Response to the reviewers on "SNP discovery by exome capture and resequencing in a pea genetic resource collection"

The authors want to thank the reviewers for their very useful comments. Our answers are in the text below.

Reviews

Reviewed by Rui Borges, 31 Dec 2022 18:25

The paper by Aubert, "SNP discovery by exome capture and resequencing in a pea genetic resource collection," aims to evaluate the genetic diversity of a collection of 240 pea (Pisum sativum L.) accessions. This paper reports on the large number of SNPs (approximately 2.3 million) obtained through whole-exome sequencing. The methods described in the paper appear to be sufficiently detailed. Additionally, I agree with the authors' conclusions and recognize the potential value of this dataset for future genome-wide association studies and breeding programs in pea.

My main concern is with the phylogenetic analysis. The authors carried out their analysis using SNPs while ignoring constant sites (lines 92-94). This could potentially impact the estimated distances based on the GTR model. It is not clear to me if this will significantly affect the tree topology (although, I suspect it will not), but it could alter the branch lengths. I recommend that the authors take steps to correct for ascertainment bias in their analyses. Furthermore, the authors chose to present the tree as a cladogram, which does not allow the reader to appreciate the branch lengths. This information could be useful in understanding the differences found between the clustering and structure methods (lines 152-53). For example, the authors should consider whether long-branch attraction could be a factor.

We agree with your comment. We now have included a phylogram replacing Figure 1. Based on this figure, we can rule out the effect of long-branch attraction on the difference between DAPC and FastStructure method.

My primary concern, however, is more fundamental. Given that the authors are working with genetic positions that are not fully sorted, can they accurately estimate a phylogenetic tree based on the GTR distance?

As you suggested below to use a coalescent approach, we used a sliding-window method for a phylogenetic inference excluding SNPs on unplaced contigs (see details below). Material and methods section has been rewritten accordingly.

I am uncertain how to interpret the estimated clades and distances in this context. Therefore, I believe it would be more appropriate to estimate a coalescent tree instead, as it provides a more appropriate description of the dataset the authors have on hand. This is of fundamental importance because I believe that some of the clades estimated in this work are likely to be used in further studies.

As mentioned above we inferred phylogenies using a coalescent approach described by Wang et al (2022). The two methods used in this paper resulted in the same clade delimitation, but their phylogenetic relationships are not totally concordant (cf supplementary figure 2 that compares the trees generated with the two methods). We briefly describe the phylogenetic

results but it is not the main scope of this data paper (The phylogeny of the germplasm will be more deeply described in subsequent papers)

There are also a few minor aspects I would like to address:

- Lines 89-90: It would be helpful if the authors could explain why a 10% threshold was chosen for filtering SNPs.

The threshold of 10% of heterozygous accessions per SNP has been chosen arbitrarily. Pea is autogamous, thus the level of heterozygosity is expected to be low. In addition, for subsequent GWAS analyses, this 10% threshold reduces biases.

- Lines 124-125: It would be beneficial to include a brief explanation of how the categories of low, moderate, and high were determined (this is only a suggestion).

The categories are determined by the SNPeff program, described in Cingolani et al. 2012. The high impact category corresponds to a major change in the protein (regrouping the start_lost, stop_gained, stop_lost, splice_acceptor_variant, and splice_donor_variant categories), the moderate category corresponds to the missense, and the low effect category is the rest. A short description has been added for the three categories

- Line 127: Is it likely that the 0.53% SNPs with a disruptive nonsense effect (or total 0.2329*0.0053) could potentially be due to sequencing errors?

We believe that the observed polymorphisms are true. Natural variation inducing nonsense mutations have been described in other species. For example, Flowers et al (The Plant Cell 27 (9):2353-2369) found nonsense mutations in 4% of genes in *Chlamydomonas reinhardtii*. In Date palms, Hazzouri et al (Nature Communications, 6:8824) found 4,162 nonsense polymorphisms affecting 3,288 genes which is comparable to the figures we obtain with a large panel.

- Line 137: It is not clear to me how the clustering analyses separated the accessions according to crop evolution. Could the authors provide further clarification on this point?

By crop evolution, we mean the transition between wild species to landraces, and then to cultivars (domestication, modern breeding)

- Lines 115-116: I found it peculiar that the cultivated winter pea fodder had only two singletons per accession. Is there a reason for this?

It has also surprised us. It could be hypothesised that closely related fodder accessions are present in the panel, so not many SNP are specific only from that accession with few singletons.

- Lines 152-153: It would be helpful if the authors could provide more specific reasoning for why they believe the differences in placement between the phylogenetic analyses and the structure/clustering analyses are due to kinship.

Indeed, that was only a hypothesis that would need proper kinship data to be verified. We removed the sentence. However, these differences could also be attributed to statistical artefacts (such as long branch attraction).

Reviewed by anonymous reviewer, 01 Jan 2023 14:03

In the manuscript "SNP discovery by exome capture and resequencing in a pea genetic resource collection", the authors performed exome sequencing to genotype 240 accessions of pea and identified a dataset of SNPs to be used for genetic diversity analysis. For this, the authors selected a large number of samples, including cultivars, landraces, and wild types with diverse geographical origins that follow the standard approach of population genetic analysis. This study is interesting and has valuable information, but some details still need to be improved. The following are some questions and suggestions for modification:

- I wish the authors could state why exome-derived SNPs were chosen instead of GBS SNPs, which are widely used to assess genetic diversity in several plant species, including those with complex genomes.

Both techniques can clearly be used. However, our previous experiences using GBS in pea did give genotyping matrixes with quite a lot of missing data. A comparative study in Picea abies (Eklöf et al, 2020, Forests) has also shown that targeted capture probes were slightly more effective than GBS to assess genetic diversity.

- In the background section, the authors should add the genome characteristics data: number of chromosomes, ploidy level.

Information has been added

- If possible, I suggest the authors indicate the sample locations in the map figure, so it will be easier to see the geographical distribution.

Unfortunately, we don't have the geographic origin for all accessions.

- In the method section, the author should provide information about the number of probes used in this study.

The probe design has been realised by Roche and we unfortunately didn't get that information from them.

- The authors should provide more details about the criteria used to select the subset of SNPs. Do you filter SNPs based on MAF? What about the threshold for linkage disequilibrium?

The PLINK option for linkage disequilibrium (--indep 50 5 2) and the MAF threshold (1%) have been added in the text.

- I suggest the authors add an analysis of maker polymorphism (PIC), genetic diversity parameters, as well as genetic differentiation (FST) among sub-populations.

This would be interesting, but we do not believe it is within the scope of a data paper

- The authors used a maximum likelihood phylogenetic tree, according to lines 94–95, but neighbor-joining phylogenetic trees are mentioned in line 134.What is the method used for the analysis of the phylogenetic tree?

Mistake, apologies .We corrected it.

- For the structure analysis, the authors should also provide information regarding the settings and parameters for software used. What threshold value of the membership coefficient was used to assign an accession to a specific group or assign an accession as admixture?

There was no threshold used to assign an accession to a specific group with Faststructure as DAPC has been used to cluster the accessions.

- In the results section, the authors should report the number of raw reads obtained from sequencing, the %map read with the reference genome, and the number of raw initial SNPs obtained.

Thanks for the suggestion, a supplementary table has been added

- The authors analyzed k values from 1 to 10. They should show a plot or the statistic that indicates which is the best value of K.

Two different values were given by ChooseK.py, one that maximises marginal likelihood which inferred K = 4, and one that determined the best model components used to explain structure, which inferred K = 5.

- Line 137: In my version, it does not have table 1.

It is meant to be Supplementary table 1, sorry for the mistake that has been corrected

- Figure1: If possible, I suggest the authors separate the DAPC plot into Figure 2 and color the branches of the phylogenetic tree according to crop evolution and cultivation types, as well as include bootstrapping supporting values in the tree.

Different phylogenetic trees have been proposed following recommendations of reviewer 1

- The manuscript should have a discussion section and should be interpreted with the results as well as discussed in relation to the present literature.

There is usually no discussion in a data paper as further papers will discuss results using these data. Does the recommender agree with that?