REVIEW 1

COMMENTS TO THE AUTHOR:

Reviewer #1: In the MS entitled "Marker imputation efficiency for Genotyping-By-Sequencing data of crop species with and without a reference genome" Authors used different tools to impute missing SNP data in Medicago sativa, and rice. The MS demonstrated very good approach to impute SNP data even independent of whole genome sequence. No doubt the efforts made by authors are valuable. I have some major concern with the MS presentation, flow of information and regarding the novelty. Here are my point-wise comments:

General Comments:

The four methods used for imputation are now very common. Even software tool like Tassel host imputations methods like FILLIN, FSFHap, Numerical, and Beagle. There are several methods and tools available for imputation and many studies demonstrating comparisons have been published.

More particularly, in species without availability of whole genome sequence methods for the imputation of SNP data have been published. For instance, Ward, et al. (Saturated linkage map construction in Rubus idaeus using genotyping by sequencing and genome-independent imputation. 2013. BMC genomics, 14(1), 2).

Even in Alfaalfa (Medicago sativa) Rocher et al. (Validation of Genotyping-By-Sequencing Analysis in Populations of Tetraploid Alfalfa by 454 Sequencing. 2015. PloS one, 10(6), e0131918.) demonstrated imputation at missing rate 0.5.

Specific comments:

In title "crop species with and without" is miss leading since authors have used two species so better to be specific "medicago and rice".

Abstract need to be more concise and informative

Introduction:

Line 38-39, GWAS can not handle missing data- need citation and reframe the sentence. Better to verify the fact.

Line 43-44 - alternative to SNP-chip- need to improve

Line 48 - '52% or up to 52%' need to verify the fact

Line 55 -57 - imputation method has been developed for SNP-Chips. I am not sure about the accuracy of statement

Line 85- rice genome recently assemble - not true

Material methods:

Since the MS specifically focused on imputation of SNP data, the details like salt tolerance or high dry matter yield is unnecessary. So better to concise portion from 96-114 in single sentence.

Similarly details about imputation methods can be concise (page 7 to 12) and more details can be provided as supplementary.

Results:

Need to avoid words like close to 100% better to be very precise.

Maximum missing allowed and maximum percentage of missing actually present are two different things and it need to be very clear.

Discussion

Subtitles are not good - for instance "Role of crop species" and "Size of the problem".

Previous publication demonstrating imputation in alfalfa, and other species without genome need to be discuss.

Also it will be better to highlight importance of percent study while discussing previous paper comparing different imputation methods.

Fig 2. and 3 - Y axis need to be zoomed to make difference visible.

I suggest to reframe the MS to make it pursuable for readers of Molecular Breeding

Reviewer #2: The study is very interesting, however, the authors should explore the results in more depth and with more creativity. the current results are shallow.11

REVIEW 2

Manuscript: Marker imputation efficiency for Genotyping-By- Sequencing data of crop species with and without a reference genome

Authors: Nelson Nazzicari, Filippo Biscarini, Paolo Cozzi, E. Charles Brummer, Paolo Annicchiarico

General comments

This manuscript deals with an important and emerging topic in plant (and animal) breeding as GBS is considered as a viable alternative to SNP array-based genotyping. These GBS data come with major

challenges, also on the imputation of missing values. This manuscript describes the application of six algorithmic approaches to data on two plant crops that are on the edges of the wide range of genotype complexity in plants. Not surprisingly, the Beagle software performs very well (superior over all other methods) when map information is available, and much worse, when map information is absent.
My main concerns on the paper are:

1. Why these two species? They have very different properties which makes it difficult to associate the differences in results to particular characteristics. For example, is ploidy level more critical than a reference genome? It might have been more insightful when including an outbred diploid species with a reference genome and similar to the rice to assess imputation accuracy with and without reference genome information.
2. A missed opportunity to study the utilization of a reference genome of a related species, as is available for alfalfa.
3. The choice of the five imputation methods. Why these?
4. Why performs Beagle still well with the unordered SNPs in the rice dataset. How severe was the re-ordering?
5. Too much attention for computational issues while this was not the main objective of the study (there was no effort to optimize implementation of the various methods, except for RF).
6. No attention to results on individual rice chromosomes. In case there were no clear differences between chromosomes, then that should be stated more clearly.

In addition, various specific comments are provided below that may be of help to further strengthen the manuscript.

Specific comments

Line 52: It would be nice if you could provide a list of associated packages, AND explain why Beagle was selected out of these methods.

Line 78: There is no development of the best method in this study, nevertheless the results of this study may identify the essential properties of such a method. Was this discussed?

Line 83: For alfalfa one could use the assembly of Medicago Trancatula – was this considered?

Line 86: Why these four general methods?

Line 109: How to interpret the 437 plants from 391 accessions as these accessions are supposed to be fully homozygous? Were there duplicates?

Line 134: Why were these values (4 and 11) were chosen? I do not see the rational of taking 4 for a heterozygote and 11 for homozygote - shouldn't these be the reverse?

Line 135-139: The reduction from 5 genotypic classes to 3 genotypic classes in alfalfa is a critical simplification step. There is no justification why this step was executed. My assumption is that this is because of the (software) methods that were used. The explanation and justification should be provided here. I am wondering whether the four general methods could be applied to the 5 genotypic classes – was this considered/studied?

Line 166: How to interpret the 'K closest markers' in case there is no map or assembly information (as here for alfalfa)?

Line 169: How was distance specified, i.e., in bp or cM? Did you try both and if so, where there differences?

Line 173-181: I wonder whether this explanation is appropriate for this audience.

Line 175: Were the eigenvectors of M'M and MM' merged to arrive at the 'first k eigenvectors'?

Line 178: Which value for k was used, was there a decision criteria used?

Line 182: For the SVD, how critical is the initial imputation step by MNI? Were other methods considered for this initial imputation?

Line 185 – 198: Explain the 'hat' for U in equation 2. Also, why has Y no hat while there is a hat on the Y in equation 3.

Line 217: Elaborate on the content of this likelihood file or provide a reference (to the manual of Beagle)?

Line 229: What was the reason to not go beyond 20% artificially induced missing values?

Line 254: It is not clear whether the 'total number of missing data' refers to the sum of 'natural' and 'artificial' missing data or only to the latter part. Please improve description.

Line 255: It may be confusing to use y in equation 4 as a discrete variable since Y is continuous in equations [2] and [3]. It seems more intuitive to use g instead of y.

Line 281/Table 1 & Table 2: Why not merge the content of these two tables? Transpose Table 1 to have the same columns. Then ADD the genotype frequencies for the four call-rate scenarios (10,20,40,70%). In that way the table may reveal shifts in frequencies when imposing stronger thresholds, which may invoke further inferences on results of imputation accuracies.

Line 282: Why not present these rates as percentages as that is probably easier to read and more consistent (with e.g., line 286).

Line 284: Figure S1 does not contain information at the 'per individual' level

Line 304: How sensitive are the results to the imposed quality filters?

Line 313/Figure 3: Suggest to maintain consistency in y-axis scale, i.e., the plots for Minor homozygous accuracy should range from 0 to 100% as well.

Line 324: the flat response seems to hold for the studied range (0.01 – 0.20). Were higher values of missing rates studied? It will be interesting to study whether there will be a critical point and whether these are different for the different genotypic classes.

Line 347: Rather extensive reporting on computation time which seems not justified as the used methods were used as these were available to the authors, i.e.., there was not dedicated efforts to speed up the implemented methods.

Line 379-381: Where was this negative impact shown? In this study or other studies – please include references.

Line 401-407: This is still rather descriptive without any postulation whether any of the methods are less or more sensitive to MAF and/or balance. The superior results of Beagle in rice are mostly driven by utilizing map/assembly information which is ignored by other methods.

Line 440: The reasoning on factor ii) is not clear – would the results for rice be less accurate when analysing the genome at once? Also, why would it simplify the imputation process?

Line 443: Did you explore the population structure of the three datasets? Would presence/absence of (severe) population structure affect imputation accuracy?

Line 450: It is too easy to state that there is additional complexity to make use of a reference genome of a related species. It would be of HIGH relevance to explore this as it may be applied in many species that lack a reference genome themselves.

Line 535-540: Too general/vague and thus not clear why this paragraph is included in the conclusion section (or even in the manuscript)!

Figure 1: Be consistent in using rates versus percentages, i.e., this Figure uses rates while on Line 311 you refer to percentages.