However, the influence of population structure on the accuracy of imputation was not formally tested in this work, and this remains an interesting topic for further investigation.

4.4 Reference genome assembly and ordered vs unordered markers

When a reference genome is available, marker position can be used in the imputation process. All methods specifically developed for the imputation of missing genotypes have been designed for markers ordered along the genome sequence, and make explicit or implicit use of position information.

In rice, where a reference genome is available, we could assess the effect of using ordered vs. unordered markers for imputation. For alfalfa there is no reference genome sequenced yet. The genome of the close relative diploid *Medicago Truncatula* is available [45] and can in principle be used. However, while the two genomes show high synteny [23], aligning on a different species (and with different ploidity) comes at the price of discarding those markers that do not align. In our case only 57.54% (Alfalfa-Med) and 57.86% (Alfalfa-PV) aligned on the *M. Truncatula* genome, compared to the 88.66% of rice markers aligning on *O. Sativa* genome. This left with 23 438 and 18 923 SNPs to be

used for analysis with ordered markers.

Among the imputation methods used in this study, Beagle and FILLIN are the only ones that make use of marker information, and were therefore tested with ordered and randomly shuffled markers. In rice dataset with ordered markers Beagle showed astounding resilience to high missingness, with imputation accuracy in the minority class >95% even in the most extreme scenarios (70% allowed, 20% artificial missing genotypes). Beagle worked substantially worse when the marker position was randomly reshuffled, with imputation accuracy in the minority class ranging from 75% (10% allowed, 1% artificial missing rates) to almost 50% (70% allowed, 20% artificial missing rates). When tested with alfalfa Beagle largerly overestimated the majority class and had very poor performances on the heterozygous class. Shuffling marker positions further worsened Beagle performance in alfalfa, and almost all missing genotypes were assigned to the majority class.

FILLIN performance in rice followed the same pattern and was substantially worse when markers were shuffled. This result confirmed that imputation methods specifically designed for ordered markers are not an option for species lacking a reference genome.

Previous studies, though not detailing a per-class breakdown of the imputation accuracy, provide interesting comparisons. In sugar beet, Biscarini et al. [4] used Beagle to impute SNP-chip markers with partial within-chromosome/scaffold alignment, finding global accuracies ranging from 84% to 80.9% with 1% to 20% missing genotypes. Huang et al. [16] tested several algorithms, including Beagle, on simulated ordered GBS rice data with missing rates in the 30-60% range, and found that Beagle gave imputation accuracy consistently higher (+15-20%) than KNNI. Finally, Swarts et al. [39] used ordered GBS maize data to compare Beagle with the FSFHap and FILLIN algorithms they developed. They mostly obtained high imputation accuracies in the three genotype

classes, providing further evidence of the added value of having a reference genome assembly.

4.5 Size of the imputation problem

Albeit this was not the main objective of the paper, and no formal effort to optimise the implementation of the various imputation methods was done, still the recorded computation times give us some interesting information. The size of the problem, besides affecting the accuracy achievable by some imputation methods, is mainly relevant with reference to the required computation resources. RFI was the most demanding algorithm, with computation times growing exponentially both with the total number of marker genotypes (m SNP $\times$ n samples) and with the number of missing genotypes to be imputed. When analyzing the complete rice dataset (all chromosomes together), RFI took a maximum of about 40 days to complete imputation, and became computationally intractable (on the available platform) for the largest missing rate scenarios. This was partially mitigated by parallelization (we used 10 CPUs in our experiments). All other imputation algorithms took much shorter times to complete imputation, even in the most challenging scenarios. Putting together imputation accuracy and computation time, we found that the best performing imputation algorithms were Beagle in rice and KNNI in alfalfa.

5 Conclusions

Imputation of missing genotypes can be an effective technique also for GBS data. The most accurate imputation methods resulted in a total amount of wrongly imputed genotypes near to zero in rice and slightly over 10% in alfalfa. The proportion of imputation errors, however, varied dramatically among genotype classes, approaching very high levels with the worst methods on the

minor homozygous class. The structure of the genome and the maturity of the reference assembly play a role in the imputation efficiency, as indicated by the greater efficiency obtained in rice compared with alfalfa. While Beagle was preferable for rice, general data imputation methods performed significantly better in the absence of a reference genome. In particular, in alfalfa RFI and KNNI showed the lowest imputation errors in all classes, with KNNI to be preferred for computational efficiency. The alignment of markers to closely related species can help, but comes at the price of discarding markers that do not align, and with general methods still performing better.

The imputation accuracy tended to be relatively robust over increasing missing rates. However, KNNI showed a somewhat lower accuracy on high missingness scenarios, probably as a consequence of the "curse of dimensionality". In terms of computation requirements, all methods proved to be tractable over all problem complexities with a standard bioinformatics-lab computation infrastructure, except RFI (whose computational requirements increased exponentially with problem size and took up to 40 days with the complete rice dataset and maximum missing rate thresholds).

The results of this paper showed that high imputation accuracies can be achieved with GBS data, and that general imputation methods are a valid option when the reference genome is not available or the alignment on the genome of a closely related species leads to losing too many markers. Additionally, we highlighted the importance of examining imputation accuracy in the different genotype classes, since the common practice of summarizing performances in a single index can be very misleading when data are unbalanced (i.e. MAF is low). Exploring ways to improve imputation accuracy in GBS data is an important research area and can help making genotyping-by-sequencing a very attractive and cost-effective technique for genomic applications in species with and without a reference genome.

Acknowledgements The rice data used in this research paper were produced within the framework of the Italian national project "RISINNOVA" (grant n. 2010-2369), financially supported by the AGER Foundation. The creation of the alfalfa data sets was funded by the projects Genomic selection in alfalfa (GENALFA) funded by the Italian Ministry of Foreign Affairs and International Cooperation in the framework of the Italy-USA scientific cooperation program, the Italian share of the FP7-ArimNet project Resilient, water- and energy-efficient forage and feed crops for Mediterranean agricultural systems (REFORMA) funded by the Italian Ministry of Agricultural and Forestry Policies.

References

1. Annicchiarico, P., Nazzicari, N., Li, X., Wei, Y., Pecetti, L., Brummer, E.C.: Accuracy of genomic selection for alfalfa biomass yield in different reference populations. BMC Genomics **16**(1), 1–13 (2015). DOI 10.1186/s12864-015-2212-y. URL http://dx.doi.org/10.1186/s12864-015-2212-y
2. Aulchenko, Y.S., Ripke, S., Isaacs, A., Van Duijn, C.M.: Genabel: an r library for genome-wide association analysis. Bioinformatics **23**(10), 1294–1296 (2007)
3. Bellman, R.: Dynamic Programming. Princeton University Press (1957)
4. Biscarini, F., Stevanato, P., Broccanello, C., Stella, A., Saccomani, M.: Genome-enabled predictions for binomial traits in sugar beet populations. BMC genetics **15**(1), 87 (2014). URL http://www.biomedcentral.com/1471-2156/15/87/
5. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001). URL http://link.springer.com/article/10.1023/A:1010933404324
6. Brøndum, R.F., Ma, P., Lund, M.S., Su, G.: Short communication: Genotype imputation within and across nordic cattle breeds. Journal of dairy science **95**(11), 6795–6800 (2012)
7. Browning, S.R., Browning, B.L.: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics **81**(5), 1084–1097 (2007). DOI 10.1086/521987. URL http://www.sciencedirect.com/science/article/pii/S0002929707638828
8. Browningr, B.: Beagle 3.3.2 (2011). URL https://faculty.washington.edu/browning/beagle/beagle_3.3.2_31Oct11.pdf
9. Crossa, J., Beyene, Y., Kassa, S., Prez, P., Hickey, J.M., Chen, C., Campos, G.d.l., Burgueo, J., Windhausen, V.S., Buckler, E., Jannink, J.L., Cruz, M.A.L., Babu, R.: Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing.

644 G3: Genes|Genomes|Genetics **3**(11), 1903–1926 (2013). DOI 10.1534/g3.113.008227.
645 URL http://www.g3journal.org/content/3/11/1903
646 10. Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S.,
647 Mitchell, S.E.: A robust, simple genotyping-by-sequencing (GBS) approach for high
648 diversity species. PLoS ONE **6**(5), e19,379 (2011). DOI 10.1371/journal.pone.0019379.
649 URL http://dx.doi.org/10.1371/journal.pone.0019379
650 11. Endelman, J.B.: Ridge regression and other kernels for genomic selection with r package
651 rrblup. Plant Genome **4**, 250–255 (2011)
652 12. Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., Buckler,
653 E.S.: TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS
654 One **9**(2), E90,346 (2014). URL http://dx.plos.org/10.1371/journal.pone.0090346
655 13. Hayes, B., Bowman, P., Chamberlain, A., Goddard, M.: Invited review: Genomic selec-
656 tion in dairy cattle: Progress and challenges. Journal of dairy science **92**(2), 433–443
657 (2009)
658 14. He, H., Garcia, E.A.: Learning from imbalanced data. Knowledge and Data Engineering,
659 IEEE Transactions on **21**(9), 1263–1284 (2009)
660 15. Hickey, J.M., Crossa, J., Babu, R., de los Campos, G.: Factors Affecting the Ac-
661 curacy of Genotype Imputation in Populations from Several Maize Breeding Pro-
662 grams. Crop Science **52**(2), 654 (2012). DOI 10.2135/cropsci2011.07.0358. URL
663 https://www.crops.org/publications/cs/abstracts/52/2/654
664 16. Huang, B.E., Raghavan, C., Mauleon, R., Broman, K.W., Leung, H.: Efficient Imputa-
665 tion of Missing Markers in Low-Coverage Genotyping-by-Sequencing Data from Multi-
666 parental Crosses. Genetics **197**(1), 401–404 (2014). DOI 10.1534/genetics.113.158014.
667 URL http://www.genetics.org/content/197/1/401
668 17. International Rice Genome Sequencing Project: The map-based se-
669 quence of the rice genome. Nature **436**(7052), 793–800 (2005). URL
670 http://www.nature.com/articles/nature03895
671 18. Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R.,
672 Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., others: Improvement of the
673 Oryza sativa Nipponbare reference genome using next generation sequence and optical
674 map data. Rice **6**(1), 4 (2013). URL http://www.biomedcentral.com/content/pdf/1939-
675 8433-6-4.pdf
676 19. Kotsiantis, S., Kanellopoulos, D., Pintelas, P., others: Handling imbalanced datasets:
677 A review. GESTS International Transactions on Computer Science and Engineering
678 **30**(1), 25–36 (2006)

679 20. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with
680 Bowtie 2. Nature methods **9**(4), 357–359 (2012). URL
681 http://www.nature.com/nmeth/journal/v9/n4/abs/nmeth.1923.html

682 21. Li, H., Durbin, R.: Fast and accurate short read alignment with burrows–wheeler trans-
683 form. Bioinformatics **25**(14), 1754–1760 (2009)

684 22. Li, X., Wei, Y., Acharya, A., Hansen, J.L., Crawford, J.L., Viands, D.R., Michaud,
685 R., Claessens, A., Brummer, E.C.: Genomic prediction of biomass yield in two
686 selection cycles of a tetraploid alfalfa breeding population. Plant Genome ((ac-
687 cepted, not published)) (2015). DOI 10.3835/plantgenome2014.12.0090. URL
688 https://www.crops.org/files/publications/tpg/first-look/plantgenome-tpg-2014-12-
689 0090.pdf

690 23. Li, X., Wei, Y., Acharya, A., Jiang, Q., Kang, J., Brummer, E.C.: A satu-
691 rated genetic linkage map of autotetraploid alfalfa (*Medicago sativa* L.) devel-
692 oped using genotyping-by-sequencing is highly syntenous with the *Medicago trun-*
693 *catula* genome. G3: Genes| Genomes| Genetics **4**(10), 1971–1979 (2014). URL
694 http://www.g3journal.org/content/4/10/1971.short

695 24. Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S.,
696 Costich, D.E.: Switchgrass genomic diversity, ploidy, and evolution: novel insights from
697 a network-based snp discovery protocol. PLoS Genet **9**(1), e1003,215 (2013). DOI
698 10.1371/journal.pgen.1003215. URL http://dx.doi.org/10.1371/journal.pgen.1003215

699 25. Ma, P., Brndum, R.F., Zhang, Q., Lund, M.S., Su, G.: Comparison of dif-
700 ferent methods for imputing genome-wide marker genotypes in Swedish and
701 Finnish Red Cattle. Journal of dairy science **96**(7), 4666–4677 (2013). URL
702 http://www.sciencedirect.com/science/article/pii/S0022030213003664

703 26. Marimont, R.B., Shapiro, M.B.: Nearest Neighbour Searches and the Curse of Di-
704 mensionality. IMA Journal of Applied Mathematics **24**(1), 59–70 (1979). DOI
705 10.1093/imamat/24.1.59. URL http://imamat.oxfordjournals.org/content/24/1/59

706 27. Nicolazzi, E.L., Biffani, S., Biscarini, F., Orozco ter Wengel, P., Caprera, A., Nazz-
707 icari, N., Stella, A.: Software solutions for the livestock genomics SNP array rev-
708 olution. Animal Genetics pp. n/a–n/a (2015). DOI 10.1111/age.12295. URL
709 http://onlinelibrary.wiley.com/doi/10.1111/age.12295/abstract

710 28. Pei, Y.F., Li, J., Zhang, L., Papasian, C.J., Deng, H.W.: Analyses and comparison of
711 accuracy of different genotype imputation methods. PloS one **3**(10), e3551 (2008). URL
712 http://dx.plos.org/10.1371/journal.pone.0003551

29. Pérez, P., de los Campos, G.: Genome-wide regression & prediction with the bglr statistical package. Genetics pp. genetics–114 (2014)
30. Perry, P.O.: bcv: Cross-Validation for the SVD (Bi-Cross-Validation) (2009). URL http://cran.r-project.org/web/packages/bcv/index.html
31. Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Snchez-Villeda, H., Sorrells, M., Jannink, J.L.: Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. The Plant Genome Journal **5**(3), 103 (2012). DOI 10.3835/plantgenome2012.06.0006. URL https://www.crops.org/publications/tpg/abstracts/5/3/103
32. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2014). URL http://www.R-project.org
33. Rocher, S., Jean, M., Castonguay, Y., Belzile, F.: Validation of genotyping-by-sequencing analysis in populations of tetraploid alfalfa by 454 sequencing. PLoS ONE **10**(6), e0131,918 (2015). DOI 10.1371/journal.pone.0131918. URL http://dx.doi.org/10.1371/journal.pone.0131918
34. Rutkoski, J.E., Poland, J., Jannink, J.L., Sorrells, M.E.: Imputation of unordered markers and the impact on genomic selection accuracy. G3: Genes| Genomes| Genetics **3**(3), 427–439 (2013). URL http://www.g3journal.org/content/3/3/427.short
35. Schwender, H.: Statistical analysis of genotype and gene expression data. Ph.D. thesis (2007). URL https://eldorado.tu-dortmund.de/handle/2003/23306
36. Schwender, H., Fritsch, A.: scrime: Analysis of High-Dimensional Categorical Data such as SNP Data (2013). URL http://cran.r-project.org/web/packages/scrime/index.html
37. Stekhoven, D.J., Bhlmann, P.: MissForestnon-parametric missing value imputation for mixed-type data. Bioinformatics **28**(1), 112–118 (2012). URL http://bioinformatics.oxfordjournals.org/content/28/1/112.short
38. Sun, Y., Wong, A.K., Kamel, M.S.: Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence **23**(04), 687–719 (2009). URL http://www.worldscientific.com/doi/abs/10.1142/S0218001409007326
39. Swarts, K., Li, H., Romero Navarro, J.A., An, D., Romay, M.C., Hearne, S., Acharya, C., Glaubitz, J.C., Mitchell, S., Elshire, R.J., Buckler, E.S., Bradbury, P.J.: Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. The Plant Genome **7**(3), 0 (2014). DOI 10.3835/plantgenome2014.05.0023. URL https://www.crops.org/publications/tpg/abstracts/7/3/plantgenome2014.05.0023

40. The 1000 Genomes Project Consortium: An integrated map of genetic variation from 1,092 human genomes. Nature **491**(7422), 56–65 (2012). DOI 10.1038/nature11632
41. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing value estimation methods for DNA microarrays. Bioinformatics **17**(6), 520–525 (2001). URL http://bioinformatics.oxfordjournals.org/content/17/6/520.short
42. VanRaden, P.M., Null, D.J., Sargolzaei, M., Wiggans, G.R., Tooker, M.E., Cole, J.B., Sonstegard, T.S., Connor, E.E., Winters, M., van Kaam, J.B.C.H.M., Valentini, A., Van Doormaal, B.J., Faust, M.A., Doak, G.A.: Genomic imputation and evaluation using high-density Holstein genotypes. Journal of Dairy Science **96**(1), 668–678 (2013). DOI 10.3168/jds.2012-5702. URL http://www.sciencedirect.com/science/article/pii/S0022030212007576
43. VanRaden, P.M., OĆonnell, J.R., Wiggans, G.R., Weigel, K.A.: Genomic evaluations with many more genotypes. Genet Sel Evol **43**(10), 10–1186 (2011). URL http://www.biomedcentral.com/content/pdf/1297-9686-43-10.pdf
44. Ward, J.A., Bhangoo, J., Fernndez-Fernndez, F., Moore, P., Swanson, J.D., Viola, R., Velasco, R., Bassil, N., Weber, C.A., Sargent, D.J.: Saturated linkage map construction in Rubus idaeus using genotyping by sequencing and genome-independent imputation. BMC genomics **14**(1), 2 (2013). URL http://www.biomedcentral.com/1471-2164/14/2
45. Young, N.D., Debell, F., Oldroyd, G.E.D., Geurts, R., Cannon, S.B., Udvardi, M.K., Benedito, V.A., Mayer, K.F.X., Gouzy, J., Schoof, H., Van de Peer, Y., Proost, S., Cook, D.R., Meyers, B.C., Spannagl, M., Cheung, F., De Mita, S., Krishnakumar, V., Gundlach, H., Zhou, S., Mudge, J., Bharti, A.K., Murray, J.D., Naoumkina, M.A., Rosen, B., Silverstein, K.A.T., Tang, H., Rombauts, S., Zhao, P.X., Zhou, P., Barbe, V., Bardou, P., Bechner, M., Bellec, A., Berger, A., Bergs, H., Bidwell, S., Bisseling, T., Choisne, N., Couloux, A., Denny, R., Deshpande, S., Dai, X., Doyle, J.J., Dudez, A.M., Farmer, A.D., Fouteau, S., Franken, C., Gibelin, C., Gish, J., Goldstein, S., Gonzlez, A.J., Green, P.J., Hallab, A., Hartog, M., Hua, A., Humphray, S.J., Jeong, D.H., Jing, Y., Jcker, A., Kenton, S.M., Kim, D.J., Klee, K., Lai, H., Lang, C., Lin, S., Macmil, S.L., Magdelenat, G., Matthews, L., McCorrison, J., Monaghan, E.L., Mun, J.H., Najar, F.Z., Nicholson, C., Noirot, C., O'Bleness, M., Paule, C.R., Poulain, J., Prion, F., Qin, B., Qu, C., Retzel, E.F., Riddle, C., Sallet, E., Samain, S., Samson, N., Sanders, I., Saurat, O., Scarpelli, C., Schiex, T., Segurens, B., Severin, A.J., Sherrier, D.J., Shi, R., Sims, S., Singer, S.R., Sinharoy, S., Sterck, L., Viollet, A., Wang, B.B., Wang, K., Wang, M., Wang, X., Warfsmann, J., Weissenbach, J., White, D.D., White, J.D., Wiley, G.B.,

782 Wincker, P., Xing, Y., Yang, L., Yao, Z., Ying, F., Zhai, J., Zhou, L., Zuber, A., Dnari,
783 J., Dixon, R.A., May, G.D., Schwartz, D.C., Rogers, J., Qutier, F., Town, C.D., Roe,
784 B.A.: The Medicago genome provides insight into the evolution of rhizobial symbioses.
785 Nature **480**(7378), 520–524 (2011). DOI 10.1038/nature10625

List of Figures

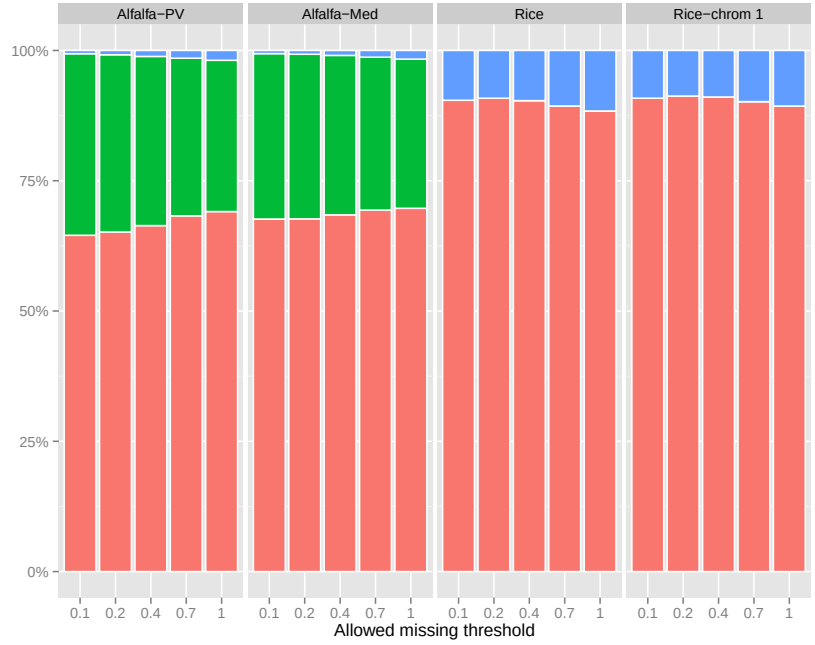

Fig. 1 Proportion of genotype classes for the different missing rate thresholds (0.1, 0.2, 0.4, 0.7 and 1 -the complete dataset), in Alfalfa-PV, Alfalfa-Med and Rice (all chromosomes and chromosome 1 only). Red bars represents the most common homozygous (AA), blue bars the least common homozygous (BB) and green bars the heterozygous (AB).

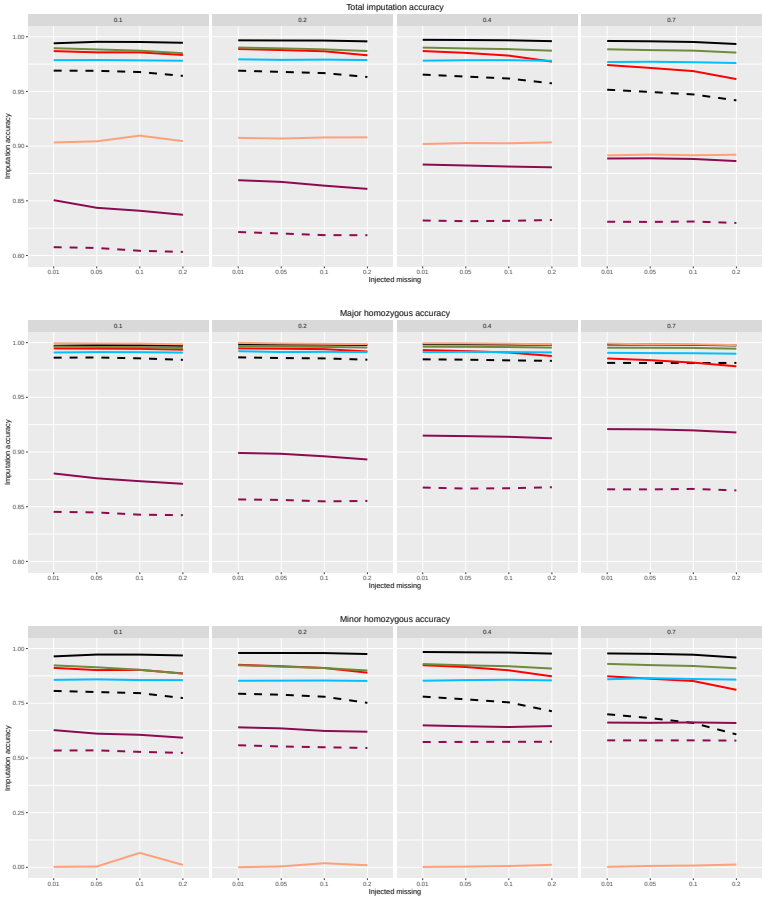

Fig. 2 imputation accuracies overall, for the major homozygous genotype (AA), and for the minor homozygous genotype (BB) in datasets consisting of 10%, 20%, 40% and 70% allowed missing data per locus (boxes) with 1%, 5%, 10% and 20% additional missing values artificially introduced (x-axis) averaged over the 12 rice chromosomes. Lines colors represent the five imputation algorithms: MNI (salmon), KNNI (red), SVDI (blue), RFI (green), Beagle with ordered markers (solid black), Beagle with unordered markers (dashed black), FILLIN with ordered markers (purple) and FILLIN with unordered markers (dashed purple). Y axis scale changes to highlight differences.

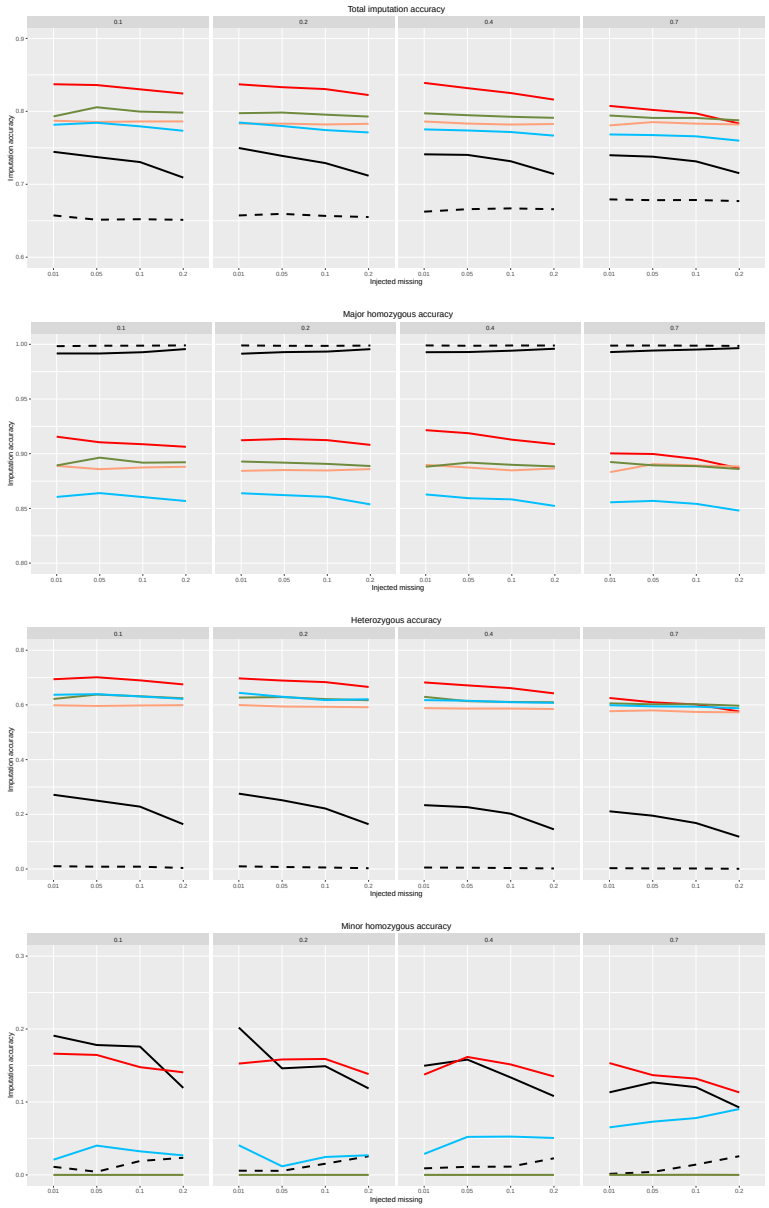

Fig. 3 imputation accuracies overall, for the major homozygous genotype (AA), for heterozygotes (AB), and for the minor homozygous genotype (BB) in datasets consisting of 10%, 20%, 40% and 70% allowed missing data per locus (boxes) with 1%, 5%, 10% and 20% additional missing values artificially introduced (x-axis) averaged two alfalfa populations (Alfalfa-Med and Alfalfa-PV). Lines colors represent the five imputation algorithms: MNI (salmon), KNNI (red), SVDI (blue), RFI (green), Beagle with ordered markers (solid black) and Beagle with unordered markers (dashed black). FILLIN was unable to impute alfalfa data and is absent from figure. Y axis scale changes to highlight differences.

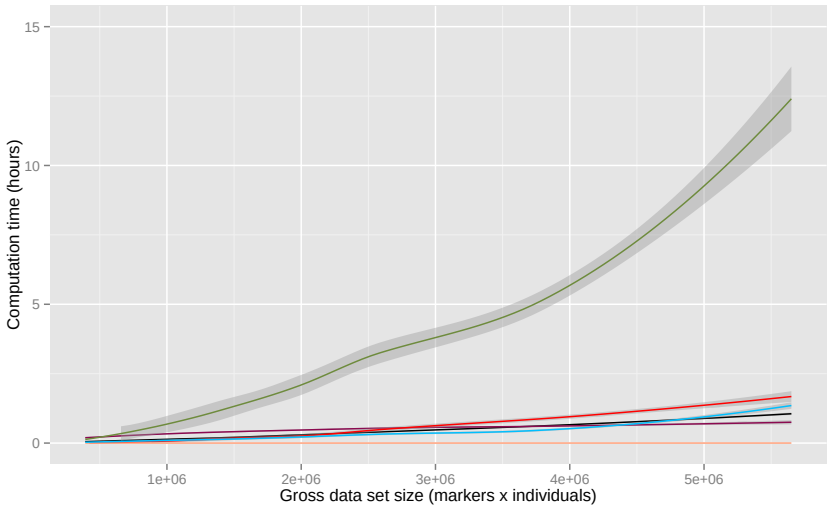

Fig. 4 computation time as a function of the total size of the imputed dataset. Lines colors represent the five imputation algorithms: MNI (salmon), KNNI (red), SVDI (light blue), RFI (green), Beagle (black) and FILLIN (purple). Plots include measures on Alfalfa-Med population, Alfalfa-PV population, and rice chromosomes 1 to 12. Complete rice datasets are omitted for readibility.

List of Tables

Table 1 Descriptive statistics of alfalfa (*Medicago sativa L.*) and rice (*Oryza sativa L.*) genotyping. MAF is average minor allele frequency; p(AA), p(AB) and p(BB) are the proportions of AA, AB and BB genotypes. Total number of markers and resulting average missing rate for all markers and for four allowed thresholds of missing rate per markers.

Allowed missing rate per marker		**Alfalfa-PV**	**Alfalfa-Med**	**Rice**
100%	**No. samples**	124	154	437
	MAF	0.1724	0.1702	0.140
	p(AA)	0.690	0.691	0.860
	p(AB)	0.274	0.276	-
	p(BB)	0.035	0.032	0.140
	Missing rate	0.666	0.596	0.534
	Markers	32 706	40 734	166 418
70%	**Missing rate**	0.325	0.29	0.364
	Markers	13 190	19 986	109 372
40%	**Missing rate**	0.161	0.142	0.198
	Markers	7 790	12 931	58 553
20%	**Missing rate**	0.076	0.074	0.099
	Markers	4 828	8 962	29 872
10%	**Missing rate**	0.046	0.045	0.049
	Markers	3 405	6 364	15 060