Xavier *et al. BMC Bioinformatics* (2016) 17:55
DOI 10.1186/s12859-016-0899-7

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Impact of imputation methods on the amount of genetic variation captured by a single-nucleotide polymorphism panel in soybeans

A. Xavier[1], William M. Muir[2] and Katy M. Rainey[1*]

Abstract

Background: Success in genome-wide association studies and marker-assisted selection depends on good phenotypic and genotypic data. The more complete this data is, the more powerful will be the results of analysis. Nevertheless, there are next-generation technologies that seek to provide genotypic information in spite of great proportions of missing data. The procedures these technologies use to impute genetic data, therefore, greatly affect downstream analyses. This study aims to (1) compare the genetic variance in a single-nucleotide polymorphism panel of soybean with missing data imputed using various methods, (2) evaluate the imputation accuracy and post-imputation quality associated with these methods, and (3) evaluate the impact of imputation method on heritability and the accuracy of genome-wide prediction of soybean traits. The imputation methods we evaluated were as follows: multivariate mixed model, hidden Markov model, logical algorithm, k-nearest neighbor, single value decomposition, and random forest. We used raw genotypes from the SoyNAM project and the following phenotypes: plant height, days to maturity, grain yield, and seed protein composition.

Results: We propose an imputation method based on multivariate mixed models using pedigree information. Our methods comparison indicate that heritability of traits can be affected by the imputation method. Genotypes with missing values imputed with methods that make use of genealogic information can favor genetic analysis of highly polygenic traits, but not genome-wide prediction accuracy. The genotypic matrix captured the highest amount of genetic variance when missing loci were imputed by the method proposed in this paper.

Conclusions: We concluded that hidden Markov models and random forest imputation are more suitable to studies that aim analyses of highly heritable traits while pedigree-based methods can be used to best analyze traits with low heritability. Despite the notable contribution to heritability, advantages in genomic prediction were not observed by changing the imputation method. We identified significant differences across imputation methods in a dataset missing 20 % of the genotypic values. It means that genotypic data from genotyping technologies that provide a high proportion of missing values, such as GBS, should be handled carefully because the imputation method will impact downstream analysis.

Keywords: Empirical Bayes, Heritability, Genomic selection, Association studies

* Correspondence: krainey@purdue.edu
[1]Department of Agronomy, Purdue University, Lilly Hall of Life Sciences, 915 W. State St., West Lafayette, Indiana 47907, USA
Full list of author information is available at the end of the article

Background

Marker-assisted selection (MAS) is a powerful tool for accelerating genetic improvement of plants and animals [1] through introgression of quantitative trait loci (QTL) and selection using genomic-enhanced breeding values (GEBV) [18, 27]. Meuwissen et al. [31] and Xu [64] first incorporated genome-wide molecular markers into the estimation of breeding values, and these so-called genomic selection (GS) methods have improved over time to handle a large number of loci [24, 59] with increased accuracy [7].

For genetic improvement and genetic studies in the post-genomic era, new genotyping platforms such as genotyping by sequencing (GBS) [10] and single-nucleotide polymorphism (SNP) arrays [45] provide a large number of molecular markers, but often with reduced genotyping accuracy [16, 52] and missing genotypes due to other factors. The implementation of such technology in crop breeding programs [18, 22, 48] helps maximize genetic gain through genomic selection [27, 56, 57] and also benefits genome wide association studies (GWAS) [17, 37, 52]. When there is a high proportion of missing data, consistent imputation may be challenging [22, 42, 53]. Accurate imputation of missing values and correction of genotyping errors is essential for eliminating gaps in genome coverage, integrating data across different arrays, and allowing robust genome-wide association mapping and prediction [21, 30, 49].

Researchers have proposed a variety of procedures for adjusting genotypic data to impute missing values and correct SNP miscalls. Most imputation algorithms, such as Hidden Markov Models [46], linear models [8, 65] and pedigree-based haplotyping [58], were designed for ordered markers which are known because there is a reference genome available. But some methods of imputation do not require information on marker order or phase [17, 44] such as k-Nearest Neighbors [55], Single Value Decomposition [36], or Random Forest [50]. These, therefore, are well-suited for *de novo* genotyping [3, 43].

Incorporating genealogical information boosts the accuracy of genotypic imputation, phasing, and the power of association analysis [16, 23]. Therefore, here we describe a new method of imputation based on the pedigree relationship matrix [62] and covariance among markers. The proposed method treats genotypes as response variables within a multivariate empirical Bayes model traditionally used in plant and animal breeding [19]. The expectation of narrow-sense heritability across a set of traits can be translated into a machine learning approach of pattern recognition to measure the amount of additive genetic information captured by the genomic information. Thus, we then compare the accuracy of each imputation method based on the amount of genetic variation that can be accounted for by a SNP panel [16].

Methods

This study employed genotypes and phenotypes from the SoyNAM project (soynam.org), downloaded on September 10th, 2014. SoyNAM is a soybean (*Glycine max*, Merr.) maturity group III nested association panel composed of 40 bi-parental crosses sharing a common parent and inbred for five generations. The panel comprises 5,596 recombinant inbred lines, phenotyped in 18 combinations of year and location, for four agronomic traits: grain yield, plant height, days to maturity, and seed protein composition. The panel was genotyped with a designed 5 k SNP chip. After removing non-segregating markers, low quality SNPs based on marker heritability (0.99) [11] and minor allele frequency (0.05) [54], we selected a set of 4,246 SNPs for this study that had 1 % of its genotypic data missing. To compare imputation performance under conditions with more missing data, we generated a second dataset by randomly deleting 20 % of genotypic data across all individuals using the prodNA function from the R package missForest [50]. If differences in imputation methods are detectable for a strict amount of 20 % missing values, these results will apply to datasets with a larger proportion of missing values.

We compared imputation approaches using a variety of methods: a multivariate mixed model (MMM); Hidden Markov Models (HMM) implemented in three software packages commonly employed in genetic studies [30]; a logical algorithm; and three non-parametric methods k-Nearest Neighbor (kNN), Single Value Decomposition (SVD), and Random Forest (RF). Raw and imputed datasets are available upon request.

Imputation methods

Multivariate mixed model

We adapted the multivariate mixed model method for imputation from Gengler et al. [13] and Yang et al. [65], based on the concept that marker inheritance will proceed in a Mendelian manner and therefore, follow the pedigree [41]. The method bases the numerator relationship matrix on the expectation of Mendelian allele inheritance with shared identity by descent (IBD). This contrasts with expectation-maximization imputation algorithms that rely on observed kinship [47, 65] and imputation through coalescent analysis [23] that attempts to recreate the pedigree.

If allele inheritance follows the pedigree, the marker should be perfectly heritable, except for Mendelian sampling error. Consequently, SNP heritability is an indicator of the gene content [11, 13], and we can estimate it by fitting the marker to a mixed model with pedigree specified as a random effect, as discussed by

Forneris et al. [11]. Thus, missing SNPs can be imputed as empirical Bayes estimates. Expanding the model to a multivariate level further improves the accuracy of prediction, “borrowing strength” from flanking markers and related individuals [16]. All markers are evaluated in a panel in a sliding window across the genome, ordered according to the genetic or physical map [9]. The method considers the marker of interest ($\mathbf{y}_j$) and its flanking markers ($\mathbf{y}_i$ and $\mathbf{y}_k$) as response variables ($\mathbf{Y} = \{\mathbf{y}_i, \mathbf{y}_j, \mathbf{y}_k\}$), coding them in an ordinal or continuous scale [51], and fitting them to a multivariate mixed model. Thus:

$$\mathbf{Y}_{n\times 3} = \mathbf{1}\boldsymbol{\mu}_{nx3} + \mathbf{Z}_{n\times n}\boldsymbol{\gamma}_{n\times 3} + \boldsymbol{\varepsilon}_{n\times 3}$$

where $\mathbf{Z}$ is the design matrices of the random effects, $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ represent the intercept and individual additive effect, and $\boldsymbol{\varepsilon}$ represents the residual term. The variances associated with the imputed marker are: $\sigma^2_{\gamma_j}$, representing the genetic variance of the marker j; σ_{ij}, the genetic covariance between marker j and the flanking marker i; σ_{ik}, the genetic covariance between marker j the flanking marker k; and $\sigma^2_{\varepsilon_j}$, the residual variance of marker j. For the given mode, variances are expressed as

$$\begin{aligned}\mathrm{Var}(\mathbf{Y}) &= \left(\mathbf{A}\otimes\sigma^2_{\gamma} + \mathbf{I}\otimes\sigma^2_{\varepsilon}\right) \\ &= \left(\mathbf{A}\times\begin{bmatrix}\sigma^2_{\gamma_i} & \sigma_{\gamma_{ij}} & \sigma_{\gamma_{ik}} \\ \sigma_{\gamma_{ij}} & \sigma^2_{\gamma_j} & \sigma_{\gamma_{jk}} \\ \sigma_{\gamma_{ik}} & \sigma_{\gamma_{jk}} & \sigma^2_{\gamma_k}\end{bmatrix} + \mathbf{I}\times\begin{bmatrix}\sigma^2_{\varepsilon_i} & 0 & 0 \\ 0 & \sigma^2_{\varepsilon_j} & 0 \\ 0 & 0 & \sigma^2_{\varepsilon_k}\end{bmatrix}\right)\end{aligned}$$

where $\mathbf{A}$ is the additive numerator relationship matrix and $\mathbf{I}$ is an identity matrix. MMM uses restricted maximum likelihood (REML) to estimate genetic covariances and then replaces missing values for the central position by the multivariate empirical Bayes estimate [66], also referred to as the best linear unbiased predictor (BLUP) [19]. The model is then incremented to the next position and repeated.

This study used the software BLUPF90 [32, 33] to compute the covariance components. However, any existing software that allows multivariate mixed models incorporating pedigree information can implement the model, such as Wombat, ASReml, or SAS. Efficient algorithms to compute mixed models are described by Zhou and Stephens [66], Legarra and Misztal [24] and VanRaden [59].

Hidden Markov models

The HMM method is commonly employed in genetics and genomics for stochastic modeling of Markov processes (such as the computation of haplotypes). Assuming ordered markers, the HMM estimates the most likely path of states (i.e. genotype) based on the transition probability of marker M^t to change state given the previous marker M^{t-1}. In genetic terms, the four possible states for a diploid organism with alleles M_1 and M_2 for locus M, are: M_1M_1, M_1M_2, M_2M_1 and M_2M_2.

This study evaluated three HMM software programs: fastPHASE [46], Beagle [4], and MaCH [26]. Beagle, fastPHASE, and MaCH implement HMM with iterative updating via Expectation-Maximization (Baum–Welch algorithm). Missing values of each chromosome were imputed separately and pedigree information was not provided. HMM is the most common method of imputation and is shown to boost power and resolution of genome-wide association studies [21, 30].

Logical algorithm

This method is implemented in the program findhap.f90 [58]. The program is computationally efficient, suited for large datasets, and becomes increasingly accurate as the proportion of pedigree data increases [38]. Thus it is advantageous for populations that pursue genotyped pedigrees.

Findhap.f90 first generates a list of possible haplotypes, then adds a genotype into the haplotype list and searches for a matching haplotype, then finds the second haplotype by subtracting the first haplotype from the genotype. It then compares each genotype to the haplotype list and imputes unknown alleles from the haplotypes. The first two iterations use population-wide genotypic data, and subsequent iterations locate matching haplotypes from pedigree data [57].

k-Nearest neighbor

kNN is a non-parametric method commonly used for prediction and classification. kNN is a memory-based learning algorithm based on voting [61]. It relies on filling missing data points with the weighted mean of the k most similar genotypes based on Euclidean distance (root sum-of-squares of differences) between standardized observations. Rutkoski et al. [44] evaluated kNN for genotypic imputation and suggested it was a promising method. We used the package knnGarden [60] with a setting of k = 10 to perform our computation.

Single value decomposition

SVD relies on orthogonal expression of the genotypic matrix [55] from the decomposition $\mathbf{M} = \mathbf{UDV}^{\mathrm{T}}$. SVD uses the most significant eigenvectors (columns of $\mathbf{U}$) to predict missing values. We performed SVD imputations by chromosome using the bi-cross-validation algorithm described in Owen & Perry [36] and implemented in package bcv [40].

Random forest

Random Forest [2] is a non-parametric method of prediction, classification, and imputation of mixed data types [50]. It establishes a combination of decision tree predictors, where the trees are bootstrapped random independent vectors that constitute training forests [2]. Imputation studies of GBS data in wheat breeding have reported promising results for RF [42–44]. We performed imputations employing random forest by chromosome to generate more informative trees and reduce computational burden, using the package missForest [50].

Comparison of methods

Imputation errors impact the reliability of breeding values [6, 43] and therefore, affect genome-wide selection. To evaluate the fitness of different imputation methods for genomic enhanced breeding values (GEBV), we calculated the intra-class correlation coefficient (ICC) for the genetic parameter, generally known as heritability, through average information restricted maximum likelihood, implemented in AI-REMLF90 with a convergence criterion of 10^{-11} [33]. The ICC is defined as

$$h^2 = \frac{\sigma_A^2}{\sigma_Y^2}$$

given the phenotypic variance $\sigma_Y^2 = \sigma_A^2 + \sigma_E^2 + \sigma_C^2 + \sigma_\varepsilon^2$; where σ_A^2 is the additive genetic variance, σ_E^2 is the variance due to environment, σ_C^2 is the variance due to microenvironment (i.e. controls), and σ_ε^2 is the residual variance. In genetic terms, ICC represents narrow-sense heritability, which is defined as the amount of genetic variation that alleles can transmit to the following generation, and it quantifies the population response to selection [28]. The rational for using narrow-sense heritability estimator as a measure of how well genetic variability is captured by the genotypic information comes from its practical application of pattern recognition in machine learning [15, 34]. Genomic relation matrices generated from a SNP panel imputed through distinct methods would yield different values of variance components. Consequently, kernels that provide enhanced genetic variance and higher heritability reflect better estimators of the genetic term [35].

To compare the effect of imputation methods on heritability, we performed analysis of variance (ANOVA) following the linear model $y_{ijkl} = \mu_{.} + \tau_i + \beta_j + \gamma_k + \varepsilon_{ijkl}$; where y_{ijkl} is the heritability of the i^{th} imputation method ($i = 1, \ldots, 8$) with j^{th} percentage of missing loci ($j = 1, 20$) of the k^{th} trait ($k = 1, \ldots, 4$) in the l^{th} observation; $\mu_{.}$ is the overall mean; τ_i represents the imputation method; β_j is the percentage of missing loci; γ_k represents the trait; and ε_{ijkl} is the residual of the ijkl observation. In this model, imputation method is the parameter of interest while blocking trait and percentage of missing. Levene's test determined the equality of variances [12]. The Shapiro-Wilk test of normality verified the normality. Tukey's honest significant difference (HSD) ($\alpha = 0.05$) grouped the imputation methods when the p-value of F test was significant (<0.05). Results shown in Fig. 1.

Imputation accuracy and prediction accuracy

This study evaluated imputation accuracy by determining the proportion of data imputed that was identical to the standard dataset [56]. For each method, we compared the imputed genotypic matrix with 20 % of the loci missing to the genotypic matrix corrected using the same method with 1 % of loci missing. Thus, the level of accuracy provides insight into both imputation method and modifications of the dataset.

To measure prediction accuracy in the context of genomic selection, we used the correlation of observed and predicted values in a five-fold cross validation divided by the square root of the heritability as described by Lehermeier et al. [25]. We used the following whole-genome regression methods [7, 14] to generate the predicted values: BayesA, BayesB, BayesCπ, Bayesian LASSO, and Bayesian Ridge Regression, as implemented by Pérez and de los Campos [39] with default settings of hyperpriors. The statistical testing of imputation method on prediction accuracy followed the same ANOVA model previously described with an additional term to accommodate the whole-genome regression method.

Results

Effect of imputation method on heritability

We found a significant association (p-value < 0.01) between heritability values (y_{ijkl}) and all terms in the statistical model: imputation method (τ_i), percentage of missing values (β_j) and agronomic trait (γ_k). The fitted model provided a coefficient of determination $R^2 = 0.95$ and a coefficient of variation CV = 6.72. The group means procedure (Fig. 1) showed that most imputation methods were similar. MMM had provided the highest across-trait heritability, although not significantly better than the other methods except for the kNN and SVD methods, which were inferior to the others, with SVD being the worst.

Post-imputation quality parameters

The average accuracy of imputation for data ranged from 79.32 to 99.60 %, (Fig. 2a). MaCH, fastPHASE, and RF were nearly identical (>99.5 %), followed by kNN (93.5 %). MMM and Beagle showed the same performance (85 %). Imputation accuracy reflects the sensitivity of imputation methods to missing data. However, lower imputation accuracy can be attributed to more severe correction of genotyping errors (e.g. SNP miscalls).

We observed a higher number of markers with full linkage disequilibrium (i.e. repeated markers) after imputation with the HMM implementations. MaCH and Beagle (Fig. 2b) also increased the number of repeated

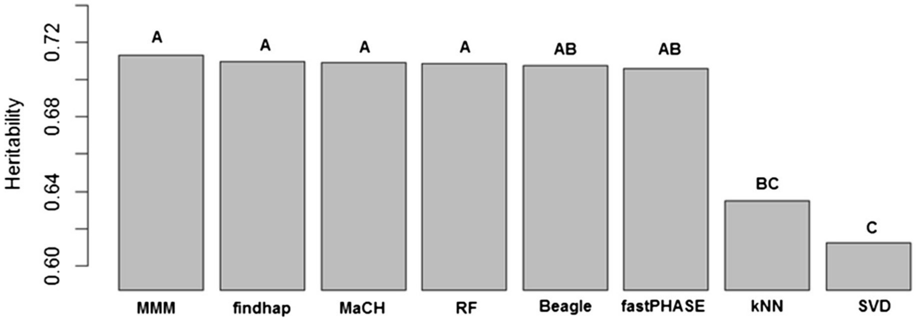

Fig. 1 Average heritability across four soybean traits using genotypic data imputed with different methods. Letters represent the statistical difference based on Tukey's HSD procedure (α = 0.05)

markers as the percentage of missing data increased. Non-parametric methods, SVD and kNN, did not provide repeated markers. Random forest and both pedigree-based methods, MMM and findhap, showed fewer repeated markers as the missing data increased.

The third measure of accuracy of imputation and data correction is the difference between the observed and expected proportion of heterozygosity in the data. The expected proportion of heterozygosity in this dataset is 6.25 % of loci given that the parents were homozygous and the plants were genotyped in the F5 generation (i.e. after 5 generations of selfing), in which the expected inbreeding coefficient is 93.75 %. The results show that the proportion of heterozygous loci tends to increase as the proportion of missing data increases (Fig. 2c). We observed this in all but two methods, Beagle and MaCH. SVD and kNN were more likely to increase the number of heterozygous loci with more missing data, especially SVD for which proportions of heterozygosity reached 24 %, meaning that heterozygous loci were assigned to most missing loci.

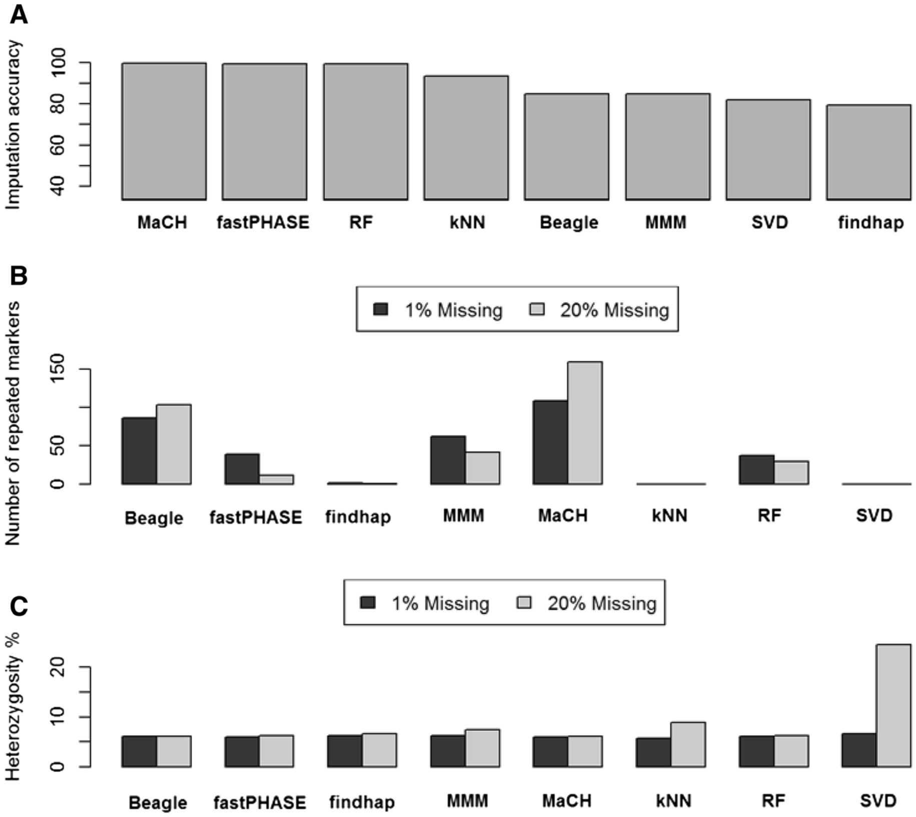

Fig. 2 Post-imputation quality parameters. **a** Imputation accuracy measured as the percentage of identical loci between datasets with 1 and 20 % missing. **b** Number of repeated markers (ie. full LD) after imputations with 1 and 20 % missing data. **c** Percentage of heterozygous loci after imputations of datasets with 1 and 20 % missing data

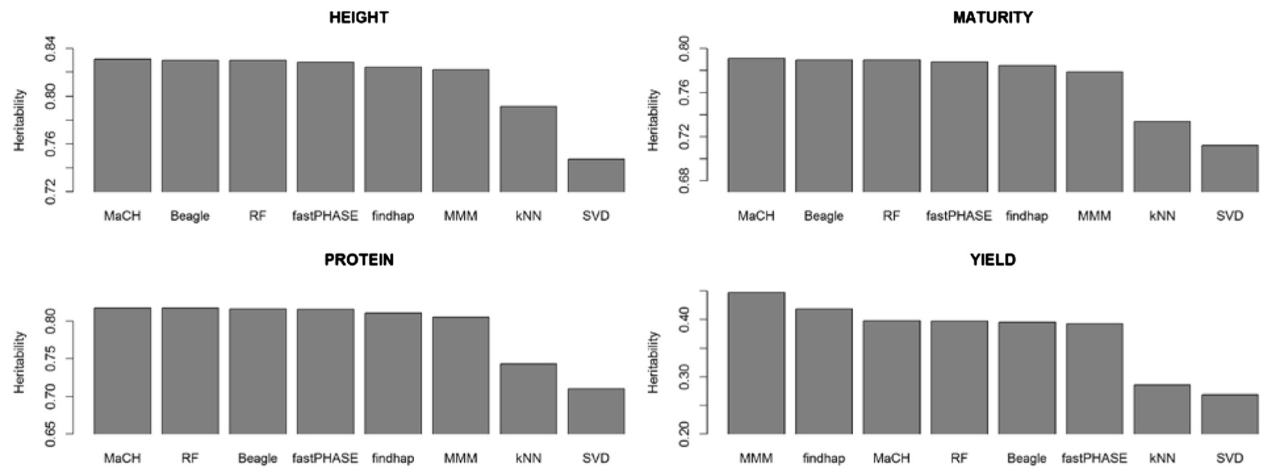

Fig. 3 Narrow-sense heritability for soybean agronomic traits computed for genotypic datasets imputed by various methods

Imputation methods and genetic variation captured

The mean heritability for traits in this study was 0.81 ± 0.04 for plant height, 0.77 ± 0.04 for days to maturity, 0.79 ± 0.06 for seed protein content, and 0.37 ± 0.08 for grain yield. Grain yield is considered minimally heritable while the other three are highly heritable. Imputation method had the greatest impact on the trait with the lowest heritability (grain yield) when compared to the impact on traits with high heritability (Fig. 3). The MMM method was superior to the other methods for minimally heritable traits while MaCH was best for highly heritable traits. The SVD and kNN methods were consistently inferior across traits.

The prediction accuracy across traits ranged from 0.606 (MMM) to 0.628 (RF), but did not differ significantly (*p*-value = 0.537). However, it appears that the genotypic data imputed with RF may have a slightly better performance. Whole-genome regression methods did not provide statistically significant differences in prediction accuracy (data not shown).

Discussion

The capture of genetic variation

Differences between imputation methods were clearly detectable on a SNP panel containing 20 % missing values, raising questions about the quality and reliability of genotypic data obtained by imputing datasets with an 80 % or greater proportion of missing values, a common scenario for GBS.

Our results indicate that genotypes imputed with methods that rely on pedigree information, such as MMM, better capture genetic variance in complex low heritability traits, while imputations performed using HMM may favor traits with high heritability. The choice of imputation method will affect downstream analysis. Thus, if the genetic architecture of the traits to be analyzed is known, it is possible to achieve more accurate results by selecting the most suitable imputation method.

The genetic architecture of the trait should also influence the choice of genomic prediction method [7], although this study found no statistically significant differences, likely

Table 1 Summary of properties of imputation methods under evaluation

Method	Better fit for low heritability traits	High imputation accuracy	Enlarges LD blocks	Corrects Miscalls	Accommodates pedigree information	Suitable for unordered markers
Beagle			X	X	X	
fastPHASE	X	X		X		
findhap					X	X
kNN						X
MaCH		X	X	X	X	
MMM	X			X	X	
RF		X				X
SVD						X

due to the similar nature of evaluated models also reported by Howard et al. [20]. Chen et al. [5] and Poland et al. [43] reported interactions between imputation method and prediction accuracy for different traits. For example, genotypic imputation through SVD showed inferior capture of genetic variance, but it did not affect prediction accuracy.

MMM: strengths and weaknesses

In this study we presented a mixed model method of imputation using pedigree information that displayed interesting properties regarding the capture of genetic variance and may have advantages for downstream analysis of complex traits. Some reported advantages of MMM include imputation of un-genotyped individuals [41] or with a large percentage of missing data [65] without losing robustness; and the identity link function of MMM allows imputation regardless of the allelic coding [51].

On the other hand, genotypic information imputed by MMM was not advantageous for analyzing highly heritable traits and it did not improve prediction accuracy. However, NAM populations have a very simple pedigree structure and more complex pedigree structures may lead to better results [29]. Another drawback was the method's computational burden because the mixed model computes each molecular marker, a limitation that can be addressed through parallel computing.

Much as in HMM, the change-of-state of a haplotype in MMM is controlled by the covariance, with flanking markers considering linkage disequilibrium (LD), and the population structure allied to the pedigree information simultaneously, while forward-backward HMM algorithms work one direction at time using just a subset of random samples [3]. Imputation methods that incorporate pedigree do not rely on samples from a reference dataset that is assumed to comprise all populations [16]. Using the genomic relationship instead of pedigree, Yang et al. [65] have reported the superiority of a similar MMM method over HMM emphasizing better incorporation of information on LD and IBD.

Choice of the imputation method

Properties of the methods are summarized in Table 1. Our results support the use of HMM and RF for a reliable representation of the genotypes, being best for analyses of highly heritable traits and a good alternative even if pedigree data are not available. In addition, imputation though MMM and findhap is preferred to analyze traits with low heritability when the pedigree is known. kNN and SVD provided consistently poor representations of the genotypes.

For GBS and other technologies where a high percentage of missing values is expected, our results support the use of MaCH, fastPHASE, and RF, methods that have shown insensitivity to the percentage of missing loci. In a similar study, [17]) observed inferior performance of RF over HMM in wheat that could be attributed to the whole-genome imputation at once as opposed to one chromosome at a time, resulting into less informative decision trees and inconsistent imputation. The imputation of each chromosome separately is also a common practice with HMM [29].

Conclusions

Quality of the imputation disturbs the genetic variance captured by the genotypic data. We were able to show that the imputation of genotypic data in the case where proportion of missing values as low as 20 % can affect the quality of the genotypic representation by the SNP panel. This results must be seen as a word of care for technologies based on low-coverage genotyping, such as GBS, where the amount of missing information commonly achieves values of 80 %. Yet, for the scenario in study it was not possible to identify significant impact of imputation method on genomic prediction. Thus, based on imputation accuracy and genetic variance captured by the SNP panel, the imputation method choice is hidden Markov models and random forest for general analysis. Pedigree-based methods of imputation were recognized to enhance the heritability of grain yield, the lowest heritable trait in this study.

Availability of supporting data

This study employed genotypes and phenotypes from the SoyNAM project (soynam.org). Data is available through the R package SoyNAM [63] and USDA-ARS soybean database SoyBase (soybase.org).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AX, WMM and KMR designed this study. AX and KMR prepared the methods and datasets. AX analyzed the datasets. AX wrote the manuscript and all authors revised and approved the final manuscript.

Acknowledgement

We thank the SoyNAM collaborators for their contributions to the experiment: Dr. Bill Beavis for experimental design, Qijan Song and Perry Cregan for genotyping, and Jim Specht and Brian Diers for creating the germplasm resource. A.X. was supported by funds from Dow AgroSciences, and the SoyNAM experiment was supported by funds from the North Central Soybean Research Program, the United Soybean Board, and Dow AgroSciences.

Author details

[1]Department of Agronomy, Purdue University, Lilly Hall of Life Sciences, 915 W. State St., West Lafayette, Indiana 47907, USA. [2]Department of Animal Science, Purdue University, Lilly Hall of Life Sciences, 915 W. State St., West Lafayette, Indiana 47907, USA.

Received: 8 October 2015 Accepted: 19 January 2016
Published online: 02 February 2016

References

1. Acquaah G. Principles of plant genetics and breeding. 2009. John Wiley & Sons. The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK.
2. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
3. Browning SR, Browning BL. High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet. 2010;86(4):526–39.
4. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084–97.
5. Chen L, Li C, Sargolzaei M, Schenkel F. Impact of Genotype Imputation on the Performance of GBLUP and Bayesian Methods for Genomic Prediction. PLoS One. 2014;9(7):e101544.
6. Dassonneville R, Brøndum RF, Druet T, Fritz S, Guillaume F, Guldbrandtsen B, et al. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. J Dairy Sci. 2011;94(7): 3679–86.
7. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013;193(2):327–45.
8. Dimauro C, Cellesi M, Gaspa G, Ajmone-Marsan P, Steri R, Marras G, et al. Use of partial least squares regression to impute SNP genotypes in Italian Cattle breeds. Gen Sel Evol. 2013;45(1):1–8.
9. Druet T, Georges M. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics. 2010;184(3):789–98.
10. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6(5):e19379.
11. Forneris NS, Legarra A, Vitezica ZG, Tsuruta S, Aguilar I, Misztal I, et al. Quality Control of Genotypes Using Heritability Estimates of Gene Content at the Marker. Genetics. 2015;199(3):675–81.
12. Gastwirth JL, Gel YR, Miao W. The impact of Levene's test of equality of variances on statistical theory and practice. Stat Sci. 2009;24(3):343–60.
13. Gengler N, Mayeres P, Szydlowski M. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal. 2007;1(1):21–8.
14. Gianola D. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics. 2013;194(3):573–96.
15. Gianola D, de los Campos G, González-Recio O, Long N, Okut H, Rosa GJ, et al. Statistical learning methods for genome-based analysis of quantitative traits. Proceedings of the 9th World Congress on Genetics Applied to Livestock Production. Vol. 14. 2010.
16. Halperin E, Stephan DA. SNP imputation in association studies. Nat Biotechnol. 2009;27(4):349–51.
17. He S, Zhao Y, Mette MF, Bothe R, Ebmeyer E, Sharbel TF, et al. Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (*Triticum aestivum* L.). BMC Genomics. 2015;16(1):168.
18. Heffner EL, Sorrells ME, Jannink JL. Genomic selection for crop improvement. Crop Sci. 2009;49(1):1–12.
19. Henderson CR. Estimation of variances and covariances under multiple trait models. J Dairy Sci. 1984;67(7):1581–9.
20. Howard R, Carriquiry AL, Beavis WD. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3. 2014;4(6):1027–46.
21. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529.
22. Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, et al. Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics. 2014;15(1):740.
23. Kimmel G, Karp RM, Jordan MI, Halperin E. Association mapping and significance estimation via the coalescent. Am J Hum Genet. 2008;83(6):675–83.
24. Legarra A, Misztal I. Technical note: computing strategies in genome-wide selection. J Dairy Sci. 2008;91(1):360–6.
25. Lehermeier C, Wimmer V, Albrecht T, Auinger HJ, Gianola D, Schmid VJ, et al. Sensitivity to prior specification in Bayesian genome-based prediction models. Stat Appl Genet Mol Biol. 2013;12(3):375–91.
26. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816–34.
27. Lorenz AJ. Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. G3. 2013;3(3):481–91.
28. Lynch M, Walsh B. Genetics and analysis of quantitative traits. Sinauer Associates, Inc. Sunderland, MA. 1998.
29. Ma P, Brøndum RF, Zhang Q, Lund MS, Su G. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. J Dairy Sci. 2013;96(7):4666–77.
30. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11(7):499–511.
31. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.
32. Misztal I. Computational techniques in animal breeding. Athens: University of Georgia; 2000. http://nce.ads.uga.edu/~ignacy/course2002/notes.pdf.
33. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH. BLUPF90 and related programs (BGF90). In: Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France, August, 2002. 2002. Session 28. (pp. 1-2). Institut National de la Recherche Agronomique (INRA).
34. Morota G, Gianola D. Kernel-based whole-genome prediction of complex traits: a review. Front Genet. 2014;5(1):363.
35. Morota G, Boddhireddy P, Vukasinovic N, Gianola D, DeNise S. Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. Front Genet. 2014;5(1):56.
36. Owen AB, Perry PO. Bi-cross-validation of the SVD and the nonnegative matrix factorization. Ann Appl Stat. 2009;3(2):564–94.
37. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet. 2012;44(6):631–5.
38. Pausch H, Aigner B, Emmerling R, Edel C, Götz KU, Fries R. Imputation of high-density genotypes in the Fleckvieh cattle population. Genet Sel Evol. 2013;45(3):10–1186.
39. Pérez P, de los Campos G. Genome-wide regression & prediction with the BGLR statistical package. Genetics. 2014;198(2):483–95.
40. Perry PO. bcv: Cross-Validation for the SVD (Bi-Cross-Validation). R package version 1.0. 2009. http://cran.r-project.org/web/packages/bcv/bcv.pdf.
41. Pimentel EC, Wensch-Dorendorf M, König S, Swalve HH. Enlarging a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture. Genet Sel Evol. 2013;45(12).
42. Poland JA, Rife TW. Genotyping-by-sequencing for plant breeding and genetics. Plant Genome. 2012;5(3):92–102.
43. Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, et al. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome. 2012;5(3):103–13.
44. Rutkoski JE, Poland J, Jannink JL, Sorrells ME. Imputation of unordered markers and the impact on genomic selection accuracy. G3. 2013;3(3):427–39.
45. Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet. 2006;123(4):218–23.
46. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78(4):629–44.
47. Schneider T. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. J Climate. 2001; 14(5):853–71.
48. Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, et al. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS One. 2013; 8(1):e54603.
49. Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 2009;5(5):e1000477.
50. Stekhoven DJ, Bühlmann P. MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–8.
51. Strandén I, Christensen OF. Allele coding in genomic evaluation. Genet Sel Evol. 2011;43(1):25.
52. Surakka I, Kristiansson K, Anttila V, Inouye M, Barnes C, Moutsianas L, et al. Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. Genome Res. 2010;20(10):1344–51.

53. Swarts K, Li H, Romero Navarro JA, An D, Romay MC, Hearne S, et al. Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. The Plant Genome. 2014;7(3):0.
54. Tabangin ME, Woo JG, Martin LJ. The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proc. 2009;3(7):S41. BioMed Central Ltd.
55. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–5.
56. van Binsbergen R. Accuracy of Imputation to Whole-Genome Sequence Data in Holstein Friesian Cattle. Genet Sel Evol. 2014;46(1):41-54.
57. VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA. Genomic evaluations with many more genotypes. Genet Sel Evol. 2011;43(1):10-21.
58. VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB, et al. Genomic imputation and evaluation using high-density Holstein genotypes. J Dairy Sci. 2013;96(1):668–78.
59. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23.
60. Wei B, Yang F, Wang X, Ge Y. knnGarden: Multi-distance based k-Nearest Neighbors. R package version 1.0.1. 2012. URL: http://cran.r-project.org/web/packages/knnGarden.
61. Wilson DR, Martinez TR. Reduction techniques for instance-based learning algorithms. Mach Learn. 2000;38(3):257–86.
62. Wright S. Coefficients of inbreeding and relationship. Am Nat. 1922;56:330–8.
63. Xavier A, Beavis WD, Specht JE, Diers B, Muir WM, Rainey KM. SoyNAM: Soybean Nested Association Mapping Dataset. R package version 1.2. 2015. URL http://CRAN.R-project.org/package=SoyNAM.
64. Xu S. Estimating polygenic effects using markers of the entire genome. Genetics. 2003;163(2):789–801.
65. Yang Y, Wang Q, Chen Q, Liao R, Zhang X, Yang H, et al. A new genotype imputation method with tolerance to high missing rate and rare variants. PLoS One. 2014;9(6):e101025.
66. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014;11(4):407–9.