Round #2

Dear managing board and reviewers

We would like to thank you for your time supporting us in the improvement of our manuscript to be recommended by PCI in Evolutionary Biology. As well as for your invaluable comments and suggestions to improve our manuscript. We have addressed all your remarks and have prepared a new version of our manuscript, see below. We hope that this new version will be satisfactory for you and be a candidate for publication in the PCI in Evolutionary Biology journal.

Sincerely,

Sergio Gabriel Olvera Vazquez, on behalf of all co-authors.

Author's Reply:

---

by Ignacio Bravo, 23 Aug 2021 21:11
Manuscript: **https://www.biorxiv.org/content/10.1101/2020.12.11.421644v2** version 2

Dear Authors

first of all, please apologise for the too long time between the reviewers' response and this answer of mine.

In the current revision of their manuscript, Olvera-Vázquez and coworkers have addressed most of the points raised during the first PCI review round. Most of the questions have been properly addressed, and I think the review process has helped clarify the message. Nevertheless, I consider that a number of questions still remain confusing in my eyes and that still require to be elucidated, as detailed below:

-Treatment of admixed individuals.

The authors were unable to assign 175 individuals (a third of the total individuals analysed) to any of the five genetic clusters. These individuals were thus not included in any further analyses. The large number of admixed individuals raises an important concern for the pertinence and interpretation of subsequent analyses. This possible caveat needs to be properly identified and a clear word of caution must be raised in the discussion and possibly in the abstract. The implication of the heterogenous distribution of the admixed individuals in the RF-ABC approach needs also to be explicitly stated.

- ➔ We agree with the reviewer that there are always cautions to take about the choice to remove the admixed individuals. We have excluded admixed individuals for summary statistic estimates and ABC inferences because we wanted to understand the relationships among the "pure" genetic groups and focus on ancient gene flow (*i.e.*, a recent history of gene flow is seen in the STRUCTURE barplots). We showed that even when removing recently admixed individuals detected with STRUCTURE, we still infer substantial gene flow among populations. Our results thus support the interpretation that gene flow plays an important role in *Dysaphis plantaginea* evolutionary history. We have tried to clarify this point in the discussion in lines 792-794 and 894-900.

I think an indication of the geographical distribution of these admixed individuals is needed. I pointed this in my previous decision and the authors answered "We believe that the map Figure 2 is already presenting those results. The Western European and Spanish genetic clusters are the most admixed, as shown in the mean membership coefficient per site". I am afraid I disagree with this answer: nothing in figure 2 indicates the geographical distribution of the admixed individuals. From data presented in FigS6 it seems that the number of admixed individuals is not evenly distributed across sampling sites. For instance, samples collected in the USA, Morocco or Romania seem to contain less admixed individuals than samples from France. (Note: it may be that the admixed individuals have been included for the analyses depicted in Figure 1, but this is unclear.) I suggest that the authors include in the pie charts in figure 2c a sixth category corresponding to the admixed individuals. I would also suggest to avoid overlapping the pie charts (the precise location is given elsewhere) and to make the individual size of each pie chart proportional to the total number of individuals analysed in the sample site.

- ➔ We have added figure S11 in supplementary material that shows the spatial distribution of the average number of admixed individuals per site. This figure shows, as stated in lines 660-662 that "Most of the admixed individuals were located in Western and Northern Europe; the spatial distribution of the mean number of admixed individuals per site is represented in Figure S11.". This was also discussed in lines 792-794. We have modified Figure 2: the size of each pie chart is proportional to the number of individuals sampled per site. In addition, we focus on the specific areas of the map which presented overlapping pie charts (*i.e.*, Europe).

The 3-D PCA is unclear (as any 3-D representation is). I would suggest to present instead two 2-D representations of PC1vsPC2 and PC1vsPC3. This may help highlight the apparently true isolation of the blue genetic cluster and may also help visualise the apparently intermediate location of admixed individuals between the green and red genetic clusters.

- ➔ We are thankful for your comment to improve the visualization of our data. We have added an extra figure S12 in the supplementary material. This new figure presents the three main components using 2D-PCA plotting. However, we still think that the three 3-D PCA is clearer than the figure S12, so we prefer to keep the 3D-PCA as the main PCA figure.

-Distribution of samples used for 16S rDNA analyses.

I understand from the answer to my comment that the authors are aware of the lack of representation of certain geographical regions in these analyses. For the sake of clarity and to avoid generalisations, I would suggest to make it explicit in the discussion what geographical locations were undersampled for metagenomics or not included with respect to the aphid genome markers.

- ➔ We have clarified the analyzed regions in lines 293-294.

Regarging isolation-by-distance analyses.

The Sp-based analyses have been performed using only the individuals allocated to one of the five genetic clusters, and the same seems to hold true for results in Fig S11. However, this analysis should probably be performed using all individuals in each orchard, including the admixed ones.

- ➔ *Sp* needs to be estimated for panmictic populations (Vekemans and Hardy, 2014). We therefore computed the *Sp* statistics within each genetic group detected with STRUCTURE, we have added an explanation in line 467 «panmictic». We used a cut-off to assign individuals to each group, thus excluding admixed individuals. The IBD tests included the admixed individuals (all samples) and allow to estimate gene flow among sites. The two approaches are therefore complementary.

Please include in the figure the results for the fit that are included in the text (F value, P value, R2). I would also recommend to perform the linear fit without the most distant sampling sites in the USA, and also probably to perform the linear fit only for the European samples. Please specify also the units for the x-axis.

➔ We draw a new figure showing the linear fit using all the individuals, removing the US sites, and using only European individuals. We specified the unit of the axes and added the F-value, *r2* value, and *P-value*. We have added this information in lines 697-700.

Reviews

Reviewed by Pedro Simões, 13 Jun 2021 21:40

I have now gone through the revised version and I am satisfied with this revised version and the answers provided by the authors to my comments. I have no further comments to add.

Thanks a lot for your feedback and comments.

➔ We thank you for all your comments and time in reviewing our manuscript.

Reviewed by anonymous reviewer, 18 Jun 2021 21:12

The authors have addressed most of my comments, and I appreciate their careful revisions. I have only a few minor suggestions:

L198: Change "America" to "North America"; (this is the instance I was referring to in previous review regarding L203.)

L203: change "alternate" to "alternates"

"We also continued utilizing the north-east and south-west because we were describing a pattern from one area to the other in lines 572-573-578." I do not insist (because the meaning is clear), but correct English usage is "northeast" and "southwest" without the hyphen.

➔ We have modified the text in the following lines: 211 ("…North America"), 216 ("…alternates"), 609-614 (we have eliminated the hyphen).

Table 1: The issues with the use of "n.s." in the table have not been addressed and are still a problem. Comment from the previous review: "Only one entry in the Sp column is marked as 'n.s.'. Does this mean all the other entries are significant? Usually significant values are marked with an asterisk or other superscript (as is done for Fis in this table), and non-significant values are left unmarked. So this Sp column and the values in the last two columns of the table being marked as n.s. or left unmarked is disorienting and unconventional, and should be changed."

➔ We tried to make this Table 1 clearer.

L842-843: abbreviate the genus name of the six listed Dysaphis species as "D." instead of spelling out each time. Same for *D. reaumuri* and *D. pyri* in L846.

L845: change "pears including," to "pears, including"

L891: change "species, are" to "species are"

➔ We attended your comments in the following lines: 951-955 (We abbreviated the genus name), 954 ("…pears, including"), and 1018 ("…species are").

References

Vekemans, X., & Hardy, O. J. (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular ecology*, *13*(4), 921-935.